Li *et al. Genome Biology* (2025) 26:402
https://doi.org/10.1186/s13059-025-03862-6

Genome Biology

METHODOLOGY Open Access

scKGBERT: a knowledge-enhanced foundation model for single-cell transcriptomics

Yang Li[1†], Guanyu Qiao[1,2†], Hongli Du[3], Xin Gao[4,5,6] and Guohua Wang[1,2*]

[†]Yang Li and Guanyu Qiao contributed equally to this work.

*Correspondence: Guohua Wang ghwang@nefu.edu.cn

Full list of author information is available at the end of the article

Abstract

Single-cell transcriptomics enables precise characterization of cellular heterogeneity, but current pre-trained models relying solely on expression data fail to capture gene associations. We present scKGBERT, a knowledge-enhanced foundation model integrating 41 M single-cell RNA-seq profiles and 8.9 M protein–protein interactions to jointly learn gene and cell representations. scKGBERT employs Gaussian attention to emphasize key genes and improve biomarker identification, achieving superior performance across gene annotation, drug response, and disease prediction tasks. scKGBERT enhances biological interpretability and offers a powerful resource for precision medicine and disease mechanism discovery.

Keywords Single-cell transcriptomics, Pre-trained language model, Pre-training model, Knowledge graph

Background

Single-cell sequencing technology has revolutionized cellular analysis by allowing data collection at the single-cell level, thereby circumventing the limitations of conventional bulk sequencing techniques that often obscure inter-cellular heterogeneity. Central to this domain is single-cell RNA sequencing (scRNA-seq), encompassing reverse transcription, amplification, and high-throughput sequencing of cellular mRNA. This approach enables detailed molecular profiling and characterization of cellular transcriptomes across diverse biological systems. As a result, scRNA-seq provides deep insights into disease etiology and plays a pivotal role in propelling precision medicine forward.

Given the exponential growth of single-cell sequencing data, there is an increasing demand for advanced computational techniques to effectively analyze and interpret this vast information. Large pre-trained language models have recently demonstrated significant success in Natural Language Processing (NLP) [1] and Computer Vision (CV) [2, 3] through large-scale pre-training and fine-tuning for specific tasks. Acknowledging the structural similarities between natural language and genetic sequences, models such as scBERT [4], Geneformer [5], and scFoundation [6] have been specifically designed for the analysis of single-cell transcriptome data, aiming to improve the understanding of cellular functions and gene expression patterns.

Despite possessing a vast parameter space and the ability to uncover inherent patterns in gene expression profiles through unsupervised learning of transcriptome sequences, current single-cell transcriptome pre-training models suffer from a significant lack of interpretability, which is crucial for deriving actionable biological insights. To overcome this limitation, we propose enhancing single-cell transcriptome pre-trained language models by integrating Protein-Protein Interaction (PPI) biological knowledge [7]. Protein interactions can significantly influence gene expression and cellular functions [8, 9], and as such, incorporating PPI network information as a collaborative knowledge source can substantially improve the biological relevance between gene and cell expression. This integration facilitates a more nuanced understanding of gene expression dynamics and intracellular regulatory mechanisms, providing deeper insights into complex biological processes. We introduce scKGBERT (Fig. 1), a knowledge-enhanced, large-scale pre-trained language model built upon 41 million single-cell transcriptome profiles. For the first time, the scKGBERT model integrates a biological knowledge graph containing 8.9 million regulatory relationships into the pre-training process, enabling simultaneous pre-training on both single-cell transcriptome sequences and biological knowledge. This novel integration allows scKGBERT to achieve state-of-the-art performance across 12 downstream tasks, demonstrating its robustness and versatility in biological research.

scKGBERT enables substantial advances in single-cell transcriptomic analysis across four critical dimensions. First, its knowledge-enhanced architecture markedly improves the annotation of both cells and genes. By integrating biological prior knowledge of gene regulation, scKGBERT enhances the decoding of gene expression patterns, enabling more accurate learning of cellular and genomic features under few-shot and zero-shot conditions. This results in improved performance on downstream bioinformatics tasks, highlighting the model's broad applicability in genomic and transcriptomic research. Second, scKGBERT demonstrates strong efficacy in resolving cancer heterogeneity and predicting drug responses. By capturing the intricate transcriptional landscape of the tumor microenvironment, the model overcomes key limitations of conventional approaches in representing cellular diversity and treatment variability. Enrichment analysis of high-confidence gene representations reveals activation of oncogenic pathways and tumor-specific regulatory circuits, enabling the prioritization of candidate therapeutic targets and informing precision oncology interventions. Third, the model significantly advances the discovery of disease-associated biomarkers. By leveraging biological priors, scKGBERT demonstrates enhanced capability in identifying key molecular determinants of disease, thereby improving interpretability and understanding of disease pathogenesis. It contributes to the development of more reliable and clinically relevant molecular signatures. Finally, scKGBERT shows strong cross-platform and cross-disease generalizability, underscoring its potential for integrative single-cell data analysis in diverse biomedical contexts. Its robustness across varied datasets and disease states makes it a promising tool for scalable, translational applications. In summary, scKGBERT offers a biologically meaningful framework that integrates knowledge-guided representation learning with a Gaussian attention mechanism, enabling accurate gene and cell annotation, robust identification of disease-associated genes, and mechanism interpretation of single-cell data. The incorporation of Gaussian attention further enhances model interpretability, providing a transparent framework to support future biological discovery, disease mechanism elucidation, and therapeutic development, particularly in oncology.

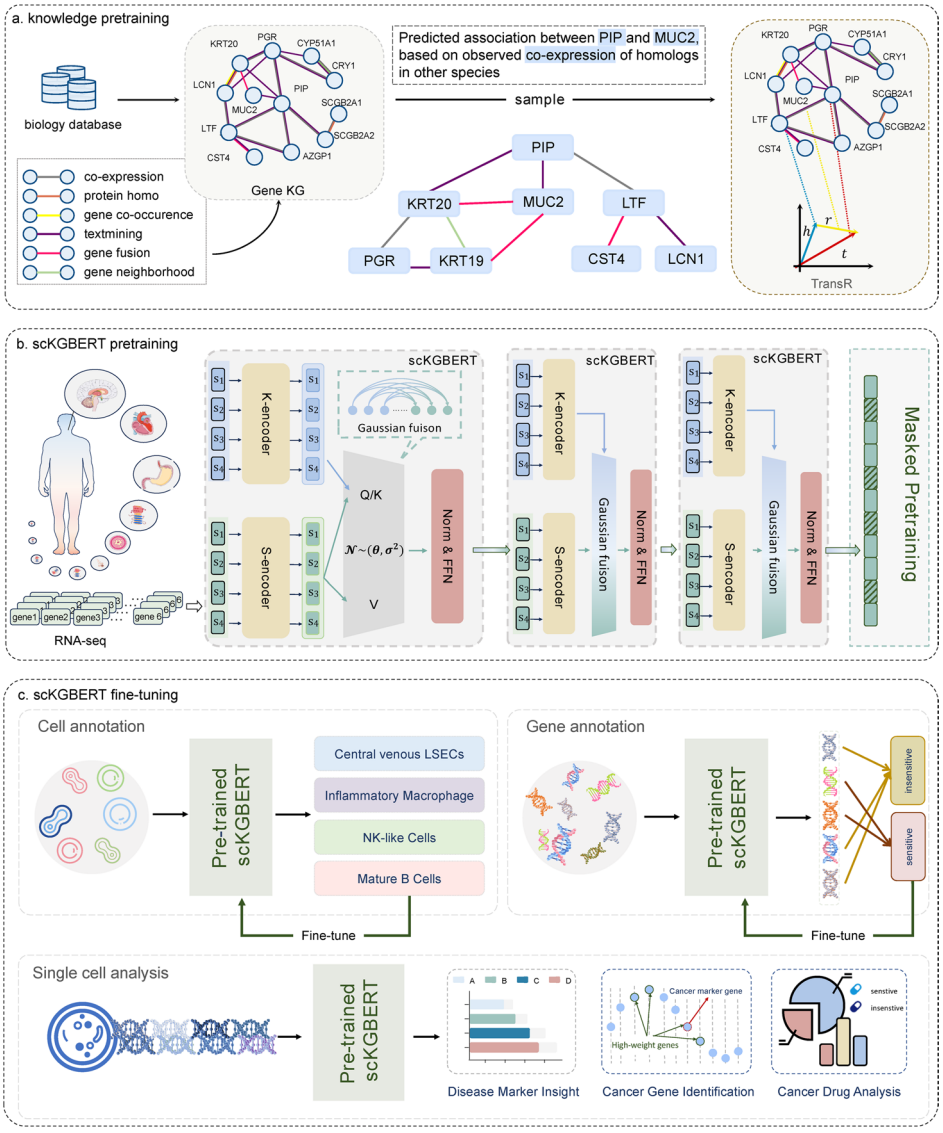

Fig. 1 An overview of the scKGBERT framework. **a** Pre-training of the biological knowledge graph. A TransR-based knowledge graph embedding approach is employed to model protein-protein interactions (PPIs), capturing both structural and semantic relationships among genes. The procedure encodes rich biological context into the initial gene embeddings, providing a foundation for modeling gene regulatory mechanisms and functional associations. **b** Pre-training phase of scKGBERT. scKGBERT is pre-trained using a masked language modeling (MLM) objective, where single-cell RNA-seq expression profiles are treated as sentences in a biological language—each gene as a word, and expression levels as contextual semantics. Unlike conventional models, scKGBERT introduces a dual-encoder architecture: a sequence encoder and a knowledge encoder, which are integrated via a Gaussian cross-attention mechanism. This mechanism highlights biologically meaningful expression patterns while incorporating regulatory priors into the learned representations. **c** Fine-tuning phase of scKGBERT. The model is adapted to downstream tasks via supervised fine-tuning, allowing the transfer of generalized knowledge to data-scarce scenarios. We benchmarked scKGBERT across 12 downstream tasks spanning gene-level and cell-level predictions, demonstrating its robust generalizability and biological interpretability across diverse single-cell applications

Results

Overview of scKGBERT

Single-cell RNA sequencing (scRNA-seq) has transformed the landscape of biology and medicine by enabling high-resolution profiling of transcriptomes at the individual cell level, thereby offering unprecedented insights into cellular heterogeneity, differentiation trajectories, and disease mechanisms. These advances have led to the discovery of

previously unrecognized cell types and enhanced our understanding of tissue organization and function. To fully harness the potential of scRNA-seq, pre-trained language models tailored to single-cell transcriptomics are emerging as powerful tools for deciphering complex biological systems at cellular resolution.

Here, we introduce scKGBERT, a knowledge-enhanced pre-trained language model designed specifically for single-cell transcriptomic data (Fig. 1). The framework consists of three core stages: (1) biological knowledge graph pre-training, (2) multimodal scKGBERT pre-training, and (3) fine-tuning on downstream tasks. Integrating structured biological knowledge is critical to improving the representation power and generalizability of large-scale pre-trained models, as supported by recent advances in biomedical NLP and systems biology [1, 7, 10, 11]. To this end, we first construct a gene-centric biological knowledge graph derived from the STRING database, which consolidates experimental evidence, computational predictions, and curated biological interactions across more than 5,000 species, encompassing over 30 billion interactions [12–14]. An unsupervised graph representation learning strategy is applied to embed this knowledge, enriching gene representations with contextual biological semantics. During the model pre-training phase, scKGBERT consumes large-scale unlabeled scRNA-seq data and leverages a modular architecture comprising three components: an RNA sequence encoder (S-encoder), a knowledge graph encoder (K-encoder), and a Gaussian cross-attention layer that fuses expression and knowledge embeddings. This dual-stream design allows the model to simultaneously capture transcriptional dynamics and regulatory context, offering a novel paradigm distinct from prior single-cell or language-based pre-training models. In alignment with the masking strategy employed in the BERT architecture, scKGBERT randomly masks 15% of gene expression tokens from single-cell transcriptomic sequences and learns to predict the masked elements using the surrounding unmasked context. It enables the model to capture contextual dependencies and regulatory relationships among genes in an unsupervised manner. Following large-scale pre-training on unlabeled single-cell RNA-seq data, scKGBERT is fine-tuned on a suite of downstream tasks spanning both gene- and cell-level resolutions, thereby facilitating its effective application to complex biological questions. The pre-training phase utilized a corpus of 41.83 million single-cell transcriptomes, encompassing 30 primary tissues and more than 60 distinct cell types, all derived from publicly available datasets. To ensure data integrity, a series of scalable quality control filters were applied to exclude cells exhibiting elevated mutation loads, transcriptomic artifacts, or evidence of cellular damage, thereby minimizing noise and enhancing training signal fidelity [5].

Evaluation of scKGBERT on gene annotation

We first evaluated the gene-level representation capability of scKGBERT through three biologically essential annotation tasks, aimed at assessing its ability to capture complex regulatory and functional patterns. Specifically, we investigated its performance on gene dosage sensitivity prediction, epigenetic marker identification, and transcription factor regulatory inference-tasks that demand a fine-grained understanding of gene regulation and interaction networks. For each task, the model was fine-tuned on dedicated benchmark datasets and systematically compared against state-of-the-art large-scale pre-trained models, including scGPT, scFoundation, and Geneformer, alongside conventional machine learning baselines. To further quantify the impact of knowledge

integration, we conducted gene association mining, demonstrating that the knowledge-augmented architecture of scKGBERT effectively captures biologically meaningful gene-gene relationships. The evaluation results underscore scKGBERT's superior capacity to decode gene function and regulatory logic, while also generating multi-omics evidence-including dosage sensitivity profiles, chromatin accessibility features, and inferred transcriptional regulatory networks-that collectively provide critical insights into the molecular underpinnings of complex diseases.

In the first task, we evaluated scKGBERT's ability to predict gene dosage sensitivity—a critical determinant of disease pathophysiology, as dosage imbalances can disrupt transcriptional regulation and impair cellular function across tissues [15, 16]. The dosage-sensitive gene labels were derived from previous studies that integrated population-scale genetic variation data and computational predictions, including LOEUF scores and haploinsufficiency models[17–19], and fine-tuned them on a dataset of 10,000 single-cell transcriptomes. scKGBERT outperformed all existing large-scale pre-trained models including scGPT, scFoundation, and Geneformer, and classical machine learning approaches in terms of AUC score (Fig. 2a), highlighting its superior capacity to capture gene dosage-related transcriptional signals. To further evaluate model robustness, we examined the classification performance for dosage-sensitive versus insensitive transcription factors (TFs) (Fig. 2c). scKGBERT outperformed all baselines, achieving the highest median F1-score (0.94) and the lowest interquartile variability, indicating both accuracy and stability. In contrast, Geneformer, while achieving a relatively high median score, showed greater dispersion across trials. Pre-trained models such as scGPT and scFoundation demonstrated competitive yet slightly lower performance, whereas traditional classifiers like SVM, Random Forest (RF), and especially Logistic Regression (LR) exhibited reduced accuracy and higher false positive rates. The consistently superior performance of scKGBERT can be attributed to its ability to integrate structured biological knowledge from protein-protein interaction networks and encode gene-gene regulatory dependencies via a Gaussian cross-attention mechanism. This architecture allows the model to learn biologically meaningful representations that are crucial for distinguishing dosage-sensitive TFs, which often play pivotal roles in developmental processes, oncogenesis, and dosage-compensated systems.

We next evaluated the performance of scKGBERT across pre-training datasets of increasing size and tissue diversity, with the goal of assessing its scalability and generalization capacity. To this end, we selected two biologically distinct cell populations (Brain Tissue Cells and Human Multi-Tissue Cells) and incrementally scaled the pre-training dataset from 1,000 to 10 million single-cell transcriptomes (Fig. 2d). Model performance improved consistently for both cell types as data volume increased. Notably, Human Multi-Tissue Cells exhibited a higher baseline ROC score and demonstrated a more pronounced performance gain compared to Brain Tissue Cells. For example, at a training scale of 1 million cells, scKGBERT achieved an ROC score of approximately 0.784 on Human Multi-Tissue Cells, whereas performance on Brain Tissue Cells plateaued at 0.661, highlighting the model's capacity to generalize more effectively in heterogeneous tissue contexts and the importance of large-scale, diverse data for robust representation learning. Confusion matrix analysis (Fig. 2f) further confirmed scKGBERT's superiority in recall and precision. The model consistently outperformed Geneformer and scFoundation [6] in identifying dosage-sensitive genes, particularly under noisy or

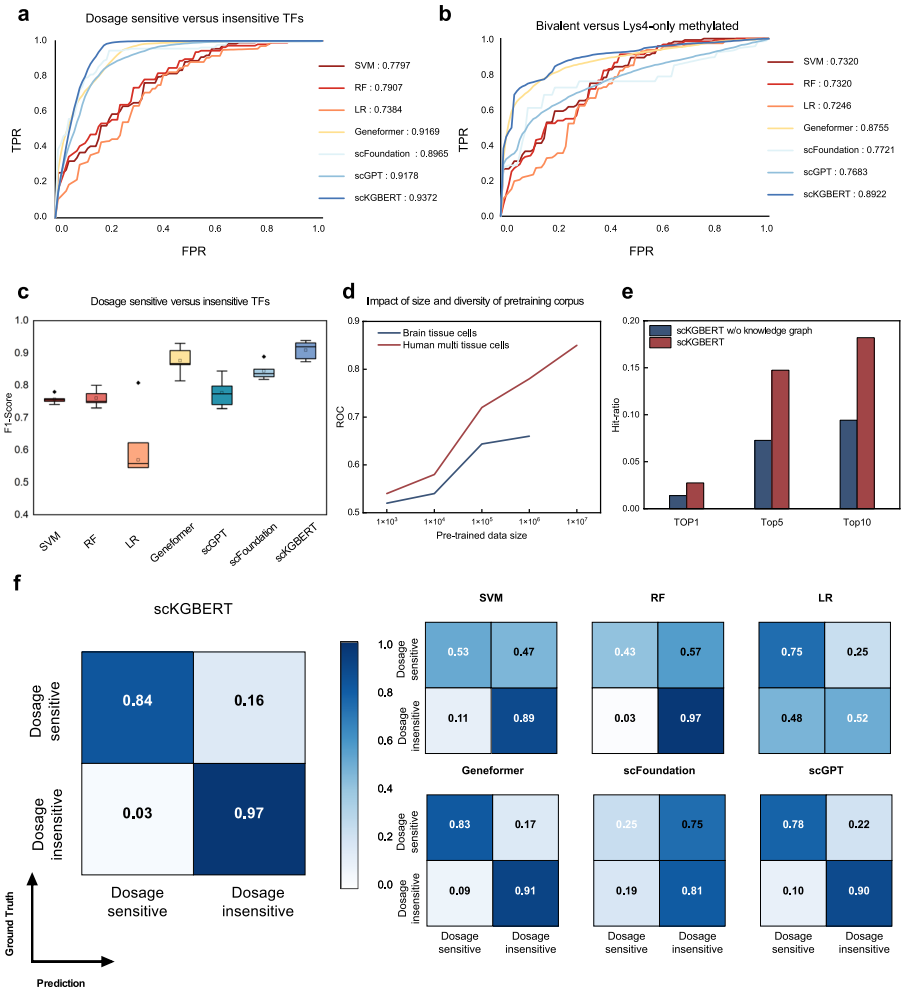

Fig. 2 scKGBERT for efficient gene annotation. **a** Gene dosage sensitivity ROC curve performance comparison. The models are fine-tuned with limited data to differentiate the dosage sensitivity of transcription factors. Compared with the traditional or current state-of-the-art methods, scKGBERT significantly outperformed other approaches. **b** Comparison of ROC curves for bivalent and Lys4-only Methylation prediction. **c** Comparison of gene dose sensitivity classification F1-scores using box plots. **d** Evaluation of the impact of pre-training data scale on the generalization performance of the scKGBERT model. We selected the "brain tissue cells" and "human multi-tissue cells" from scKGBERT-41M to examine the impact on the model's generalization capability. **e** To compare knowledge-enhanced (scKGBERT) and non-knowledge-enhanced (scKGBERT w/o knowledge graph) pre-trained models, the bar chart shows the ratio of similar marker genes within the top *k*. The x-axis represents the Top *k* (the top 1, 5, 10 most similar genes), and the y-axis represents their proportion in the total dataset. **f** Confusion matrices of the six comparison methods and scKGBERT model

heterogeneous expression conditions. It achieved lower false positive rates and higher specificity, supporting its robustness in biologically complex and imbalanced scenarios. In The superior performance of scKGBERT in distinguishing dosage-sensitive from insensitive transcription factors suggests that the model captures fine-grained regulatory cues that underlie gene dosage effects. These effects are often implicated in developmental disorders and cancer, where precise transcriptional control is critical. By integrating prior knowledge from protein-protein interaction networks and modeling gene-gene dependencies via cross-attention, scKGBERT effectively reveals regulatory signatures that may be overlooked by conventional or purely data-driven models. This highlights its potential utility in uncovering transcriptional vulnerabilities and therapeutic targets in dosage-sensitive disease contexts.

In the second task, we evaluated scKGBERT's ability to predict genes carrying bivalent chromatin marks, a key epigenetic signature involved in transcriptional poising during early development. The model was fine-tuned on a dataset comprising 15,000 embryonic stem cells from the panglaoDB database [20], with annotations distinguishing bivalently marked genes from those marked solely by H3K4me3 [21]. Chromatin state labels distinguishing bivalent genes from H3K4me3-only marked genes were obtained through ChIP-seq experiments in embryonic stem cells[22]. ROC curve analysis (Fig. 2b and Additional file 1: Fig. S1a-b) revealed that scKGBERT achieved the highest overall performance, with a steeper curve and higher true positive rate (TPR) across the full range of false positive rates (FPR), compared to all baseline models. Geneformer also performed competitively, though with slightly reduced accuracy. In contrast, traditional machine learning models such as SVM and Random Forest (RF) showed only moderate predictive power, and Logistic Regression (LR) underperformed significantly-failing to distinguish subtle epigenetic differences. Notably, scGPT and scFoundation exhibited improved performance over classical baselines, but remained inferior to scKGBERT in both sensitivity and specificity. The enhanced performance of scKGBERT is likely due to its integration of prior regulatory knowledge and Gaussian cross-attention mechanism, which together enable the model to capture complex gene-regulatory relationships embedded in chromatin context. These findings demonstrate that scKGBERT not only excels at identifying bivalent epigenetic marks, but also captures the nuanced chromatin-state distinctions critical for understanding developmental gene regulation. Its ability to differentiate between poised and active promoter states further illustrates its potential for decoding chromatin dynamics in stem cell systems and early embryogenesis.

In the final task, we tested scKGBERT's capacity to predict long-range versus short-range transcription factors (TFs) during the differentiation of induced pluripotent stem cells (iPSCs) into cardiomyocytes. Regulatory range annotations were obtained from enhancer–promoter interaction data and chromatin conformation analyses, which identified short-range and long-range transcription factor targe[23]. Using a dataset comprising 12,000 single-cell transcriptomes collected across differentiation stages [24], we fine-tuned the model to classify TFs exclusively based on transcriptomic patterns, without incorporating external genomic modalities such as ChIP-seq or chromatin conformation data. Despite the limited input, scKGBERT significantly outperformed all baseline models, including traditional classifiers and other large-scale pre-trained models (Additional file 1: Fig. S1b-d). Competing models frequently approached random performance, underscoring the difficulty of this task. In contrast, scKGBERT successfully learned to differentiate TFs based on their regulatory range, capturing subtle transcriptional signatures embedded in expression dynamics. The result highlights scKGBERT's ability to infer regulatory hierarchies and spatial coordination within gene regulatory networks using only single-cell RNA-seq data. Such capability offers a scalable and data-efficient framework for deciphering transcriptional control, particularly in scenarios where multi-omics data are unavailable or cost-prohibitive.

Additionally, we conducted ablation experiments to investigate the contribution of the knowledge graph in enhancing scKGBERT's performance. By comparing scKGBERT with its variant excluding the knowledge graph (scKGBERT w/o knowledge graph), we evaluated the models' ability to acquire single-cell representations and identify marker genes in single cells. Using the Muraro dataset, particularly focusing on Alpha-cells,

we evaluated the cosine similarity of gene embeddings. The results (Fig. 2e) showed that the model incorporating the knowledge graph nearly doubled the performance in detecting the top relevant genes compared to the variant without the graph. This finding indicates that the knowledge graph significantly enhances the model's ability to recognize subtle gene associations, providing a rich framework of background knowledge that facilitates better understanding of complex biological signals. The high degree of relevance in gene embeddings contributed to scKGBERT's enhanced sensitivity to differences between cells, resulting in exceptional performance in cell annotation tasks. These findings further validate scKGBERT's robustness in accurately identifying gene associations, enriching the exploration of disease mechanisms and supporting biotechnological advancements.

Cell annotation and cross-tissue generalization of scKGBERT

In multicellular organisms, individual cells exhibit distinct phenotypes, surface markers, and transcriptional programs that underpin their diverse functional roles in development and physiology [25]. Single-cell annotation, which includes analyzing transcriptomic profiles at single-cell resolution, is essential for accurately identifying and classifying cell types [26, 27]. This process enables the provision of deep insights into cellular states, lineage differentiation, metabolic regulation, and signaling dynamics. To evaluate scKGBERT's ability to encode cell-specific transcriptional features and functional identities, we fine-tuned the model on cell annotation tasks and evaluated its alignment with expert-curated single-cell labels. This evaluation served to determine how effectively the model captures biologically relevant patterns in gene expression and reflects the underlying heterogeneity within single-cell transcriptomes.

To evaluate the representation capacity of scKGBERT at the cell level, we performed cell annotation tasks across ten tissue types and compared its performance with several state-of-the-art models (Fig. 3a). scKGBERT achieved the highest average accuracy (0.9516), outperforming all baselines across diverse tissues including brain, kidney, immune, and spleen. In contrast, scDeepSort [28] and CaSTLe [29], which rely on supervised learning, showed lower generalizability, particularly in complex tissues such as spleen and placenta. scDeepSort performed the worst overall (0.5854), while CaSTLe reached a moderate 0.8867. Among pre-trained models, Geneformer performed well in liver, lung, and pancreas (0.9466 average), but exhibited less consistency than scKGBERT. scGPT (0.9165) and scFoundation (0.8654) also showed competitive yet variable performance. The accompanying heatmap highlights scKGBERT's consistent performance across all tissue types, while other models showed sharp drops in specific conditions. These results confirm that scKGBERT's self-supervised, knowledge-enhanced pre-training enables robust, label-efficient annotation and generalization across transcriptionally diverse tissues.

To further assess the quality of scKGBERT's learned representations, we visualized its performance on kidney and other tissue-derived cell types (Fig. 3b; Additional file 1: Fig. S2-3). UMAP and heatmap projections of the model's output revealed well-separated clusters corresponding to distinct cell types, highlighting its ability to capture meaningful transcriptional signatures. The sharp inter-cluster boundaries suggest that scKGBERT not only accurately identifies cell types, but also effectively encodes higher-order relationships and transcriptional heterogeneity across cell populations. This

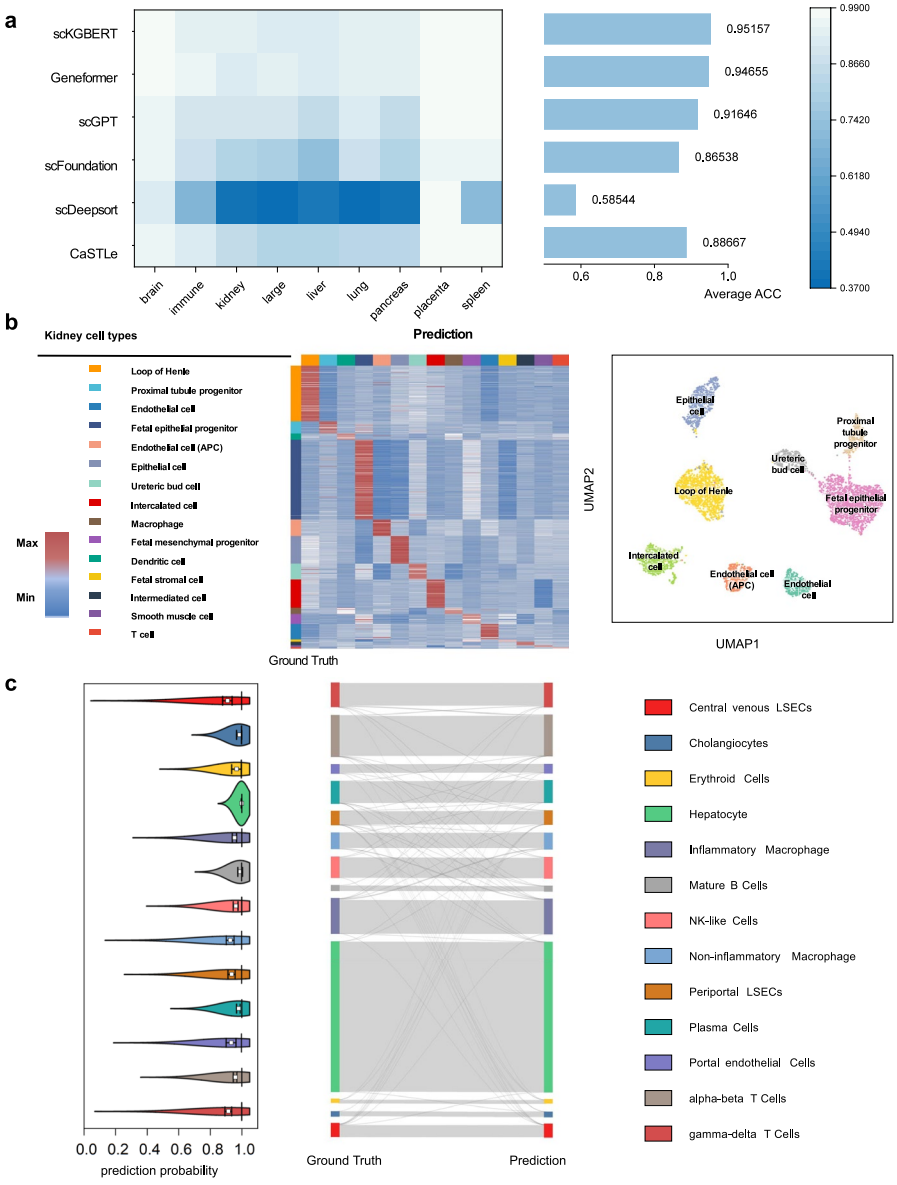

Fig. 3 Cell annotation using scKGBERT. **a** Performance comparison of scKGBERT on cell annotation tasks. The left heatmap shows the accuracy (ACC) of different methods across multiple datasets, and the right bar plot displays the average ACC of each method over all datasets. **b** UMAP visualization of scKGBERT annotations on the kidney cell dataset. Cell types are clearly separated, indicating the model's ability to capture biologically meaningful transcriptional structures. **c** Cell annotation performance on the Zheng68k dataset. scKGBERT maintains high consistency across large-scale single-cell data, further supporting its robustness and scalability for cell-type identification

high-dimensional structure, preserved in a low-dimensional embedding space, provides compelling evidence for the model's capacity to learn biologically coherent and discriminative representations. Collectively, these results position scKGBERT as a powerful tool for unsupervised cell classification, tissue deconvolution, and broader biomedical discovery in single-cell transcriptomics.

We also performed lightweight fine-tuning on a range of cell and cell subtype annotation tasks. Across four publicly available single-cell datasets, scKGBERT consistently outperformed more than ten leading methods (Fig. 3c; Additional file 1: Fig. S4a–b), including scBERT [4], scNym [30], SciBet [31], Seurat [32], SingleR [33], CellID (both

CellID_cell and CellID_group) [34], and scmap (scmap_cell and scmap_cluster) [35]. Remarkably, even under minimal fine-tuning restricted to the final classification layer, scKGBERT demonstrated robust and consistent performance across all datasets. These results highlight the model's strong transferability and data efficiency, confirming its utility for scalable and accurate annotation in diverse single-cell contexts.

Transcriptomic datasets generated using different sequencing technologies often exhibit platform-specific variations in sensitivity, accuracy, and gene expression profiles. To assess the robustness of scKGBERT to such variability, we evaluated its domain adaptability across sequencing platforms using a cross-platform setting. Specifically, the model was fine-tuned on Drop-seq single-cell data and evaluated on both Drop-seq and DroNc-seq single-nucleus data from a study of iPSC differentiation into cardiomyocytes (Fig. 4a–b) [24]. Despite being trained exclusively on Drop-seq data, scKGBERT maintained high annotation accuracy when applied to DroNc-seq, demonstrating strong cross-platform generalization. Notably, cell embeddings remained well-clustered by cell type, regardless of the underlying sequencing technology. These findings demonstrate scKGBERT's robustness to batch effects and platform-specific signal noise, positioning

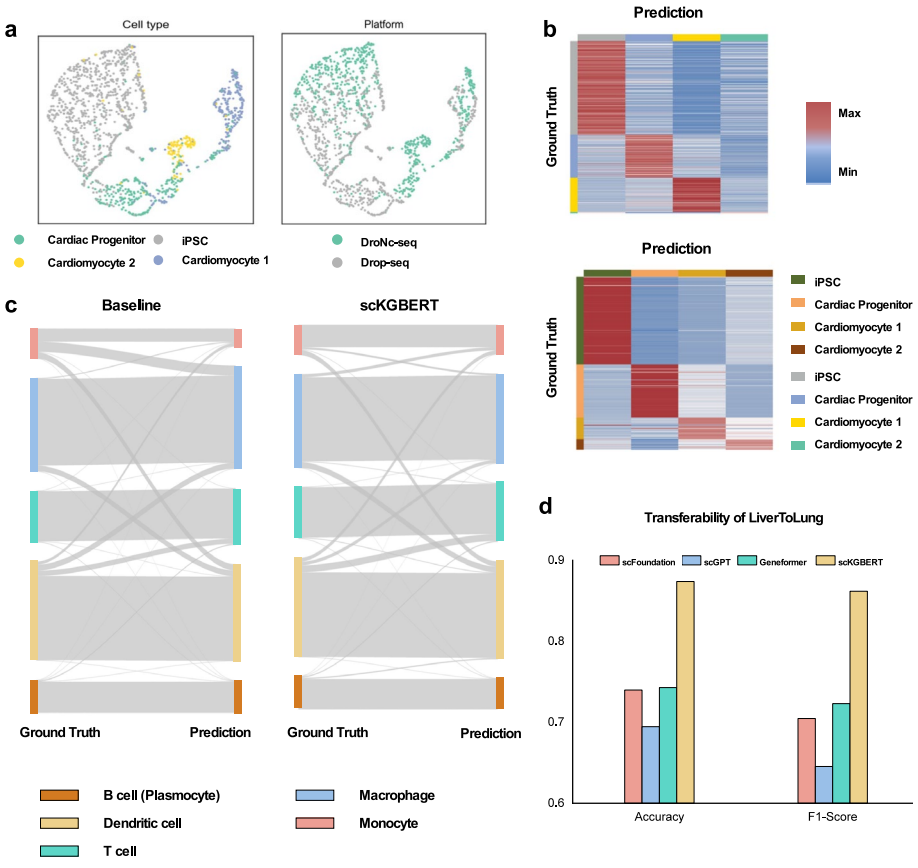

Fig. 4 Domain adaptability of scKGBERT. **a** Fine-tuning scKGBERT with limited data obtained from the Drop-seq platform, we can examine the visualization results for annotated Drop-seq (which did not participate in the fine-tuning) and cells obtained from the DroNc-seq platform. **b** Cross-platform annotation results from scKGBERT, trained on Drop-seq and tested on both Drop-seq and DroNc-seq datasets, illustrating consistent labeling across sequencing technologies. **c** Cross-tissue cell annotation results, where scKGBERT was fine-tuned on liver cells and applied to predict lung cell types. **d** Quantitative comparison of annotation performance across models in the cross-tissue setting, showing scKGBERT achieves superior accuracy and F1-score

it as a versatile model for integrative single-cell analysis across heterogeneous transcriptomic datasets.

To assess the cross-tissue generalization capacity of scKGBERT, we performed transfer learning experiments in which the model was fine-tuned on liver cells and applied to annotate five distinct lung cell types (Fig. 4d). Despite pronounced transcriptional and functional divergence between tissues, scKGBERT demonstrated clear superiority, achieving the highest accuracy (0.88) and F1-score (0.86) among all evaluated models, including Geneformer, scGPT, and scFoundation. The performance margin was especially notable in F1-score, indicating more consistent and reliable predictions under tissue-shifted conditions. This robust cross-tissue performance can be attributed to scKGBERT's integration of structured biological priors via a dedicated knowledge graph, which enhances its ability to capture tissue-specific regulatory logic and subtle phenotypic cues. Unlike conventional pre-trained models that rely solely on data-driven representation learning, the knowledge-guided approach empowers scKGBERT with greater sensitivity to inter-tissue functional heterogeneity and improves its capacity to preserve cell-type identity across biological contexts.

To further dissect model predictions, we visualized annotation outcomes using a Sankey diagram (Fig. 4c), which showed that baseline models exhibited notable errors in macrophage classification-a cell type characterized by high heterogeneity and context-dependent transcriptional programs. While large-scale pre-trained models such as Geneformer leverage extensive data, their lack of embedded biological priors likely limits their ability to distinguish such nuanced cellular states. In contrast, scKGBERT accurately resolved macrophage-specific transcriptional signatures, demonstrating its capacity to recognize context-driven expression shifts. These findings highlight scKGBERT's potential as a generalizable and biologically informed framework for cell annotation across complex, heterogeneous tissue landscapes.

In summary, scKGBERT achieves state-of-the-art performance in cell annotation, cross-platform robustness, and cross-tissue generalization. By integrating biological knowledge with self-supervised learning, it offers a scalable and biologically informed framework for single-cell transcriptomic analysis and biomedical discovery.

scKGBERT enhances cancer drug response prediction

Cancer remains a leading cause of mortality worldwide, driven by complex dysregulation of genetic and molecular pathways. The advent of single-cell transcriptomics has revolutionized cancer research by enabling high-resolution characterization of tumor heterogeneity and microenvironmental interactions. In parallel, the integration of transcriptome-based drug screening data offers new opportunities for optimizing precision oncology strategies. To explore the utility of scKGBERT in this context, we applied the model to predict drug responses, identify key marker genes, and assess interpretability. Specifically, we leveraged single-cell drug response datasets for Erlotinib, Cisplatin, and Docetaxel, each containing high-resolution annotations of drug sensitivity and resistance at the single-cell level [36]. We fine-tuned scKGBERT on these datasets and benchmarked its performance against Geneformer [5] and GenePT [37]. As shown in Fig. 5a, scKGBERT outperformed Geneformer across all three drug tasks and surpassed GenePT on multiple datasets, achieving significantly higher AUC scores — including an

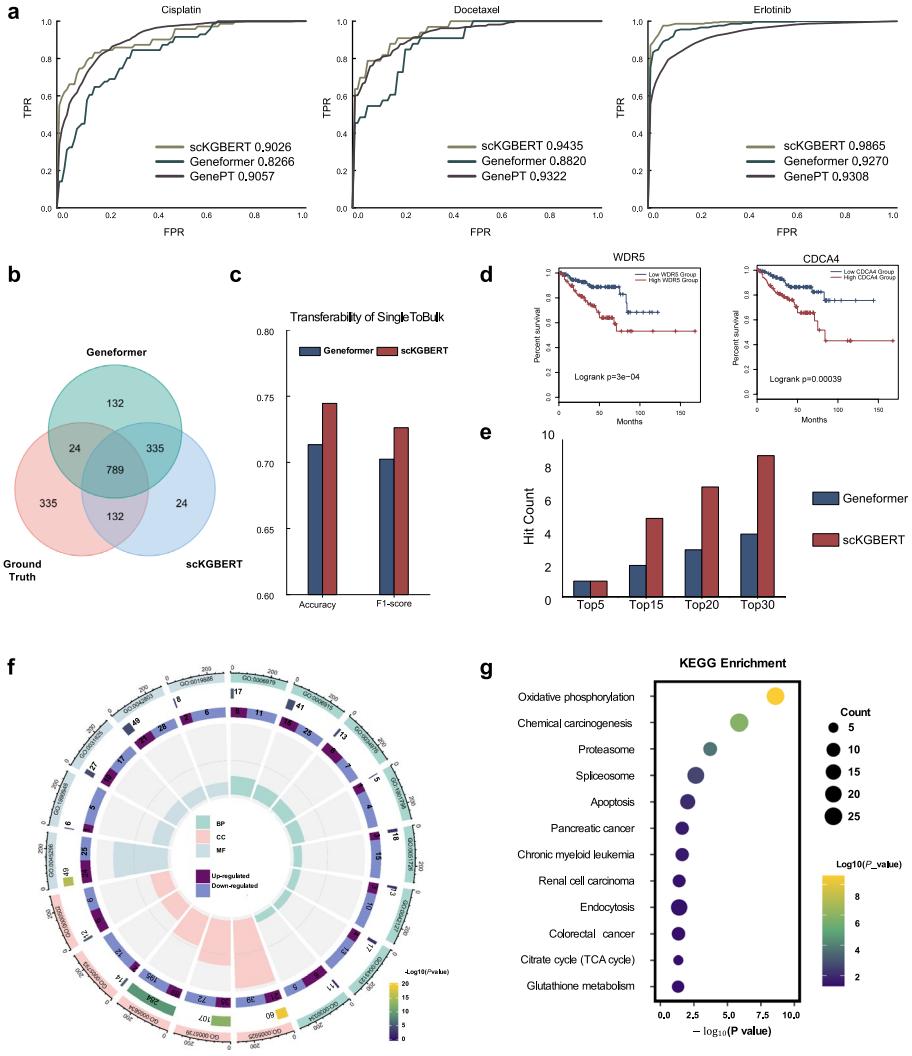

Fig. 5 Cancer drug response prediction by scKGBERT. **a** ROC curves showing the predictive performance of scKGBERT and baseline models for drug sensitivity classification across three compounds: Erlotinib, Cisplatin, and Docetaxel. **b** Venn diagram illustrating the overlap in correctly predicted samples between scKGBERT and Geneformer, comparing predictions made on bulk RNA-seq data after fine-tuning on single-cell datasets. **c** Quantitative evaluation of cross-modality prediction, with accuracy and F1-score of scKGBERT applied to bulk-level data after training on single-cell profiles. **d** Kaplan-Meier survival analysis based on model-inferred drug resistance-associated gene weights, showing significant stratification of patient outcomes. **e** Bar plot showing the number of significantly prognostic genes among the top-N weighted genes identified by each model. **f** Gene Ontology (GO) enrichment analysis of differentially weighted genes identified in tumor epithelial cells, highlighting biological processes relevant to tumor progression. **g** KEGG pathway enrichment analysis of cancer-associated genes, showing significant enrichment in pathways related to cell adhesion, oxidative stress response, and metabolic reprogramming

8% improvement on the cisplatin task. These results highlight the model's ability to capture fine-grained transcriptomic features for predicting therapeutic responses.

To further evaluate the generalization capability of scKGBERT across data modalities, we fine-tuned the model using single-cell RNA-seq data specific to Docetaxel and applied it to bulk RNA-seq data for cross-platform prediction. As shown in Fig. 5c, scKGBERT achieved a 3% improvement in accuracy (ACC) and a higher F1-score compared to the competitive baseline model Geneformer, demonstrating its robustness in transferring predictive power across single-cell and bulk expression data. To demonstrate the

improvement in prediction coverage, we visualized the overlap between predicted and true cell labels using a Venn diagram (Fig. 5b). The number of correctly predicted cells shared between Geneformer and the ground truth is 813, whereas scKGBERT shares 921 correctly predicted cells with the ground truth, achieving an increase of 108 correctly classified cells compared to Geneformer. These findings confirm the model's ability to effectively analyze single-cell profiles and identify gene signatures linked to drug response.

Building on this, we further assessed the model's ability to identify drug resistance–related genes by fine-tuning scKGBERT to extract gene importance scores for predicting drug-resistant phenotypes. The identified gene set was validated using the TCGA-PRAD cohort. In the survival analysis, patients were divided into high-expression and low-expression groups for each gene based on the median expression level: those with expression values above the median were assigned to the high-expression group, and those below to the low-expression group. As shown in Fig. 5d, among the top 30 high-weight genes identified by scKGBERT, several genes were significantly associated with poorer overall survival. Notably, previous studies have reported that WDR5 is highly expressed in prostate cancer (PCa) and is correlated with advanced clinicopathological features and poor prognosis [38]. We also counted the number of significantly prognostic genes within the top N weighted genes, as illustrated in Fig. 5e. In contrast to Geneformer, scKGBERT showed a substantial increase in the number of prognostic genes as N increased, which may be attributed to the incorporation of gene regulatory knowledge into the model. These findings not only highlight the predictive capability of scKGBERT but also reveal its potential for uncovering clinically relevant biomarkers of drug resistance, offering important biological insights that may be overlooked by conventional models.

To further evaluate the model's capacity to capture tumor cell heterogeneity, we curated multiple tumor epithelial cell datasets and fine-tuned scKGBERT to distinguish malignant epithelial cells from their normal counterparts based on single-cell transcriptomic profiles. Gene-level attention weights were extracted from the model's Gaussian attention mechanism, enabling interpretable identification of genes contributing to the classification. Differential analysis showed 727 high-weight genes closely associated with tumor states. Functional enrichment analysis using KEGG and GO databases indicated that these genes are significantly enriched in cancer-relevant biological pathways ($P < 0.005$, Fig. 5f-g), including cell adhesion, reactive oxygen species (ROS) response, and metabolic reprogramming. For example, cadherins, known regulators of epithelial–mesenchymal transition (EMT), are directly implicated in tumor invasion and metastasis [39]. Tumor cells also adapt to elevated ROS levels in oxidative stress environments by modulating sulfur metabolism and antioxidant systems, processes that play pivotal roles in tumor initiation and progression [40]. Additionally, enhanced oxidative phosphorylation has been linked to the activation of oncogenic signaling and malignant proliferation within the tumor microenvironment [41]. These results highlight scKGBERT's ability to uncover biologically meaningful, cancer-associated transcriptional programs and provide mechanistic insights into tumor evolution and progression.

These findings demonstrate the broad applicability and high accuracy of the scKGBERT model in drug screening and cancer research. By capturing tumor-specific molecular signatures and offering biologically interpretable predictions, the model provides

mechanistic insights into the complex pathology of cancer and the diverse responses of tumor cells to therapeutic agents. This interpretability supports a deeper understanding of drug-tumor interactions and their effects on cell survival and proliferation, offering a valuable framework for guiding the development of personalized cancer therapies.

Insights into heart disease marker genes from scKGBERT

Heart failure is an increasingly critical concern in public health, characterized by impaired cardiac contractile function and complex clinical manifestations. Previous studies have largely focused on transcriptomic or proteomic alterations in bulk cardiac tissue, often lacking a comprehensive, cell-type-resolved analysis of gene expression dynamics and disease-specific markers. scKGBERT addresses this gap by enabling precise characterization of heart failure at single-cell resolution, effectively distinguishing disease-specific transcriptomic signatures while capturing the cellular heterogeneity across diverse cardiac cell populations. This approach offers a novel biological framework for investigating heart failure pathogenesis, providing insights into how distinct cell types respond to pathological stress and contribute to disease progression.

We collected nearly 560,000 single-cell RNA sequencing (scRNA-seq) data [42] from the left ventricles of 11 dilated cardiomyopathy (DCM) hearts, 15 hypertrophic cardiomyopathy (HCM) hearts, and 16 non-failing (NF) hearts to test whether the scKGBERT model could differentiate gene anomalies in normal human cells from those in disease cells. By fine-tuning scKGBERT, we aimed to differentiate non-failing (normal) myocardial cells from DCM and HCM disease cells. The proposed scKGBERT model employs a multi-layer architecture, with each layer featuring a weight allocation strategy that integrates sequence and knowledge encoding through a Gaussian cross-attention mechanism. The Gaussian cross-attention mechanism assigns higher weights to significant genes and lower weights to less critical ones, thereby amplifying the differentiation among genes. It enhances the learning of more accurate contextual representations of both genes and cells, thereby improving the model's interpretability. On this basis, we constructed a list of marker genes for DCM and HCM disease cells and used Gaussian weights from the scKGBERT model's prediction process as an indicator to evaluate the importance of these genes. Notably, these marker genes generally received higher weights in the fine-tuned scKGBERT model (as shown in Fig. 6a-b), demonstrating the model's accuracy in capturing gene expression patterns in diseased cells. For example, mutations in the TNN gene can lead to structural or functional abnormalities in MyBP-C (myosin-binding protein C) [43, 44], affecting the normal contractile function of the myocardium. During this process, the weights generated by the fine-tuned scKGBERT model were significantly higher than those in normal cells, highlighting the model's ability to learn specific disease biomarkers, their interactions, and roles in biological pathways, thereby offering a more comprehensive understanding of disease mechanisms.

Furthermore, we performed differential analyses on the gene weights assigned by scKGBERT in samples from the three distinct populations. Figure 6c present the top five most differentially weighted genes in each group. We performed Reactome pathway enrichment analysis on the genes associated with hypertrophic cardiomyopathy (HCM) (Fig. 6d), identifying multiple relevant signaling pathways. Notably, Notch1-related pathways were enriched, which have been previously reported to be involved in cardiac morphogenesis, valve development, and ventricular remodeling, further highlighting

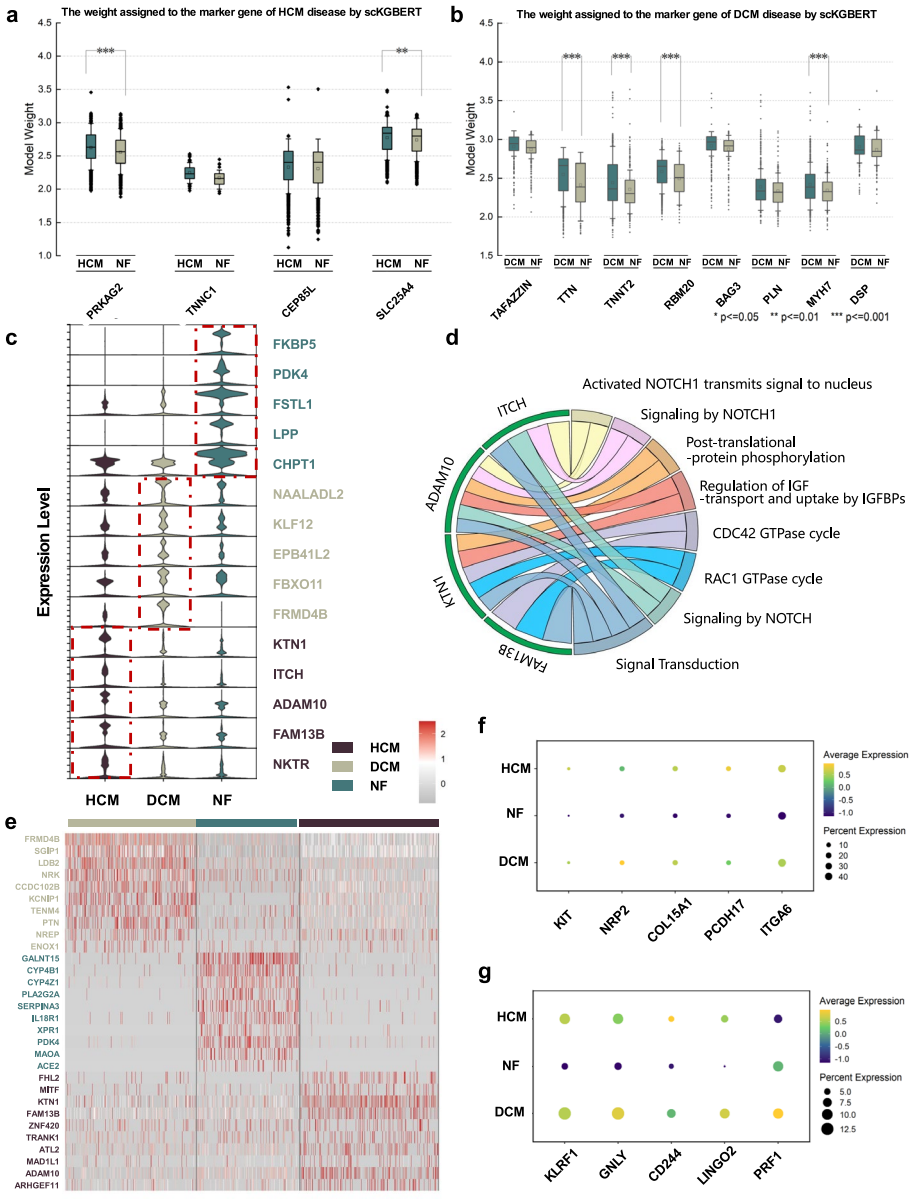

Fig. 6 Insights into disease marker genes by scKGBERT. **a** Boxplot showing the distribution of model-inferred gene weights for HCM-related marker genes in hypertrophic cardiomyopathy (HCM) patients versus non-failing (NF) controls. **b** Boxplot comparing gene weight distributions for DCM-associated markers between dilated cardiomyopathy (DCM) patients and NF controls. **c** Violin plot depicting the top five differentially weighted genes in HCM, DCM, and NF groups. **d** Reactome pathway enrichment analysis of high-weight genes identified in HCM samples. **e** Heatmap of differentially expressed genes in fibroblasts across disease and control conditions. **f** Average model-derived expression weights of endothelial cell markers (KIT, NRP2, COL15A1, PCDH17, ITGA6) across NF, HCM, and DCM groups. **g** Average model-derived expression weights of lymphocyte markers (KLRF1, GNLY, CD244, PRF1) across NF, HCM, and DCM conditions

the critical role of Notch1 in embryonic development and heart formation, as well as its importance in cardiac stem cell proliferation and differentiation [45]. Inhibition of Notch1 signaling could lead to life-threatening dilated cardiomyopathy, emphasizing that scKGBERT can not only validate known disease-related genes and pathways but also expand our perspectives in researching cardiac disease biomarkers and therapeutic targets.

Fibroblasts play an essential role in the heart, responsible for synthesizing the extracellular matrix (ECM) and collagen, ensuring the normal structure and function of the heart [46]. In cardiac diseases such as HCM and DCM, abnormalities in fibroblast function are closely related to pathological myocardial fibrosis, which leads to damage to the structure and function of the heart. We analyzed the gene expression of fibroblasts and presented a weight analysis generated by scKGBERT for these cell populations in Fig. 6e. The results showed that, in the fibroblasts of patients with HCM, the top ten upregulated genes are mainly concentrated in the regulation pathway of the CDC42 GTPase activity cycle. Interestingly, these findings have been verified in a recent study [47]. The research indicates that CDC42 is related to the pathogenesis of HCM, and experiments based on mouse models have shown that the specific knockout of the CDC42 gene in the heart under stress leads to more significant cardiac hypertrophy. Two weeks and eight weeks after stimulation, these mice progress to heart failure at a faster rate.

At the same time, previous research observed that in DCM and HCM hearts compared to NF hearts, the expression of markers for certain endothelial and lymphoid cells, such as KIT, NRP2, and KLRF1 [42], was significantly increased. To validate these observations, we focused on the endothelial and lymphoid cell populations and examined the expression levels of these known marker genes within each respective cell type. As shown in Fig. 6f-g, the expression levels were indeed elevated, consistent with the findings of prior studies. This consistency highlights the effectiveness of scKGBERT in capturing disease-relevant gene expression patterns at the cell-type level. By delving deeper into scKGBERT's predictions, we have captured not only the changes in marker genes but also unveiled potential mechanisms behind the progression of the disease. Particularly in cardiomyopathy model studies, genes and related biological pathways highlighted by the model's recognized Gaussian weights not only reflect the close relationship between specific cardiac diseases and specific genes and biological processes but also emphasize the importance of heterogeneity at the cellular level and the complexity of intercellular communication in the development of heart diseases. Therefore, the scKGBERT model provides a powerful tool that can deepen our understanding of the mechanisms behind heart failure and other cardiac diseases and offer valuable insights for the future diagnosis and treatment of heart diseases.

Discussion

The integration of biological knowledge into large-scale single-cell foundation models represents an important step toward interpretable and biologically grounded artificial intelligence in transcriptomics. In this study, we developed scKGBERT, a knowledge-enhanced foundation model that unifies single-cell RNA-seq data and protein–protein interaction networks within a dual-encoder framework. Our results demonstrate that this design substantially improves the model's ability to capture gene regulatory dependencies and cellular heterogeneity, leading to superior performance in gene annotation, cell classification, and drug response prediction. Importantly, the incorporation of Gaussian cross-attention allows scKGBERT to highlight biologically meaningful genes and enhance interpretability without increasing model complexity.

Compared with existing models such as scGPT, Geneformer, and scFoundation, scKGBERT shows clear advantages in generalization and biological interpretability. The model effectively identifies dosage-sensitive and epigenetically regulated genes, resolves

subtle transcriptional differences among cell types, and generalizes across tissues and sequencing platforms. Its robust performance in cancer and cardiovascular disease analysis further demonstrates its translational potential. By mapping learned representations to known pathways and regulatory circuits, scKGBERT bridges computational predictions with mechanistic biological understanding, providing a foundation for precision medicine applications.

Future work will extend this framework to multi-omics integration, including chromatin accessibility, proteomics, and spatial transcriptomics, to further enhance model comprehensiveness. Additionally, incorporating literature-based biomedical knowledge and causal inference strategies may strengthen interpretability and hypothesis generation. Overall, scKGBERT establishes a new paradigm for knowledge-guided single-cell modeling, offering both computational scalability and deep biological insight.

Conclusions

In this study, we present scKGBERT, a knowledge-enhanced foundation model for single-cell transcriptomics. To our knowledge, it is the first unified pre-training framework that concurrently integrates RNA expression profiles and structured biological priors-captured through a protein–protein interaction (PPI) knowledge graph-within a single modeling architecture. By employing a dual-encoder design coupled with a Gaussian cross-attention mechanism, scKGBERT learns biologically informed and interpretable representations of genes and cells, enabling robust characterization of transcriptional landscapes and gene regulatory networks at single-cell resolution.

scKGBERT demonstrates superior generalization and interpretability across a broad spectrum of downstream tasks, including gene annotation, cell type classification, cancer drug response prediction, and disease biomarker discovery. In gene-level tasks, the model effectively captures dosage sensitivity, epigenetic regulation, and long-range transcriptional effects without relying on multi-omics input. At the cellular level, it achieves state-of-the-art performance in cell annotation tasks across tissues and sequencing platforms, and exhibits strong cross-tissue transferability-highlighting its capacity to resolve complex cell states in noisy or heterogeneous environments. Critically, scKGBERT's knowledge integration confers significant advantages over existing models, as demonstrated by ablation studies showing improved marker gene prioritization and disease-specific signal detection. In cancer and cardiovascular disease applications, the model accurately identifies clinically relevant gene signatures, maps them to known signaling pathways, and links gene-level perturbations to patient outcomes. These capabilities underscore its translational potential for precision medicine and therapeutic target discovery. Furthermore, scKGBERT achieves these results with significantly fewer parameters-just one-tenth of some leading models-emphasizing its efficiency without compromising performance. It makes scKGBERT not only a biologically principled and interpretable model, but also a computationally scalable one.

Looking ahead, scKGBERT lays the foundation for next-generation single-cell models capable of integrating multi-omic and cross-modal inputs, including sequence, epigenetics, imaging, and literature-based knowledge. By expanding the biological context within which transcriptomic data are interpreted, future iterations will enable deeper insights into cell fate decisions, disease progression, and therapeutic response-paving the way toward a unified and comprehensive framework for single-cell systems biology.

Methods

scKGBERT architecture and pretraining

The scKGBERT model consists of two primary components. The first involves knowledge construction and representation. This phase utilizes additional biological knowledge to construct a knowledge graph, followed by a knowledge embedding method to learn the knowledge representation. The second component is scKGBERT pre-training. This process draws an analogy between single-cell RNA-seq data and natural language. Gene expression sequences from individual cells are treated as sentences, with each gene or transcript akin to a word in this biological language. The expression levels of genes represent the "meaning" of these sentences.

scKGBERT consists of two parallel Masked Language Models [48]: a Sequence Encoder (S-Encoder) and a Knowledge Encoder (K-Encoder), both following the classical BERT paradigm. Each encoder is composed of multiple transformer blocks, incorporating self-attention mechanisms and feed-forward neural networks. To construct representative and standardized input sequences, we adopt the rank-based gene selection strategy proposed by Geneformer [5]. Specifically, genes are first ranked based on their expression levels within each single-cell transcriptome, and the top 2048 genes are selected to represent the cell. These selected genes are then encoded into fixed-length input sequences. During pretraining, we apply a BERT-style random masking strategy, where approximately 15% of the input genes are randomly masked. The model is trained to predict the masked genes from the context, enabling it to learn the contextual dependencies and structure among genes. The Gaussian cross attention mechanism integrates additional biological knowledge into the sequence information, effectively mapping the two representations into a unified feature space. It is designed to reduce feature redundancy, enhance weight differentiation among genes, and highlight crucial information in the input sequence, thereby reducing computational complexity [49]. It also prevents the oversmoothing of embeddings, particularly in cases of longer input sequences. The MLMs train the model to predict masked or missing labels in the sequence, focusing on understanding the contextual relationships between genes and inferring missing expressions.

Knowledge construction and representation

Domain knowledge in biology is essential for enhancing large pre-trained models in single-cell RNA sequencing. We incorporate the STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) as an external knowledge graph. The STRING database offers a comprehensive protein-protein interaction (PPI) network, crucial for understanding cellular protein interactions. It allows pre-trained models to identify significant protein interactions in single-cell RNA sequencing analysis. Additionally, STRING enhances biological pathway analysis, enabling the identification of pathways and functional modules within gene expression data. These capabilities provide deeper insights into cellular states at the single-cell level, where distinct gene expression patterns are observed. The interaction data from STRING are also vital for identifying biomarkers and genes associated with specific diseases or biological traits, playing a key role in disease diagnosis and treatment planning.

To fully utilize the fine-grained semantic relationships between genes in the existing knowledge base, we use the STRING [12] (Search Tool for the Retrieval of Interacting

Genes/Proteins) database to acquire external knowledge. STRING is a powerful database for studying protein-protein interactions. It provides data on established and predicted protein interactions, including both direct (physical) and indirect (functional) connections [12].

In the knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, $\mathcal{E}$ represents the set of gene entities, $\mathcal{R}$ denotes the set of relations, and $\mathcal{T}$ comprises triples (*h*, *r*, *t*). Each triple indicates that the head gene entity *h* is related to the tail gene entity *t* via relation *r*. For each gene entity *e*, its initial knowledge embedding is transformed from a one-hot vector to a *d*-dimensional vector $\mathbf{e}$ using a transformation matrix $\mathbf{W} \in \mathbb{R}^{V \times d}$, where V is the complete vocabulary containing all gene tokens. When *e* acts as the head entity *h*, and when *e* acts as the tail entity *t*. For each gene relation *r*, we introduce a transformation matrix $\mathbf{M}^{(r)} \in \mathbb{R}^{d \times k}$. This matrix captures the specific characteristics of the relation and projects entity embeddings $\mathbf{h}$ and $\mathbf{t}$ from the entity space to a relation-specific space, as illustrated by the following equation:

$$\mathbf{h}^{(r)} = \mathbf{h}\mathbf{M}^{(r)}$$

$$\mathbf{t}^{(r)} = \mathbf{t}\mathbf{M}^{(r)}$$

We then define a scoring function to evaluate the possibility of a triple (*h*, *r*, *t*) using the L1 or L2 distance:

$$S(r, h, t) = \|\mathbf{h}^{(r)} + \mathbf{r} - \mathbf{t}^{(r)}\|_p$$

where the representation of head entity gene *h* and the tail entity gene *t* are both transformed into the relation space *r*, $\mathbb{R}^k$.

To learn the knowledge embeddings of genes and transformation matrices, we use a margin-based loss function, to encourage the model to assign higher scores to true triples (*h*, *r*, *t*) and lower scores to corrupted triples (h, r, t'), where t' represents a negative sample. Finally, we obtain the updated *d*-dimensional vector $\mathbf{e}$ as the initial knowledge representation for each gene, which can then be sequentially input into the knowledge encoder.

Sequence encoder and knowledge encoder

scKGBERT includes a sequence encoder and a knowledge encoder, both adhering to the masked language model paradigm. Specifically, for a given gene sequence from a single cell $\{g_1, \ldots, g_n\}$ and its corresponding entity sequence $\{e_1, \ldots, e_n\}$, where *n* is the length of the input RNA-sequence, the sequence encoder first combines the token embedding and positional embedding for each gene to create its input embedding. It then computes the syntactic features $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ as follows:

$$\mathbf{s}_1, \ldots, \mathbf{s}_n = \text{S-Encoder}(\{g_1, \ldots, g_n\}), \tag{1}$$

Similarly, we take the representation of the knowledge entity $\{e_1, \ldots, e_n\}$ learned from the pre-trained TransR model as the input, and through the knowledge encoder `K-Encoder`, we obtain the knowledge representation of the gene.

$$\mathbf{h}_1, \ldots, \mathbf{h}_n = \text{K-Encoder}(\{e_1, \ldots, e_n\}), \tag{2}$$

where both $\mathrm{S\text{-}Encoder}(\cdot)$ and $\mathrm{K\text{-}Encoder}(\cdot)$ are multi-layer bidirectional Transformer encoders.

We employ two independent and parallel encoders: the S-Encoder for capturing contextual information from gene sequences and the K-Encoder for encoding prior knowledge. After processing, we obtain the contextual representation matrix $\mathbf{H}_{seq} \in \mathbb{R}^{n \times d}$ and the knowledge representation matrix $\mathbf{H}_{kg} \in \mathbb{R}^{n \times d}$, where d denote the respective hidden dimensions of the S-Encoder and K-Encoder. No parameter sharing is applied between the two encoders.

Gaussian cross attention mechanism

To better integrate sequence-based contextual features with knowledge-based relational embeddings, we propose a Gaussian-normalized mechanism. Traditional cross-attention mechanisms compute attention scores using a dot-product between queries and keys, followed by softmax normalization. While widely used, this approach often produces overly smoothed weights and lacks sensitivity to score magnitude variation across samples. These issues can weaken the contrast in attention patterns, particularly when dealing with heterogeneous information such as transcriptomic sequences and symbolic knowledge.

To address this, we modify the attention formulation by incorporating Gaussian normalization prior to softmax. Let $\mathbf{H}_{seq}$ and $\mathbf{H}_{kg}$ denote the hidden representations from the sequence encoder and knowledge graph encoder, respectively. We first concatenate the two inputs and project them into Q, K, and V as follows:

$$\begin{aligned} \mathbf{Q} &= \mathbf{W}_Q[\mathbf{H}_{seq} \,|\, \mathbf{H}_{kg}], \\ \mathbf{K} &= \mathbf{W}_K[\mathbf{H}_{seq} \,|\, \mathbf{H}_{kg}], \\ \mathbf{V} &= \mathbf{W}_V \mathbf{H}_{seq} \end{aligned} \tag{3}$$

The raw Gaussian fuison score matrix is computed using the dot product between $\mathbf{Q}$ and $\mathbf{K}$:

$$S = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \tag{4}$$

Instead of applying softmax directly to S, we first perform Gaussian normalization:

$$\mu = \mathrm{mean}(S), \quad \sigma = \mathrm{std}(S), \quad Gaussian_{score} = Act(\frac{S - \mu}{\sigma}) \tag{5}$$

To ensure non-negativity and stabilize the attention distribution, we apply an activation function (denoted as $Act(\cdot)$) to transform negative values in the normalized attention scores into a small positive constant, ensuring that all attention weights remain non-negative. By doing so, the softmax operation can be omitted while still producing meaningful attention weights. To maintain a scale comparable with standard attention mechanisms, the resulting scores can optionally be normalized across each row. The fused representation is then computed as:

$$\mathbf{H}_{fusion} = Gaussian_{score} \cdot \mathbf{V} \tag{6}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable weight matrices, and the operations $\mathrm{mean}(\cdot)$ and $\mathrm{std}(\cdot)$ are computed over all elements of the attention score matrix S for each individual sample.

Unlike kernel-based Gaussian attention, which models attention weights using a Gaussian probability density over input distances, our method preserves the dot-product attention structure while incorporating Gaussian statistical normalization. By enhancing the contrast of attention weights and mitigating over-smoothing effects, the approach remains lightweight and computationally efficient. Since normalization is performed independently for each sample, attention weights are not directly comparable across samples. Nevertheless, within each sample, the method substantially improves the sharpness and interpretability of attention patterns. The proposed mechanism is consistently applied across all fusion layers in the model.

Pre-training and fine tuning

Overall, scKGBERT is divided into two stages: pre-training and fine-tuning for downstream tasks. During the self supervised pre-training, a certain percentage (usually 15%) of tokens in each sequence is randomly selected and replaced with a [mask] token. Then, the modified sequence and its masked tokens are fed into a multi-layer transformer encoder. Each Transformer layer captures contextual information from both directions (left and right context) around each token. It then predicts the original values of the masked tokens based on their context. While this prediction is not used directly in sequence encoding but helps the model learn better representations during pre-training. The objective function for the model's pre-training process is as follows:

$$L_{Pre-training} = -\sum_{i=1}^{N_{Cell}} \sum_{j=1}^{N_{MLM}} y_{ij} log(\hat{y}_{ij}) \tag{7}$$

where N_{Cell} represents the total number of cells, N_{MLM} denotes the number of masked genes in the masked language model, and y_{ij} and $\hat{y}_{ij}$ represent the ground-truth and prediction results for the genes, respectively. The optimal hyperparameters during pre-training include an initial learning rate of 10^{-3}, using the AdamW optimizer, with the weight decay set at 10^{-10}.

Experiments setting

The scKGBERT model consists of three stacked scKGBERT blocks, each comprising two parallel BERT blocks and a Gaussian fusion module. The initial embedding dimension is set to 256, with a hidden layer size of 512, and a 4-head multi-head attention mechanism. A dropout rate of 0.05 is applied to prevent overfitting. Model training was performed using the AdamW optimizer, with an initial learning rate of 5e-5 and a weight decay of 1e-6. The model was trained for 2 epochs with a batch size of 16, balancing computational efficiency and model performance. All training was conducted on two NVIDIA A100 GPUs.

System analysis

Model interpretability and marker gene identification

In the scKGBERT model, the Gaussian cross-attention mechanism produces a Gaussian weight matrix representing the expression weight of each gene for every single-cell transcriptome sequence. Specifically, we average the weight matrices of the three-layer scKGBERT model to obtain a comprehensive weight representation. Subsequently, we further average along the columns to obtain a vector representing the relative importance of each gene in the cell. We utilize this vector as a surrogate for gene expression to explore the interpretation methods of the model, including differential gene analysis and functional enrichment analysis. Our model not only can identify genes with significant differential expression, but also can discover important biological processes. In particular, compared with healthy people, our model can identify the disease marker genes and accurately reflect the characteristics of abnormal gene expression. We also try to capture the associations between genes without fine-tuning the model. The results show that the model can effectively capture the interactions between landmark genes. These findings demonstrate that our model not only facilitates superior interpretation of biological data, but also contributes valuable tools and insights to the field of biomedical research.

Gene-level tasks

To evaluate scKGBERT's effectiveness in gene-level tasks, such as gene annotation, we fine-tuned the pre-trained scKGBERT model and compared it with three machine learning based methods and three advanced pre-training models, including Geneformer, scFoundation and scGPT [50]. Initially, we ranked and sorted the target single-cell expression data based on rank-value criteria, selecting the top 2048 genes to represent cell features effectively. These genes were then converted into gene sequences, serving as input for the model. For gene annotation, scKGBERT was fine-tuned, specifically by adding one or two MLP layers for classification tasks. It is important to highlight that the layers of scKGBERT have been frozen, and superior gene annotation performance can be achieved solely by fine-tuning one or two subsequently added MLP layers, which demonstrates scKGBERT possesses robust task transferability.

Cell-level tasks

To evaluate the performance of scKGBERT on cell-level tasks, including drug response prediction and cell annotation, we select twelve traditional methods and two advanced pre-training models (Geneformer and scBERT) as baselines. We feed single-cell gene tokens into the pre-trained scKGBERT model and take the average of all gene embeddings in one cell as the cell representation. During fine-tuning, we add a multilayer perceptron (MLP) to scKGBERT and tune the MLP and the last two layers of the scKGBERT module. The reason is that cell-level annotation tasks usually have high heterogeneity and complex behaviors, which requires a deeper network structure to distinguish cells from each other and capture the relationship between cells. In addition, to verify the cross-platform generalization ability of our model, we fine-tune it on Drop-seq data and test it on DroNc-seq sequences. The results show that scKGBERT can achieve consistent performance on different tasks.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03862-6.

Additional file 1: The file contains detailed descriptions of marker gene insights from scKGBERT, dataset information, baseline method summaries, and supplementary figuresillustrating extended analyses of gene annotation, cell annotation, and model performance

Acknowledgements
We thank the anonymous reviewers for their constructive suggestions.

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
G.W. conceived the work, assisted in algorithm design and implementation. Y.L. and G.Q. contributed to algorithm design, implementation, and computational experiment analysis. Y.L. drafted the initial manuscript. G.W. and X.G. contributed to manuscript revision. L.D. contributed to data acquisition and processing.

Funding
This work was partially supported by the National Natural Science Foundation of China Original Exploration Program [62450112], the National Key R&D Program of China [2024YFF1206602], the National Science Fund for Distinguished Young Scholars [62225109], the National Natural Science Foundation of China [32470687], and the Fundamental Research Funds for the Central Universities [2572025JT05]. Corresponding author: Guohua Wang (E-mail: ghwang@nefu.edu.cn).

Data availability
All data utilized in this study is publicly available, and its usage is thoroughly explained in the Methods section. The Panglao dataset was obtained from https://panglaodb.se/ [20]. The published Zheng68K dataset was downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets (SRP073767) [51] under the "Fresh 68 K pbmc" section. The nine datasets for single-cell annotation across different human tissues can be downloaded from Hugging Face at https://huggingface.co/datasets/ctheodoris/Genecorpus-30M [5]. The pancreas dataset was obtained from the GEO database at https://www.ncbi.nlm.nih.gov/geo/ (Baron: GSE84133 [52], Muraro: GSE85241 [53], Xin: GSE81608 [54], MacParland: GSE115469 [55]). The Segerstolpe dataset was downloaded from the ArrayExpress database at https://www.ebi.ac.uk/arrayexpress (ID: E-MTAB-5061) [56]. Marker genes for cells were obtained from https://xteam.xbio.top/CellMarker/ [57]. Additionally, this paper provides all processed data in https://huggingface.co/datasets/catly/scKGBERT-41M [58]. All relevant data and source codes including the pre-trained model, can be downloaded from https://huggingface.co/catly/scKGBERT [59].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China
[2]Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China
[3]School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China
[4]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[5]Center of Excellence for Smart Health (KCSH), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[6]Center of Excellence on Generative AI, King Abdullah University of Science and Technology (KAUST), Makkah 23955-6900, Kingdom of Saudi Arabia

Received: 6 May 2025 / Accepted: 5 November 2025
Published online: 24 November 2025

References

1. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv. 2023;56(2):1–40.
2. He K, Girshick R, Dollár P. Rethinking imagenet pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. Seoul, South Korea. Los Alamitos, CA, USA: IEEE Computer Society; pp. 4918–27.
3. Jia Y, Liu J, Chen L, Zhao T, Wang Y. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. Brief Bioinform. 2023;25(1). bbad464 . https://doi.org/10.1093/bib/bbad464.

4. 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(10):852–66.
5. 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(7965):616-24.
6. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024;21(8):1481–91.
7. Zhang Z, Zhao L, Wang J, Wang C. A hierarchical graph neural network framework for predicting protein-protein interaction modulators with functional group information and hypergraph structure. IEEE J Biomed Health Inform. 2024;28(7):4295–305.
8. Zhang Z, Zou Q, Wang C, Wang J, Zhao L. Improving protein-protein interaction modulator predictions via knowledge-fused language models. Inf Fusion. 2025;123:103227. https://doi.org/10.1016/j.inffus.2025.103227.
9. Qiao G, Wang G, Li Y. Causal enhanced drug-target interaction prediction based on graph generation and multi-source information fusion. Bioinformatics. 2024;40(10):btae570.
10. Yang KD, Belyaeva A, Venkatachalapathy S, Damodaran K, Katcoff A, Radhakrishnan A, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat Commun. 2021;12(1):31.
11. Fortelny N, Bock C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol. 2020;21(1):1–36.
12. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(D1):D447–52.
13. Al-Harazi O, El Allali A, Colak D. Biomolecular databases and subnetwork identification approaches of interest to big data community: an expert review. OMICS J Integr Biol. 2019;23(3):138–51.
14. Jia Y, Dong H, Li L, Wang F, Juan L, Wang Y, et al. xQTLatlas: a comprehensive resource for human cellular-resolution multi-omics genetic regulatory landscape. Nucleic Acids Res. 2025;53(D1):D1270–7.
15. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.
16. Rice AM, McLysaght A. Dosage-sensitive genes in evolution and disease. BMC Biol. 2017;15(1):1–10.
17. Shihab HA, Rogers MF, Campbell C, Gaunt TR. Hipred: an integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017;33(12):1751–7.
18. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
19. Ni Z, Zhou XY, Aslam S, Niu DK. Characterization of human dosage-sensitive transcription factor genes. Front Genet. 2019;10:1208.
20. Franzén O, Gan LM, Björkegren JL. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019:baz046.
21. Aich M, Chakraborty D. Role of lncRNAs in stem cell maintenance and differentiation. Curr Top Dev Biol. 2020;138:73–112.
22. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26.
23. Chen CH, Zheng R, Tokheim C, Dong X, Fan J, Wan C, et al. Determinants of transcription factor regulatory range. Nat Commun. 2020;11(1):2472.
24. Selewa A, Dohn R, Eckart H, Lozano S, Xie B, Gauchat E, et al. Systematic comparison of high-throughput single-cell and single-nucleus transcriptomes during cardiomyocyte differentiation. Sci Rep. 2020;10(1):1535.
25. Sana TR, Janatpour MJ, Sathe M, McEvoy LM, McClanahan TK. Microarray analysis of primary endothelial cells challenged with different inflammatory and immune cytokines. Cytokine. 2005;29(6):256–69.
26. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20(5):257–72.
27. Wu AR, Wang J, Streets AM, Huang Y. Single-cell transcriptional analysis. Annu Rev Anal Chem. 2017;10:439–62.
28. Shao X, Yang H, Zhuang X, Liao J, Yang P, Cheng J. Scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. Nucleic Acids Res. 2021;49(21):e122–e122.
29. Lieberman Y, Rokach L, Shay T. CaSTLe-classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. PLoS One. 2018;13(10):e0205499.
30. Kimmel JC, Kelley DR. Semisupervised adversarial neural networks for single-cell classification. Genome Res. 2021;31(10):1781–93.
31. Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, et al. SciBet as a portable and fast single cell type identifier. Nat Commun. 2020;11(1):1818.
32. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87.
33. Aran D. Single-cell RNA sequencing for studying human cancers. Annu Rev Biomed Data Sci. 2023 ;6(1):1–22. https://doi.org/10.1146/annurev-biodatasci-020722-091857.
34. Cortal A, Martignetti L, Six E, Rausell A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nat Biotechnol. 2021;39(9):1095–110.
35. Kiselev VY, Yiu A, Hemberg M. Scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15(5):359–62.
36. Chen J, Wang X, Ma A, Wang QE, Liu B, Li L, et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. Nat Commun. 2022;13(1):6494.
37. Chen Y, Zou J. Simple and effective embedding model for single-cell biology built from chatgpt. Nat Biomed Eng. 2025;9(4):483–93.
38. Zhou Q, Chen X, He H, Peng S, Zhang Y, Zhang J, et al. WD repeat domain 5 promotes chemoresistance and programmed death-ligand 1 expression in prostate cancer. Theranostics. 2021;11(10):4809.
39. Yu W, Yang L, Li T, Zhang Y. Cadherin signaling in cancer: its functions and role as a therapeutic target. Front Oncol. 2019;9:989.
40. Hayes JD, Dinkova-Kostova AT, Tew KD. Oxidative stress in cancer. Cancer Cell. 2020;38(2):167–97.
41. Ashton TM, McKenna WG, Kunz-Schughart LA, Higgins GS. Oxidative phosphorylation as an emerging target in cancer therapy. Clin Cancer Res. 2018;24(11):2482–90.

42. Chaffin M, Papangeli I, Simonson B, Akkad AD, Hill MC, Arduini A, et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature. 2022;608(7921):174–80.
43. Letavernier E, Zafrani L, Perez J, Letavernier B, Haymann JP, Baud L. The role of calpains in myocardial remodelling and heart failure. Cardiovasc Res. 2012;96(1):38–45.
44. Nica AC, Ongen H, Irminger JC, Bosco D, Berney T, Antonarakis SE, et al. Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. Genome Res. 2013;23(9):1554–62.
45. Zhang W, Liu J, Wu Z, Fan G, Yang Z, Liu C. Notch1 is involved in physiologic cardiac hypertrophy of mice via the p38 signaling pathway after voluntary running. Int J Mol Sci. 2023;24(4):3212.
46. Manabe I, Shindo T, Nagai R. Gene expression in fibroblasts and fibrosis: involvement in cardiac hypertrophy. Circ Res. 2002;91(12):1103–13.
47. Maillet M, Lynch JM, Sanna B, York AJ, Zheng Y, Molkentin JD. Cdc42 is an antihypertrophic molecular switch in the mouse heart. J Clin Invest. 2009;119(10):3079–88.
48. Jiang Z, Xu FF, Araki J, Neubig G. How can we know what language models know? Trans Assoc Comput Linguist. 2020;8:423–38.
49. Niu C, Zhang J, Wang G, Liang J. Gatcluster: Self-supervised gaussian-attention network for image clustering. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Cham, Switzerland: Springer; 2020. pp. 735–51.
50. 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(8): 1470-1480. https://doi.org/10.1038/s41592-024-02201-0.
51. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):14049.
52. 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(4):346–60.
53. 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(4):385–94.
54. 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(4):608–15.
55. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018;9(1):4383.
56. Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24(4):593–607.
57. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. Cell Marker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. 2019;47(D1):D721–8.
58. Li Y, Qiao G, Du H, Gao X, Wang G. scKGBERT-41M: A Knowledge-Enhanced Foundation Model for Single-Cell Transcriptomics. Hugging Face. 2025. https://doi.org/10.57967/hf/6742.
59. Li Y, Qiao G, Du H, Gao X, Wang G. scKGBERT: A Knowledge-Enhanced Foundation Model for Single-Cell Transcriptomics. Hugging Face. 2025. https://doi.org/10.57967/hf/6741.

Publisher's Note