He *et al. Genome Biology* (2025) 26:300
https://doi.org/10.1186/s13059-025-03779-0

Genome Biology

RESEARCH

Open Access

scKAN: interpretable single-cell analysis for cell-type-specific gene discovery and drug repurposing via Kolmogorov-Arnold networks

Haohuai He[1], Zhenchao Tang[2], Guanxing Chen[3], Fan Xu[1], Yao Hu[1], Yinglan Feng[1], Jibin Wu[1,4], Yu-An Huang[5*], Zhi-An Huang[3*] and Kay Chen Tan[1]

*Correspondence:
yuanhuang@nwpu.edu.cn;
huang.za@cityu-dg.edu.cn

[1] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR, China
[2] Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, China
[3] Department of Computer Science, City University of Hong Kong (Dongguan), Dongguan 523000, China
[4] Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[5] School of Computer Science, Northwestern Polytechnical University, Xi'an 710000, China

Abstract

Background: Analysis of single-cell RNA sequencing (scRNA-seq) data has revolutionized our understanding of cellular heterogeneity, yet current approaches face challenges in efficiency, interpretability, and connecting molecular insights to therapeutic applications. Despite advances in deep learning methods, identifying cell-type-specific functional gene sets remains difficult.

Results: In this study, we present scKAN, an interpretable framework for scRNA-seq analysis with two primary goals: accurate cell-type annotation and the discovery of cell-type-specific marker genes and gene sets. The key innovation is using the learnable activation curves of the Kolmogorov-Arnold network to model gene-to-cell relationships. This approach provides a more direct way to visualize and interpret these specific interactions compared to the aggregated weighting schemes typical of attention mechanisms. This architecture achieves superior performance in cell-type annotation, with a 6.63% improvement in macro F1 score over state-of-the-art methods. Additionally, it enables the systematic identification of functionally coherent cell-type-specific gene sets. We demonstrate the framework's translational potential through a case study on pancreatic ductal adenocarcinoma, where gene signatures identified by scKAN led to a potential drug repurposing candidate, whose binding stability was supported by molecular dynamics simulations.

Conclusions: Our work establishes scKAN as an efficient and interpretable framework that effectively bridges single-cell analysis with drug discovery. By combining lightweight architecture with the ability to uncover nuanced biological patterns, our approach offers an interpretable method for translating large-scale single-cell data into actionable therapeutic strategies. This approach provides a robust foundation for accelerating the identification of cell-type-specific targets in complex diseases.

Keyword: Single-cell analysis, Kolmogorov-Arnold networks, Drug repurposing, Marker gene discovery, Interpretable AI

Background

Single-cell technologies have revolutionized our understanding of biological processes and human diseases by offering unprecedented insights into cellular heterogeneity at high resolution [1, 2]. Cell-type annotation is a fundamental component of this revolution, which identifies distinct cell populations across tissues, developmental stages, and organisms. This process provides a critical foundation for understanding cellular functions and gene regulation in health and disease states [3–5]. Cell-type annotation fundamentally relies on identifying unique gene expression signatures, which are also known as marker genes, that distinguish one cell population from another. Beyond serving as identifiers, a subset of these genes is also functionally critical for the cell's identity and survival within its specific context [6]. The systematic identification of these essential genes is a primary objective. These genes likely encode tissue-specific modulators of key cellular functions and represent potential therapeutic targets, particularly in cancer [7].

Cell-type annotation methods rely on single-cell gene expression data to distinguish cell populations. Traditional methods like Seurat [8], while widely used, often require extensive preprocessing steps and manual marker selection, limiting their scalability and automation [9]. To address these limitations, machine learning approaches have emerged and demonstrated significant progress. For example, CellTypist [10] leveraged supervised learning techniques to identify 101 distinct cell types from over a million cells. Deep learning methods have further propelled the field. ACTINN [11] employed neural networks to achieve rapid and accurate cell type identification. MetaTiME [12] integrated single-cell gene expression to characterize meta-components, enabling the identification of heterogeneous cell types and states in the tumor microenvironment. Meanwhile, Cellcano [13] introduced a supervised learning framework that effectively mitigates the distributional shift between reference and target data, improving annotation accuracy.

Recently, transformer-based large language models (LLMs) with attention mechanisms have shown immense promise for single-cell analysis. TOSICA [14] pioneered one-shot annotation, adapting to new cell types with minimal training examples while maintaining interpretability through biologically understandable entities. scBERT [15] introduced a BERT-inspired pretraining approach to capture gene–gene interactions from large-scale unlabeled scRNA-seq data. Furthermore, LangCell [16] was developed to construct unified representations of single-cell data and natural language during pretraining, enabling zero-shot cell identity understanding. Geneformer [17] advanced context-aware attention-based learning with self-supervised pretraining on about 30 million single-cell transcriptomes. More recently, scGPT [18], trained on over 33 million cells, established itself as a foundation model capable of diverse downstream applications, including cell-type annotation, multi-batch integration, and gene network inference. Meanwhile, GeneCompass [19] further contributed by integrating prior biological knowledge across species. However, despite their potential, these LLM-based models often demand substantial computational resources for training, require frequent fine-tuning for new datasets, and struggle to provide cell-type-specific interpretability of gene functions and interactions [20].

These limitations present three key technical challenges in single-cell analysis. First, despite extensive pretraining, these models require considerable fine-tuning to achieve acceptable accuracy on new data. Second, the adopted attention mechanisms compute gene representations by weighing information from all other genes in the input sequence, a process that learns a "global context." While powerful, the global nature of the attention mechanism can make it challenging to directly isolate and interpret gene interactions that are specific to a single-cell type from the learned representations. Third, current methods operate in isolation from downstream applications such as drug discovery, creating a gap between single-cell analysis and practical therapeutic development. Addressing these challenges requires a method to directly identify cell-type-specific genes and connect these molecular findings to drug repurposing applications.

To address these challenges, we developed scKAN, an interpretable deep learning framework for analyzing single-cell transcriptomic data. The primary analytical goal of scKAN is to perform accurate cell-type annotation while simultaneously identifying the cell-type-specific marker genes. To achieve this, scKAN takes a single-cell gene expression matrix as input and produces two outputs: predictive cell-type labels for each cell and a set of interpretable parameters that quantify gene-cell relationships.

This architecture, which integrates knowledge distillation with a Kolmogorov-Arnold network (KAN), is designed to overcome prior methods' limitations systematically. First, the knowledge distillation component mitigates the need for extensive fine-tuning by efficiently transferring knowledge from a large pre-trained model into the lightweight scKAN, ensuring computational efficiency. Second, the KAN component provides cell-type-specific interpretability by using its learnable activation curves to model gene-cell interactions directly, moving beyond the "global context" limitations of transformer models. Finally, the interpretable outputs serve as a natural bridge to downstream applications, providing biologically informed gene signatures that can be used directly for therapeutic target discovery and drug repurposing.

Unlike traditional drug discovery approaches that rely on differential expression analysis, scKAN's integration of cell-type-specific gene importance scores with activation curve patterns provides a novel framework for identifying druggable targets. Notably, scKAN enables the discovery of potential therapeutic targets that might be overlooked by conventional methods, particularly those with moderate expression levels but high functional significance. These findings led to the discovery of potential therapeutic targets, demonstrating scKAN's value in translational research. Together, these results highlight scKAN's capability to advance the methodological framework of single-cell analysis and its practical applications in disease research.

In this paper, we present the complete scKAN framework and demonstrate its effectiveness through extensive experiments. We first show that scKAN achieves superior cell-type annotation accuracy compared to state-of-the-art (SOTA) methods across multiple benchmark datasets. We then validate its core innovation: the ability to identify biologically meaningful, cell-type-specific gene sets, and marker genes. Finally, we showcase the practical utility of our framework in a case study on pancreatic ductal adenocarcinoma (PDAC), where scKAN successfully identifies potential drug targets, leading to a potential drug repurposing candidate. The remainder of this paper details the scKAN architecture, presents these comparative and validation results, and discusses the broader implications of our findings for translational research.

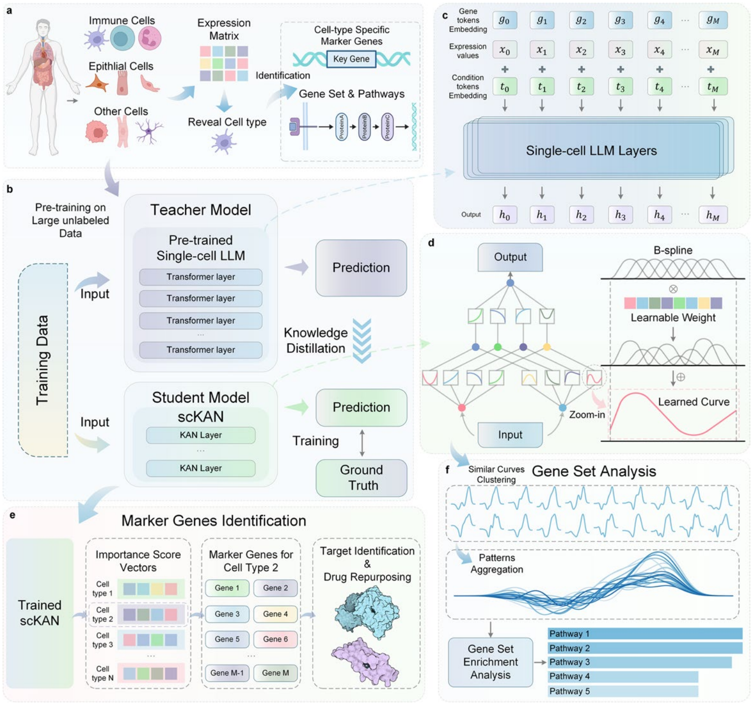

Fig. 1 Overview of the scKAN framework for single-cell analysis and marker gene identification. **a** Schematic illustration of the single-cell analysis workflow, showing the extraction of different cell types from biological samples, followed by generation of expression matrices and identification of cell-type-specific marker genes and pathways. **b** Knowledge distillation framework architecture consisting of two main components: a teacher model (pre-trained single-cell LLM with multiple transformer layers trained on large unlabeled datasets) and a student model (scKAN with multiple KAN layers). The framework enables knowledge transfer through prediction-based distillation while incorporating ground truth information during training. **c** Structure of single-cell LLM layers, showing the integration of gene tokens, expression values, and condition embeddings to generate cell embeddings (output). **d** Detailed neural network architecture of scKAN, illustrating the hierarchical structure of KAN layers where connections between nodes are represented by learnable B-spline activation functions instead of traditional weights. The right panel shows how these B-spline functions are learned and adjusted during training to capture complex relationships in the data. **e** Marker gene identification process using trained scKAN, progressing from importance score vectors for different cell types to specific marker gene identification, ultimately supporting target identification and drug repurposing applications. **f** Gene set analysis workflow showing the clustering of similar activation curves from trained scKAN, pattern aggregation, and subsequent pathway enrichment analysis to identify cell-type-specific pathways

Results

Overview of the scKAN framework

The workflow and model architecture of scKAN are shown in Fig. 1. scKAN aims to achieve accurate cell annotation and identify marker genes and gene sets for specific cell types. Figure 1a provides a schematic overview of the biological context, highlighting the existence of diverse cell types within organisms and the role of single-cell technologies in generating gene expression matrices for individual cells. scKAN leverages these matrices to decipher cell types and identify marker genes and potential pathways.

As illustrated in Fig. 1b–d, the core architecture of scKAN employs a knowledge distillation strategy. A pre-trained LLM based on the transformer architecture [21] serves as the teacher model, guiding a KAN-based module as the student model [22]. The development of scKAN involves two steps: first, fine-tuning an LLM that has been pre-trained on large-scale unlabeled cellular data on specific datasets; second, training the student model through knowledge distillation to enable it to integrate the teacher model's prior knowledge with ground truth cell type information. In addition to the distillation process, we incorporate unsupervised learning objectives to enhance the discriminative power of learned feature representations [23]. Specifically, we employ a self-entropy loss [24] to prevent over-concentration on dominant cell types and maintain sensitivity to rare cell populations. Furthermore, leveraging Cauchy–Schwarz divergence [25], a modified deep divergence-based clustering (DDC) loss [26, 27] optimizes the relationships between hidden features and cluster assignments, ensuring alignment with ideal cell-type distributions. This combined loss function, integrating distillation and unsupervised components, is designed to guide the model toward learning generalizable representations across different cell types.

To obtain a pre-trained LLM with extensive single-cell prior knowledge, we employ the SOTA single-cell analysis foundation model, scGPT [18], as the teacher model. This transformer-based model has been extensively pre-trained on over 33 million cells, capturing representation patterns of various human cell types, including pancreatic and blood cells. As shown in Fig. 1c, it uses a gene encoder to encode gene IDs, applies binning to expression values to obtain expression embeddings, and incorporates condition embeddings for specific genes. By integrating these embedding inputs through multiple transformer layers, this pretraining imbues the teacher model with a rich understanding of diverse human cell types for annotation.

We implement scKAN using multiple KAN layers as the student model to learn from the teacher model while acquiring annotation capabilities on labeled cellular data. Following the Kolmogorov-Arnold representation theorem, the KAN model learns activation function curves for edges, rather than weights as in traditional multilayer perceptrons (MLPs). As shown in Fig. 1d, these curves, fitted using B-splines, capture underlying representation patterns, establishing latent representation connections between cells and genes.

After training, the edge scores in KAN, initially used for pruning, indicate the significance of activation function curves between nodes. In our framework, we adapt these scores to quantify the learned contribution of each gene to the classification of a specific cell type. We subsequently validate this approach by demonstrating that genes with high importance scores are significantly enriched for known cell-type-specific markers and differentially expressed genes. As shown in Fig. 1e, these importance scores are used to identify marker genes for specific cell types. These serve as inputs for downstream analysis, such as target identification of human diseases and functional characterization for specified cell types. Combined with drug-target affinity prediction methods, this workflow is designed to support drug repurposing studies.

Additionally, the similarity among learned activation function curves provides insights into gene co-expression patterns within specific cell types. As shown in Fig. 1f, similar curves are clustered to reveal functionally related gene sets. Subsequent gene set enrichment analysis is applied to these sets to identify cell-type-specific pathways.

Ablation studies validate the essential components of scKAN

To systematically evaluate the contribution of each component in scKAN, we conducted comprehensive ablation studies by creating three model variants: "W/o Teacher" removes the teacher model to assess the impact of knowledge distillation, "W/o Cluster" eliminates the clustering loss to evaluate the importance of maintaining cell-type-specific feature representations, and "Replaced by MLP" substitutes the KAN module with a MLP architecture to examine the advantages of our KAN-based design. These variants were designed to validate our architectural choices and understand the role of each component in achieving optimal performance.

To ensure a fair comparison, we tested these variants on multiple datasets: peripheral blood mononuclear cells 10X (PBMC) [28] dataset with 9631 single cells across 19 cell types, Muto2021 kidney dataset (Muto2021) [29] containing 19,985 single cells, human pancreas dataset [14] (hPancreas) with 14,818 single cells, and myeloid dataset [30] (Mye) with 13,178 cells. These datasets include various cell types, such as immune and kidney cells. Detailed information about these datasets is provided in the "Dataset" subsection of the "Methods" section.

We evaluated these variants using accuracy and macro F1 scores across all four datasets. These two metrics were chosen as they provide complementary insights: accuracy reflects the overall classification performance, while macro F1 score better captures the model's performance on imbalanced cell populations. Detailed information about these metrics is provided in Additional file 1: Note S1.

The ablation results in Table 1 demonstrate the importance of each component. "W/o Teacher" led to consistent performance drops across all datasets, particularly noticeable decreases in macro F1 scores. On the PBMC dataset, the macro F1 score dropped from 0.894 to 0.869, while on hPancreas, it decreased from 0.917 to 0.871, confirming the value of knowledge transfer from the pre-trained language model. Similarly, "W/o Cluster" reduced performance, especially in the macro F1 scores of more challenging datasets. For instance, on the hPancreas dataset, removing the clustering loss led to a decrease in the macro F1 score from 0.917 to 0.892, highlighting its role in maintaining robust cell-type representations. When replacing KAN with MLP, we observed an interesting trade-off. The MLP variant achieved higher accuracy on the Mye dataset, reaching 0.747 compared to scKAN's 0.728. However, it showed consistently lower macro F1 scores across datasets. This discrepancy likely stems from MLP's tendency to optimize for majority cell types at the expense of rare populations, particularly evident in the complex and imbalanced Mye dataset. Moreover, replacing KAN with MLP eliminates the interpretability advantages provided by KAN's learned activation curves and importance scores, losing the ability to gain biological insights from the model's decision-making process.

These ablation studies demonstrate that each component plays a crucial role in scKAN's overall performance and functionality. The knowledge distillation framework provides essential prior knowledge, the clustering loss helps maintain robust cell-type-specific features, and the KAN architecture offers a unique combination of competitive performance and biological interpretability. Alternative architecture like MLP might achieve comparable accuracy in specific scenarios. However, our design choices offer

Table 1 Ablation study evaluating the contribution of key components in scKAN. The performance metrics include accuracy and macro F1 scores with standard deviations (shown in parentheses) across four datasets: PBMC, Muto2021, hPancreas, and Mye. "W/o Teacher" represents the model without knowledge distillation, "W/o Cluster" removes the clustering-based loss functions, and "Replace by MLP" substitutes the KAN module with a traditional multilayer perceptron. The "Interpretable" column indicates whether the model variant maintains the inherent ability to provide interpretable insights into cell-gene relationships

Setting	Interpretable	PBMC		Muto2021		hPancreas		Mye	
		Accuracy (std)	Macro F1 (std)	Accuracy (std)	Macro F1 (std)	Accuracy (std)	Macro F1 (std)	Accuracy (std)	Macro F1 (std)
scKAN (full)	**Yes**	**0.922 (0.003)**	**0.894 (0.004)**	**0.969 (0.001)**	**0.959 (0.008)**	**0.985 (0.001)**	**0.917 (0.052)**	0.728 (0.003)	**0.585 (0.011)**
W/o Teacher	**Yes**	0.911 (0.019)	0.869 (0.026)	0.965 (0.004)	0.952 (0.007)	0.982 (0.003)	0.871 (0.071)	0.717 (0.015)	0.566 (0.024)
W/o Cluster	**Yes**	0.916 (0.004)	0.879 (0.018)	0.966 (0.004)	0.949 (0.013)	0.984 (0.002)	0.892 (0.041)	0.713 (0.009)	0.576 (0.013)
Replace by MLP	No	0.920 (0.005)	0.876 (0.013)	0.962 (0.004)	0.946 (0.008)	0.984 (0.002)	0.880 (0.062)	**0.747 (0.007)**	0.574 (0.012)

Bold indicates the best performance

comprehensive benefits: balanced performance across cell types and enhanced biological interpretability. These advantages strongly support the current architecture of scKAN.

scKAN achieves superior performance in cell-type annotation

To evaluate the performance of scKAN in cell-type annotation tasks, we compared it with multiple baseline models: Geneformer [17], Tosica [14], and scGPT [18], which are LLM-based single-cell foundation models; Cellcano [13], a deep learning method employing knowledge distillation; Celltypist [10], a machine learning-based method; and Seurat [8], a classical single-cell analysis tool. Detailed information about these baseline methods is provided in the "Baseline" subsection of the "Methods" section.

Figure 2 shows the performance comparison between scKAN and baseline methods across these datasets. We used classical multiple classification metrics, including accuracy, macro precision, macro recall, and macro F1 score, to assess model performance comprehensively.

scKAN consistently outperformed all baseline methods across all datasets and metrics, including the single-cell foundation models such as scGPT, Geneformer, and Tosica. This demonstrates the superior performance of scKAN while also underscoring the limitations of current single-cell foundation models, which still face challenges in cell-type annotation despite their versatility. Specifically, the experiment results are shown in Fig. 2, Additional file 1: Fig. S1, and Table S1. On average, scKAN achieved a 1.06% improvement in accuracy and a 6.63% improvement in macro F1 score compared to the second-best models across all datasets. Notably, on the PBMC dataset, scKAN achieved its highest accuracy improvement of 2.95% over the second-best model, scGPT. For the hPancreas dataset, scKAN demonstrated its most significant macro F1 score improvement, with a 9.77% increase over scGPT, reaching a score of 0.917. These consistent improvements across different datasets demonstrate that scKAN is a reliable and effective tool for cell-type annotation tasks.

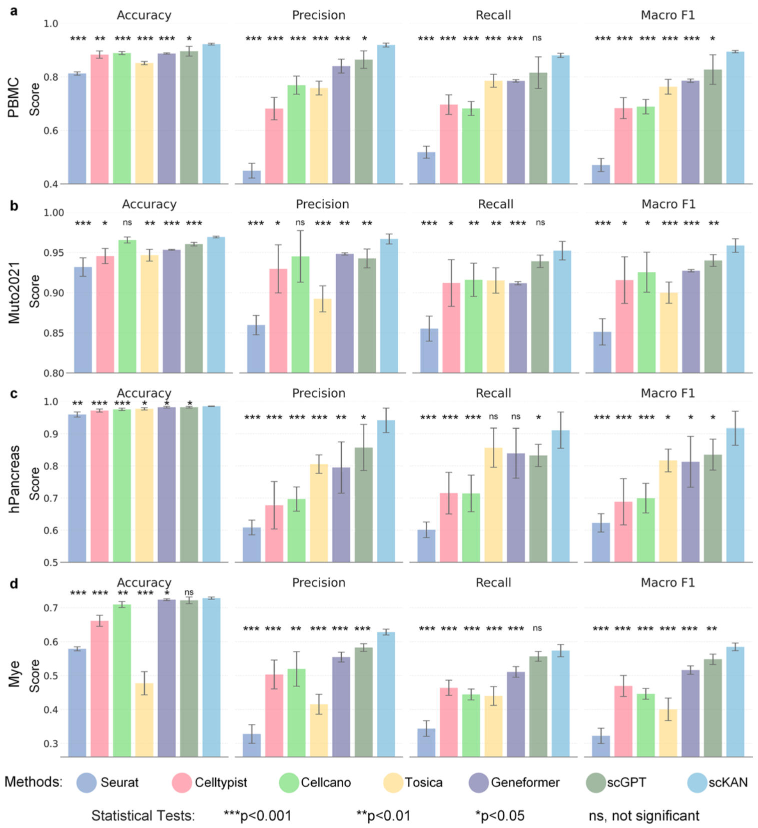

Fig. 2 Comparative performance evaluation of scKAN against baseline methods across multiple datasets using fivefold cross-validation. **a**–**d** Comparison of test set performance in cell-type annotation across four datasets (PBMC, Muto2021, hPancreas, and Mye) using four metrics: accuracy, macro precision, macro recall, and macro F1 score. Results are averaged across five test sets from fivefold cross-validation, with error bars representing standard deviation. Baseline methods include traditional tools (Seurat), machine learning approaches (Celltypist), deep learning methods (Cellcano), and foundation LLMs for single-cell analysis (Tosica, Geneformer, and scGPT). scKAN consistently outperforms baseline methods across different datasets and metrics, with significant improvements in accuracy and macro F1 scores. The statistical significance of scKAN's performance compared to each baseline method was evaluated using an independent t-test. The significance level for each pairwise comparison is indicated above the respective baseline method's bar: *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$; ns indicates that the difference is not significant

Statistical analysis further validated these results. Using independent *t*-tests, we found significant improvements in the comprehensive metrics of accuracy and macro F1 score. For the PBMC dataset, both macro F1 and accuracy showed significant improvements ($p < 0.05$) compared with the second-best methods. In Muto2021, while accuracy differences were insignificant, the macro F1 score showed significant improvement ($p < 0.05$). The hPancreas dataset showed significant improvements in both accuracy and macro F1 ($p < 0.05$), and the Mye dataset showed significant improvement in macro F1 score ($p < 0.01$) with no significant difference in accuracy. These statistical results further confirm that scKAN provides significantly better and more stable performance in cell-type annotation than existing methods.

The superior performance of scKAN can be attributed to three key design choices. First, knowledge distillation effectively transfers rich biological representations from the large teacher model, providing a strong inductive bias. Second, KAN architecture's learnable activation functions capture complex, non-linear gene-cell relationships more effectively than conventional networks. Finally, the auxiliary unsupervised losses mitigate class imbalance by enhancing sensitivity to rare cell types, directly contributing to the significant gains observed in the macro F1 score.

Beyond performance advantages, scKAN demonstrates remarkable computational efficiency. When evaluated on the Muto2021, scKAN achieves a 15.8-fold reduction in GPU memory consumption and 5.4-fold faster training speed compared to scGPT, while requiring only 2.3% of the parameters (Additional file 1: Table S2). This substantial improvement in computational efficiency makes scKAN more accessible for widespread adoption in the research community.

We also provided a detailed visualization analysis to demonstrate scKAN's performance further. As illustrated in Fig. S1, the confusion matrices and UMAP visualizations across four datasets demonstrate scKAN's consistent and accurate performance across different cell types. In the PBMC dataset, the confusion matrix shows high diagonal values, indicating accurate classification across all 19 cell types, with robust performance in identifying major immune cell populations. The UMAP visualization of PBMC shows that the predicted cell-type distributions closely match the annotated labels, with clear boundaries between different cell populations. Similar high-quality results were observed in Muto2021, where the model accurately distinguished between different cell types, including CNT, DCT, ENDO, and LEUK populations. For the hPancreas dataset, scKAN successfully identified pancreatic cell types, including alpha, beta, and ductal cells, with the UMAP plot showing nearly identical clustering patterns between predicted and annotated labels. In the more challenging Mye dataset with its complex hierarchy of immune cell subtypes, scKAN maintained reliable performance, accurately distinguishing between closely related myeloid cell populations as shown in both the confusion matrix and UMAP visualization. These visual results comprehensively demonstrate scKAN's robust capability in cell-type annotation across diverse tissue types and cellular compositions. Overall, these detailed comparisons and analyses establish scKAN as a reliable and superior tool for single-cell annotation tasks.

scKAN maintains robust performance in cross-study and cross-disease settings

To evaluate scKAN's performance under more realistic conditions, we conducted cross-study experiments on the hPancreas dataset. Following Tosica experimental setup [14], we divided the hPancreas dataset into non-overlapping reference and query sets from completely independent studies. The reference set contained 10,600 single-cell samples from GSE84133 and GSE85241, covering 14 cell types, while the query set included 4218 previously unseen single-cell samples from three distinct studies (GSE81608, E-MTAB-5061, and GSE86473), representing 11 cell types. This setup ensures a strict evaluation of the model's generalization ability across independent datasets with no sample overlap.

As shown in Fig. 3 and Additional file 1: Table S3, in this cross-study setting, scKAN achieved 97.42% accuracy and a macro F1 score of 0.734, showing improvements of 1.01% and 2.03% over the second-best models, Tosica and scGPT, respectively. As shown in Fig. 3a–d, the pie charts clearly illustrate scKAN's superior performance across all evaluation metrics compared to other methods. Even under this more challenging

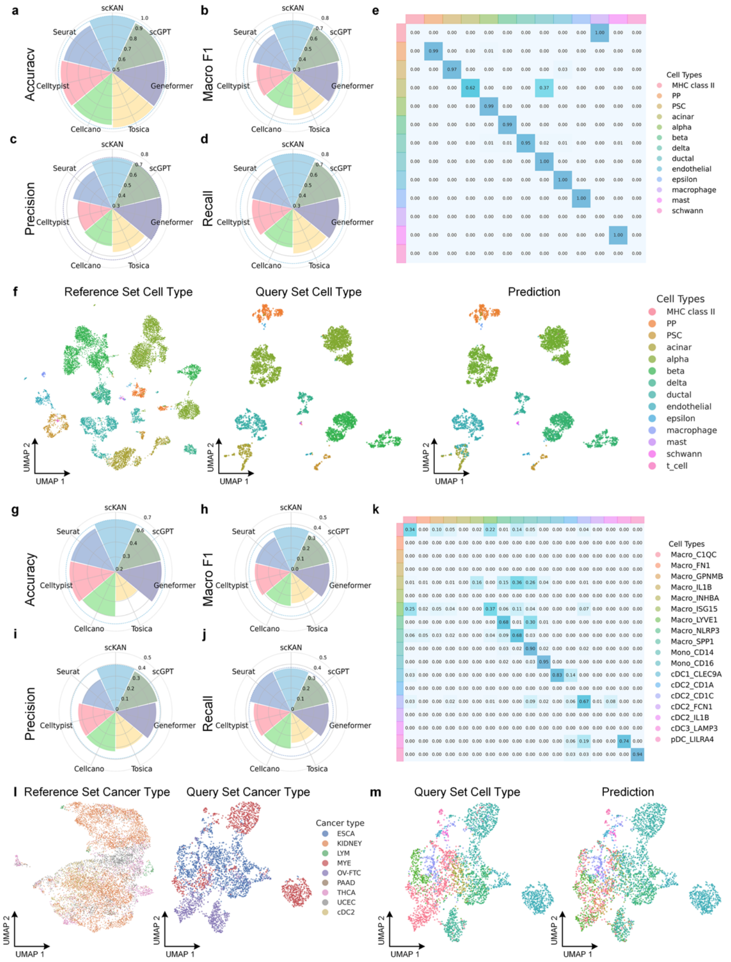

Fig. 3 Performance evaluation of scKAN in cross-study and cross-disease settings. **a**–**d** Comparative performance metrics (accuracy, macro F1, precision, and recall) across different methods in the cross-study pancreas dataset experiment. The pie charts show scKAN's superior performance compared to baseline methods. **e** Confusion matrix showing cell-type classification performance across 13 pancreatic cell types in the cross-study setting. **f** UMAP visualization of pancreas data showing reference set cell types (left), query set cell types (middle), and scKAN predictions (right), demonstrating consistent cell-type distribution patterns. **g**–**j** Performance metrics displayed as pie charts for the cross-disease Mye dataset experiment, showing scKAN's maintained advantage in this challenging setting. **k** Confusion matrix demonstrating classification performance across 17 myeloid cell types in the cross-disease setting. **l** UMAP visualization comparing reference set (left) and query set (right) cancer-type distributions. **m** UMAP visualization of query set cell types (left) and corresponding scKAN predictions (right), showing preserved cell-type distribution patterns despite the cross-disease setting

cross-study scenario, scKAN maintained high performance with accuracy close to 98%, demonstrating strong generalization ability. The confusion matrix (Fig. 3e) reveals high diagonal values for most cell types, indicating accurate classification performance. The inability to effectively annotate MHC class II cells was also observed across all baseline models due to this severe data imbalance. The UMAP visualization (Fig. 3f) further supports these findings, showing consistent cell-type distributions between reference and query sets, with predicted labels closely matching the actual cell types. The high degree of concordance between the UMAP visualization of the query set cell types and the scKAN predictions (Fig. 3f) indicates the model's strong generalization capability and high annotation accuracy.

We further tested scKAN under a more challenging cross-disease setting using the Mye dataset. The pie charts (Fig. 3g–j) demonstrate that in this cross-disease scenario, scKAN achieved an accuracy of 63.84% and a macro F1 score of 0.373, surpassing the second-best model, scGPT, by 4.48% and 7.44%, respectively. As illustrated in Fig. 3l–m, we selected six cancer types for the reference set (9748 cells) and three for the query set (3430 cells). Although overall performance decreased in this more challenging setting, scKAN maintained its advantage over baseline methods across all metrics, including macro F1 score.

The confusion matrix (Fig. 3k) shows the classification results across different cell types, revealing a block-diagonal pattern that indicates strong classification performance within related cell-type groups. The model achieved high accuracy (> 0.6) for most cell types, with lower performance only in C1QC, INHBA, and LYVE1 classifications, similar to the hPancreas dataset, primarily due to limited sample sizes for these cell types. The UMAP visualizations (Fig. 3l–m) demonstrate that despite the challenging cross-disease setting, scKAN successfully maintained the overall structure of cell-type distributions between reference and query sets, with predicted labels showing strong concordance with actual cell types. These cross-study and cross-disease results demonstrate that scKAN can effectively learn intrinsic patterns in single-cell data, enabling accurate cell-type annotation across different experimental conditions.

scKAN enhances cell-type-specific gene set and pathway discovery

After validating the model's cell annotation capabilities, we investigated its ability to identify biologically meaningful gene sets. scKAN learns gene characteristics across different cell types by distilling knowledge from single-cell LLMs, implicitly encoding cell-type-level gene expression patterns within gene-cell activation curves, thereby enabling gene set identification.

First, as shown in Fig. 4a, we visualized gene programs clustered by these activation curves from scKAN in the PBMC dataset, revealing diverse expression patterns across cell types. Analysis of the gene programs showed that scKAN captures complex expression patterns that extend beyond simple, uniform biological programs. Furthermore, Fig. 4b demonstrates that genes with similar scKAN curves within gene programs showed high similarity, particularly when contrasted with the randomly selected activation curves shown in Fig. 4c.

Specifically, the gene programs encompassed various functionally distinct programs. For instance, as illustrated in Fig. 4d, cluster 37 contained several T cell-related genes. These included the CD8A and CD8B genes, which work together to form the CD8 complex essential for T cell recognition [31]. The cluster also contained GZMA, a crucial effector molecule in CD8 + T cells [32], along with KLRK1, which is an activation receptor in both NK and T cells [33]. Additionally, ITGB1 was identified as playing a vital role in T cell adhesion and migration processes [34, 35]. The signaling-related components of this cluster included RGS2, which regulates G-protein signaling, and RAB31, which mediates intracellular transport [36, 37]. Notably, the activation curves of these functionally related genes exhibited high similarity.

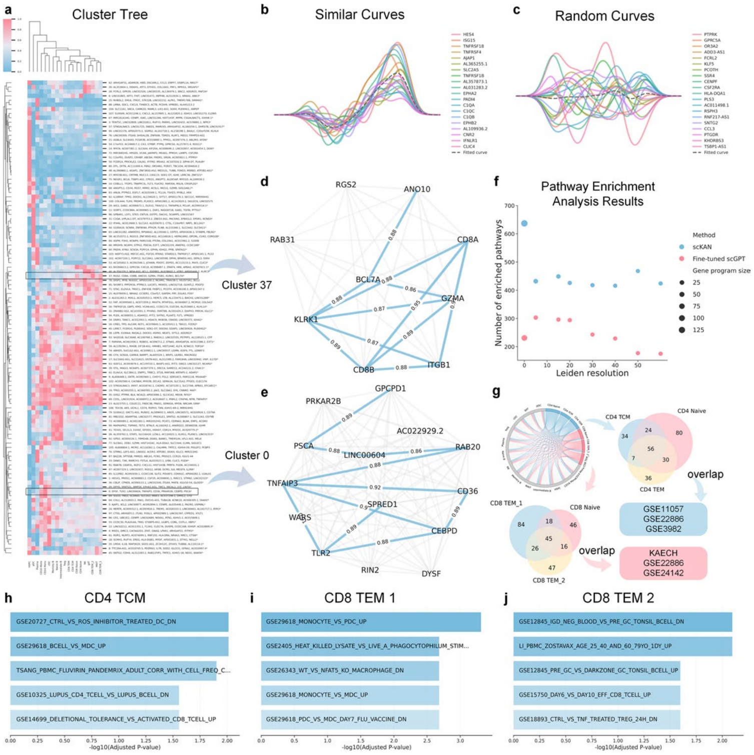

Fig. 4 Gene set identification and pathway analysis capabilities of scKAN. **a** Hierarchical clustering of genes based on the similarity of their learned scKAN activation curves. The heatmap displays the corresponding expression levels of these genes across different cell types. The cluster numbers are the numerical labels assigned to the resulting gene groups (gene programs), which are highlighted by the dendrogram on the left. **b** Example of similar activation curves from a functionally related gene cluster, demonstrating coherent patterns. **c** Random gene activation curves are shown for comparison, illustrating the contrast with functionally related curves. Gray dashed lines in b and c show the mean fitted curves. **d**–**e** Network visualization of two representative gene clusters: (**d**) T cell-related gene cluster (cluster 37) and (**e**) inflammation-related gene cluster (cluster 0), with edge weights indicating similarity scores between gene activation curves. **f** Comparison of pathway enrichment analysis results between scKAN and scGPT across different Leiden resolutions, showing the number of enriched pathways (dot size indicates gene program size). **g** Venn diagram and circular visualization showing pathway overlap between CD4 and CD8 T cell populations, with data source labels (GSE11057, GSE3982, GSE22886, KAECH) indicated. h–j Top 5 enriched pathways (− log10 adjusted *p*-value) for specific T cell populations: (**h**) CD4 TCM, (**i**) CD8 TEM 1, and (**j**) CD8 TEM 2, demonstrating cell-type-specific pathway enrichment

Similarly, cluster 0, shown in Fig. 4e, revealed a distinct group of genes involved in inflammation and immune response pathways. TLR2 emerged as a key component, a critical receptor for pathogen recognition and initiation of innate immunity [38]. The cluster also included TNFAIP3, a central regulator of inflammatory responses, alongside CD36, which functions as an important immune cell surface receptor [39]. CEBPD was also identified within this cluster, playing a crucial role in the transcriptional regulation of inflammatory genes [40]. The functional coherence of this cluster was further supported by the high similarity scores (> 0.88) among the scKAN-derived curves for these genes, providing additional validation of the model's effectiveness.

To quantitatively evaluate scKAN's ability to resolve biologically meaningful gene programs, we conducted a comparative analysis against scGPT [18], a SOTA single-cell LLM. We first clustered the learned representations from both models into gene programs using the Leiden algorithm at various resolution parameters. Subsequently, we performed Gene Set Enrichment Analysis [41] (GSEA) on the resulting gene clusters.

As shown in Fig. 4f, scKAN consistently identifies a greater number of gene clusters and, more importantly, a larger number of unique enriched biological pathways across all tested resolutions. This quantitative improvement is not merely numerical; it reflects a deeper and more granular biological understanding. The advantage stems from scKAN's architecture, where activation curves are learned to model specific gene-to-cell-type relationships. This inherent cell-type-specific representation allows the model to capture subtle, context-dependent gene co-expression patterns that may be averaged out or obscured within the more global embeddings produced by scGPT.

By demonstrating this superiority across a range of Leiden resolutions, we confirm that enhanced performance is a robust feature of scKAN's representations. The ability to resolve more functionally coherent gene sets is biologically significant because it enables the dissection of broad biological functions into more specific sub-pathways active in distinct cellular states. This higher-resolution view is critical for uncovering nuanced regulatory mechanisms and generating more precise, testable hypotheses.

Further pathway analysis, as illustrated in Fig. 4g, revealed overlapping pathways identified by scKAN across different cell types. Within the CD4 cell populations (CD4 TCM, CD4 Naïve, and CD4 TEM), we identified several significantly enriched pathways characteristic of T cell development and function, with 56 pathways shared among all three subsets. Notably, pathways from the GSE11057 collection distinguished memory CD4 + T cells from peripheral blood mononuclear cells, capturing memory T cell states with characteristic gene expression patterns [42]. We also identified GSE22886 pathways [43], revealing molecular distinctions between naïve CD4 + T cells and monocytes, while GSE3982 pathways [44] highlighted the unique molecular features of central memory CD4 + T cells through comparison with B cells.

Similarly, within CD8 cell populations (CD8 TEM_1, TEM_2, and naïve), we observed enrichment in pathways associated with CD8 + T cell differentiation and activation, with 45 pathways shared among all three populations. The KAECH pathways [45] were particularly informative, capturing the dynamic transcriptional changes during CD8 + T cell activation and the transition from naïve to effector memory states. GSE22886 pathways [43] highlighted CD8 + T cell-specific molecular signatures through comparisons with B cells and monocytes, while GSE24142 pathways [46] revealed programs involved in early T cell development. All enriched pathways and their biological implications are provided in Additional file 1: Table S4.

Finally, as demonstrated in Fig. 4h–j, Additional file 1: Fig. S2, and Table S5, the top enriched pathways identified by scKAN curves for specific cell types are well supported by previous literature. For CD4 TCM, the top enriched pathway reveals its critical role in modulating DC function, where CD4 TCM cells may influence DC antigen presentation through cytokine-mediated regulation of oxidative stress responses [47]. CD8 TEM1-associated pathway (GSE29618_MONOCYTE_VS_PDC_UP) demonstrates its involvement in regulating both monocyte and plasmacytoid dendritic cell functions, potentially impacting their antigen presentation capabilities through cytokine-mediated activation [48]. The CD8 TEM2-enriched first-rank pathway indicates its role in modulating B cell activation states and immunoglobulin gene expression, particularly in germinal center

B cell development [49]. Additional cell-type-specific pathways were also identified, as shown in Additional file 1: Fig. S2: NK cells exhibited enrichment in BCG vaccine-induced bactericidal activity pathways [50], while naïve B cells showed association with vaccine-induced PBMC responses [51], and cDCs demonstrated enrichment in pathways related to vaccine-mediated immune responses [52]. To validate the robustness of this finding, we conducted the enrichment analysis using varying gene set sizes. We confirmed that "GSE29618_MONOCYTE_VS_PDC_UP" consistently ranks as a top pathway, underscoring the stability of our approach (Additional file 1: Note S2, Additional file 1: Fig. S3).

Collectively, these comprehensive analyses validate scKAN's effectiveness in interpretable gene set identification and its capability for cell-type-specific characterization.

scKAN discovers reliable cell-type-specific marker genes

Identifying cell-type-specific marker genes is crucial for understanding cellular identity and function. Building on its ability to interpret gene-cell relationships, scKAN also provides a novel approach for marker gene discovery through its importance scores. To systematically evaluate scKAN's capabilities, we conducted a comprehensive analysis using a two-part validation strategy. Computationally, we assessed whether the top-ranked genes identified by scKAN's importance scores were statistically significant as differentially expressed genes and benchmarked this capability against a leading foundation model. Biologically, we confirmed the relevance of these marker genes by reviewing existing literature and visualizing their cell-type-specific expression patterns.

We performed differential expression analysis on the PBMC dataset encompassing 19 distinct cell types. We leveraged a well-trained scKAN model for each cell type to identify the top 20 candidate marker genes and evaluated their expression patterns using Wilcoxon rank-sum tests. Figure 5a presents a comprehensive visualization where dot size represents the magnitude of expression fold change, color indicates regulatory direction, and black borders denote statistical significance (adjusted p-value < 0.05). The analysis revealed that scKAN's top 10 marker genes generally exhibited more pronounced differential expression patterns. However, some cell types, such as HSPCs, showed less significant differentiation. Markers ranked 10–20 displayed more varied patterns, including genes with modest expression differences. This dual capability of identifying highly expressed and subtly regulated marker genes provides opportunities to discover novel regulatory mechanisms in cell-type-specific gene expression.

As shown in Fig. 5b–e and Additional file 1: Fig. S4, detailed examination through volcano plots further validates scKAN's capacity to identify both strongly and subtly differentially expressed genes. To assess the biological relevance of these computationally derived markers, we performed a subsequent literature review on the top-ranked genes. This analysis confirmed that many genes prioritized by scKAN are well-established markers, providing strong biological validation for our model's discovery capability. Key examples include SEL1L3 in naïve B cells, which has been previously established as a crucial regulator of B cell development [53]; CST7 in NK cells, known for its role in cytotoxic functions [53]; BACH2 in CD8 naïve cells, essential for T cell homeostasis [54]; SERPINA1 in CD16 monocytes, involved in inflammatory responses [55, 56]; and PPIB in plasma cells, critical for antibody production [53]. UMAP visualizations

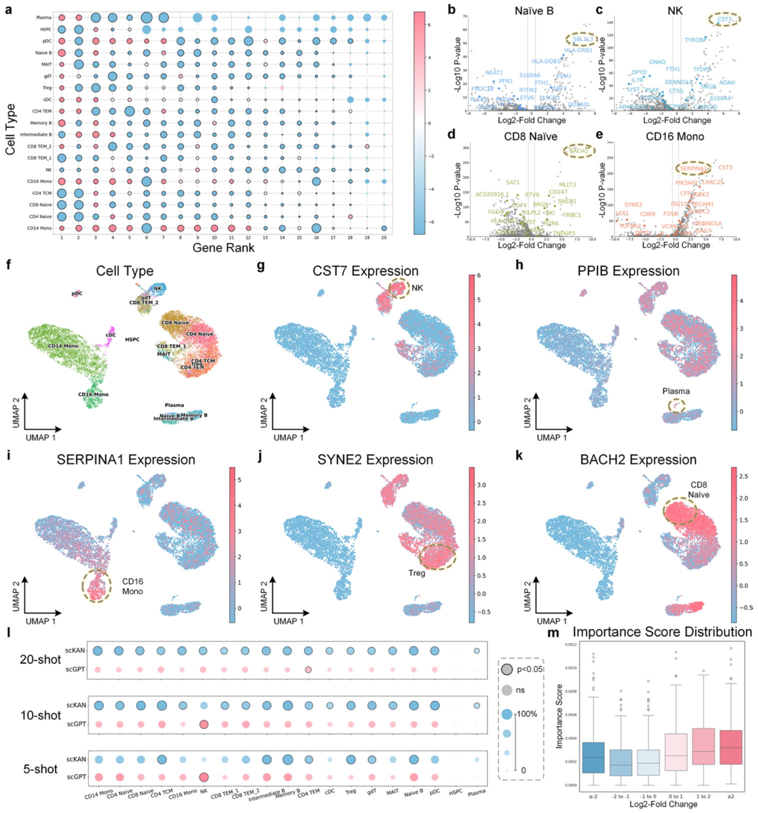

Fig. 5 Marker gene identification and validation using scKAN. **a** Dot plot showing expression patterns of the top 20 marker genes identified by scKAN across 19 cell types. Dot size indicates expression fold change magnitude, color represents regulatory direction, and black borders denote statistical significance (adjusted *p*-value < 0.05). **b**–**e** Volcano plots highlight differentially expressed genes in selected cell populations: (**b**) Naïve B cells, (**c**) NK cells, (**d**) CD8 Naïve cells, and (**e**) CD16 Mono cells, with marker genes highlighted in dashed circles. **f** UMAP visualization of cell type distributions in the PBMC dataset. **g**–**k** Expression patterns of validated marker genes shown on UMAP plots: CST7 (**g**), PPIB (**h**), SERPINA1 (**i**), SYNE2 (**j**), and BACH2 (**k**), with color intensity indicating expression level and dashed circles highlighting cell-type-specific expression. **l** Comparative analysis of marker gene identification performance between scKAN and scGPT under different few-shot settings (20-shot, 10-shot, 5-shot), with dot size indicating the proportion of differentially expressed genes and black borders showing statistical significance ($p < 0.05$). **m** The distribution of scKAN importance scores across different log2-fold change ranges, showing the relationship between importance scores and expression differences

(Fig. 5f–k and Additional file 1: Fig. S5) demonstrate the distinct spatial distribution of these markers, confirming their cell-type-specific expression patterns.

To benchmark scKAN's interpretable marker gene identification capabilities, we conducted a systematic comparison with scGPT, a SOTA single-cell foundation model, under 20-shot, 10-shot, and 5-shot settings. We evaluated the top 20, top 10, and top 5 potential marker genes for each cell type. Figure 5l illustrates the proportion of differentially expressed genes captured by each model across different cell types, with black-bordered circles indicating statistical significance determined by hypergeometric testing. scKAN demonstrated superior performance in capturing differentially expressed genes across most cell types, significantly enriching the overall differentially expressed gene pool. The only exception was NK cells under 10-shot and 20-shot conditions, where scGPT showed comparable performance. These statistical results strongly support scKAN's effectiveness in capturing cell-type-specific gene expression patterns.

To further conduct a rigorous evaluation of the biological significance of the identified marker genes, we performed a comprehensive validation for three distinct cell types: Naïve B cells, CD8 Naïve T cells, and NK cells. This analysis involved three components: assessing the overlap with previously validated markers, performing functional enrichment analysis using the Gene Ontology (GO) database, and conducting pathway enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. The results confirmed the biological significance of the identified genes. For instance, the top genes for Naïve B cells were enriched in "Antigen Receptor-Mediated Signaling," those for CD8 Naïve T cells in "T Cell Activation," and for NK cells in "Natural Killer Cell Mediated Immunity." These results provide strong evidence that the model discovers meaningful biological signals. This validation's detailed methodology and results are presented in Additional file 1: Note S3 and Additional file 1: Tables S6–S12.

Moreover, Fig. 5m shows the relationship between scKAN-derived importance scores and log2-fold change. While a positive trend exists, confirming that the importance score effectively captures strong differential expression signals, the distribution also reveals that numerous genes with high importance scores exhibit only modest fold changes. This demonstrates that the scKAN framework can identify functionally relevant genes beyond what a simple differential expression threshold would capture.

Additionally, to quantitatively validate the interpretability of scKAN, we benchmarked its explanation fidelity against other interpretable machine learning methods, including scGPT, random forest, and L1-regularized logistic regression. We employed a standard explainable AI evaluation technique where the fidelity of an explanation is measured by the predictive power of the top-K genes it identifies. A simple proxy classifier was trained using only these gene sets. The results showed that the gene sets identified by scKAN consistently outperformed those from other models. Specifically, for any given number of top genes (top-K), a simple proxy classifier trained on the scKAN-identified set achieved higher accuracy and macro F1 scores. This superior performance in the fidelity test provides quantitative evidence that scKAN generates more faithful and informative explanations. A detailed description of this assessment is available in Additional file 1: Note S4 and Additional file 1: Table S13.

These comprehensive analyses demonstrate that scKAN provides a robust and interpretable approach for marker gene identification, capturing both strong and subtle cell-type-specific signals. The model's effectiveness, validated through statistical and biological approaches, establishes its utility for marker gene discovery in single-cell analysis.

scKAN enables systematic drug repurposing for PDAC treatment

PDAC remains one of the most lethal malignancies, with a 5-year survival rate below 8.5% [57]. Despite extensive research efforts, therapeutic options remain limited, highlighting the urgent need for novel drug development approaches [58]. This urgent clinical need motivated us to explore scKAN's potential in interpretable drug discovery through a systematic drug repurposing study targeting PDAC.

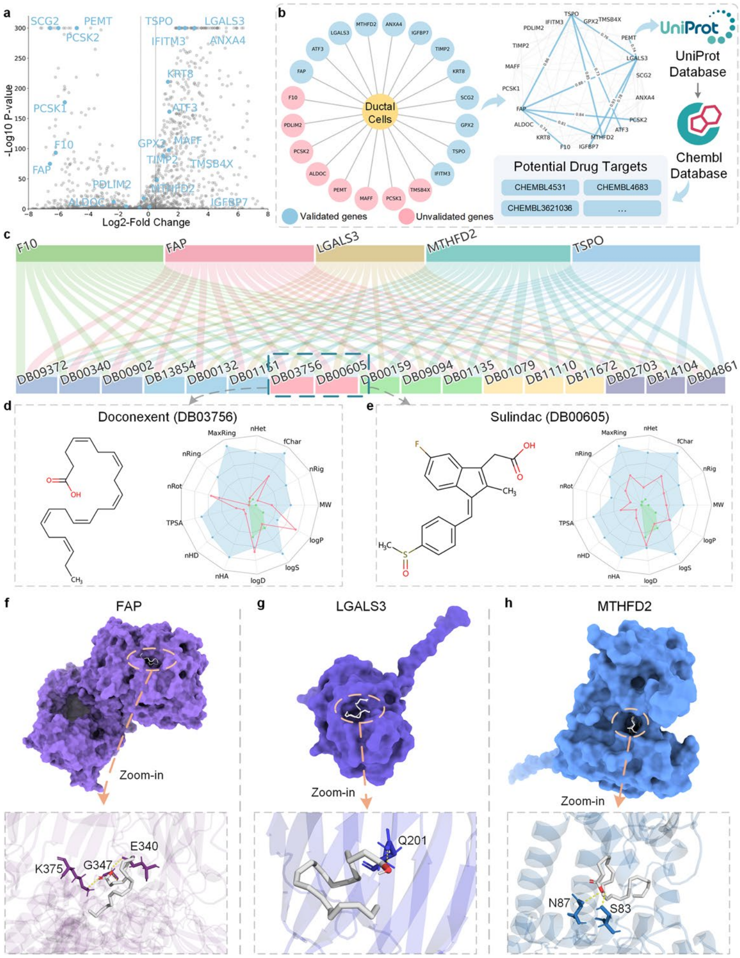

Fig. 6 Drug repurposing analysis for PDAC treatment using scKAN and molecular verification. **a** Volcano plot showing differentially expressed genes in ductal cells. Potential marker genes identified by scKAN are highlighted in blue. **b** Network showing scKAN-identified marker genes for ductal cells, with previously validated genes in blue and unvalidated genes in pink. Similarity graph constructed from scKAN activation curves showing relationships between those marker genes, followed by potential drug target identification through UniProt and drug-target affinity data retrieval from the ChEMBL database. **c** Sankey diagram showing predicted binding affinity relationships between potential protein targets and FDA-approved drugs, with top candidates DB03756 (Doconexent) and DB00605 (Sulindac) highlighted in pink. **d**–**e** 2D structures and ADMET radar plots for top drug candidates: d Doconexent (DB03756) and e Sulindac (DB00605). f–h Molecular docking results showing binding poses of Doconexent with three target proteins: (**f**) FAP, (**g**) LGALS3, and (**h**) MTHFD2, with surface representations (top) and detailed binding pocket views (bottom)

We designed a comprehensive workflow for drug repurposing for PDAC. Initially, we employed scKAN trained on the hPancreas dataset to identify ductal cell-specific marker genes. Figure 6a highlights these markers in the volcano plot, showing diverse differential expression patterns consistent with our previous analyses. Notably, literature validation revealed that 12 out of the top 20 marker genes had been previously implicated in pancreatic cancer or specifically PDAC [59–62] (Fig. 6b). Among the top 10 genes, 8 were validated, representing a significantly high validation rate [63–70], further underscoring scKAN's reliability in identifying cell-type-specific markers. Detailed information about supporting work for those potential genes is shown in Additional file 1: Table S14.

To refine this initial list of markers and prioritize the most functionally coherent targets, we integrated a gene set identification method from scKAN into the workflow. Instead of relying on individual marker genes, we leveraged scKAN's learned activation curves to analyze the functional relationships among the top-ranked markers. By constructing a similarity graph based on these curves, we identified a highly interconnected and functionally coordinated gene set comprising FAP, F10, IGFBP7, MTHFD2, ATF3, PCSK2, LGALS3, PEMT, and TSPO. This robust gene set, representing a core functional signature of ductal cells as identified by scKAN, formed the basis for our subsequent drug repurposing screen. We then mapped these genes to their corresponding protein targets using the UniProt database [71] and searched for existing drug-target interaction data in the ChEMBL database [72]. This systematic database integration revealed that only five targets (F10, FAP, LGALS3, MTHFD2, and TSPO) had corresponding drug-target datasets available.

To ensure the robustness of our predictions, we first selected a drug-target affinity (DTA) prediction model by benchmarking several SOTA methods. As detailed in Additional file 1: Table S15, NHGNN [73] demonstrated superior performance over other models, including MMFA-DTA [74] and AttentionMGT [75], across all evaluation metrics. Based on this result, we employed the NHGNN model, which we fine-tuned on our collected datasets. We then performed a virtual screen of 2509 FDA-approved drugs against the five PDAC-associated protein targets. This comprehensive screening produced a ranked list of all candidates based on an aggregate prediction score reflecting their binding affinities across the target set. Screening 2509 FDA-approved drugs against these targets (Fig. 6c) identified Doconexent (DB03756) and Sulindac (DB00605) as the top two candidates, ranking first and second, respectively, based on combined prediction scores. The detailed predicted binding affinities for the top-ranked candidates are provided in Additional file 1: Table S16. Notably, Doconexent has been demonstrated to induce apoptosis in pancreatic cancer cells through multiple mechanisms. It promotes the efflux of glutathione, accumulating intracellular reactive oxygen species (ROS) that subsequently activate caspase cascades, resulting in ROS-mediated apoptosis [76]. Furthermore, Doconexent has been observed to cause DNA fragmentation, activate caspase-3, and increase the Bax/Bcl-2 ratio in PANC-1 cells. Other studies also report that Doconexent and eicosapentaenoic acid can trigger ROS accumulation and caspase-8-dependent cell death [77]. Besides, Sulindac has been previously demonstrated to have therapeutic potential in pancreatic cancer treatment [78], providing external validation for our computational approach and supporting scKAN's utility in identifying clinically relevant drug targets.

To evaluate our identified candidate's drug-likeness and potential pharmacological properties, we employed ADMETLab 3.0 [79] to assess their absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. The 2D structural visualization and ADMET analysis results are shown in Fig. 6d–e, revealing that both molecules mostly satisfy ADMET criteria, falling within acceptable ranges for most parameters. Only Doconexent showed elevated nRot, LogD, and LogP values, while both molecules exhibited lower than optimal LogS values, suggesting potential for further molecular optimization while maintaining drug-like properties.

To further investigate Doconexent, our top-ranked candidate that outperformed the literature-supported Sulindac in virtual screening, we sought to characterize its potential binding mechanisms through molecular docking. We employed CB-Dock2 [80], a cavity-detection-guided docking approach, to examine the binding interactions between Doconexent and PDAC's potential protein targets. Crystal structures for these targets were obtained from the RCSB Protein Data Bank [81] for FAP (PDB ID: 1Z68 [82]), LGALS3 (PDB ID: 1KJL [83]), MTHFD2 (PDB ID: 5TC4 [84]), and F10 (PDB ID: 4Y6D [85]). For TSPO, due to the absence of an available experimental structure, we generated a predicted structure using the AlphaFold3 [86] (AF3) web server. It is important to note that simulations based on computationally predicted models carry inherent uncertainties. Therefore, the results for the TSPO target should be considered preliminary and require further validation with an experimentally determined structure.

The molecular docking analysis yielded high CB-Dock scores (> 6) across all protein targets, indicating strong binding potential. The docked complexes and detailed binding interfaces are visualized in Fig. 6f–h and Additional file 1: Fig. S6, where the upper panels show the surface representation of protein–ligand complexes, and the lower panels present detailed views of the binding pockets. Multiple hydrogen bond formations were observed within these binding pockets, further supporting the stability of these protein–ligand interactions. These structural analyses provide molecular-level evidence supporting Doconexent's potential therapeutic application in PDAC.

This comprehensive drug repurposing workflow for PDAC demonstrates scKAN's potential as a computational strategy for drug target discovery, providing a foundation for future experimental validation studies.

scKAN-identified drug candidate shows stable binding to potential targets

Following our molecular docking analyses, we extended our investigation through molecular dynamics (MD) simulations to evaluate the binding persistence and stability of the predicted Doconexent-target complexes. These simulations were designed to assess the maintenance of key protein–ligand interactions under physiological conditions, providing insights into the stability of binding modes identified from molecular docking. We performed 100 ns all-atom MD simulations at 310.15 K and 1 atm pressure, with detailed protocols in the "Methods" section.

The initial (gray) and final (blue) conformations of the five protein–ligand complexes are shown in Fig. 7 (left panels). Structural comparison before and after simulation revealed high overall conformational conservation, with major variations primarily observed in the terminal residues of the target proteins (highlighted by orange dashed circles). Notably, Doconexent maintained its position within the initial binding pockets across all five complexes throughout the 100 ns simulations, suggesting stable and effective protein–ligand interactions.

Trajectory analysis through root mean square deviation (RMSD) and root mean square fluctuation (RMSF) calculations (Fig. 7, middle and right panels) provided detailed insights into the binding stability. As shown in Fig. 7a, the FAP-Doconexent complex showed rapid initial RMSD increases for both ligand and backbone, stabilizing around 0.2–0.25 nm throughout the simulation, with the ligand showing

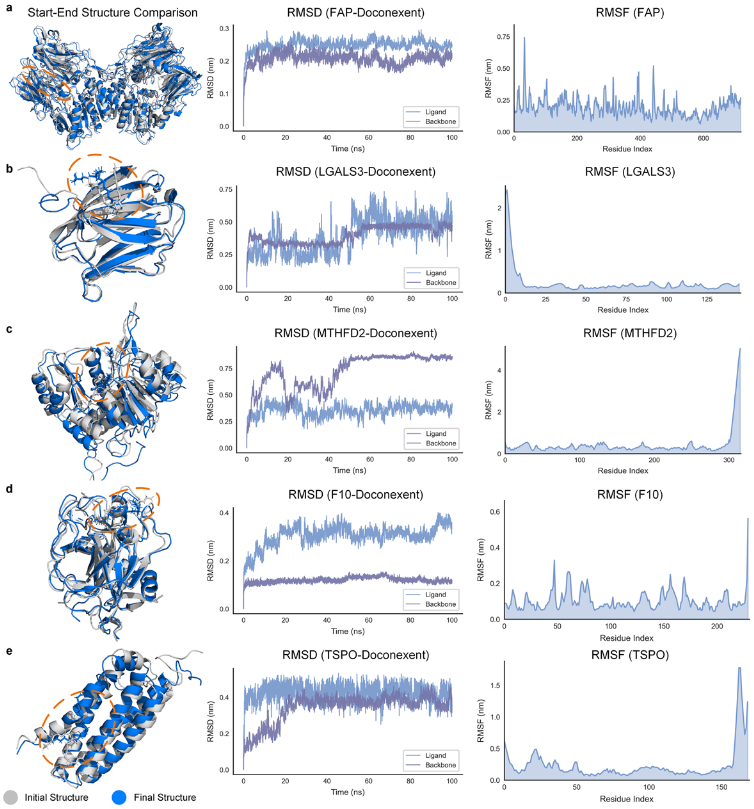

Fig. 7 MD simulations assessing the stability of Doconexent bound to five target proteins. **a**–**e** The figure presents results from 100 ns MD simulations for Doconexent complexed with (**a**) FAP, (**b**) LGALS3, (**c**) MTHFD2, (**d**) F10, and (**e**) TSPO. For each complex, the left panel shows a structural alignment of the initial (gray) and final (blue) conformations, with the binding pocket highlighted by an orange circle. The middle panel plots the RMSD for the protein backbone and the ligand, which indicates the conformational stability of the system over the simulation period. The right panel displays the RMSF per residue, which measures the flexibility of different regions within the protein. All simulations were conducted at 310.15 K and 1 atm pressure. The initial structures for FAP, LGALS3, MTHFD2, and F10 were obtained from the RCSB Protein Data Bank. The structure for TSPO was predicted using AlphaFold3

slightly higher fluctuations than the backbone, indicating achievement of a stable equilibrium state.

LGALS3 and MTHFD2 complexes with Doconexent (Fig. 7b–c) demonstrated distinct behaviors. The LGALS3 complex showed initial stability followed by increased ligand RMSD fluctuations (0.2–0.3 nm) after 40 ns, while its backbone maintained relatively stable RMSD around 0.2 nm. The MTHFD2 complex exhibited a gradual increase in ligand RMSD, reaching approximately 0.4 nm, with backbone RMSD stabilizing around 0.3 nm. Their RMSF analyses revealed localized flexibility, with LGALS3 showing distinct peaks in N-terminal residues (0–15, ~ 2.5 nm) and MTHFD2 displaying significant C-terminal mobility (residues 300–320, ~ 5 nm). These terminal region fluctuations likely reflect the absence of structural constraints in the crystal structures rather than instability in the protein–ligand binding interface.

The F10-Doconexent complex (Fig. 7d) showed gradual increases in ligand RMSD, reaching around 0.35 nm, while its backbone maintained relatively stable RMSD values around 0.15 nm. Despite these variations, the final binding pose closely resembled the initial configuration, confirming stable pocket retention. The RMSF profile showed moderate fluctuations distributed across various regions of the protein.

The TSPO-Doconexent complex (Fig. 7e) exhibited stable backbone RMSD around 0.2 nm, while the ligand RMSD showed larger fluctuations around 0.25–0.3 nm. The RMSF analysis revealed a notable peak around residues 150–170, which likely stems from lower confidence predictions in these regions of the AF3-generated structure, rather than inherent binding instability.

Collectively, our MD simulation analyses provide multiple supporting evidence for the stable binding of Doconexent to these five PDAC-associated targets. The maintenance of binding poses throughout the simulations and the convergence of RMSD values demonstrate stable protein–ligand interactions. Notably, the observed structural fluctuations were primarily localized to peripheral regions, while the core binding interfaces remained stable, as revealed by RMSF analyses. These computational findings provide a theoretical foundation supporting Doconexent's further development as a potential PDAC therapeutic lead compound.

Discussion

In this study, we introduced scKAN, a knowledge distillation framework that addresses key challenges in single-cell analysis through three major innovations: efficient knowledge transfer from single-cell LLMs, interpretable cell-type-specific gene set identification through activation curves, and cell-type-specific marker gene discovery via importance scores. Our comprehensive evaluation across four diverse datasets demonstrated scKAN's superior performance in cell-type annotation, while its application to PDAC drug discovery validated its practical utility in translational research.

The superior performance of scKAN, marked by an average 1.06% increase in accuracy and a notable 6.63% in macro F1 score over SOTA models, is not just a marginal gain. This improvement, coupled with significantly reduced resource requirements, demonstrates that our knowledge distillation approach successfully captures essential biological signals without the computational overhead of full-scale LLMs. Moreover, the interpretability provided by KAN's learnable univariate functions offers unique insights into cell-gene relationships that are more biologically intuitive than transformer-based attention mechanisms.

The biological insights derived from scKAN's gene set identification capability have demonstrated both established and novel biological relationships. Our analysis identified functionally related gene clusters enriched in known biological pathways, particularly in immune cell development and pancreatic cell differentiation. These findings were further validated through pathway enrichment analysis, showing significant overlap with established molecular signatures. Notably, scKAN's ability to capture subtle gene–gene relationships through activation curve similarities revealed previously unrecognized functional connections, suggesting potential new regulatory mechanisms in cell-type-specific contexts.

scKAN's marker gene identification capability demonstrated remarkable accuracy across diverse cell types, combining both statistical robustness and biological relevance.

The model successfully captured both strongly and subtly differentially expressed genes, outperforming SOTA methods when selecting limited candidate markers. The model's importance scores strongly correlate with differential expression patterns while capturing more complex regulatory relationships, providing a more nuanced understanding of cell-type-specific gene expression programs. This capability was particularly valuable in identifying both canonical markers and novel candidates, as demonstrated by our validation of key markers such as SEL1L3 in naïve B cells and CST7 in NK cells.

The application of scKAN to PDAC research is intended to demonstrate its potential for translational impact. By identifying ductal cell-specific markers that led to a promising drug candidate, Doconexent, we illustrate a tangible outcome of the model's interpretability. It is important, however, to frame this as a foundational proof-of-concept rather than a complete, end-to-end drug discovery platform. Our primary objective is to show that the high-fidelity marker genes generated by scKAN can serve as a direct and effective starting point for therapeutic hypothesis generation. This represents an early but significant attempt to bridge the gap between cell-type annotation and drug discovery, an application space that remains largely underexplored by computational models designed principally for annotation accuracy. While the current workflow is not fully integrated and relies on external SOTA methods for drug-target affinity prediction, it validates the unique potential of an interpretable model like scKAN to translate abstract single-cell insights into clinically relevant and testable hypotheses, setting the stage for future, more comprehensive investigations.

From a methodological perspective, scKAN advances single-cell analysis by introducing a framework that balances computational efficiency with biological interpretability. This efficiency is demonstrated by a 15.8-fold reduction in GPU memory usage and 5.4-fold faster training compared to the teacher model scGPT, while maintaining robust performance. Furthermore, this robustness is evident across datasets with vastly different feature dimensions. While the feature space was harmonized to 2000 highly variable genes across all datasets, the model's robust performance was consistent across datasets of varying biological complexity, cellular composition, and original sequencing depth. This consistency suggests that the observed performance gains are attributable to the intrinsic properties of the scKAN framework. The integration of knowledge distillation with interpretable architecture provides a template for future developments in computational biology, potentially extending beyond single-cell analysis to other high-dimensional biological data types. While our current implementation utilizes scGPT as the teacher model, the framework's design allows for seamless integration of more advanced LLMs.

Despite these promising results, several limitations and opportunities for future development warrant discussion. First, while scKAN demonstrates potential in cell-type annotation and marker gene identification, its original KAN framework architecture currently limits its application in more complex single-cell analysis tasks such as perturbation response prediction and multi-modal data integration. Second, the comparable performance between KAN networks and MLPs on specific datasets suggests room for architectural refinement to enhance representational capacity while maintaining interpretability. Third, although scKAN effectively identifies potential therapeutic targets for specific diseases, the current drug repurposing workflow still relies on external methods for compound lead screening, indicating the need for an end-to-end solution from gene signatures to disease-specific drug candidates. Future work should address these limitations by extending the KAN framework with graph-based architectures for complex analysis tasks, incorporating multimodal learning capabilities for integrated omics analysis, and developing an end-to-end solution from target identification to lead compound screening.

Conclusions

Overall, the broader impact of scKAN extends beyond its immediate technical contributions. By providing an interpretable and efficient framework for single-cell analysis, it empowers researchers to extract meaningful biological insights from increasingly large and complex datasets. The success in PDAC drug discovery suggests that cell-type-specific therapeutic targeting could improve treatment efficacy in precision medicine for various diseases, including cancer, autoimmune disorders, and neurodegenerative diseases. Moreover, the framework's ability to identify novel gene–gene relationships and drug targets could accelerate the drug discovery process, potentially reducing the time and cost of therapeutic development.

Methods

Kolmogorov-Arnold networks

The architecture of KAN [22] leverages the Kolmogorov-Arnold representation theorem, which states that any continuous multivariate function f can be represented as a composition of a finite number of continuous univariate functions:

$$f(x) = f(x_1, \cdots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right) \quad (1)$$

where $\phi_{q,p} : [0{,}1] \to \mathbb{R}$, $\Phi_q : \mathbb{R} \to \mathbb{R}$. Unlike traditional neural networks, such as MLPs that use fixed activation functions, KAN uses learnable activation functions on the network edges. In this design, each weight parameter in KAN can be replaced by a univariate function, which is typically parameterized using spline functions. This approach provides high flexibility and can model complex functions with fewer parameters, which improves model interpretability.

Specifically, KAN uses multiple layers of learnable activation functions as network edges. In contrast to MLPs, KAN does not apply nonlinear transformations when aggregating function outputs:

$$KAN(x) = (\Phi_L \circ \Phi_{L-1} \circ \cdots \circ \Phi_1 \circ \Phi_0)(x) \quad (2)$$

The activation functions are characterized through spline functions:

$$\phi(x) = spline(x) = \textstyle\sum_i c_i B_i(x) \quad (3)$$

where c_i are learnable parameters and $B_i(x)$ are B-spline basis functions defined on a grid with density controlled by a hyperparameter interval G. Using a larger G allows more control over the spline and higher precision, but requires learning more parameters. The flexibility of splines enables the network to adaptively model complex relationships in data by adjusting their shapes. It minimizes approximation errors and enhances the network's ability to learn subtle patterns from high-dimensional datasets.

In addition to spline functions, previous studies [87] have shown that the radial basis function (RBF) can also be used to approximate the activation function $\phi(x)$ in KAN:

$$\phi(x) = \sum_i w_i \omega(\|x - c_i\|) \tag{4}$$

where w_i are learnable weights and ω is the radial basis function that depends on the distance between input x and center c_i. The Gaussian function is the most common choice for RBFs:

$$\omega(r) = exp\left(-\frac{r^2}{2h^2}\right) \tag{5}$$

where r is the radial distance and h is a hyperparameter that controls the spread of the function.

Single-cell large language model

In this study, we use scGPT [18], a SOTA foundation model for single-cell analysis, as the teacher model. We aim to transfer the representation patterns learned by scGPT, which was pre-trained on 33 M single-cell data, into the lightweight scKAN model. scGPT consists of multiple self-attention transformer layers that process tokenized genes and their expression values to extract features.

Specifically, for single-cell data containing N cells and M genes, the expression matrix can be represented as follows: $X \in \mathbb{R}^{N \times M}$, where each element x_{ij} represents the RNA molecule read count for scRNA-seq data or chromatin accessibility of a peak region for scATAC-seq data. The model then obtains gene embeddings through gene tokens:

$$emb_g = embedding\left(t_g\right) = embedding\left(\left[id\left(g_0\right), id\left(g_1\right), id\left(g_2\right), \cdots, id\left(g_M\right)\right]\right) \tag{6}$$

Additionally, the expression values are divided into B continuous intervals through binning:$[b_k, b_{k+1}]$, where $k \in \{1,2,\ldots,B\}$. The model then obtains expression embeddings through binning value tokens:

$$x_i = \begin{cases} k, & ifX_i > 0andX_i \in [b_k, b_{k+1}] \\ & 0, ifX_i = 0 \end{cases} \tag{7}$$

$$emb_x = embedding(X) = embedding([x_0, x_1, x_2, \cdots, x_M]) \tag{8}$$

The model also incorporates conditional tokens containing various meta-information related to individual genes, such as cell batch (represented by batch tokens). We use input vectors with the same dimension as the input genes to represent position-aware conditional tokens and perform embedding. This embedding is represented as:

$$emb_t = embedding(T) = embedding([t_0, t_1, t_2, \cdots, t_M]) \tag{9}$$

The embeddings are then concatenated and fed into transformer layers to extract high-dimensional features:

$$H = Transformer\left(\left[emb_g + emb_x + emb_t\right]\right) \tag{10}$$

For cell-type annotation tasks, cell-level features can be obtained through the "[CLS]" token, and class predictions are made using an MLP:

$$Output_{Teacher} = MLP\left(H_{[cls]}\right) \tag{11}$$

Distillation learning framework

To enhance the efficiency of cell-type annotation, we propose a knowledge distillation framework with domain-aware learning principles that transfers the cell-type annotation capability from a pre-trained LLM (teacher model) to a lightweight model (student model, scKAN). We design a comprehensive loss function system comprising three components to ensure effective knowledge transfer and maintain annotation accuracy.

First, we implement a knowledge distillation loss to transfer the learned representations from the teacher model to the student model. This loss combines the conventional cross-entropy loss with a soft target distribution from the teacher model:

$$\mathcal{L}_{dist} = \mathcal{L}_{CE}\left(y_s, y_{true}\right) + \alpha\mathcal{L}_{\mathcal{KL}}\left(softmax\left(\tfrac{z_s}{T}\right), softmax\left(\tfrac{z_t}{T}\right)\right) \tag{12}$$

where z_s and z_t denote the logits from student and teacher models, respectively. T is the temperature parameter that controls the softness of the probability distribution, and α balances the contribution of the distillation term.

Second, we introduce a self-entropy loss to maintain model robustness. In cell-type annotation, models may sometimes over-concentrate on certain dominant cell types while neglecting rare cell populations. It leads to a "trivial solution" in which the model's predictions are heavily biased toward specific cell types. To prevent this, we employ a self-entropy loss that encourages a balanced distribution of predictions across different cell types:

$$\mathcal{L}_{sce} = -\sum_{\mathrm{k=1}}^{\mathrm{K}} \overline{\mathrm{p_k}}\log(\overline{p_k}) \tag{13}$$

where $\overline{p_k}$ represents the mean probability of class k across the batch, and K is the number of cell types. This loss term ensures that the model maintains sensitivity to all cell types, including rare populations, by penalizing highly skewed prediction distributions.

Third, we incorporate a DDC loss to enhance the quality of learned cell-type representations. The DDC loss was initially proposed with three components: feature-assignment consistency ($\mathcal{L}_1$), assignment orthogonality ($\mathcal{L}_2$), and feature-target consistency ($\mathcal{L}_3$). In our adaptation, we modify the original DDC loss by removing the $\mathcal{L}_2$ term, which enforces orthogonality between different cluster assignments through L2 regularization. We omit this term because, in our cell-type annotation task, the orthogonality constraint is inherently satisfied through the knowledge distillation process, where the teacher model already provides well-separated cell-type representations. Therefore, our modified DDC loss focuses on two essential components:

$$\mathcal{L}_{\lceil\lceil\rfloor} = \mathcal{L}_1 + \mathcal{L}_3 \tag{14}$$

The first component, $\mathcal{L}_1$, measures the consistency between hidden features and cluster assignments using Cauchy–Schwarz divergence:

$$\mathcal{L}_1 = \sum\nolimits_{i=1}^{k-1}\sum\nolimits_{j>i}^{k} \frac{\alpha_i^T K \alpha_j}{\sqrt{\alpha_i^T K \alpha_j \alpha_i^T K \alpha_j}} \tag{15}$$

Here, K is the kernel matrix computed from hidden features using a Gaussian kernel:

$$\boldsymbol{K_{ij}} = \exp\left(-\frac{|\boldsymbol{h_i} - \boldsymbol{h_j}|^2}{2\sigma^2}\right) \tag{16}$$

where h_i represents the hidden features of sample i, and σ is the Gaussian kernel bandwidth determined based on the median distance in the batch. The second component, $\mathcal{L}_3$, optimizes the relationship between cluster assignments and ideal one-hot distributions using a similar formulation with an exponential distance matrix.

$$\mathcal{L}_3 = \sum\nolimits_{i=1}^{k-1}\sum\nolimits_{j>i}^{k} \frac{m_i^T K m_j}{\sqrt{m_i^T K m_i m_j^T K m_j}} \tag{17}$$

where $m_i = \left[m_{q,i}\right] \in R^n$ and $m_{q,i} = exp\left(-\|\alpha_q - e_i\|^2\right)$. $e_i \in R^k$ is a vector denoting the i th corner of the simplex, representing the i th cluster.

The complete training objective combines these three losses to simultaneously achieve knowledge transfer, maintain prediction balance, and optimize cell-type-specific feature representations:

$$\mathcal{L}_{total} = \mathcal{L}_{dist} + \mathcal{L}_{\mathrm{sce}} + \mathcal{L}_{ddc} \tag{18}$$

This comprehensive loss function design enables our lightweight model to effectively learn from the teacher model while maintaining robust and balanced cell-type annotation capabilities.

Dataset

PBMC dataset [28]

The PBMC dataset was generated from a healthy donor using the 10 × Genomics single-cell multiome ATAC + gene expression platform (Cell Ranger ARC 1.0.0). It comprises 9631 cells, comprising 19 major immune cell populations, with an initial feature set of 29,095 genes. The dataset is available at the 10X Genomics official website. Cell-type annotations were used as provided by the original 10 × Genomics resource. In this study, we focused on analyzing the scRNA-seq component of this multiome dataset.

Muto2021 dataset [29]

This dataset, accessed from the Gene Expression Omnibus database under accession number GSE151302, represents a multi-omics study of adult human kidney cells, combining single-nucleus ATAC-seq (snATAC-seq) and RNA-seq (snRNA-seq) data. It provides detailed insights into kidney cell heterogeneity, focusing on proximal tubule epithelial cells. The dataset contains 19,985 cells with 13 distinct cell types identified, spanning 27,146 genes. The cell-type annotations were adopted from the original publication associated with the GEO accession number. In this study, we focused on analyzing the RNA-seq component of this multi-omics dataset.

Human pancreas dataset [14]

This dataset integrates scRNA-seq data from five independent studies of human pancreas cells accessible from the Gene Expression Omnibus (GEO) database [88]. The data was strategically split into reference and query sets based on data sources. The reference set (10,600 cells) consists of two studies: Baron et al. [89] (GSE84133) and Muraro et al. [90] (GSE85241). The query set (4218 cells) comprises three studies: Xin et al. [91] (GSE81608), Segerstolpe et al. [92] (E-MTAB-5061), and Lawlor et al. [93] (GSE86473). The reference set encompasses 13 distinct cell populations, while the query set contains 11 cell types. This dataset was preprocessed by Cui et al. [18], resulting in an initial feature set of 3000 genes. Cell-type labels for all studies were used as provided in their original publications.

Mye dataset [30]

The Mye dataset was retrieved from the GEO database under accession number GSE154763, spanning nine cancer types. The reference set contains 9748 cells from six cancer types: KIDNEY, LYM, PAAD, THCA, UCEC, and cDC2. The query set includes 3430 cells from three types of cancer: ESCA, MYE, and OV-FTC. This dataset was preprocessed by Cui et al. [18], resulting in an initial feature set of 3000 genes. The dataset mainly focuses on myeloid cell populations across different cancer contexts. Cell type and cancer-type labels were used as processed and provided by the original scGPT study.

Data pre-processing

To ensure a rigorous evaluation and prevent information leakage from the test data into the training process, all data processing and feature engineering steps were conducted following a strict protocol where parameters were learned exclusively from the training (or reference) set and subsequently applied to the validation and test (or query) sets. The input for each analysis was a gene expression matrix with dimensions of cells by genes.

The processing pipeline, implemented using the Scanpy library, proceeded as follows. First, we identified the top 2000 highly variable genes (HVGs) using only the training set data, based on the "seurat_v3" method. The data matrices for all sets were then reduced to these 2000 HVGs. Following this, we performed normalization; specifically, the gene counts for each cell were normalized such that the total count per cell equaled the median of the total counts calculated across cells in the training set. This was followed by a log-transformation using $\log(X + 1)$. Next, the data underwent standardization, which was performed on a per-gene basis. The mean and variance for each of the 2000 genes were calculated across the cells in the training set, and these learned parameters were then used to scale the validation and test sets. For dimensionality reduction, principal component analysis (PCA) was performed on the scaled training data to compute the first 10 principal components, and the resulting transformation was applied to the other data splits. Finally, for UMAP visualization, a neighborhood graph was constructed on the PCA representation using the cosine metric to define cell-to-cell similarity.

Baseline

Seurat [8]

A widely used analysis pipeline that follows standard preprocessing steps, performs clustering analysis, and assigns cell-type labels to clusters using established marker gene expressions.

CellTypist [10]

An annotation method integrating a reference database of immune cell types from multiple public datasets. The approach implements regularized linear models with stochastic gradient descent for automated cell-type identification. Through analysis of immune populations across different tissues, CellTypist captures both tissue-specific and shared features of immune cell subsets.

Cellcano [13]

A supervised learning framework designed for scATAC-seq data analysis. The method uses a two-stage learning strategy to reduce the impact of data distribution differences between reference and query datasets. The approach performs well across multiple cell typing tasks while maintaining computational efficiency.

TOSICA [14]

A transformer-based model that incorporates biological knowledge into cell-type annotation. TOSICA provides accurate cell-type labels and biological interpretations by analyzing pathway and regulon information. The method demonstrates effectiveness in identifying cell states during disease progression.

Geneformer [17]

A deep learning model that learns from an extensive collection of single-cell transcriptomes using attention mechanisms. The approach captures gene regulatory relationships through self-supervised training and allows adaptation to specific biological tasks with small amounts of data.

scGPT [18]

A large-scale generative model that processes cellular gene expression patterns using transformer architecture. Like language processing systems, scGPT learns the relationships between genes across millions of cells. The model supports multiple downstream tasks through transfer learning, including cell annotation and data integration.

Training settings

We employed fivefold cross-validation for a robust evaluation of model performance on the PBMC, Muto2021, hPancreas, and Mye datasets. To prevent any data leakage between training and testing phases, we implemented a strict data partitioning protocol for each fold. Specifically, 20% of the cells were first randomly held out for each of the fivefolds as an independent test set. The remaining 80% of the cells were then further partitioned into a training set (70% of the total data) and a validation set (10% of the total data). All hyperparameter tuning and model selection were conducted using only the training and validation sets. This procedure ensures that the test set for each fold remains entirely unseen during model training and hyperparameter tuning. All reported performance metrics are the means and standard deviations computed across the five independent test sets.

The training consists of two main stages: LLM fine-tuning and knowledge distillation to scKAN. In the first stage, the LLM was fine-tuned for 10 epochs with a learning rate of 5e-5. The number of bins for continuous value discretization was set to 51. The Adam optimizer was employed with an epsilon value of 1e-4. A StepLR scheduler was implemented to adjust the learning rate during training with a predefined interval and decay ratio.

For the knowledge distillation stage, scKAN was trained for 10 epochs with a higher learning rate of 5e-4. The knowledge distillation process used a temperature of 10 and combined soft and hard targets with weights of 0.5. All experiments were conducted on NVIDIA RTX 3090 GPUs.

The model performance was evaluated using multiple metrics: accuracy, macro F1, precision, and recall. These metrics were chosen to comprehensively assess overall annotation accuracy and performance across individual cell types. All reported results represent the mean and standard deviation across the fivefold cross-validation. Detailed descriptions of the training settings and hyperparameter configurations can be found in Additional file 1: Note S5.

Gene set identification using scKAN

To identify cell-type-specific gene sets, we developed a systematic approach leveraging the trained scKAN model's interpretable architecture. We employed a two-layer scKAN structure to establish a direct mapping between cell types and gene expression patterns, enabling the extraction of cell-gene activation curves. This process consists of three main steps: curve extraction, gene set optimization, and enrichment analysis.

First, we extracted the learned curves from the final layer of scKAN for each cell type using the "plot_curve" method. These curves represent the learned relationships between gene expression and cell type. Specifically, for each gene i and cell type j, we obtained a curve by sampling 400 points within the range "[−2-2 h, 2 + 2 h]," where h is the grid interval of the radial basis functions.

We then employed a beam search algorithm to identify optimal gene sets for each cell type. The objective function was defined as the sum of pairwise cosine similarities between the curves of genes within a set:

$$S = \sum_{i,j \in G} \frac{cos(C_i, C_j)}{|G| * (|G| - 1)} \tag{19}$$

where G represents a candidate gene set, Ci and Cj are the curves for genes i and j, and $|G|$ is the size of the gene set. The beam search maintained the top-k partial solutions at each step, where k was set to 20, and the final gene set size was limited to 20 genes per cell type.

The identified gene sets were subsequently analyzed using GSEA implemented through the GSEApy package [94]. We focused on immune-related pathways using the immune gene set from the molecular signatures database [41, 95]. This analysis provided insights into the immunological processes and functions associated with cell-type-specific gene expression patterns.

Identification of cell-type-specific marker genes

To identify markers that characterize different cell types, we utilized the attribution mechanism initially designed for network pruning in KAN. The computation is implemented using the original KAN architecture to preserve the network's attribution capabilities. The attribute function generates edge scores through backward propagation from the output layer to the input layer. The scores are initialized to an identity matrix representing cell type outputs. Then, the function computes successive transformations through node-to-subnode and subnode-to-edge connections, producing attribution scores reflecting gene-cell-type associations. Detailed implementation of the attribution mechanism can be found in the original paper [22].

These attribution scores are represented as a matrix of dimensions $N \times M$, where N represents the number of cell types and M represents the number of genes. Each element (i,j) in this matrix quantifies the contribution strength of gene j to cell type i. The resulting edge scores provide a quantitative measure of gene importance for each cell type, enabling the identification of cell-type-specific marker genes. Higher attribution scores indicate stronger associations between genes and specific cell types, revealing potential marker genes for each cell population [22].

ADMET property evaluation

To evaluate the drug-likeness and potential pharmacological properties of our screened lead compounds, we performed comprehensive absorption, distribution, metabolism, excretion, and toxicity (ADMET) analyses using ADMETlab 3.0 [79], a widely used web platform for systematic ADMET property prediction.

Multiple physicochemical properties were assessed for each compound. These include molecular weight (MW), which influences drug distribution and elimination; partition coefficient (LogP) and distribution coefficient (LogD), which reflect lipophilicity and drug partitioning within the body; and aqueous solubility (LogS), which is crucial for oral bioavailability. We also examined structural features, including the number of hydrogen bond acceptors (nHA) and donors (nHD), which affect solubility and receptor binding; topological polar surface area (TPSA), which correlates with drug transport properties; and the number of rotatable bonds (nRot), which influences oral bioavailability. Additional structural characteristics such as the number of rings (nRing), maximum ring size (MaxRing), number of heteroatoms (nHet), formal charge (fChar), and number of rigid bonds (nRig) were analyzed to evaluate drug-receptor interactions and overall pharmacokinetics.

These comprehensive predictions provide valuable insights into the compounds' potential as drug candidates and help identify possible limitations in their pharmacological profiles.

Molecular docking analysis

To investigate potential interactions between screened compounds and target proteins, we performed molecular docking simulations using CB-Dock2 [80]. This advanced protein–ligand blind docking tool integrates cavity detection, docking, and homologous template fitting. We set the maximum number of binding cavities for each protein–ligand pair to 5 to ensure comprehensive coverage of potential binding sites while maintaining computational efficiency.

CB-Dock2 first detects potential binding cavities on the protein surface using a curvature-based approach. It then performs molecular docking by combining structure-based and template-based docking strategies. The structure-based docking employs AutoDock Vina to sample ligand conformations within the detected cavities, while the template-based approach leverages known protein–ligand complex structures to guide the docking process. Multiple binding poses were generated and ranked for each cavity based on their binding scores. The docking results provide insights into the potential binding modes and interaction patterns between the screened compounds and their target proteins. The top-ranked protein–ligand complexes were subsequently used as initial configurations for MD simulations to evaluate the stability of these interactions.

Molecular dynamics simulation

We employed MD simulations to investigate protein–ligand complex dynamics. The system was prepared using the OpenFF Toolkit and OpenMM framework [96]. The protein was parameterized using the AMBER ff14SB force field, while the ligand parameters were derived using the GAFF-2.11 force field. The complex was solvated in an explicit TIP3P water box extending 10 Å from the solute surface. Counter-ions (Na + and Cl-) were added to neutralize the system and achieve physiological ionic strength.

The system underwent energy minimization followed by a two-stage equilibration protocol: a 50,000-step NVT equilibration at 310.15 K using a Langevin thermostat (collision frequency of 1.0 ps^{-1}), followed by a 50,000-step NPT equilibration at 1 atm pressure using a Monte Carlo barostat. Production MD simulations were performed with a 2 fs time step, with all bonds involving hydrogen atoms constrained using the SHAKE algorithm. The temperature was maintained at 310.15 K using a Langevin thermostat. The production run was extended to 100 ns, and coordinates were saved every 1000 steps for analysis.

All simulations were performed using OpenMM [96] on GPU-accelerated computing resources with periodic boundary conditions and particle mesh Ewald for long-range electrostatic interactions.

Trajectory analysis was performed using the MDTraj [97] library to evaluate system stability and conformational changes. We calculated the RMSD of the ligands and binding site residues and the RMSF of the protein backbones. The RMSD of molecular coordinates indicates conformational changes over time and is calculated as:

$$\mathrm{RMSD} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\mathrm{r_i}(\mathrm{t}) - \mathrm{r_i^{ref}}\right|^2} \tag{20}$$

where n is the number of atoms, $r_{\mathrm{i}}(t)$ is the position of atom i at time t, and r_{i}^{ref} is the position of atom i in the initial complex structure obtained from molecular docking.

The RMSF measures the fluctuation of each residue around its average position during the simulation, calculated as:

$$\mathrm{RMSF} = \sqrt{\frac{1}{\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}|\mathrm{r_i}(\mathrm{t}) - \langle \mathrm{r_i}\rangle|^2} \tag{21}$$

where T is the total number of time steps in the simulation, $r_{\mathrm{i}}(t)$ is the position of residue i at time t, and $\langle r_{\mathrm{i}}\rangle$ is the average position of residue i over the entire simulation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03779-0.

Additional file 1. Contains supplementary texts, tables and figures.

Acknowledgements

The authors thank the editors and anonymous reviewers for their Sinsightful comments and suggestions, which helped improve the quality of the manuscript.

Peer review information

Qin Ma and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

H.H.: Methodology, Conceptualization, Software, Validation, Writing—original draft, Visualization. Z.T.: Conceptualization, Formal analysis, Investigation. G.C.: Visualization, Investigation. F.X.: Data Curation, Formal analysis. Y.H.: Formal analysis, Data Curation. Y.F.: Validation, Formal analysis. J.W.: Supervision, Funding acquisition, Resources. Y.H.: Writing—Review & Editing, Funding acquisition. Z.H.: Writing—Review & Editing, Funding acquisition, Resources, Supervision. K.T.: Writing—Review & Editing, Supervision, Funding acquisition.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62572413, 62202399, U21A20512, and 62472353); the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515012944 and 2024A1515011984); the Fundamental Research Funds for the Central Universities (Grant No. G2023KY05102); the Research Grants Council of the Hong Kong SAR (Grant No. C5052-23G, PolyU15229824, PolyU15218622, and PolyU15215623); The Hong Kong Polytechnic University (Project IDs: P0053758, P0051130, and P0052694); and the City University of Hong Kong (Dongguan) New Faculty Start-up Fund (Project IDs: B01040000138).

Data availability

All datasets used are obtained from public data repositories. The Mye dataset is publicly accessible from the GEO database using accession number GSE154763 [98]. The processed human pancreas dataset was retrieved from https://github.com/JackieHanLab/TOSICA [99]. The PBMC dataset is available at https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k [28]. The Muto2021 dataset is available at GEO with the accession number GSE151302 [100]. The codebase for this study is publicly available at Github (https://github.com/hehh77/scKAN) [101] and Zenodo (https://doi.org/10.5281/zenodo.16938598) [102] under the permissive Apache 2.0 Licence.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 April 2025 Accepted: 9 September 2025
Published online: 24 September 2025

References

1. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer. 2017;17:557–69.
2. Ramachandran P, Matchett KP, Dobie R, Wilson-Kanamori JR, Henderson NC. Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. Nat Rev Gastroenterol Hepatol. 2020;17:457–72.
3. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science. 2018. https://doi.org/10.1126/science.aaq1723.
4. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566:496–502.
5. Zhao X, Wu S, Fang N, Sun X, Fan J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. Brief Bioinform. 2020;21(5):1581–95.
6. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature. 2002;418:387–91.
7. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350:1096–101.
8. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177:1888-1902.e21.
9. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nat Rev Nephrol. 2020;16:408–21.
10 Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376:eabl5197.
11. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics. 2020;36:533–8.
12. Zhang Y, Xiang G, Jiang AY, Lynch A, Zeng Z, Wang C, et al. Metatime integrates single-cell gene expression to characterize the meta-components of the tumor immune microenvironment. Nat Commun. 2023;14:2634.
13. Ma W, Lu J, Wu H. Cellcano: supervised cell type identification for single cell ATAC-seq data. Nat Commun. 2023;14:1864.
14. Chen J, Xu H, Tao W, Chen Z, Zhao Y, Han J-DJ. Transformer for one stop interpretable cell type annotation. Nat Commun. 2023;14:223.
15. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell. 2022;4:852–66.
16. Zhao S, Zhang J, Wu Y, Luo Y, Nie Z. LangCell: Language-cell pre-training for cell identity understanding. arXiv [q-bio.GN]. 2024. Available from: http://arxiv.org/abs/2405.06708.
17. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618:616–24.
18. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. ScGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21:1470–80.
19. Yang X, Liu G, Feng G, Bu D, Wang P, Jiang J, et al. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. Cell Res. 2024;34:830–45.
20. Brunner G, Liu Y, Pascual D, Richter O, Ciaramita M, Wattenhofer R. On identifiability in transformers. arXiv [cs.CL]. 2019. Available from: http://arxiv.org/abs/1908.04211.
21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30. Available from: https://proceedings.neurips.cc/paper/7181-attention-is-all.
22. Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, et al. KAN: Kolmogorov-Arnold networks. arXiv [cs.LG]. 2024. Available from: http://arxiv.org/abs/2404.19756.
23. Tang X, Zhang J, He Y, Zhang X, Lin Z, Partarrieu S, et al. Explainable multi-task learning for multi-modality biological data analysis. Nat Commun. 2023;14:2546.
24. Dang Z, Deng C, Yang X, Wei K, Huang H. Nearest neighbor matching for deep clustering. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2021. Available from: https://doi.org/10.1109/cvpr46437.2021.01348.
25. Vikjord VV, Jenssen R. Information theoretic clustering using a k-nearest neighbors approach. Pattern Recognit. 2014;47:3070–81.
26. Kampffmeyer M, Løkse S, Bianchi FM, Livi L, Salberg A-B, Jenssen R. Deep divergence-based approach to clustering. Neural Netw. 2019;113:91–101.
27. Trosten DJ, Lokse S, Jenssen R, Kampffmeyer M. Reconsidering representation alignment for multi-view clustering. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2021. Available from: https://doi.org/10.1109/cvpr46437.2021.00131.
28. PBMC from a Healthy Donor, Single Cell Multiome ATAC Gene Expression Demonstration Data by Cell Ranger ARC 1.0.0. 10X Genomics. 2020. https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k.
29. Muto Y, Wilson PC, Ledru N, Wu H, Dimke H, Waikar SS, et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. Nat Commun. 2021;12:2190.
30. Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. Cell. 2021;184:792-809.e23.
31. Kioussis D, Ellmeier W. Chromatin and CD4, CD8A and CD8B gene expression during thymic differentiation. Nat Rev Immunol. 2002;2:909–19.
32. Munitic I, Evaristo C, Sung HC, Rocha B. Transcriptional regulation during CD8 T-cell immune responses. Adv Exp Med Biol. 2010;684:11–27.
33. Han Y, Zhang M, Li N, Chen T, Zhang Y, Wan T, et al. KLRL1, a novel killer cell lectinlike receptor, inhibits natural killer cell cytotoxicity. Blood. 2004;104:2858–66.

34. Ostermann G, Weber KSC, Zernecke A, Schröder A, Weber C. JAM-1 is a ligand of the beta(2) integrin LFA-1 involved in transendothelial migration of leukocytes. Nat Immunol. 2002;3:151–8.
35. Wu J, Wang W, Li Z, Ye X. The prognostic and immune infiltration role of ITGB superfamily members in non-small cell lung cancer. Am J Transl Res. 2022;14:6445–66.
36. Masuho I, Balaji S, Muntean BS, Skamangas NK, Chavali S, Tesmer JJG, et al. A global map of G protein signaling regulation by RGS proteins. Cell. 2020;183:503-521.e19.
37. Wei D, Zhan W, Gao Y, Huang L, Gong R, Wang W, et al. Rab31 marks and controls an ESCRT-independent exosome pathway. Cell Res. 2021;31:157–77.
38 Mogensen TH. Pathogen recognition and inflammatory signaling in innate immune defenses. Clin Microbiol Rev. 2009;22:240–73 Table of Contents.
39. Vereecke L, Beyaert R, van Loo G. The ubiquitin-editing enzyme A20 (TNFAIP3) is a central regulator of immunopathology. Trends Immunol. 2009;30:383–91.
40. Chen Y, Zhang J, Cui W, Silverstein RL. CD36, a signaling receptor and fatty acid transporter that regulates immune cell metabolism and fate. J Exp Med. 2022;219. Available from: https://doi.org/10.1084/jem.20211314.
41. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
42. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS ONE. 2009;4:e6098.
43. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun. 2005;6:319–31.
44. Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, Grumont RJ, et al. Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat Immunol. 2006;7:274–83.
45. Kaech SM, Hemby S, Kersh E, Ahmed R. Molecular and functional profiling of memory CD8 T cell differentiation. Cell. 2002;111:837–51.
46. Belyaev NN, Biró J, Athanasakis D, Fernandez-Reyes D, Potocnik AJ. Global transcriptional analysis of primitive thymocytes reveals accelerated dynamics of T cell specification in fetal stages. Immunogenetics. 2012;64:591–604.
47. Künzli M, Masopust D. CD4+ T cell memory. Nat Immunol. 2023;24:903–14.
48. van der Leun AM, Thommen DS, Schumacher TN. CD8+ T cell states in human cancer: insights from single-cell analysis. Nat Rev Cancer. 2020;20:218–32.
49. Chen Y, Xu Z, Sun H, Ouyang X, Han Y, Yu H, et al. Regulation of CD8+ T memory and exhaustion by the mTOR signals. Cell Mol Immunol. 2023;20:1023–39.
50. Hoft DF, Blazevic A, Selimovic A, Turan A, Tennant J, Abate G, et al. Safety and immunogenicity of the recombinant BCG vaccine AERAS-422 in healthy BCG-naïve adults: a randomized, active-controlled, first-in-human phase 1 trial. EBioMedicine. 2016;7:278–86.
51. Gaucher D, Therrien R, Kettaf N, Angermann BR, Boucher G, Filali-Mouhim A, et al. Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med. 2008;205:3119–31.
52. Chen X, Hu X, Chen F, Yan J. Expansion and polarization of human γδT17 cells *in vitro* from peripheral blood mononuclear cells. Bio-Protoc. 2024;14:e4914.
53. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, Stewart BJ, Ferdinand JR, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science. 2018;361:594–9.
54. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell. 2017;169:1342-1356.e16.
55. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356. Available from: https://doi.org/10.1126/science.aah4573.
56. Pallett LJ, Swadling L, Diniz M, Maini AA, Schwabenland M, Gasull AD, et al. Tissue CD14+CD8+ T cells reprogrammed by myeloid cells and modulated by LPS. Nature. 2023;614:334–42.
57. Latenstein AEJ, van der Geest LGM, Bonsing BA, Groot Koerkamp B, Haj Mohammad N, de Hingh IHJT, et al. Nationwide trends in incidence, treatment and survival of pancreatic ductal adenocarcinoma. Eur J Cancer. 2020;125:83–93.
58. Garrido-Laguna I, Hidalgo M. Pancreatic cancer: from state-of-the-art treatments to promising novel therapies. Nat Rev Clin Oncol. 2015;12:319–34.
59. Noguchi K, Konno M, Koseki J, Nishida N, Kawamoto K, Yamada D, et al. The mitochondrial one-carbon metabolic pathway is associated with patient survival in pancreatic cancer. Oncol Lett. 2018; Available from: https://doi.org/10.3892/ol.2018.8795.
60. Yang D, Sun X, Moniruzzaman R, Wang H, Citu C, Zhao Z, et al. Genetic deletion of galectin-3 inhibits pancreatic cancer progression and enhances the efficacy of immunotherapy. Gastroenterology. 2024;167:298–314.
61. Kha M-L, Hesse L, Deisinger F, Sipos B, Röcken C, Arlt A, et al. The antioxidant transcription factor Nrf2 modulates the stress response and phenotype of malignant as well as premalignant pancreatic ductal epithelial cells by inducing expression of the ATF3 splicing variant ΔZip2. Oncogene. 2019;38:1461–76.
62. Lo A, Li C-P, Buza EL, Blomberg R, Govindaraju P, Avery D, et al. Fibroblast activation protein augments progression and metastasis of pancreatic ductal adenocarcinoma. JCI Insight. 2017;2. Available from: https://doi.org/10.1172/jci.insight.92232.
63. Wang P, Pan Y, Zhang Y, Chen C, Hu J, Wang X. Role of interferon-induced transmembrane protein family in cancer progression: a special focus on pancreatic cancer. Med Oncol. 2024;41:85.
64. Cohen AS, Li J, Hight MR, McKinley E, Fu A, Payne A, et al. Tspo-targeted PET and optical probes for the detection and localization of premalignant and malignant pancreatic lesions. Clin Cancer Res. 2020;26:5914–25.
65. Wu X, Yu R, Yang M, Hu Y, Tang M, Zhang S, et al. Integrated analysis of glutathione metabolic pathway in pancreatic cancer. Front Cell Dev Biol. 2022;10:896136.
66. Liang Y, Chen W, Tang Y, Chen M. Identification of the genetic association between type-2-diabetes and pancreatic cancer. Biochem Genet. 2023;61:1143–62.

67. Nordgård O, Lapin M, Tjensvoll K, Oltedal S, Edland KH, Neverdahl NB, et al. Prognostic value of disseminated tumor cells in unresectable pancreatic ductal adenocarcinoma: a prospective observational study. BMC Cancer. 2022;22:609.
68. Roy R, Zurakowski D, Wischhusen J, Frauenhoffer C, Hooshmand S, Kulke M, et al. Urinary TIMP-1 and MMP-2 levels detect the presence of pancreatic malignancies. Br J Cancer. 2014;111:1772–9.
69. An W, Ben Q-W, Chen H-T, Zheng J-M, Huang L, Li G-X, et al. Low expression of IGFBP7 is associated with poor outcome of pancreatic ductal adenocarcinoma. Ann Surg Oncol. 2012;19:3971–8.
70. Wei B, Guo C, Liu S, Sun M-Z. Annexin A4 and cancer. Clin Chim Acta. 2015;447:72–8.
71. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45:D170–6.
72. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52:D1180–92.
73. He H, Chen G, Chen CY-C. NHGNN-DTA: a node-adaptive hybrid graph neural network for interpretable drug-target binding affinity prediction. Bioinformatics. 2023;39. Available from: https://doi.org/10.1093/bioinformatics/btad355.
74. Chen G, He H, Lv Q, Zhao L, Chen CY-C. MMFA-DTA: multimodal feature attention fusion network for drug-target affinity prediction for drug repurposing against SARS-CoV-2. J Chem Theory Comput. 2024; Available from: https://doi.org/10.1021/acs.jctc.4c00663.
75. Wu H, Liu J, Jiang T, Zou Q, Qi S, Cui Z, et al. AttentionMGT-DTA: a multi-modal drug-target affinity prediction using graph transformer and attention mechanism. Neural Netw. 2024;169:623–36.
76. Park M, Kim H. Anti-cancer mechanism of docosahexaenoic acid in pancreatic carcinogenesis: a mini-review. J Cancer Prev. 2017;22:1–5.
77. Park M, Lim JW, Kim H. Docoxahexaenoic acid induces apoptosis of pancreatic cancer cells by suppressing activation of STAT3 and NF-κB. Nutrients. 2018;10:1621.
78. Xie C-K, Liao C-Y, Lin H-Y, Wu Y-D, Lu F-C, Huang X-X, et al. Sulindac (K-80003) with nab-paclitaxel and gemcitabine overcomes drug-resistant pancreatic cancer. Mol Cancer. 2024;23:215.
79. Fu L, Shi S, Yi J, Wang N, He Y, Wu Z, et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Res. 2024;52:W422-31.
80. Liu Y, Yang X, Gan J, Chen S, Xiao Z-X, Cao Y. CB-dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting. Nucleic Acids Res. 2022;50:W159–64.
81 Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB Protein Data Bank (Rcsb.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res. 2023;51:D488-508.
82. Aertgeerts K, Levin I, Shi L, Snell GP, Jennings A, Prasad GS, et al. Structural and kinetic analysis of the substrate specificity of human fibroblast activation protein alpha. J Biol Chem. 2005;280:19441–4.
83. Sörme P, Arnoux P, Kahl-Knutsson B, Leffler H, Rini JM, Nilsson UJ. Structural and thermodynamic studies on cation-Pi interactions in lectin-ligand complexes: high-affinity galectin-3 inhibitors through fine-tuning of an arginine-arene interaction. J Am Chem Soc. 2005;127:1737–43.
84. Gustafsson R, Jemth A-S, Gustafsson NMS, Färnegårdh K, Loseva O, Wiita E, et al. Crystal structure of the emerging cancer target MTHFD2 in complex with a substrate-based inhibitor. Cancer Res. 2017;77:937–48.
85. Convery MA. Factor Xa complex with GTC000101. Worldwide Protein Data Bank; 2015. Available from: https://doi.org/10.2210/pdb4y6d/pdb.
86. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
87. Li Z. Kolmogorov-Arnold networks are radial basis function networks. arXiv [cs.LG]. 2024. Available from: http://arxiv.org/abs/2405.06721.
88. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
89. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst. 2016;3:346-360.e4.
90. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst. 2016;3:385-394.e3.
91. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, et al. RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metab. 2016;24:608–15.
92. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24:593–607.
93. Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, Sivakamasundari V, et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. Genome Res. 2017;27:208–22.
94. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023. https://doi.org/10.1093/bioinformatics/btac757.
95. Tang Z, Chen G, Chen S, He H, You L, Chen CY-C. Knowledge-based inductive bias and domain adaptation for cell type annotation. Commun Biol. 2024;7:1440.
96. Eastman P, Galvelis R, Peláez RP, Abreu CRA, Farr SE, Gallicchio E, et al. OpenMM 8: molecular dynamics simulation with machine learning potentials. J Phys Chem B. 2024;128:109–16.
97. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophys J. 2015;109:1528–32.
98. Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. Gene Expression Omnibus. 2021. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154763. Cited 2021 Feb 4.

99. Chen J, Xu H, Tao W, Chen Z, Zhao Y, Han J-DJ. Transformer for one stop interpretable cell type annotation. Github. 2023. Available from: https://github.com/JackieHanLab/TOSICA. Cited 2023 Jan 14.
100. Muto Y, Wilson PC, Humphreys BD. Single cell transcriptional and chromatin accessibility profiling on the human adult kidneys. Gene Expression Omnibus. 2021. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151302. Cited 2021 Feb 21.
101. He H, Tang Z, Chen G, Xu F, Hu Y, Feng Y, et al. scKAN: interpretable single-cell analysis for cell-type-specific gene discovery and drug repurposing via Kolmogorov-Arnold networks. Github. 2025. Available from: https://github.com/hehh77/scKAN. Cited 2025 Feb 2.
102. He H, Tang Z, Chen G, Xu F, Hu Y, Feng Y, et al. scKAN: interpretable single-cell analysis for cell-type-specific gene discovery and drug repurposing via Kolmogorov-Arnold networks. Zenodo. 2025. Available from: https://doi.org/10.5281/zenodo.16938598. Cited 2025 Aug 25.

Publisher's Note