scE^{2}TM improves single-cell embedding interpretability and reveals cellular perturbation signatures

Hegang Chen[1,2], Yuyin Lu[1], Yifan Zhao[3], Zhiming Dai[1], Fu Lee Wang[4], Qing Li[5], Yanghui Rao[1,*] and Yue Li[2,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China.
[2]School of Computer Science, McGill University, Montreal, Canada
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, USA.
[4]School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China.
[5]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China.
[*]Correspondence to raoyangh@mail.sysu.edu.cn, yueli@cs.mcgill.ca

Abstract

Single-cell RNA sequencing technologies have revolutionized our understanding of cellular heterogeneity, yet computational methods often struggle to balance performance with biological interpretability. Embedded topic models have been widely used for interpretable single-cell embedding learning. However, these models suffer from the potential problem of interpretation collapse, where topics semantically collapse towards each other, resulting in redundant topics and incomplete capture of biological variation. Furthermore, the rise of single-cell foundation models creates opportunities to harness external biological knowledge for guiding model embeddings. Here, we present scE^{2}TM, an external knowledge-guided embedded topic model that provides a high-quality cell embedding and interpretation for scRNA-seq analysis. Through embedding clustering regularization method, each topic is constrained to be the center of a separately aggregated gene cluster, enabling it to capture unique biological information. Across 20 scRNA-seq datasets, scE^{2}TM achieves superior clustering performance compared with seven state-of-the-art methods. A comprehensive interpretability benchmark further shows that scE^{2}TM-learned topics exhibit higher diversity and stronger consistency with underlying biological pathways. Modeling interferon-stimulated

PBMCs, scE^{2}TM simulates topic perturbations that drive control cells toward stimulated-like transcriptional states, faithfully mirroring experimental interferon responses. In melanoma, scE^{2}TM identifies malignant-specific topics and extrapolates them to unseen patient data, revealing gene programs associated with patient survival.

1 Introduction

Advances in single-cell sequencing technologies have revolutionized our ability to study biological systems with unprecedented resolution [1, 2]. A central computational task in analyzing single-cell RNA sequencing (scRNA-seq) data is unsupervised clustering, which enables *de novo* cell-type identification and has revealed previously uncharacterized cellular states [3]. In recent years, deep learning methods have attracted significant attention in the single-cell community due to their remarkable ability to process sparse, noisy, and high-dimensional data [4]. However, the inherent complexity of multi-layer nonlinear architectures, activation functions, and large parameter spaces often makes these deep learning models difficult to understand and verify [5]. This lack of interpretability poses a significant challenge to verifying whether proposed models can capture actual biological mechanisms and to translating computational findings into biological insights.

Interpretable deep learning has become a promising approach to address challenges in analyzing biological data. A model is deemed interpretable when its learned parameters can directly link input features with latent variables or target results [6], allowing researchers to gain biological insights without needing extensive *post hoc* analyses [7–9]. Recently, topic modeling has been successfully applied in single-cell genomics [6, 10, 11]. Originally developed for mining textual data [12–14], topic modeling's appeal lies in its ability to efficiently handle high-dimensional sparse data while providing interpretability by associating input features with latent factors or outcomes through discovered topics. In the context of single-cell omics analysis, a series of studies have integrated deep learning-based embedding techniques with topic modeling, leading to the development of various single-cell embedded topic models [11, 15–20]. scETM [6] adapted the embedded topic modeling [21], utilizing interpretable linear decoders to learn highly interpretable gene and topic embeddings. Building upon this, d-scIGM [18] enhances the model's capacity to capture latent data signals by employing deeper neural network architectures, thereby better modeling the dependencies between topics and genes. SPRUCE [17] combines a topic modeling approach with cell-cell interactions to simulate informational crosstalk between cell-cells while identifying cell types.

Despite this progress, existing embedded topic models tend to suffer from interpretation collapse, where the optimization disproportionately weights frequent elements during trainnig, causing topic embeddings to converge and lose diversity [22, 23]. scRNA-seq data mirrors in an important way: just as text has a long-tailed word distribution, scRNA-seq data is usually dominated by highly expressed housekeeping genes that lack cell-type-specificity. This creates conditions where interpretation collapse can occur. Interpretation collapse generally manifests in two ways. First, models over-emphasize highly expressed genes, leading to redundant identification of common gene programs while missing diverse biological signals in rare cell

types. Second, the learned topics may not always align well with actual cell types. In other words, if the topic assignments do not align well with established cell-type annotations, it suggests that the models are failing to capture a biologically meaningful representation of cellular identity. Current interpretability assessments for most single-cell embedding topic models rely on qualitative analysis of model-derived topics [24, 25]. Such an approach carries inherent risks of bias and subjective interpretation, making it difficult to systematically determine whether model optimization preserves interpretation quality. Meanwhile, recent advances in single-cell foundation models trained on massive databases [26, 27] have demonstrated robust performance across numerous downstream tasks, offering a promising source of diverse knowledge as supervisory signals for embedding and clustering. The rich knowledge of single-cell foundation models can effectively guide embedding and clustering by providing external supervision beyond the constraints of individual datasets.

We present **scE^{2}TM** (**s**ingle-**c**ell **E**xternal knowledge-guided **E**mbedding clustering regularization **T**opic **M**odel), a deep generative model for large-scale single-cell transcriptomic analysis (Fig. 1). scE^{2}TM introduces two key innovations to overcome interpretation collapse. First, **E**mbedding **C**lustering **R**egularization (**ECR**) [23] encourages topic diversity by regularizing topic embeddings as clustering centers and gene embeddings into clustering samples, with cluster soft assignments modeled through Optimal Transport (OT) and enhances representational learning of distinct biological processes. Second, scE^{2}TM integrates external knowledge from single-cell foundation models through a **C**ross-**v**iew **E**ncoder (**CVE**), shifting from internal guidance paradigms to scalable external knowledge integration. To systematically evaluate scE^{2}TM, we employed ten quantitative interpretability metrics that, combined with clustering metrics, enable rigorous assessment of topic model performance. Benchmarking across 20 datasets revealed that scE^{2}TM mitigates interpretation collapse while existing methods suffer. Critically, clustering performance and interpretability are mostly uncorrelated–demonstrating that previous models sacrifice bioloical insights while optimizing clustering accuracy. scE^{2}TM achieved state-of-the-art performance in both aspects. We then showcase real-world applications of scE^{2}TM in three case studies. In a human pancreas scRNA-seq dataset, scE^{2}TM captured insulin biosynthesis pathways and programs reflecting key cellular identities. Topic-based perturbations in interferon (IFN)-stimulated PBMCs successfully transformed cellular states, validating capacity to simulate cellular responses to exogenous stimuli. In a melanoma scRNA-seq dataset, scE^{2}TM-derived topics generalized to TCGA cohorts with topic activity significantly correlating with patient survival, demonstrating clinical relevance. Collectively, these results establish scE^{2}TM's as a framework for extracting mechanistically insightful and clinically actionable patterns from single-cell transcriptomics.

2 Results

2.1 Overview of scE^{2}TM and interpretability evaluation benchmarks

We present scE^{2}TM, a single-cell embedded topic model designed for modeling scRNA-seq data, together with comprehensive framework for quantitatively evaluating model interpretability.

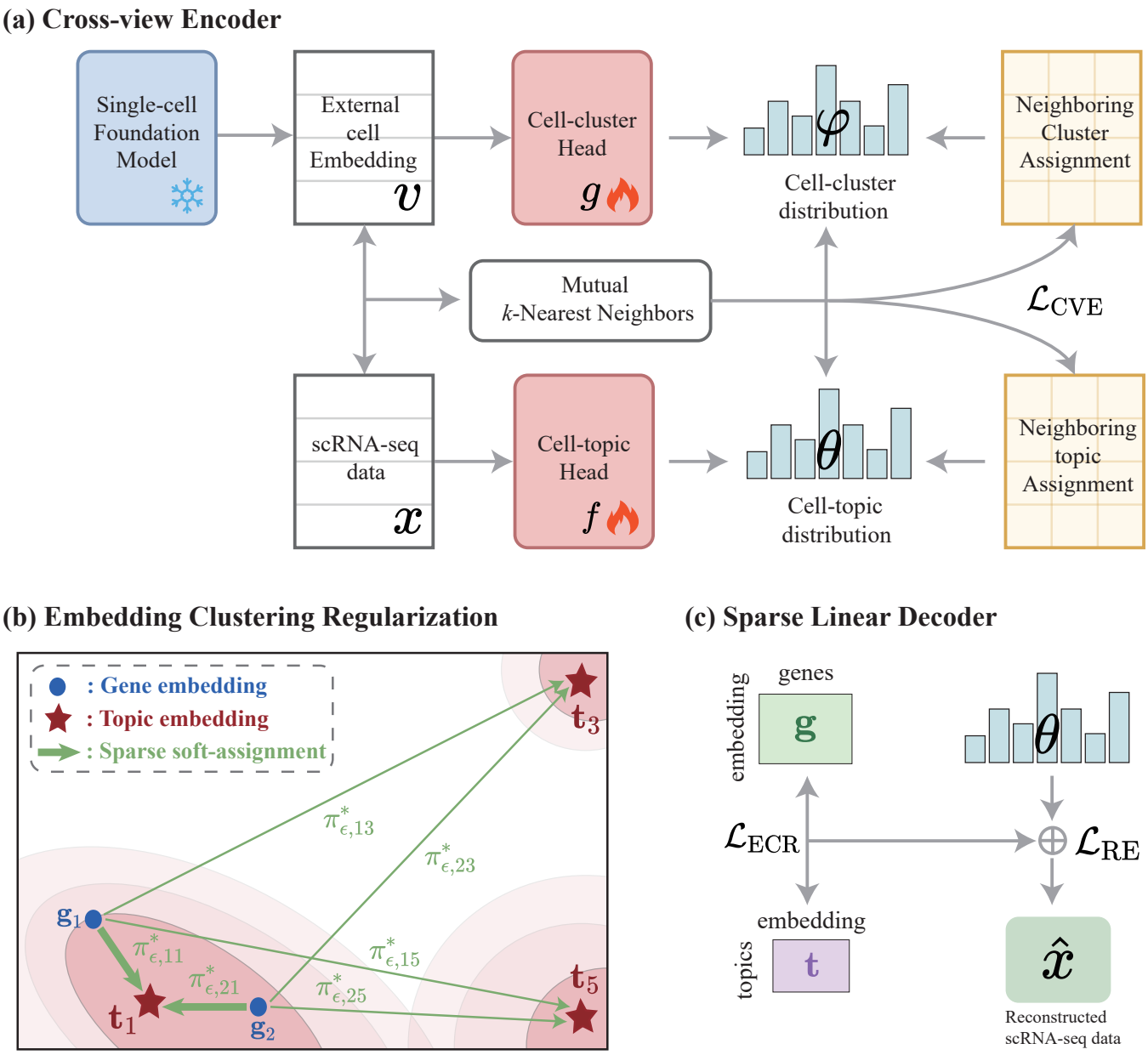

Figure 1: Schematic overview of scE^{2}TM. (a) Cross-view encoder. This encoder integrates single-cell expression data with embeddings extracted from a single-cell foundation model. To combine these two perspectives, cluster and topic heads are trained based on the mutual neighborhood information by encouraging consistent clustering assignments of mutual nearest neighbors of the corresponding cells. (b) Embedding clustering regularization module. ECR clusters gene embeddings $\mathbf{g}_j$ ($\bullet$) as samples and topic embeddings $\mathbf{t}_k$ ($\star$) as centers with soft-assignment $\pi^*_{\epsilon,jk}$. For instance, ECR pushes $\mathbf{g}_1$ and $\mathbf{g}_2$ close to $\mathbf{t}_1$ and away from $\mathbf{t}_3$ and $\mathbf{t}_5$. (c) Sparse linear decoder. Decoder learns topic embeddings and gene embeddings as well as sparse topic-gene dependencies during reconstruction, thereby ensuring model interpretability.

While existing single-cell topic models have provided valuable foundations for the field, three opportunities remain for advancing the state-of-the-art: integrating external biological knowledge from foundation models to enhance performance, addressing topic interpretation collapse where models may generate overlapping topics, and establishing systematic methods for quantifying interpretation quality.

To address these challenges, scE^{2}TM provides three innovations. First, we introduce a Cross-View Encoder (CVE) which systematically transfers knowledge from single-cell foundation models into the embedding space. The CVE aligns the internal transcriptomic features (scRNA-seq transcriptome, $\mathbf{x}$) with external embedding ($\mathbf{v}$) derived from a foundation model through cross-view mutual distillation which integrates different viewpoints through mutual distillation of neighborhood information. By enforcing consistent topic and cluster assignments between corresponding cells and their neighbors across views (Fig. 1a), CVE effectively incorporates contextual biological knowledge learned from millions of cells into the model's representation learning process. Second, we implement Embedding Clustering Regularization (ECR) [23] to prevent interpretation collapse and enhance topic diversity. ECR solves an Optimal Transport (OT) problem that establishes soft cluster assignments by linking topic embeddings with the gene embeddings (Fig. 1b). This formulation compels each topic embedding to be unique cluster center aggregating unique gene embeddings, ensuring topics remain distinct and span diverse semantic spaces. Cell representations are then reconstructed using the distribution of learned topics and the optimized topic-gene dependencies between topics and genes facilitated by ECR (Fig. 1c). Third, we conduct a systematic interpretability assessment framework comprising ten quantitative metrics to evaluate topics from multiple perspectives, such as alignment with cell types, intrinsic coherence and diversity, and the capacity to capture biological pathways.

2.2 Robust and ablation analysis of scE^{2}TM

We systematically assessed the robustness of scE^{2}TM by conducting extensive sensitivity analyses on key hyperparameters based on clustering and interpretability metrics (Supplementary Fig. 1; Section 4.2, 4.4). To integrate different viewpoints, CVE employs mutual distillation of neighborhood information. We observed stable clustering and interpretability performance across various numbers of nearest neighbors, suggesting the model's robustness to the choice of neighborhood size (Supplementary Fig. 1a). Subsequently, we analyzed the impact of the regularization term weights in ECR on performance. While increasing the weights induces some changes in the performance of topics and the related pathways for topic mining, the overall stability is maintained (Supplementary Fig. 1b). Finally, we visualized how λ (i.e., regularization weight used in ECR) influences the topic-gene geometry on the human pancreas data via UMAP [28] (Supplementary Fig. 1c). As λ increases, the topic distribution becomes more discrete, with genes confined to the neighborhood of the topic.

To further investigate the contribution of each component in scE^{2}TM, we performed ablation studies across four scenarios: (i) omitting the embedding clustering regularization module (without ECR), (ii) excluding the cross-view encoder (without CVE), (iii) substituting the cross-view distil-

lation strategy with embedded connectivity (with EC), and (iv) utilizing cell embeddings derived from a foundational model scGPT for clustering. Supplementary Table 1 presents the average metrics for these four ablation scenarios of scE^{2}TM. The ECR ablation notably decreases the diversity of interpretations, indicating its essential role in preventing interpretation collapse. The ECR module achieves this by minimizing the repetition of key genes across topics, ensuring geometric compactness between topic embeddings and their related gene embeddings. Over-representation analysis (ORA), concentrating on the top genes, mirrors the topic metrics such as Topic Coherence (TC) versus ORA$_N$ and Topic Diversity (TD) versus ORA$_U$. Unlike ORA, GSEA shows different results when the ECR module is removed. Without the ECR, the distance between a topic and its top gene embedding is not constrained, increasing the repetition rate among top genes and reducing TD and ORA$_U$. Conversely, GSEA, which considers all genes, reveals that without ECR, genes are more evenly distributed in the embedding space. This means that even though the repetition rate of top genes in each topic rises, they still represent only a small part of the entire gene set, leading to contrasting outcomes. This difference underscores how various interpretability metrics capture different facets of biological coherence.

Eliminating CVE significantly worsens clustering performance, with ARI decreasing by 7.2% and NMI by 4.8%, whereas straightforward knowledge incorporation (using EC) offers only slight improvements. These findings validate the necessity of the proposed cross-view distillation framework for effectively integrating external knowledge with scRNA-seq data. Furthermore, the scGPT foundation model shows inferior performance in zero-shot scenarios, with ARI dropping by 25.0% and NMI by 19.8% compared to scE^{2}TM, underscoring the importance of adapting to specific tasks [29]. Together, ablation studies reveal that scE^{2}TM consistently performs well across crucial hyperparameters and that both its ECR and CVE are crucial elements, jointly preventing interpretation failures and successfully merging knowledge from foundation models to achieve enhanced clustering and interpretability.

2.3 scE^{2}TM provides high-quality cell embedding

We evaluated scE^{2}TM against seven leading single-cell methods across 20 scRNA-seq datasets that span multiple species and diverse single-cell profiling technologies, with dataset sizes ranging from a few hundred to over 100,000 cells. scE^{2}TM substantially outperforms all state-of-the-art methods, achieving an overall Adjusted Rand Index (ARI) gain of 11.3% (p-value = 1.2e-2) and Normalized Mutual Information (NMI) gain of 8.7% (p-value = 2.1e-2) over the best-performing competitor, scVI-LD (Fig. 2a; Supplementary Fig. 2). Methods such as SPRUCE and d-scIGM show pronounced performance variability across datasets (Fig. 2a, error bars), indicating sensitivity to dataset-specific characteristics. In contrast, scE^{2}TM demonstrates robustness and consistent performance, with standard deviations of 0.082 and 0.066 for ARI and NMI, respectively. Notably, scE^{2}TM surpasses the cell network-integrated method, scLEGA, by 18.1% (p-value = 5.2e-4) in ARI and 14.1% (p-value = 4.6e-3) in NMI. This advantage stems from scE^{2}TM's external knowledge-guided embedding strategy and CVE framework, which more effectively leverages biological prior knowledge compared to scLEGA's dependence on internally built cell

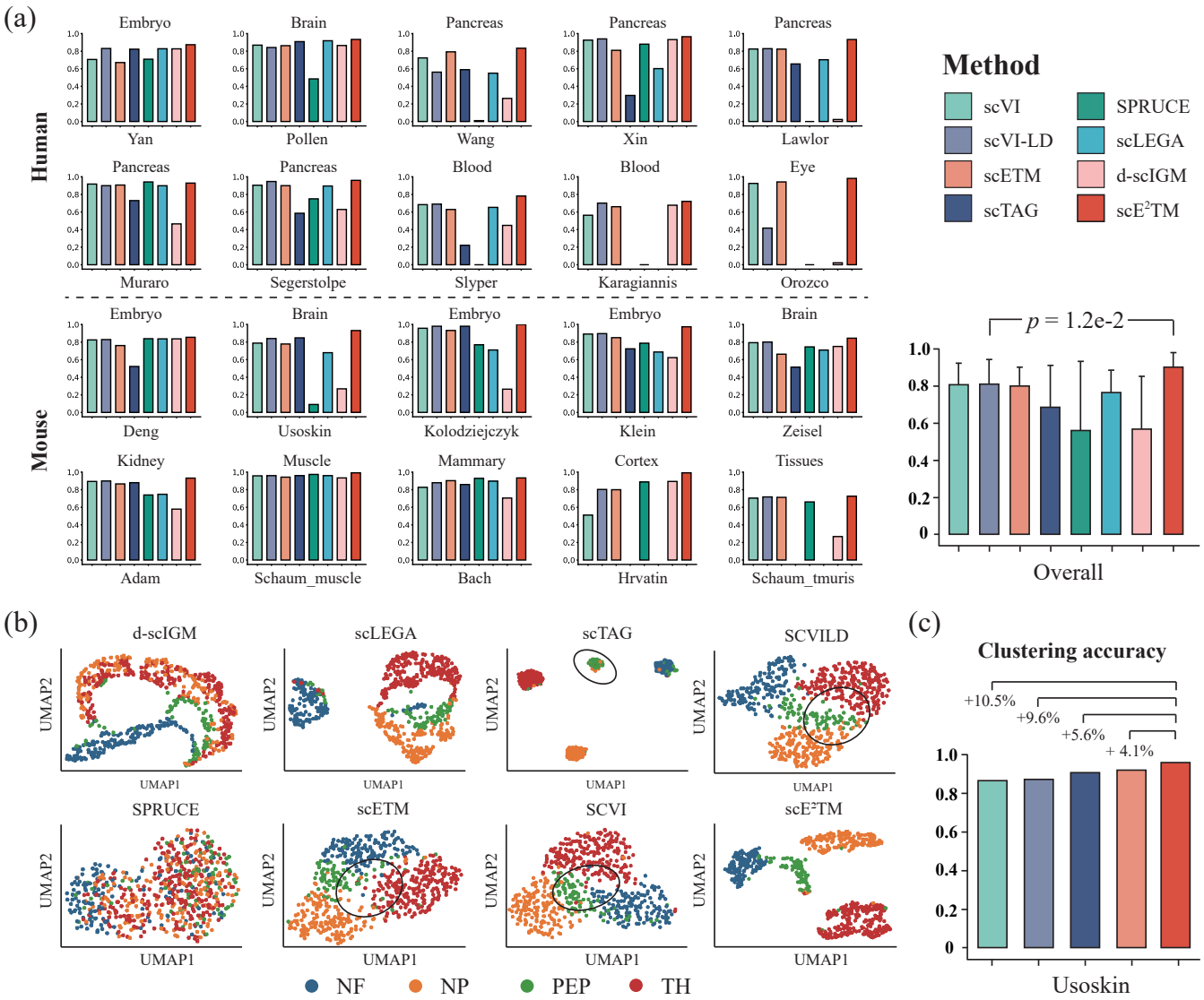

Figure 2: Clustering and visualization. (a) Evaluating clustering performance using ARI. The 20 panels on the left show the ARI values for the different methods on individual datasets, and the right panel shows the average ARI values and standard deviations, with statistical significance tested by pairwise Mann-Whitney U-test. We do not report the results of scTAG and scLEGA on some large scRNA-seq datasets (Karagiannis, Orozco, Hrvatin, and Schaum_tmuris) because they ran out of memory. (b) UMAP visualization. Two-dimensional embeddings of all benchmark models on the Usoskin dataset illustrate that scE^{2}TM better captures biological variation, clearly separating distinct cell populations, including peptidergic nociceptors (PEP), non-peptidergic nociceptors (NP), neurofilament (NF), and tyrosine hydroxylase (TH). (c) Cell-type identification accuracy. Comparison of the identification accuracies for different cell types, with the percentages indicating the degree of accuracy improvement achieved by scE^{2}TM relative to the baseline.

networks. We also assessed the computational efficiency across three datasets of different sizes (Supplementary Table 2). The results indicate that scTAG and scLEGA based on cellular networks achieve fast runtimes but prohibitively high GPU memory usage. Moreover, scVI and scVI-LD show more efficient speed and resource utilization, but both models exhibit poor interpretability.

In contrast, scE^{2}TM achieves optimal clustering and interpretability while maintaining competitive computational efficiency.

We visualized cell embeddings from scE^{2}TM and competing methods on the Usoskin dataset [30], which has well-defined cell types. scE^{2}TM's embeddings effectively capture biological variation across distinct conditions, whereas methods like scETM, scVI, and scVI-LD fail to distinguish overlapping populations (Fig. 2b, black circles). We further quantified per-cell-type clustering accuracy when the predicted cluster number matches the true cell type count. scE^{2}TM achieves the highest mean accuracy, further validating its embedding quality (Fig. 2c). Together, scE^{2}TM delivers robust, high-quality cell embeddings that consistently capture biologically meaningful structures across diverse datasets, outperforming existing state-of-the-art methods.

2.4 Quantitative benchmark of model interpretability

Recent studies [25, 31, 32] demonstrated the inherent risks of bias and cognitive limitations when relying primarily on qualitative methods to assess interpretability, underscoring the need for quantifying the biological interpretability. We employed ten quantitative metrics (Section 4.4) to comprehensively evaluate the interpretability of single-cell embedded topic models: Interpretation Purity (IP), Topic Diversity (TD), Topic Coherence (TC), Topic Quality (TQ), Number of pathways enriched by topic ORA (ORA$_N$), Uniqueness of pathways enriched by topic ORA (ORA$_U$), ORA Quality (ORA$_Q$), Number of pathways enriched by topic GSEA (GSEA$_N$), Uniqueness of pathways enriched by topic GSEA (GSEA$_U$), and GSEA Quality (GSEA$_Q$). scE^{2}TM achieves the best

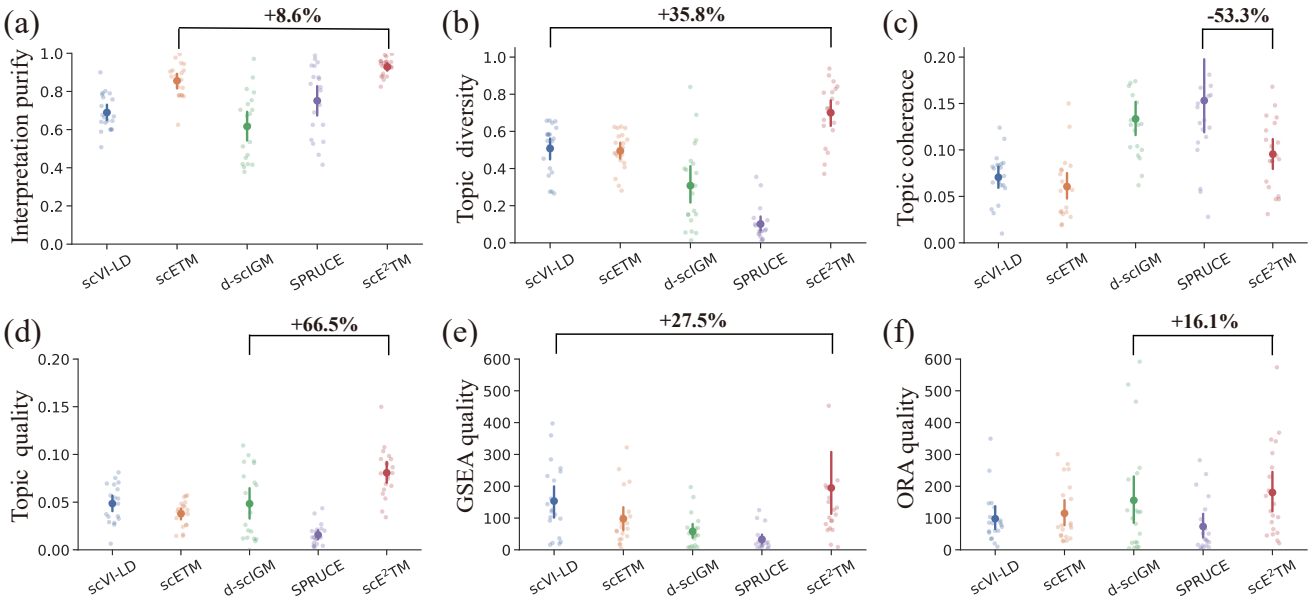

Figure 3: Interpretability metrics comparison. Quantitative assessment of interpretability using six core metrics—Interpretation Purity (IP), Topic Diversity (TD), Topic Coherence (TC), Topic Quality (TQ), GSEA quality (GSEA$_Q$), and ORA quality (ORA$_Q$)—across 20 scRNA-seq datasets. scE^{2}TM consistently achieves the highest interpretability scores, substantially outperforming baseline models.

interpretable performance overall, substantially outperforming the second best method in terms of IP (8.6% performance gain), TQ (66.5% performance gain), $GSEA_Q$ (27.5% performance gain) and ORA_Q (16.1% performance gain) (Fig. 3; Supplementary Fig. 3). scVI-LD and d-scIGM exhibit low interpretation purity (Fig. 3a), even compared to SPRUCE which has the lowest topic quality (Fig. 3d), indicating that their interpretations do not reflect the cell type information well and may not accurately capture real biological signals. d-scIGM and SPRUCE showed low topic

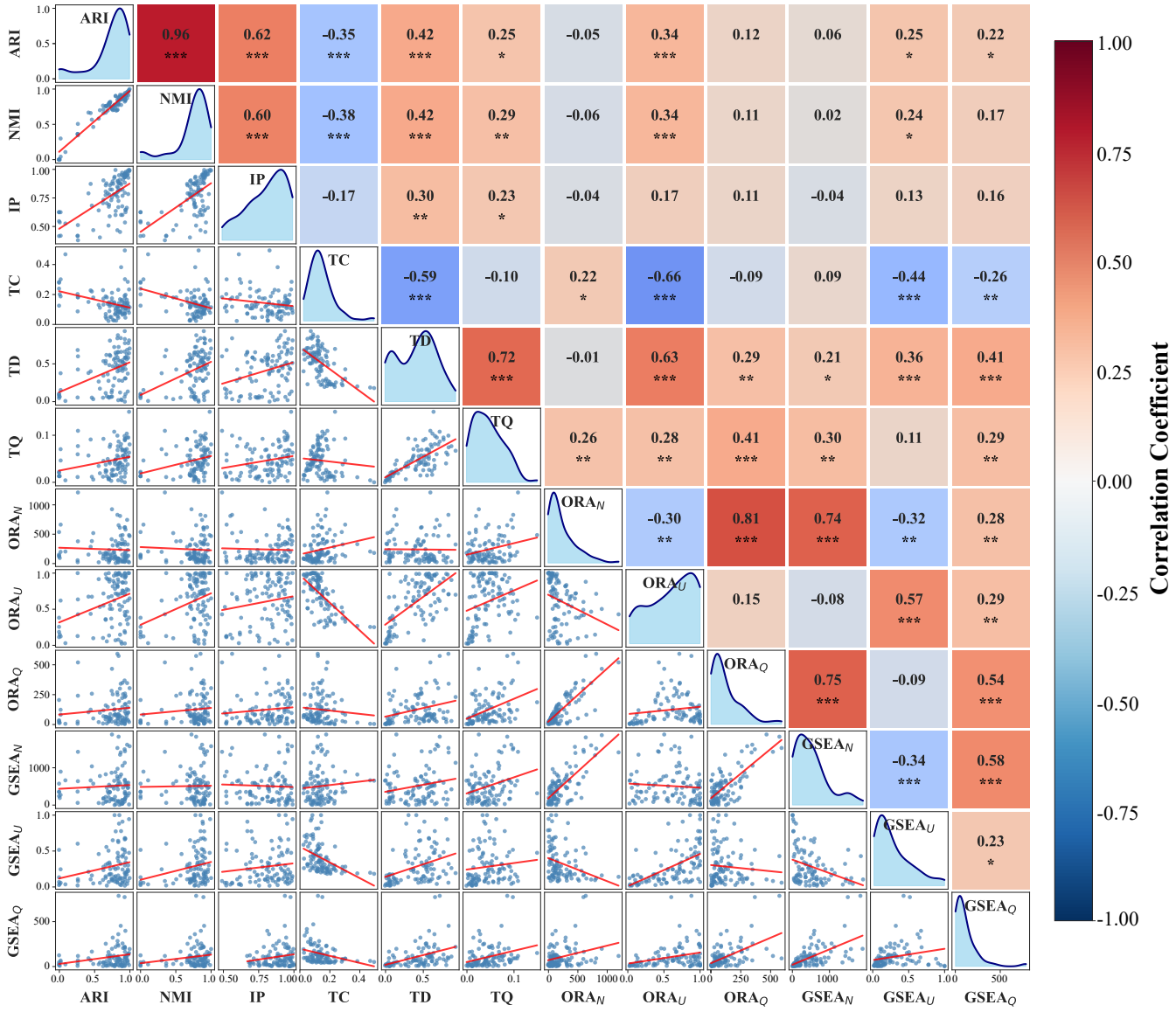

Figure 4: Correlation between interpretability and clustering metrics. Pairwise scatter plots illustrate the relationships between each pair of metrics, while kernel density estimates along the diagonal depict their marginal distributions. Each point corresponds to the metric scores of one model-dataset pair, covering five single-cell embedded topic models applied to twenty datasets. Pearson correlation coefficients (r) and their statistical significance are summarized in the accompanying heatmap, where each cell reports the r value together with its significance level ($p < 0.05$: $*$, $p < 0.01$: $**$, $p < 0.001$: $***$). Together, these visualizations provide a comprehensive assessment of the inter-metric associations.

diversity (Fig. 3b), with values close to 0.1 in some datasets, while their topic coherence scores are relatively high (Fig. 3c). This indicates that these methods focus on only a few coherent topics and cannot comprehensively represent single-cell data. Quantitative analyses of GSEA and ORA output reveal that d-scIGM- and SPRUCE-derived topics are enriched for a large number of repetitive pathways (Supplementary Fig. 3). These results reveal that existing single-cell embedded topic models suffer from interpretation collapse—characterized by low topic diversity, poor cell-type alignment, and repetitive pathway enrichment—which scE^{2}TM effectively mitigates through superior performance across all interpretability metrics.

To understand the relationships among our interpretability metrics and their associations with existing clustering benchmarks, we analyzed Pearson's correlation between all evaluation metrics (Fig. 4). Interpretation purity strongly correlates with clustering metrics (ARI and NMI; $r > 0.6$, p-value $<$ 0.001) . This is expected as both metrics assess cell type coherence. However, topic quality does not correlate with clustering scores. For example, although scVI-LD achieved the second-best topic quality and clustering performance, it ranked second lowest in interpretation purity (Fig. 2a and Fig. 3a,d). Moreover, TQ and topic enrichment (i.e., ORA$_Q$, and GSEA$_Q$) show weak association with the clustering metrics, with all correlations falling below 0.3. This indicates that high clustering performance does not imply biologically meaningful interpretations. Topic Quality is positively correlated with the quality of the topic enrichment analyses (ORA$_Q$ and GSEA$_Q$). Consistent with this observation, ORA and GSEA reveal similar trends. Lastly, we found negative correlation between Topic Coherence and Topic Diversity as well as between the number of pathways (ORA$_N$ / GSEA$_N$) and uniqueness of topic enrichment (ORA$_U$ / GSEA$_U$), which is consistent with previous work [33]. Overall, these quantitative results demonstrate scE^{2}TM's superior interpretability across multiple complementary metrics and reveal a clear discrepancy between clustering performance and biologically meaningful interpretability.

2.5 scE^{2}TM identifies distinct topics linked to pancreatic function

We evaluated scE^{2}TM's biological discovery capabilities through analysis of a human pancreas scRNA-seq dataset [28]. UMAP visualization of gene and topic embeddings shows that scE^{2}TM learns geometrically coherent representations where topics form distinct hubs, surrounded by functionally related genes (Fig. 5a). This clustering pattern, induced by ECR approach, contrasts with methods lacking regularization, which produce diffuse, overlapping gene-topic associations. By visualizing topics 12 and 69 and their associated top-10 genes, we observe a strong co-localization between genes comprising the "`REACTOME_INSULIN_PROCESSING`" and "`VANGURP_PANCREATIC_BETA_CELL`" pathways and their corresponding enriched topics (Fig. 5b).

We investigated the top ten genes for each scE^{2}TM topic and calculated their frequency of occurrence across topics. Our results show the ten most frequent genes, including INS, CEACAM7, and UCA1 (Supplementary Table 3), all experimentally validated or reported in the literature as significantly associated with pancreatic function or related disease mechanisms [34–36]. Pathway enrichment analysis identifies many enriched pathways closely related to pancreatic function for each topic (Benjamin-Hochberg (BH) false discovery rate $<$ 0.05; Supplementary

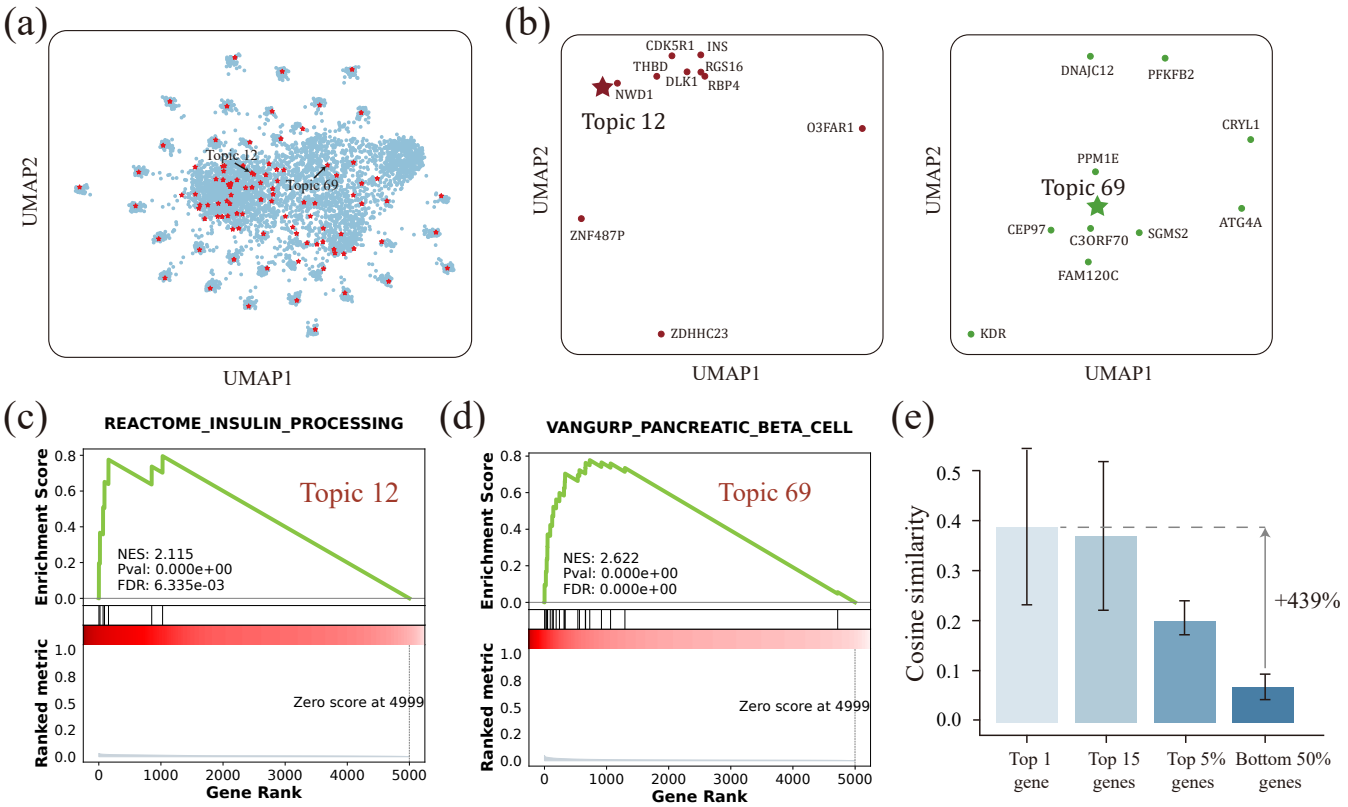

Figure 5: Analysis of scE^{2}TM topic and gene embeddings on human pancreas scRNA-seq data. (a) Global embedding structure. UMAP visualization shows the global distribution of genes (•) and topics (⋆) in the learned embedding space. (b) Local topic-gene structure. Visualization of topics 12 and 69 with their top-10 genes. (c) Gene Set Enrichment Analysis (GSEA). Leading-edge analysis of the "`REACTOME_INSULIN_PROCESSING`" pathways using Topic 12. The running sum enrichment score is calculated by GSEA. Topic 12 is significantly enriched in "`REACTOME_INSULIN_PROCESSING`" pathway (GSEA permutation test BH adjusted q-value < 0.05). (d) Topic 69 is significantly enriched in the "`VANGURP_PANCREATIC_BETA_CELL`" pathway. (e) Topic-gene similarity. Comparison of cosine similarity between each topic and its top-h genes.

Fig. 4). Enrichment of "`REACTOME_INSULIN_PROCESSING`" and "`VANGURP_PANCREATIC_BETA_CELL`" highlights the model's ability to capture key pancreatic functions, including insulin biosynthesis and cell-specific molecular characteristics (Fig. 5c,d). Cosine similarity analysis quantifies the geometric structure of the embedding space: genes with higher topic weights display greater embedding similarity to their corresponding topics (Fig. 5e). The analysis of human pancreas data establishes that scE^{2}TM recovers biologically coherent topic-gene structures in the embedding space and successfully identifies key pancreatic functional genes and pathways, validating its capability to uncover mechanistically relevant biology from single-cell transcriptomes.

2.6 Topic-based perturbation reveals distinct interferon response in cell-state transitions

We applied scE^{2}TM to the PBMC dataset from Kang *et al* [37], which includes single-cell RNA-seq data from lupus patient PBMCs treated with IFN-β, a type I IFN (Section 4.6). scE^{2}TM effectively separates known cell types and treatment conditions while identifying several topics specifically associated with IFN stimulation (Fig. 6a,b). Differential expressed topic analysis identifies topics 59 and 76 (IFN-specific topic) as significantly up-regulated in IFN-stimulated cells versus control cells (BH-adjusted q-value < 0.05; Fig. 6c; Supplementary Fig. 5). Topic 76 exhibits the highest activation levels in FCGR3A+Mono cells. The top genes of topic 76 are significantly enriched for for both FCGR3A+Mono-specific up-regulate genes (Supplementary Fig. 6a) and type I IFN response pathway components (Supplementary Fig. 6b).

We performed a topic-guided perturbation analysis (Section 4.7) to investigate whether perturbing relevant topics (i.e., gene expression programs) could induce intended cellular state changes. We identified 104 Differential Expressed Genes (DEGs) significantly up-regulated in IFN-stimulated versus control cells, which serve as a reference gene set. Using the IFN-specific topics 76 and 59, we extracted cell-topic distributions from control cells and performed *in silico* stimulation by increasing the intensities of these two topics. The decoder computes the expected gene expression of both perturbed and unperturbed control cells, revealing 75 DEGs between these conditions, 26 of which overlap with the reference DEGs (Fig. 6d and Supplementary Fig. 7). Among these, the most strongly upregulated overlapping gene, CH25H, has been identified as an IFN-dependent gene with potent antiviral activity [38]. Permutation testing shows that this overlap is significantly exceeds expected levels from perturbing random topics (Fig. 6e), confirming the specificity of IFN-related topic perturbations.

Perturbation strengths modulates the magnitude of cell-state transitions. As the intensity of IFN-specific topic perturbation increases, a greater proportion of control cells become identified as IFN-stimulated (Fig. 6f). To visualize these transitions at the population level, we focus on CD4T cells, which represent the major cell type in this dataset. We projected the reconstructed representations of control and IFN-stimulated cells into a UMAP space, and then mapped the perturbed control cells into the same space. With increasing IFN-specific topic stimulation, the perturbed control cells progressively converge toward the actual IFN-stimulated cells (Fig. 6g). In contrast, perturbation with a set of non-IFN-specific topics that showed low correlation with IFN and were matched in number to the IFN-specific topics resulted in cells that remained clustered with unperturbed control cells (Fig. 6h). Moreover, perturbing topics 76 and 59 individually did not lead the perturbed control cells to shift noticeably closer to the actual IFN-stimulated cells (Supplementary Fig. 8). These perturbation experiments establish that scE^{2}TM-identified topics capture biologically relevant gene programs that can be leveraged to computationally simulate treatment-induced cell-state transitions.

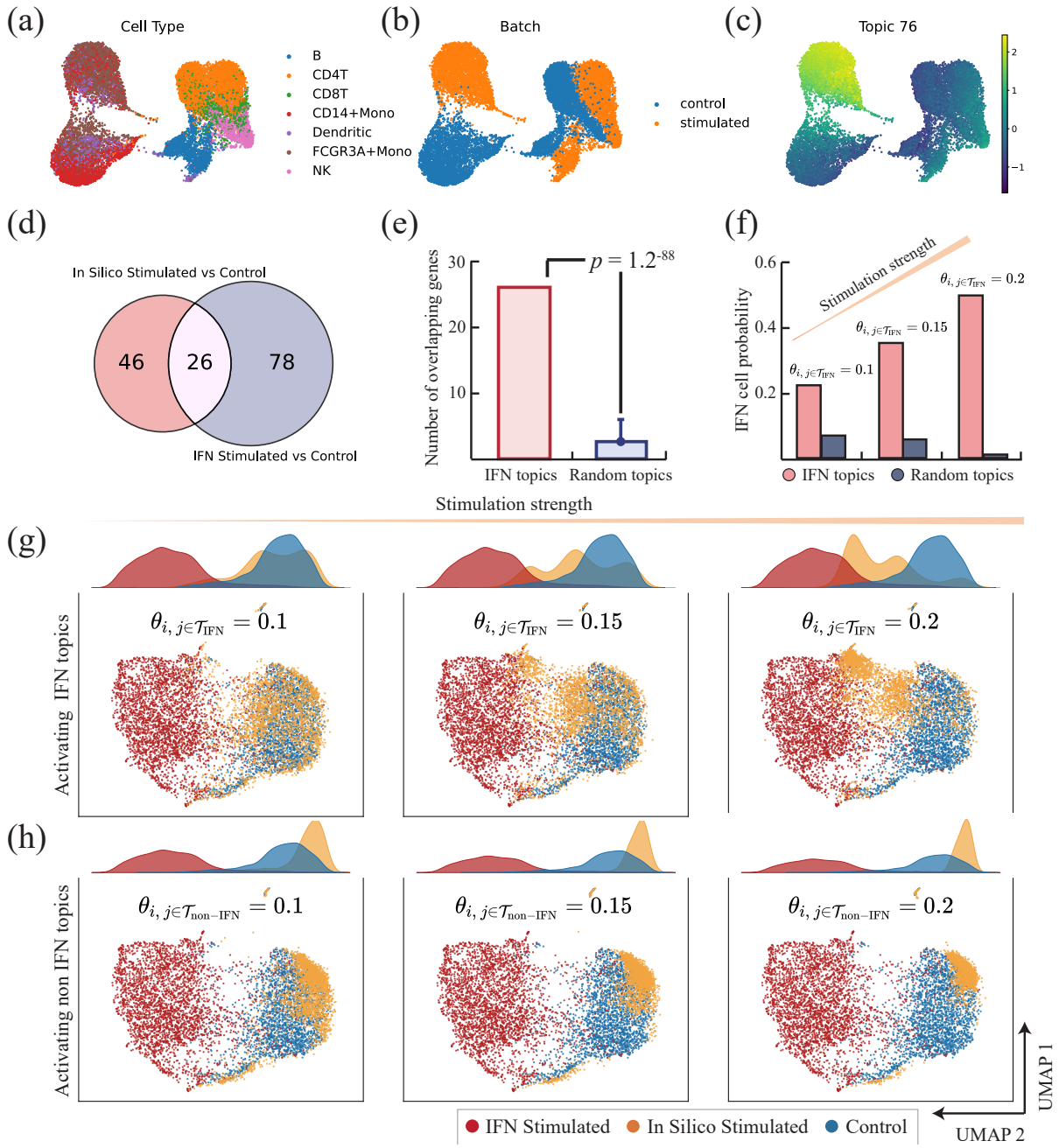

Figure 6: scE^{2}TM predicts the response to IFN treatment. (a-c) Cell embedding and topic activity. UMAP visualization of PBMCs, colored by cell type (left), treatment condition (middle), and topic activity (right). (d) Consistency between simulated topic stimulation and experimental IFN response. The venn diagram showing the overlap between genes up-regulated by in silico stimulation (IFN-specific topic intensity set to 0.2 in control cells) and those up-regulated by IFN-β treatment. (e) Statistical results confirms that the observed overlap is significantly greater than expected from random topic perturbations. (f) Probability of control group cells being classified as stimulated after perturbation with IFN-specific topics versus random topics. The dataset is split into training and test sets, and the classification probabilities are generated by a Support Vector Machine (SVM) classifier on the test set. $\theta_{i,j}$ denotes the topic intensity of cell i on topic j, and $\mathcal{T}_{\text{IFN}}$ denotes the set of IFN-specific topics. (g) Cell representations as a function of perturbation strength for the IFN-specific topics. UMAP for the reconstructed CD4T cells in the control (green), IFN-stimulated (blue), and in silico stimulated (yellow) are shown under weak and strong perturbation strength. The corresponding density plots are displayed to illustrate the proportions of the three cell conditions. (h) Cell representations as a function of perturbation strength for the non-IFN-specific topics.

2.7 Topic-based perturbation shifts melanoma cells towards normal state

We applied scE^{2}TM to a melanoma scRNA-seq dataset [18, 39] to identify gene topics associated with malignancy. The dataset comprised 4,645 malignant, immune, and stromal cells isolated from 19 melanoma patients, including ten metastases to lymphoid tissues, eight distant-site metastases, and one primary acral melanoma. Differential topic analysis reveals 33 malignant-specific topics in comparison with normal cell states (Supplementary Fig. 9). These topics exhibit high selectivity for cell-type marker genes and strong discriminative power for distinguishing cell types (Fig. 7). We performed topic-based perturbation by reducing the topic scores for malignant-specific topics and computed the expected gene expression for perturbed and unperturbed malignant cells using the decoder. DEG analysis identifies 428 genes that are significantly upregulated in reconstructed unperturbed malignant cells compared with perturbed malignant cells (Supplementary Fig. 10). As reference, we also identified DEGs using the observed malignant and normal cells. The two differential expressed gene sets show significant overlaps (Fig. 8a).

To assess whether the perturbed malignant cells shift towards the normal state, we trained

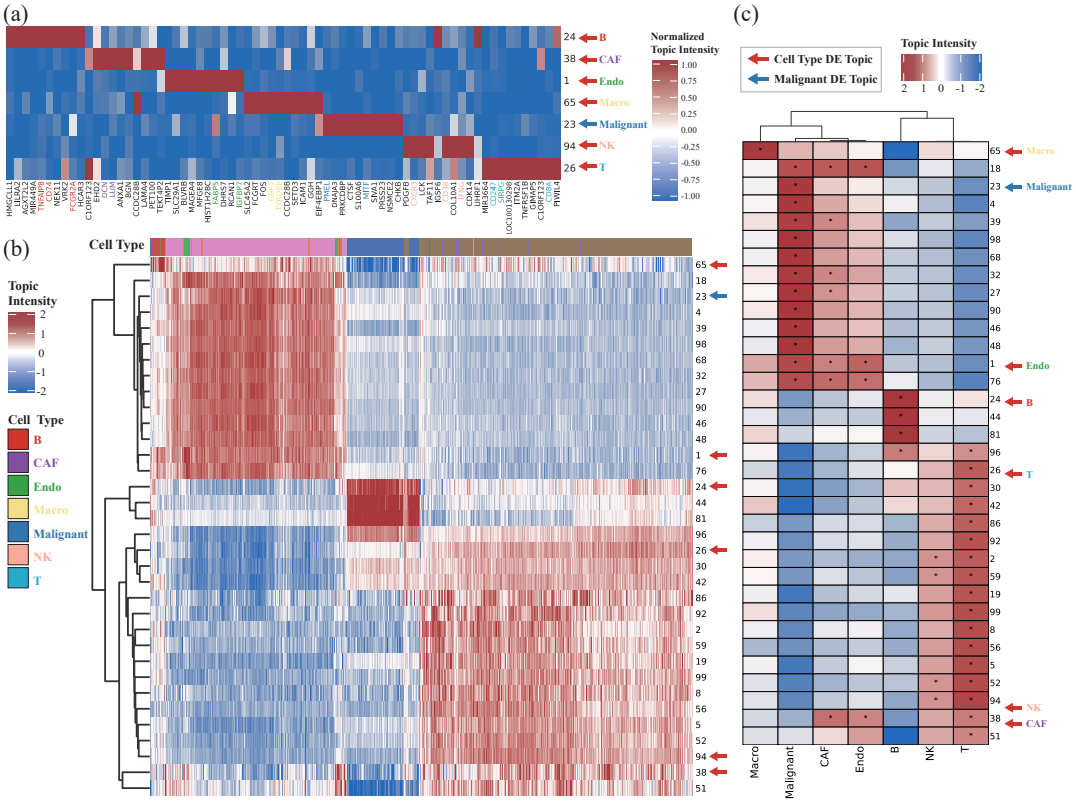

Figure 7: Topic activity and cell-state associations revealed by scE^{2}TM in melanoma. (b) Gene-topic heatmap. For visualization purposes, we normalized topic values by dividing them by the maximum absolute value within the same topic. (c) Differential expression analysis by topic. Colors indicate mean differences between cell groups with a specific label (cell type or state) and those without that label. Asterisks denote BH-adjusted q-values < 0.01. The topic strengths displayed here represent Gaussian averages prior to softmax application. Only selected topics with absolute values > 50 relative to the sum of all cells are shown.

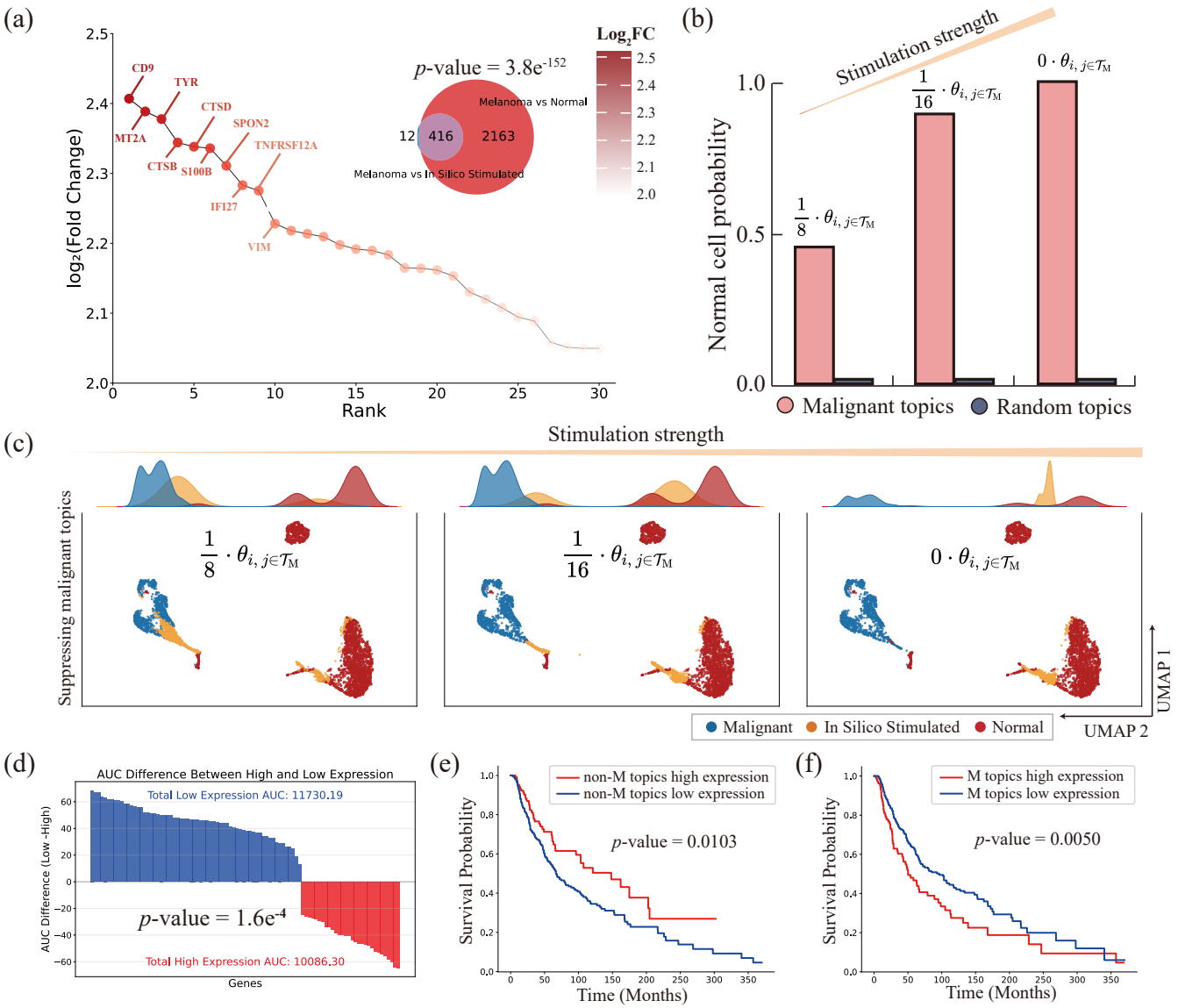

Figure 8: Topic-driven perturbation responses and prognostic associations in melanoma. (a) Consistency between simulated topic stimulation and malignant cell response. The venn diagram illustrates the overlap between genes up-regulated in malignant cells relative to in silico stimulation (malignant-specific topic intensity in malignant cells set to 0) and genes up-regulated relative to normal cells. The p-value indicates significance relative to random topic perturbation. The hockey stick plot displays the overlapping genes sorted by degree of upregulation. (b) Classification Shift. The probability of malignant cells being classified as malignant after perturbing malignant-specific topics versus random topics. (c) Perturbed cell-state visualization. UMAP projections and density maps of reconstructed cells under different stimulus intensities. $\mathcal{T}_{\mathrm{M}}$ denotes the set of malignant-specific topics. (d) Gene survival analysis. Gene-level survival analysis of the 416 overlapping genes shown in (a) using an unseen TCGA melanoma cohort. Each bar represents the AUC difference between patient groups stratified by high versus low gene expression levels. (e) Topic survival analysis. Survival analysis for non-malignant-specific (non-M) topics in the TCGA melanoma cohort. (f) Survival analysis for malignant-specific (M) topics.

an SVM classifier to predict cell states (malignant or normal) (Section 4.9). As perturbation intensity increases by progressive suppression of malignant-specific topics, a greater proportion of malignant cells become classified as normal (Fig. 8b), indicating that these topics are critical for maintaining the malignant phenotype and might represent key drivers of malignant transformation. To visualize these transitions, we projected perturbed malignant cells into UMAP space. With increasing perturbation of the malignant-specific topics, the perturbed cells gradually converge toward the normal cell cluster (Fig. 8c), whereas perturbing non-malignant-specific topics topics results in cells remaining clustered among the original malignant cells (Supplementary Fig. 11). Furthermore, we found that perturbing the top-10 malignant-specific topics induced detectable shifts in cellular states, and that these shifts became more pronounced as the number of perturbed malignant-specific topics increased (Supplementary Fig. 12). This observation indicates that the malignant-specific topics identified by scE^{2}TM represent key transcriptional axes of tumor progression, and perturbing them reverses malignant cells toward normal phenotypes.

To assess the clinical relevance of our analysis, we further analyzed melanoma patient data from The Cancer Genome Atlas (TCGA) [40]. The up-regulated genes identified in the melanoma dataset show a significant negative correlation with survival rates in TCGA melanoma cohort patients (Fig. 8d; p-value = 1.6e-4), confirming that perturbation of the target indeed impacts key signaling pathways in melanoma biology. We then applied the melanoma-trained scE^{2}TM model to the TCGA melanoma gene expression data and examined the associations between the previously identified malignant-specific topics and patient outcomes. The expression of malignant-specific topics shows significant associations with survival (Fig. 8e,f), demonstrating that scE^{2}TM-identified topics capture clinically relevant gene programs that generalize across independent patient cohorts. Together, scE^{2}TM reveals malignancy-specific topics and can be robustly extended to independent TCGA melanoma studies. These findings suggest that interpretable topic characterization holds promise as a bridge linking single-cell transcriptomics to cancer clinical prognosis prediction.

3 Discussion

Identifying marker genes for previously uncharacteristic cell types and disease-specific transcriptional states remains a key challenge in scRNA-seq analysis. Topic models present an effective framework that can identify interpretable gene-topic distributions and characterize cell by these topics. However, existing topic models suffer from the lack of external knowledge and interpretation collapse, where many topics convey redundant information. To this end, we present scE^{2}TM, a novel single-cell embedding topic model through two technical contributions.

First, we integrate external knowledge from a pretrained foundation model scGPT through a novel cross-view distillation strategy. This approach effectively coordinates cellular neighborhood information across different representations, encourages consistent topic assignment between cells and their corresponding neighbors in the embedding space, thereby integrating rich contextual information while maintaining computational efficiency. Our results demonstrate that this

strategy achieves both superior clustering performance and enhanced scalability, enabling direct knowledge transfer from pretrained foundation models without requiring joint clustering or label transfer operations.

Second, to address interpretation collapse, we adopt embedding cluster regularization (ECR), which treats topics as clustering centroids and genes as samples. Through optimal transport, the ECR ensures that each topic serves as the anchors of a distinct gene cluster. Quantitative experiments show scE^{2}TM's substantial improvement over benchmark models, with increases of 8.6% in interpretation purity, 66.5% in topic quality, 27.5% in GSEA quality, and 16.1% in ORA quality. Qualitative experiments further confirm that scE^{2}TM topics capture cell type-specific and functionally relevant biological signals. Our correlation analysis between clustering and interpretability metrics reveals that high clustering performance does not guarantee biologically meaningful interpretations, underscoring the importance of dedicated interpretability assessment.

Beyond benchmarking interpretability, our topic-based perturbation experiments in PBMC and melanoma datasets demonstrate scE^{2}TM's potential to model cell-state transitions *in silico*. By perturbing topic activations associated with IFN response or malignant programs, scE^{2}TM accurately reconstructed trajectories from stimulated to resting immune states and from tumor to normal-like melanoma cells, respectively. These findings highlight that topic perturbations capture causal relationships between gene programs and cellular phenotypes. In practical terms, this framework can support *in silico* hypothesis testing—predicting how cells might respond to immune stimulation, drug treatment, or oncogenic suppression—thus offering a powerful tool for precision medicine and therapeutic target discovery in real-world clinical settings.

Future extensions of our study will focus on three directions. First, we will collect tissue- or cell type-specific pathway datasets to enable a more accurate evaluation of the model's interpretability. Although scE^{2}TM identifies topics enriched for diverse and coherent biological pathways, several interpretability metrics (e.g., ORA$_Q$ and GSEA$_Q$) are evaluated against all known pathways and therefore cannot determine how well these pathways reflect the underlying cellular characteristics. Collecting biological pathway datasets with tissue- or cell type–specific information is a promising direction for improving our interpretability evaluation benchmarks. Nevertheless, we anticipate that diversity and consistency will remain valid assessment criteria. Secondly, existing single-cell embedding topic models generate global topic-gene dependencies for entire datasets, potentially overlooking cell-specific characteristics. Developing hierarchical topic structures that generate topics at multiple granularities would enable more refined interpretations focused on individual cell features [41]. Finally, with the increasing prevalence of multimodal single-cell assays, extending scE^{2}TM and our interpretability metrics to multimodal data would enable more comprehensive biological signal capture across modalities [42–44].

In summary, scE^{2}TM represents an external knowledge-guided single-cell embedded topic model with robust clustering performance and enhanced interpretability. We identified critical limitations in the current evaluation framework and introduced a suite of quantitative metrics to address them. We envision that these contributions will advance our understanding of cellular biology and establish foundations for future developments in interpretable single-cell analysis.

4 Methods

4.1 scRNA-seq datasets

Twenty scRNA-seq datasets were used to conduct experiments, including human and mouse data. For all single-cell benchmarking datasets, we performed the same preprocessing steps using Scanpy package (v.1.9.8). Specifically, we used "`scanpy.pp.log1p()`" to log-transform the counts of each cell, and then use "`scanpy.pp.highly_variable_genes()`" to select the top-5000 highly variable genes. Details of all single-cell data are provided in Supplementary Table 4.

In addition, we used two scRNA-seq datasets for the topic-based perturbation experiments. The first dataset by Kang *et al* [37] consists of two groups of PBMCs, one control and one stimulated with IFN-β. We applied the same preprocessing steps as in previous work [45], ultimately yielding a dataset containing 18,868 cells. Raw data are available at GSE96583. The second dataset were obtained from Tirosh *et al* [39] (GSE72056) for melanoma study. We utilized the dataset processed by Chen *et al* [18], which removed all samples of indeterminate type and resulted in a final dataset of 4,092 cells comprising seven distinct cell types.

4.2 Clustering benchmark metrics

For evaluating cell clustering, we used standard metrics, namely the Normalized Mutual Information (NMI) and the Adjusted Rand Index (ARI). All embedding maps were visualized using UMAP. We clustered the cells using Louvain's algorithm. Furthermore, to better assess the clustering performance of cell embeddings, we experimented with multiple resolution values and reported the results with the highest ARI of each method, following previous studies [18, 46]. Moreover, we assessed model interpretability using the proposed metrics: IP, TC, TD, TQ, ORA$_N$, ORA$_U$, ORA$_Q$, GSEA$_N$, GSEA$_U$, and GSEA$_Q$.

4.3 Baselines methods

For the single-cell embedded topic model, we used scVI-LD [47], scETM [6], SPRUCE [17], and d-scIGM [18] as baselines. The scVI-LD method provided interpretability by determining the relationship between cell representative coordinates and gene weights through matrix factorization, where the cell representative coordinates are considered as topics. Notably, d-scIGM utilized hierarchical prior knowledge to bootstrap the model, which may lead to data leakage in the interpretability evaluation. To prevent that, we used the version without prior knowledge described in the original article.

All experiments on the benchmark models were performed using publicly available code. We used the optimal hyperparameters reported in their original papers. For the models that utilize information about the topology of the cellular network, we used scTAG [48] and scLEGA [49] as baselines. In addition, scVI [50] was also used as a baseline model, which is a classical single-cell deep embedding method that has demonstrated excellent overall performance in

previous comprehensive evaluations [46]. More detailed descriptions of the relevant methods can be found in Supplementary Note 1.

4.4 Quantitative assessment of interpretability

4.4.1 Topic coherence

Topic Coherence (TC) quantifies the semantic consistency of a topic by measuring the statistical co-occurrence patterns of its characteristic terms. For a given topic, TC evaluates whether its top-associated terms frequently co-appear in relevant contexts within an external corpus, indicating a coherent semantic representation. This metric strongly correlates with human interpretability judgments in traditional topic modeling [51]. We adapt TC to single-cell biology by leveraging biological pathways as semantic units, using the GASE database [52] as the external knowledge base. Specifically, we suggest that genes assigned to the same topic collaboratively participate in common biological pathways. Formally, TC is defined as

$$\mathrm{TC} = \frac{1}{K}\sum_{k=1}^{K}\frac{2}{h(h-1)}\sum_{i=1}^{h}\sum_{j=i+1}^{h} F\left(\mathbf{g}_i^{(k)}, \mathbf{g}_j^{(k)}\right), \tag{1}$$

where $\left\{\mathbf{g}_1^{(k)}, \ldots, \mathbf{g}_h^{(k)}\right\}$ denotes the top-h most likely genes in the k-th topic. And the correlation of word pairs is estimated as follows:

$$F\left(\mathbf{g}_i, \mathbf{g}_j\right) = \frac{\log \frac{P(\mathbf{g}_i, \mathbf{g}_j)}{P(\mathbf{g}_i)P(\mathbf{g}_j)}}{-\log\left(P\left(\mathbf{g}_i, \mathbf{g}_j\right)\right)}, \tag{2}$$

where $P\left(\mathbf{g}_i, \mathbf{g}_j\right)$ is the probability that genes $\mathbf{g}_i$ and $\mathbf{g}_j$ co-occur in the same biological process, and $P\left(\mathbf{g}_i\right)$ is the marginal probability of gene $\mathbf{g}_i$. TC ranges continuously on [0, 1], with higher values indicating superior coherence.

4.4.2 Topic diversity

Topic Diversity (TD) quantifies the distinctness of biological mechanisms captured by different topics. Existing single-cell embedded topic models often suffer from interpretation collapse, i.e., the model mines many overlapping topics. Therefore, topic diversity should be one of the important metrics for assessing interpretability. Formally, TD is defined as follows:

$$\mathrm{TD} = \frac{|\mathrm{Unique}(M)|}{|M|}, \tag{3}$$

where $M = \bigcup_{k=1}^{K}\left\{\mathbf{g}_1^{(k)}, \ldots, \mathbf{g}_h^{(k)}\right\}$ denotes the gene pool containing the top-h genes corresponding to all topics. $|M|$ and $|\mathrm{Unique}(M)|$ denote the number of all genes and the number of unique genes, respectively. TD ranges continuously over [0, 1], where $\mathrm{TD} \to 0$ indicates high redundancy and $\mathrm{TD} \to 1$ reflects maximal diversity.

4.4.3 Topic quality

There is a mutual trade-off between topic coherence and topic diversity in the text mining domain [33], and we suggest that this phenomenon also exists in the single-cell domain. Since some important genes correspond to multiple gene programs, these genes may appear in different topics at the same time, which would result in a higher TC score but a slightly reduced TD score. In addition, specific gene programs may be captured repeatedly by topics, which would further exacerbate the above problem. Therefore, we define the product of topic diversity and topic coherence as the overall Topic Quality (TQ) following Dieng *et al* [33], i.e.,

$$\mathrm{TQ} = \mathrm{TC} \times \mathrm{TD}, \tag{4}$$

which ranges from 0 to 1, with higher values indicating superior balance between coherent biological pathways and diverse functional coverage.

4.4.4 Interpretation purity

Interpretation Purity (IP) captures the intuition that cells assigned to the same topic should share common biological characteristics. Specifically, it tests the hypothesis that dominant topic assignments should align with known cell type annotations at a coarse-grained level, reflecting shared functional states. Adapted from the cluster purity metric, IP quantifies this alignment as:

$$\mathrm{IP}(\Omega, \mathbf{Y}) := \frac{1}{n} \sum_{k=1}^{K} \max_{j} \left| \Omega_k \cap \mathbf{Y}_j \right|, \tag{5}$$

where $\Omega = \{\Omega_k \mid k = 1, \ldots, K\}$ denotes the topic assignments with $\Omega_k = \{i \mid \arg\max_j \theta_{ij} = k,\ i = 1, \ldots, n\}$ representing cells assigned to topic k. Besides, $\mathbf{Y} = \{\mathbf{Y}_j \mid j = 1, \ldots, J\}$ represents J ground truth cell types with $\mathbf{Y}_j$ containing cells of type j. And n is the total number of cells. The metric ranges within $[0, 1]$, where higher values indicate stronger agreement between topics and biological cell states, reflecting more meaningful cellular interpretations.

4.4.5 Quality of topic over-representation analysis

Complementing the cell-level assessment of interpretation purity, which measures the alignment of topics with biological cell states, we now evaluate the functional relevance of topics at the gene and pathway level. To validate the biological mechanisms captured by the mined topics, we perform Over-Representation Analysis (ORA) that statistically evaluates whether top topic-associated genes cluster significantly in known biological pathways. This quantifies how effectively topics capture interpretable biological mechanisms. Specifically, we propose three metrics that assess different dimensions of this functional enrichment quality: (i) number of pathways enriched by topic ORA (ORA_N), which measures the breadth of distinct biological functions captured across all topics; (ii) uniqueness of pathways enriched by topic ORA (ORA_U), which quantifies the avoidance of pathway repetition across topics; and (iii) quality of topic ORA (ORA_Q) balances

the assessment in both coverage and uniqueness. The formal definitions of these metrics are as follows:

$$\mathrm{ORA}_N = |\mathrm{Unique}(P_O)|, \mathrm{ORA}_U = \frac{|\mathrm{Unique}(P_O)|}{|P_O|}, \mathrm{ORA}_Q = \mathrm{ORA}_N \times \mathrm{ORA}_U, \tag{6}$$

where P_O contains the topic enrichment pathways obtained via ORA. $|P_O|$ and $|\mathrm{Unique}(P_O)|$ denote the number of all pathways and the number of unique pathways, respectively. Similar to TQ, ORA_Q is defined as the product of ORA_N and ORA_U.

4.4.6 Quality of topic gene set enrichment analysis

While ORA evaluates pathway enrichment using thresholded top genes, this approach requires arbitrary selection of gene cutoffs [6, 18]. To overcome this limitation, we conduct GSEA, a threshold-free method that evaluates enrichment across the entire ranked list of genes sorted by their association strength with each topic. GSEA computes a running enrichment score that avoids discrete gene selection, providing more robust pathway associations (method details in Supplementary Note 2). Following our ORA evaluation framework, we adapt three analogous metrics (GSEA_N, GSEA_U, GSEA_Q) to quantify biological interpretability:

$$\mathrm{GSEA}_N = |\mathrm{Unique}(P_G)|, \mathrm{GSEA}_U = \frac{|\mathrm{Unique}(P_G)|}{|P_G|}, \mathrm{GSEA}_Q = \mathrm{GSEA}_N \times \mathrm{GSEA}_U, \tag{7}$$

where P_G contains the topic enrichment pathways obtained via GSEA.

4.5 scE^{2}M method details

4.5.1 Integrating external embedding via cross-view knowledge distillation

scE^{2}TM is a Variational AutoEncode (VAE) framework. The topic-based encoder maps the input sample into the latent Gaussian distribution. The reparameterized sample is then transformed via softmax function to represent the cell-topic latent variable $\boldsymbol{\theta}$. To integrate internal embedding learned directly from the scRNA-seq training data with the external knowledge provided by the single-cell foundation model, we propose a cross-view distillation strategy. The key idea is that the same cell from different perspectives should be mapped to the same topic, and that consistent topic assignment between each cell and the corresponding neighboring cell in another perspective is encouraged by integrating the cell's neighborhood information. Specifically, let $\mathbf{x}$ and $\mathbf{v}$ be the internal and external cell embedding representations, respectively. We introduce a cell-topic head $f(\cdot) : \mathbf{x} \to \boldsymbol{\theta} \in \mathbb{R}^K$ and a cell-cluster head $g(\cdot) : \mathbf{v} \to \boldsymbol{\varphi} \in \mathbb{R}^K$, where K is the number of target

topics or clusters*. For $i \in \{1, \ldots, n\}$ cells, we denote the soft topic/cluster assignment of cell i as

$$\boldsymbol{\mu}, \log \boldsymbol{\sigma} = f(\mathbf{x}_i) = \mathrm{MLP}\left(\mathbf{x}_i; \mathbf{W}_f\right) \tag{8}$$

$$\mathbf{z}_i \sim \boldsymbol{\mu} + \boldsymbol{\sigma}\mathcal{N}(0, \mathbf{I}) \tag{9}$$

$$\boldsymbol{\theta}_i = \mathrm{Softmax}\left(\mathbf{z}_i\right), \tag{10}$$

$$\boldsymbol{\varphi}_i = g\left(\mathbf{v}_i\right) = \mathrm{Softmax}\left(\mathrm{MLP}\left(\mathbf{v}_i; \mathbf{W}_g\right)\right) \tag{11}$$

where the softmax function $\exp(z_k)/\sum_k \exp(z_k)$ is applied after the MLP function to produce the probabilistic topic and cluster assignments from the two views, respectively.

To encourage topic consistency between internal and external perspectives of the same cell, we define a cross-view assignment consistency loss:

$$\mathcal{L}_{\mathrm{CON}} = -\log \sum_{i=1}^{n} \boldsymbol{\theta}_i^{\top} \boldsymbol{\varphi}_i, \tag{12}$$

which is minimized for similar embedding between $\boldsymbol{\theta}_i$ and $\boldsymbol{\varphi}_i$.

To utilize the cell neighborhood information of different views, we first obtain the nearest neighbors set of $\mathbf{x}_i$ and $\mathbf{v}_i$, denoted as $\mathcal{N}\left(\mathbf{x}_i\right)$ and $\mathcal{N}\left(\mathbf{v}_i\right)$, respectively. For randomly sampled neighbors $\mathbf{x}_i^{\mathcal{N}} \in \mathcal{N}\left(\mathbf{x}_i\right)$ and $\mathbf{v}_i^{\mathcal{N}} \in \mathcal{N}\left(\mathbf{v}_i\right)$, their corresponding topic assignments are defined as

$$\boldsymbol{\theta}_i^{\mathcal{N}} = f\left(\mathbf{x}_i^{\mathcal{N}}\right), \quad \boldsymbol{\varphi}_i^{\mathcal{N}} = g\left(\mathbf{v}_i^{\mathcal{N}}\right). \tag{13}$$

Let $\boldsymbol{\theta}_{ik}$, $\boldsymbol{\theta}_{ik}^{\mathcal{N}}$, $\boldsymbol{\varphi}_{ik}$, $\boldsymbol{\varphi}_{ik}^{\mathcal{N}}$ be the assignment of the k-th topic/cluster, the loss terms of the cross-view neighborhood consistency are defined as follows:

$$\mathcal{L}_{\mathrm{NEI}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (L_{ik}^{\mathbf{x}\to\mathbf{v}} + L_{ik}^{\mathbf{v}\to\mathbf{x}}), \tag{14}$$

$$L_{ik}^{\mathbf{x}\to\mathbf{v}} = -\log \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{\varphi}_{ik}, \boldsymbol{\theta}_{ik}^{\mathcal{N}}\right)/2\right)}{\sum_j \exp\left(\mathrm{sim}\left(\boldsymbol{\varphi}_{ik}, \boldsymbol{\theta}_{ij}^{\mathcal{N}}\right)/2\right) + \sum_{j\neq k} \exp\left(\mathrm{sim}\left(\boldsymbol{\varphi}_{ik}, \boldsymbol{\varphi}_{ij}\right)/2\right)}, \tag{15}$$

$$L_{ik}^{\mathbf{v}\to\mathbf{x}} = -\log \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{\theta}_{ik}, \boldsymbol{\varphi}_{ik}^{\mathcal{N}}\right)/2\right)}{\sum_j \exp\left(\mathrm{sim}\left(\boldsymbol{\theta}_{ik}, \boldsymbol{\varphi}_{ij}^{\mathcal{N}}\right)/2\right) + \sum_{j\neq k} \exp\left(\mathrm{sim}\left(\boldsymbol{\theta}_{ik}, \boldsymbol{\theta}_{ij}\right)/2\right)} \tag{16}$$

Notably, $\mathcal{L}_{\mathrm{NEI}}$ encourages consistent topic assignment between a cell and its corresponding neighbor in another view, where $L_{ik}^{\mathbf{x}\to\mathbf{v}}$ aligns the internal feature of the cell to the neighborhood of its external features, and $L_{ik}^{\mathbf{v}\to\mathbf{x}}$ vice versa. In addition, the second terms of the denominator in the definitions of $L_{ik}^{\mathbf{x}\to\mathbf{v}}$ and $L_{ik}^{\mathbf{v}\to\mathbf{x}}$ can minimize inter-topic similarity, resulting in more discriminative topics. As a classical neighbor search algorithm, the k-nearest neighbor algorithm is susceptible to "`dimensionality catastrophe`" and pivot effect, so we choose mutual kNN to find the set of nearest neighbors of cells [53]. In practice, we set the number of nearest neighbors $|\mathcal{N}\left(\mathbf{x}_i\right)| = |\mathcal{N}\left(\mathbf{v}_i\right)|$ = 15 for $i \in \{1, \ldots, n\}$ for all datasets.

*Clusters and topics are used to mainly distinguish the internal and external embedding, respectively.

To prevent all samples from collapsing into a few topics, we introduce an entropy regularization term to stabilize the training process, which is defined as follows:

$$\mathcal{L}_{\mathrm{REG}} = -\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\boldsymbol{\theta}_{ik}\log\boldsymbol{\theta}_{ik} + \boldsymbol{\varphi}_{ik}\log\boldsymbol{\varphi}_{ik}\right). \tag{17}$$

Finally, we derive the overall loss function of our cross-view distillation strategy as follows:

$$\mathcal{L}_{\mathrm{CVE}} = \mathcal{L}_{\mathrm{CON}} + \mathcal{L}_{\mathrm{NEI}} - \alpha\cdot\mathcal{L}_{\mathrm{REG}}. \tag{18}$$

4.5.2 Embedding clustering regularization (ECR)

To mitigate the potential interpretation collapse problem, we extend the ECR approach of Wu *et al* [23] to the single-cell domain to model the relationship between topics and genes. In our context, each topic serves as a centroid of a distinct gene cluster. We adopt ECR to model the soft assignment of genes to topics using a transport scheme defined specifically for the optimal transport problem. Specifically, K topic embeddings serve as cluster centroids and V gene embeddings represent the data points. Let n_k denote the number of genes assigned to topic embedding $\mathbf{t}_k$, and $s_k = n_k/V$ represent the corresponding cluster size proportion. The vector $\mathbf{s} = (s_1,\ldots,s_K)^\top \in \Delta_K$ summarizes all cluster size proportions. We initially assume uniform distribution over teh cluster size $n_k = V/K$, $\mathrm{s} = (1/K,\ldots,1/K)^{\mathrm{T}}$, which is shown to avoid trivial solutions with empty clusters [23,54]. For the learning objective, we define two discrete measures of gene embedding $\mathbf{g}_m$ and topic embedding $\mathbf{t}_k$: $\gamma = \sum_{m=1}^{V}\frac{1}{V}\delta_{\mathbf{g}_m}$ and $\phi = \sum_{k=1}^{K} s_k\delta_{\mathbf{t}_k}$, where δ_z denotes the Dirac unit mass on z. ECR formulates the entropic regularized optimal transport between γ and ϕ as

$$\begin{aligned}\underset{\boldsymbol{\pi}\in\mathbb{R}_+^{V\times K}}{\arg\min}\mathcal{L}_{\mathrm{OT}_\varepsilon}(\gamma,\phi) &= \sum_{m=1}^{V}\sum_{k=1}^{K}\mathrm{D}\left(\mathbf{g}_m,\mathbf{t}_k\right)\pi_{mk} + \sum_{m=1}^{V}\sum_{k=1}^{K}\varepsilon\pi_{mk}\left(\log\left(\pi_{mk}\right)-1\right)\\ &\text{s.t.}\quad \boldsymbol{\pi}\mathbf{1}_K = \frac{1}{V}\mathbf{1}_V \quad\text{and}\quad \boldsymbol{\pi}^\top\mathbf{1}_V = \frac{1}{K}\mathbf{1}_K\end{aligned} \tag{19}$$

where the first term is the original optimal transport problem and the second term with the hyperparameter ε is entropy regularization. Eq. (19) is to find the optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$ that minimizes the total cost of transporting weight from gene embeddings to topic embeddings. We measure the transport cost between gene $\mathbf{g}_m$ and topic $\mathbf{t}_k$ by Euclidean distance: $\mathrm{D}\left(\mathbf{g}_m,\mathbf{t}_k\right) = C_{mk} = \|\mathbf{g}_m - \mathbf{t}_k\|^2$, and the transport cost matrix is denoted as $\mathbf{C}\in\mathbb{R}^{V\times K}$. The two constraints in Eq. (19) restrict the weight of each gene embedding $\mathbf{g}_m$ as $\frac{1}{V}$ and each topic embedding $\mathbf{t}_k$ as $\frac{1}{K}$, where $\mathbf{1}_K$ ($\mathbf{1}_V$) is a $K(V)$ dimensional column vector of ones. Besides, π_{mk} denotes the transport weight from $\mathbf{g}_m$ to $\mathbf{t}_k$ and $\boldsymbol{\pi}\in\mathbb{R}_+^{V\times K}$ is the transport plan that includes the transport weight of each gene embedding to fulfill the weight of each topic embedding. To meet these constraints, ECR models clustering soft assignments with the optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$, i.e., the soft-assignment of

$\mathbf{g}_m$ to $\mathbf{t}_k$ is the transport weight between them, $\pi^*_{\varepsilon,mk}$. The formula is defined as follows:

$$\mathcal{L}_{\text{ECR}} = \sum_{m=1}^{V} \sum_{k=1}^{K} \|\mathbf{g}_m - \mathbf{t}_k\|^2 \pi^*_{\varepsilon,jk} \tag{20}$$

where

$$\pi^*_\varepsilon \leftarrow \text{sinkhorn}(\gamma, \phi, \varepsilon) \approx \underset{\boldsymbol{\pi} \in \mathbb{R}_+^{V \times K}}{\arg\min} \mathcal{L}_{\text{OT}_\varepsilon}(\gamma, \phi) \tag{21}$$

The Sinkhorn algorithm [55] is used to compute π^*_ε, which can be efficiently executed on GPUs.

4.5.3 Sparse linear decoder

We compute the expected gene expression by taking the dot product of the cell-topic matrix $\boldsymbol{\theta}_i$ and gene-topic matrix $\mathbf{B} \in \mathbb{R}^{V \times K}$: $\hat{\mathbf{x}}_i \propto \boldsymbol{\theta}_i \mathbf{B}^\top$. Prior work has typically modeled $\mathbf{B}$ as the product of topic and gene embeddings [6]. Here we propose to directly derive $b_{mk} \in \mathbf{B}$ from the ECR-guided embedding distance between the gene m and topic k:

$$b_{mk} = \frac{\exp(-\|\mathbf{g}_m - \mathbf{t}_k\|^2 / \tau)}{\sum_{j=1}^{K} \exp(-\|\mathbf{g}_m - \mathbf{t}_j\|^2 / \tau)} \tag{22}$$

where τ is the temperature hyperparameter.

The categorical likelihood for cell i and gene m is defined as in the original ETM:

$$p(x_{im} | \boldsymbol{\theta}_i, \mathbf{B}) = \prod_{\ell=1}^{n_i} r_{im}^{[\ell=m]} = r_{im}^{x_{im}} \tag{23}$$

where n_i is total number of transcripts in cell i (e.g., Unique Molecular Identifier or UMI counts) and the categorical rate or the expected transcriptional rate for gene m in cell i is defined as:

$$r_{im} = \frac{\exp(\boldsymbol{\theta}_i \mathbf{b}_m^\top)}{\sum_m \exp(\boldsymbol{\theta}_i \mathbf{b}_m^\top)} \tag{24}$$

4.5.4 Overall loss function:

The VAE loss function is defined as

$$\mathcal{L}_{\text{VAE}} = \frac{1}{n} \sum_{i=1}^{n} -\mathbf{x}_i^\top \log\left(\text{Softmax}\left(\boldsymbol{\theta}_i \cdot \mathbf{B}^\top\right)\right) + \text{KL}\left[q\left(\boldsymbol{\theta}_i \mid \mathbf{x}_i\right) \| p\left(\boldsymbol{\theta}_i\right)\right] \tag{25}$$

The first term is the reconstruction error or negative log likliheood. The second term is the Kullback-Leibler (KL) divergence between the variational distribution and the standard Gaussian

prior distribution:

$$\begin{aligned}
\mathrm{KL}[q(z|x)||p(z)] &= \mathbb{E}_{q(z|x)}\left[\log q(z|x)\right] - \mathbb{E}_{q(z)}\left[\log p(z)\right] && (26)\\
&= \sum_k \left[-\frac{1}{2}\log\{\sigma_k^*\}^2 - \frac{1}{2} - \frac{1}{2}\log(2\pi)\right] - \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\{\mu_k^*\}^2 - \frac{1}{2}\{\sigma_k^*\}^2\right] && (27)\\
&= -\frac{1}{2}\left[\log\{\sigma_k^*\}^2 - \{\mu_k^*\}^2 - \{\sigma_k^*\}^2 + 1\right] && (28)
\end{aligned}$$

Throughout the training process, cell embedding learning and topic modeling are co-optimized. We define the overall loss function of scE^{2}TM as the combination of $\mathcal{L}_{\mathrm{VAE}}$ (Eq. (25)), $\mathcal{L}_{\mathrm{CVE}}$ (Eq. (18)), and $\mathcal{L}_{\mathrm{ECR}}$ (Eq. (20)), that is,

$$\mathcal{L} = \mathcal{L}_{\mathrm{VAE}} + \mathcal{L}_{\mathrm{CVE}} + \lambda \cdot \mathcal{L}_{\mathrm{ECR}}, \tag{29}$$

where λ is the weight hyperparameter. Algorithm 1 outlines the training workflow of scE^{2}TM.

Algorithm 1 Training algorithm for scE^{2}TM

Input: observed single-cell gene expression $\mathbf{X} = \{\mathbf{x}_i\}$; cell embeddings from scGPT $\mathbf{V} = \{\mathbf{v}_i\}$.
Hyperparameter: number of epochs n_e; weight of the ECR loss λ.
Output: topic embeddings $\mathbf{T} = \{\mathbf{t}_k\}$; gene embeddings $\mathbf{G} = \{\mathbf{g}_m\}$; cell-topic matrix $\Theta = \{\boldsymbol{\theta}_i\}$; gene-topic distributions $\mathbf{B}$.

for 1 **to** n_e **do**
  for each mini-batch from $\mathbf{X}$ and $\mathbf{V}$ **do**
    $\boldsymbol{\theta}_i, \boldsymbol{\varphi}_i \leftarrow \mathrm{Encode}(\mathbf{x}_i, \mathbf{v}_i)$ ▷ Cross-view encoder
    Compute $\mathbf{B}$ by Eq. (22)
    $\hat{\mathbf{x}}_i \leftarrow \mathrm{Decode}(\boldsymbol{\theta}_i, \mathbf{B})$ ▷ Sparse linear decoder
    Compute $\mathcal{L}_{\mathrm{VAE}}$ by Eq. (25)
    ▷ CVE
    Sample neighbors $\mathbf{x}_i^{\mathcal{N}} \in \mathcal{N}(\mathbf{x}_i)$ and $\mathbf{v}_i^{\mathcal{N}} \in \mathcal{N}(\mathbf{v}_i)$
    $\boldsymbol{\theta}_i^{\mathcal{N}}, \boldsymbol{\varphi}_i^{\mathcal{N}} \leftarrow \mathrm{Encode}(\mathbf{x}_i^{\mathcal{N}}, \mathbf{v}_i^{\mathcal{N}})$ ▷ Cross-view encoder
    Compute $\mathcal{L}_{\mathrm{CVE}}$ by Eq. (18)
    ▷ ECR
    $C_{mk} = \|\mathbf{g}_m - \mathbf{t}_k\|^2, \forall m, k$
    $\boldsymbol{\pi}_\epsilon^* \leftarrow \mathrm{Sinkhorn}(\mathbf{C})$
    Compute $\mathcal{L}_{\mathrm{ECR}}$ by Eq. 20
    ▷ Parameter Inference
    Compute $\mathcal{L} = \mathcal{L}_{\mathrm{VAE}} + \mathcal{L}_{\mathrm{CVE}} + \lambda \cdot \mathcal{L}_{\mathrm{ECR}}$
    Update $\mathbf{T}$, $\mathbf{G}$, Θ, $\mathbf{B}$ with a gradient step

4.5.5 Software implementation

Our scE^{2}TM is implemented in Python and the core model is based on the PyTorch (v.1.9.8) framework. Throughout the training process, we set the dimension size of each hidden layer to 200 and used $\tanh()$ as the activation function. The optimization of scE^{2}TM is performed using

the RMSprop optimizer with 500 iterations. For datasets with more than 10,000 cells, we uniformly use a batch size of 2048, otherwise it is set to 512. For λ, following the setup of Wu *et al* [23], its value is set to 20 or 100 for different datasets. The weight α of $\mathcal{L}_{\text{REG}}$ is set to 5. Furthermore, for fair comparison, we followed previous work [6, 18] in setting the number of topics K to 100 for all single-cell embedded topic models. When calculating TC, TD, TQ, ORA$_N$, ORA$_U$, and ORA$_Q$, we set the number of top genes h for each topic to 10. All experiments, except for the embedding extraction of the foundation model, were conducted in a Python environment equipped with an Nvidia RTX 3090 GPU and 128G RAM.

4.6 Differential gene expression and topic analysis

For differential gene expression analysis, we compute the log fold-change in expression between cells with and without a given label. Statistical significance is assessed using the Mann-Whitney U test, as implemented in "`scanpy.tl.rank_genes_groups()`". P-values are derived from the U statistic using a normal approximation with tie correction, and multiple testing correction is performed using the BH method.

For differential topic analysis, we aim to identify topics significantly associated with specific cell types or disease states. For each topic-cell label pair, we compute the difference in mean topic activity between cells with and without the label. Statistical significance is evaluated using the same Mann-Whitney U test framework described above, followed by BH correction for multiple comparisons. A topic is considered specific to a cell type or condition if its BH-adjusted q-value is below 0.01 and the mean difference for cell type-related topics exceeds 1. We use the CellMarker [56] database to examine the overlap between top genes in cell type-specific differentially expressed topics and known cell type markers.

4.7 Topic-based *in-silico* perturbations experiments

In the IFN-β PBMC perturbation experiment, we increased the treatment-specific topics among untreated cells and observe their changes in term of the expected gene expression output from the decoder. For each control cell, we set the topic scores to 0.1, 0.15, and 0.2. To study perturbations effects of the malignant-specific topics, we decrease the topic scores in the malignant cells by setting them to 1/8, 1/16, and 0 of their original values. This simulated varying degrees of malignant topic suppression. As negative control experiments for both case studies, we applied two types of controls. First, we randomly selected the same number of non-condition-specific topics and repeated the analysis (random topics). Second, we selected an equal number of topics that showed the lowest correlation with the condition (non-malignant or non-IFN specific topics).

4.8 Survival analysis

We extracted clinical and scRNA-seq data of melanoma patients from the TCGA database using the `cgdsr` R software package. Following an existing protocol [18], patient samples were divided

into two groups corresponding to low (25%) and high (75%) expression of the target gene. We ran Kaplan-Meier algorithm implemented in the survival R package to estimate the survival functions of the two groups. The same analysis was performed for the malignant-specific topics.

4.9 Categorization of perturbed cells

We divided the data into training and testing sets in a 7:3 ratio, which were used to train and test a binary SVM classifier to predict treatment or condition label of a cell. The reported results represent the average of 10 independent runs.

4.10 Code availability

The scE^{2}TM model and all the code necessary for reproducing our results is publicly available via GitHub at `https://github.com/nbnbhwyy/scE2TM`. To ensure reproducible results, all analyses are performed using the same fixed random seed. An archived version will be deposited in the Zenodo database upon acceptance.

5 Acknowledgments

Y.L. is supported by Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare (CRC-2021-00547) and CIHR Project Grant (202503PJT-540722-GMX-CFAA-255237). Y.R. is supported by the National Natural Science Foundation of China (62372483).

6 Competing interests

The authors declare that they have no competing interests.

References

1. Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

2. Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

3. Tian Tian, Jie Zhang, Xiang Lin, Zhi Wei, and Hakon Hakonarson. Model-based deep embedding for constrained clustering analysis of single cell rna-seq data. *Nature communications*, 12(1):1873, 2021.

4. Qin Ma and Dong Xu. Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology*, 23(5):303–304, 2022.

5. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

6. Yifan Zhao, Huiyu Cai, Zuobai Zhang, Jian Tang, and Yue Li. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications*, 12(1):5261, 2021.

7. Frederick Klauschen, Jonas Dippel, Philipp Keyl, Philipp Jurmeister, Michael Bockmayr, Andreas Mock, Oliver Buchstab, Maximilian Alber, Lukas Ruff, Grégoire Montavon, et al. Toward explainable artificial intelligence for precision pathology. *Annual Review of Pathology: Mechanisms of Disease*, 19(1):541–570, 2024.

8. Alex Tseng, Avanti Shrikumar, and Anshul Kundaje. Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics. In *NeurIPS*, pages 1913–1923, 2020.

9. Yifeng Tao, Xiaojun Ma, Drake Palmer, Russell Schwartz, Xinghua Lu, and Hatice Ulku Osmanbeyoglu. Interpretable deep learning for chromatin-informed inference of transcriptional programs driven by somatic alterations across cancers. *Nucleic Acids Research*, 50(19):10869–10881, 2022.

10. Lakshmipuram Seshadri Swapna, Michael Huang, and Yue Li. Gtm-decon: guided-topic modeling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes. *Genome Biology*, 24(1):190, 2023.

11. Yimin Fan, Yu Li, Jun Ding, and Yue Li. Gfetm: Genome foundation-based embedded topic model for scatac-seq modeling. In *International Conference on Research in Computational Molecular Biology*, pages 314–319. Springer, 2024.

12. HeGang Chen, Pengbo Mao, Yuyin Lu, and Yanghui Rao. Nonlinear structural equation model guided gaussian mixture hierarchical topic modeling. In *ACL*, pages 10377–10390, 2023.

13. Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. Self-supervised topic taxonomy discovery in the box embedding space. *Transactions of the Association for Computational Linguistics*, 12:1401–1416, 2024.

14. Ruihao Chen, Hegang Chen, Yuyin Lu, Yanghui Rao, and Chunjiang Zhu. Supervised neural topic modeling with label alignment. *Transactions of the Association for Computational Linguistics*, 13:249–263, 2025.

15. Allen W Lynch, Christina V Theodoris, Henry W Long, Myles Brown, X Shirley Liu, and Clifford A Meyer. Mira: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nature methods*, 19(9):1097–1108, 2022.

16. Yichen Zhang, Mohammadali Sam Khalilitousi, and Yongjin P Park. Unraveling dynamically encoded latent transcriptomic patterns in pancreatic cancer cells by topic modeling. *Cell Genomics*, 3(9), 2023.

17. Sishir Subedi and Yongjin P Park. Single-cell pair-wise relationships untangled by composite embedding model. *Iscience*, 26(2), 2023.

18. Hegang Chen, Yuyin Lu, Zhiming Dai, Yuedong Yang, Qing Li, and Yanghui Rao. Comprehensive single-cell rna-seq analysis using deep interpretable generative modeling guided by biological hierarchy knowledge. *Briefings in Bioinformatics*, 25(4):bbae314, 2024.

19. Nour El Kazwini and Guido Sanguinetti. Share-topic: Bayesian interpretable modeling of single-cell multi-omic data. *Genome Biology*, 25(1):55, 2024.

20. Chengwei Zhong, Kok Siong Ang, and Jinmiao Chen. Interpretable spatially aware dimension reduction of spatial transcriptomics with stamp. *Nature Methods*, pages 1–12, 2024.

21. Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

22. He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. Neural topic model via optimal transport. In *ICLR*, 2021.

23. Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. Effective neural topic modeling with embedding clustering regularization. In *ICML*, pages 37335–37357. PMLR, 2023.

24. Valerie Chen, Muyu Yang, Wenbo Cui, Joon Sik Kim, Ameet Talwalkar, and Jian Ma. Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nature Methods*, 21(8):1454–1461, 2024.

25. Manoj M Wagle, Siqu Long, Carissa Chen, Chunlei Liu, and Pengyi Yang. Interpretable deep learning in single-cell omics. *Bioinformatics*, page btae374, 2024.

26. Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024.

27. Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024.

28. Yue J Wang, Jonathan Schug, Kyoung-Jae Won, Chengyang Liu, Ali Naji, Dana Avrahami, Maria L Golson, and Klaus H Kaestner. Single-cell transcriptomics of the human endocrine pancreas. *Diabetes*, 65(10):3028–3038, 2016.

29. Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

30. Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145–153, 2015.

31. Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods*, 20(9):1355–1367, 2023.

32. Manqi Zhou, Hao Zhang, Zilong Bai, Dylan Mann-Krzisnik, Fei Wang, and Yue Li. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods*, 3(8), 2023.

33. Zhicheng Lin, HeGang Chen, Yuyin Lu, Yanghui Rao, Hao Xu, and Hanjiang Lai. Hierarchical topic modeling via contrastive learning and hyperbolic embedding. In *COLING*, pages 8133–8143, May 2024.

34. Sreenath S Andrali, Megan L Sampley, Nathan L Vanderford, and Sabire Özcan. Glucose regulation of insulin gene expression in pancreatic β-cells. *Biochemical Journal*, 415(1):1–10, 2008.

35. Deepak Raj, Maria Nikolaidi, Irene Garces, Daniela Lorizio, Natalia M Castro, Sabrina G Caiafa, Kate Moore, Nicholas F Brown, Hemant M Kocher, Xiaobo Duan, et al. Ceacam7 is an effective target for car t-cell therapy of pancreatic ductal adenocarcinoma. *Clinical Cancer Research*, 27(5):1538–1552, 2021.

36. Ping Chen, Daiwei Wan, Dingcheng Zheng, Qi Zheng, Feng Wu, and Qiaoming Zhi. Long non-coding rna uca1 promotes the tumorigenesis in pancreatic cancer. *Biomedicine & pharmacotherapy*, 83:1220–1226, 2016.

37. Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94, 2018.

38. Su-Yang Liu, Roghiyh Aliyari, Kelechi Chikere, Guangming Li, Matthew D Marsden, Jennifer K Smith, Olivier Pernet, Haitao Guo, Rebecca Nusbaum, Jerome A Zack, et al.

Interferon-inducible cholesterol-25-hydroxylase broadly inhibits viral entry by production of 25-hydroxycholesterol. *Immunity*, 38(1):92–105, 2013.

39. Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

40. Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

41. Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18, 2024.

42. Manqi Zhou, Hao Zhang, Zilong Bai, Dylan Mann-Krzisnik, Fei Wang, and Yue Li. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods*, 3(8):100563, 2023.

43. Kyle Coleman, Amelia Schroeder, Melanie Loth, Daiwei Zhang, Jeong Hwan Park, Ji-Youn Sung, Niklas Blank, Alexis J Cowan, Xuyu Qian, Jianfeng Chen, et al. Resolving tissue complexity by multimodal spatial omics modeling with miso. *Nature methods*, pages 1–9, 2025.

44. Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J Theis, and Bo Wang. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.

45. Lucas Seninge, Ioannis Anastopoulos, Hongxu Ding, and Joshua Stuart. Vega is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature communications*, 12(1):5684, 2021.

46. Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

47. Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

48. Zhuohan Yu, Yifu Lu, Yunhe Wang, Fan Tang, Ka-Chun Wong, and Xiangtao Li. Zinb-based graph embedding autoencoder for single-cell rna-seq interpretations. In *AAAI*, volume 36, pages 4671–4679, 2022.

49. Zhenze Liu, Yingjian Liang, Guohua Wang, and Tianjiao Zhang. sclega: an attention-based deep clustering method with a tendency for low expression of genes on single-cell rna-seq data. *Briefings in Bioinformatics*, 25(5):bbae371, 2024.

50. Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053, 2018.

51. Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *NAACL*, pages 530–539, 2014.

52. Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

53. Ayush Dalmia and Suzanna Sia. Clustering with umap: Why and how connectivity matters. *arXiv preprint arXiv:2108.05525*, 2021.

54. Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. *Advances in neural information processing systems*, 22, 2009.

55. Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *NeurIPS*, pages 2292–2300, 2013.

56. Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. *Nucleic acids research*, 51(D1):D870–D876, 2023.