scBERT: a Large-scale Pretrained Deep Langurage Model for Cell Type Annotation of Single-cell RNA-seq Data

Wenchuan Wang*, Fan Yang*, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu#, Jianhua Yao#

Author Information

Affiliations

SJTU-Yale Joint Center for Biostatistics and Data Science, Department of Bioinformatics and Biostatistics, School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China

Wenchuan Wang, Hui Lu

AI Lab, Tencent, Shenzhen, Guangdong Province, 518000, China

Wenchuan Wang, Fan Yang, Yuan Fang, Duyu Tang, Junzhou Huang, Jianhua Yao

Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

Yuan Fang

Center for Biomedical Informatics, Shanghai Engineering Research Center for Big Data in Pediatric Precision Medicine, Shanghai Children's Hospital, Shanghai, China

Hui Lu

Contributions

Wenchuan Wang and Fan Yang contributed equally to this work.

Corresponding author

Correspondence to: Hui Lu (huilu@sjtu.edu.cn) and Jianhua Yao (jianhuayao@tencent.com)

Abstract

Reliable cell type annotation is a prerequisite for downstream analysis of single-cell RNA sequencing data. Existing annotation algorithms typically suffer from improper handling of batch effect, lack of curated marker gene lists, or difficulty in leveraging the latent gene-gene interaction information. Inspired by large scale pretrained langurage models, we present a pretrained deep neural network-based model scBERT (*s*ingle-*c*ell *B*idirectional *E*ncoder *R*epresentations from *T*ransformers) to overcome the above challenges. scBERT follows the state-of-the-art paradigm of *pre-train* and *fine-tune* in the deep learning field. In the first phase of scBERT, it obtains a general understanding of gene-gene interaction by being pre-trained on huge amounts of unlabeled scRNA-seq data. The pre-trained scBERT can then be used for the cell annotation task of unseen and user-specific scRNA-seq data through supervised fine-tuning. Extensive and rigorous benchmark studies validate the superior performance of scBERT on various tasks, including cell type annotation, novel cell type discovery, as well as investigation of gene-gene interactions. Thus, scBERT enjoys the advantages of improved generalization and interpretability than existing annotation tools.

Introduction

Since its invention, single-cell RNA sequencing (scRNA-seq) has been extensively used for the characterization of complex tissues and organisms at single-cell level[1–3] , which has revolutionized transcriptomic studies. Accurate cell type annotation on scRNA-seq is the most critical step before downstream analysis for biological and medical researches[4]. Currently, cell type annotation methods can be categorized mainly into three types: annotation using marker gene sets, correlation-based methods, and annotation by supervised classification[5].

Cluster-then-annotate is the commonly-used manner when annotating cell types using maker genes[6]. Manually curated marker genes identified from literature and scRNA-seq datasets are employed to assign cell types for each cluster derived from unsupervised learning[5]. However, selecting the marker genes depends on researchers' prior knowledge, and is therefore prone to bias and errors[7]. In addition, maker

genes for interested cell types are not always available, and novel cell types don't have marker gene sets yet. Besides, most cell types are determined by a set of genes instead of a single marker gene[8]. Without a proper method to integrate the expression information of multiple marker genes, it is difficult to guarantee a unified and accurate cell type assignment for each cluster[9,10]. For example, some automatic annotation methods are built upon the hypothesis that marker genes should have high expression in cells. However, even some well-documented marker genes do not have high expression in all the cells in the corresponding cell types[11]. Therefore, the absence or fluctuation of the expression of these marker genes might significantly affect the preciseness of this type of annotation methods.

Instead of relying on a spot of marker genes, correlation-based methods measure the correlation of gene expression profiles between the query samples and reference dataset[5]. These methods are potentially affected by the batch effect across platforms and experiments[12]. Although there exist batch-effect correction methods, it is still challenging to distinguish true biological diversity from technical differences and thus preserve important biological variations[13]. Meanwhile, the commonly-used similarity measures (such as cosine similarity, Spearman's correlation, and Pearson correlation) may not be robust or efficient for measuring the distance between two sets of high-dimensional sparse scRNAseq data[14].

Annotation by supervised/semi-supervised classification follows the classic paradigm in machine learning that recognizes patterns in the gene expression profiles and then transfers the labels from labelled datasets to unlabeled datasets[5]. This type of methods is more widely used recently due to its robustness to noise and variability of data, and its independence of artificially selected marker genes. Nevertheless, most of these methods need to perform highly-variable-gene (HVG) selection and dimensionality reduction before inputting the data into the classifier due to their limited model capacity [15–19]. However, HVGs are variable across different batches and different datasets, hindering their generalization ability across cohorts[16]. Dimensionality reduction like PCA may make the model lose high-dimensional information and therefore loss gene-level independent interpretability. Furthermore, the parameter settings of HVG selection and PCA in these methods are far to reach a consensus and inevitably introduce additional artificial bias for

performance evaluation[15–19]. Given that the HVGs are selected based on the expression variance across the whole dataset where the dominant cell types account for the most variance, there is the risk of overlooking the key genes of rare cell types. Considering only the HVGs will ignore co-occurrence and biological interaction of genes (especially HVG and non-HVG), which are indicated to be useful for cell type annotation[20]. Besides, simple classifiers such as fully connected networks are not able to capture gene-gene interactions. Therefore, a new method with improved pattern recognition ability is required to overcome the issue of under-fitting to large-scale datasets.

Recently, a growing number of deep learning-based methods have been applied to scRNA-seq data analyses and achieved superior performance[21–23]. BERT (the Bidirectional Encoder Representations from Transformers) is a state-of-the-art Transformer-based language representation learning model. It has made breakthrough progress in the fields of natural language processing due to the powerful self-attention mechanism and long-range information integration capability introduced by Transformer layers[24,25]. BERT introduced the new paradigm of pre-train and then fine-tune, which enables the use of large-scale unlabled data to improve the generalization ability of the model. Inspired by the exciting progress mentioned above, we developed scBERT (*s*ingle-*c*ell *B*idirectional *E*ncoder *R*epresentations from *T*ransformers) model for the cell annotation task of scRNA-seq data. With innovative designs, scBERT formulates the expression profile of each single cell into embeddings for genes (like word embeddings in BERT) for representation learning. Following the pre-train and fine-tune paradigm, we first validated the power of applying self-supervised learning on large-scale unlabeled scRNA-seq data to improve the model's generalization ability and to overcome the batch effect. Extensive benchmark studies also indicated that scBERT can provide robust and accurate cell type annotations. We then explored the interpretability of the model, and showed that scBERT can distinguish the contribution of genes. To the best of our knowledge, scBERT pioneered the application of Transformer architectures in scRNA-seq data analysis with innovatively designed embeddings for genes.

Results

The scBERT algorithm.

We developed a novel deep learning-based model named scBERT to predict cell types for scRNA-seq data (Fig. 1). As implemented in the BERT model, we first pre-trained the scBERT model with millions of unlabled scRNA-seq data via self-supervised learning across sequencing platforms and experiments. Pre-training enables the model to learn general syntax of gene-gene interactions, which helps to remove the batch effects across datasets and further to improve the generalization ability of the model (Fig. S1). The pre-trained scBERT can then be used to annotate cell types for user's scRNA-seq data after being finetuned with reference scRNA-seq data. Finetuning ensures that the output embedding for each gene encodes context information that is more relevant to the transcriptional profiles of the reference dataset. To annotate a query cell, scBERT computes the probability for the provided cell to be any cell type labelled in the reference dataset. Of note, if there is no cell type with significantly high probability, the query cell would be labeled as 'unknown' to prevent incorrect assignment and to allow novel cell type discovery.

The scBERT model has some innovative designs compared to the original BERT model to unleash the power of BERT in the cell type annotation task. First, we replaced the positional embedding for a word in the standard BERT model with gene embedding for a gene to inject the gene-gene interaction information of a given cell into the model (like semantic information of a sentence). Gene embeddings were obtained from the pre-trained gene2vec[26]. The genes that share co-expression patterns retain closer representations, where distributed representation of genes have been proved to be useful for capturing gene-gene interactions[26]. The word embedding in the standard BERT model was substituted with expresson embedding through binning gene expression values. Second, we replaced the Transformer encoder used in BERT with Performer[27]. Compared to Transformer, Performer allows longer input length which makes it possible for each gene to attend to all other genes to gather relevant information when generating the representation of the gene. With Performer, scBERT keeps the full gene-level interpretation, abandons the use of HVGs and dimensionality reduction, and lets discriminative genes and useful interaction come to

124 the surface by themselves. Thus, scBERT allows for the discovery of gene expression patterns that
125 account for cell type annotation in an unbiased data-driven manner.

126 Evaluating cell type annotation performance and robustness by intra-dataset cross-
127 validation.

128 We first estimated the performance of scBERT against alternative approaches on 6 scRNA-seq datasets,
129 comprehensively considering the diversity of data size and data complexity[28]. For comparison, marker
130 gene-based methods (SCINA, Garnett, scSorter), correlation-based methods (Seurat v4, SingleR, scmap-
131 cell, scmap-cluster, Cell-ID(c), Cell-ID(g)), and machine learning methods (SciBet, scNym) are used. For
132 each of the datasets, we applied the 5-fold cross-validation strategy and repeated the procedure 5 times to
133 avoid the influence of random results on the conclusion. To avoid bias introduced by marker selection, we
134 obtained well-documented marker lists associated with well-defined cell types from CellMarker[29]. We
135 compared the cell type annotation performance of each method at cell-level and cell-type-level using the
136 metrics of accuracy and macro F1 score, respectively. Our results show that scBERT surpassed the
137 performance of the alternative methods in both accuracy and macro F1 score on all datasets (Fig. 2a, Fig.
138 S2). For the Zheng 68K dataset from human peripheral blood mononuclear cells (PBMC) (Fig. 2b-c, Fig.
139 S4), the performance of all methods does not reach 0.9 (both accuracy and macro F1 score). This may be
140 due to the high similarity between some cell populations (i.e., CD8+ cytotoxic T cells and
141 CD8+/CD45RA+ naive cytotoxic T cells), which increased the difficulty for classification. Still and all,
142 scBERT showed better performance than other methods (scBERT macro F1 score = 0.691, accuracy =
143 0.759; best macro F1 score of other methods = 0.659, accuracy = 0.704). Furthermore, compared to
144 alternative methods, scBERT achieved the highest macro F1 scores (0.788 vs. 0.617, P-value=9.025e-05;
145 0.801 vs. 0.724, P-value= 2.265e-05) for the CD8+ cytotoxic T cells and the CD8+/CD45RA+ T cells
146 that are highly similar and hard to distinguish in previous studies[30].
147 To explore whether the number of cells of a reference dataset affects the performance of scBERT, we
148 constructed a series of fine-tuning datasets from the Zheng 68K dataset by uniformly subsampling it

149 proportionally from 10% to 90% (Fig. 2d). On the smallest fine-tuning dataset (10% cells from Zheng
150 68K dataset), scBERT outperformed most of the alternative methods except SciBet, which achieved a
151 slightly better performance. Overall, the performance of scBERT improved rapidly as the reference cell
152 number increased. scBERT beated all other models in terms of both the macro F1 score and accuracy
153 when using 90% of data. We noted that the performance of machine learning-based methods benefited
154 from the increasing size of reference datasets while the performance of correlation-based methods only
155 showed slightly changes, suggesting that machine learning models and deep learning models have better
156 learning capabilities compared to correlation-based methods.
157 We next tested the robustness of scBERT when the distributions of cells across multiple cell types were
158 severely biased. In this study, four PBMC cell types from the Zheng 68K dataset, CD8+ cytotoxic T cells,
159 CD19+ B cells, CD34+ cells, CD8+/CD45RA+ naive cytotoxic cells, with transcriptomic similarity
160 between each pair, were selected for class-imbalanced tests. At the fine-tuning stage, cells of the four
161 types were randomly sampled with the numbers 10000, 100, 10000, 100 as the reference data. At the
162 inference stage, 100 cells per cell type formed the query data. scBERT performed the most accurate cell
163 type prediction among the cell type annotation methods, with accuracy reaching 0.840 and macro F1
164 score reaching 0.826. From the confusion matrices (Fig. 2e), we can see that Seurat v4 is prone to
165 misidentify CD8+ cytotoxic T cells to CD8+/CD45RA+ naive cytotoxic cells. In contrast, scBERT
166 performed the least misclassification rate even though high similarity exists between the two cell
167 populations. In addition, SingleR misclassified all CD19+ B cells as other cell populations, because of the
168 small population of CD19+ B cells. Overall, the results indicate that scBERT is robust to class-
169 imbalanced reference datasets.

170 Cell type annotation across scRNA-seq datasets from different experiments.

171 In real-world circumstances, the reference dataset and query dataset are always sourced from multiple
172 studies, and even different sequencing platforms. As reported in previous studies, the batch effects existed
173 in inter-dataset scenarios can lead to poor performance on cell type annotation. Here, we benchmarked

174 scBERT and alternative methods by employing a leave-one-dataset-out strategy with four human
175 pancreas datasets generated by distinct sequencing techniques (Baron[31], Muraro[32], Segerstolpe[33], and
176 Xin[34]) (Fig. 3a). For each experiment, we used three of the datasets as a reference for fine-tuning on a
177 classifier and then predicted the labels of cells in the fourth leave-out dataset. Considering the difference
178 in cell populations among these datasets, all datasets were aligned, retaining only four kinds of pancreas
179 islet cells (alpha cell, beta cell, delta cell, and gamma cell) that are common in these datasets. scBERT
180 showed a slight cut over several advanced methods like scNym and outperformed other popular methods
181 including Cell-ID (Fig. 3b-d, Fig. S4). Compared to the intra-dataset tests above, the performance of
182 scBERT on inter-dataset tests was still relatively equivalent. scBERT correctly annotated most cells (>
183 97%) in the Muraro dataset and over 99% in the other three datasets, demonstrating the superb and stable
184 performance of our method in cross-cohort tasks.

185 Discovery of novel cell types in the query dataset.

186 As mentioned above, the cell types that exist in the reference dataset may not cover all cell types present
187 in the query dataset in most tasks. To evaluate the ability of scBERT in detecting unseen cell types, we
188 assessed scBERT performance on the MacParland dataset[35] from human liver tissue by removing several
189 cell populations during the scBERT training process. In this experiment, we took four immune cells for
190 novel cell type simulation, which were absent from other liver datasets. To comprehensively consider the
191 overall prediction accuracy and the sensitivity to unseen cell populations, metrics calculated on both
192 novel and known cell types were employed as the performance indicators. Not surprisingly, scBERT
193 performed the best on novel cell types using the default threshold, and further achieved the first-class
194 performance on the known cell types (Fig. 4a). Taken together, these results suggest that scBERT can
195 correctly discover novel cell types that are not present in original reference datasets and remain accurate
196 prediction performance of other cell types in the meantime. Furthermore, the output probability of
197 scBERT models can serve as the confidence score that indicates how confident scBERT assigns cell type
198 labels, according to its reliable results in our experiments (Fig. 4b).

Investigating scBERT model interpretability.

Finally, we conducted a comprehensive interpretability analysis to explore the key genes for decision-making since scBERT models were built on the self-attention mechanism and all genes' representation remained in the end of our workflow. We integrated all the attention matrices into one matrix by taking element-wise average across all attention matrices in multi-head multi-layer Performers. In this average attention matrix, each value $A(i,j)$ represented how much attention from gene i was paid to gene j. In oder to focus on the importance of genes to each cell, we summed the attention matrix along columns into an attention-sum vector, and its length is equal to the number of genes. In this way, we could obtain the top attention genes corresponding to a specific cell type compared to other cell types. Here, we took the Muraro dataset as an illustration and top attention gene lists were produced for the four kinds of pancreas islet cells (Fig. 5a). We found that top attention genes included markers of specific cell types reported by previous studies like LOXL4 for alpha cell and ADCYAP1 for beta cell[36]. On the other hand, almost all the top attention genes except markers were identified as differentially expressed genes using DESeq[37], and these genes could be potential markers (Fig. 5c, Fig. S5). For instance, SCD5 has not been reported as a cell type specific marker for beta cells, but in a GWAS study, a novel loci for type 2 diabetes susceptibility was fine-mapped to a coding variant of SCD[38]. The results demonstrated that scBERT could facilitate understanding the cell type annotated by our model and provide some support for further biological findings.

We also performed enrichment analysis for the top 50 attention gene lists using various gene-set libraries and the results revealed there were some interesting relationships between top enriched terms and the corresponding cell types (Fig. 5b). In particular, with the cell type associated gene-set library from PanglaoDB, the top one enriched term for each type always hits the true cell population. As another example, insulin secretion and AMPK signal pathway, the top two enriched KEGG pathways in beta cells, are vital to beta cell function. Meanwhile, we extracted the cell representation provided by scBERT

224 and found that it's more distinguishable for cell type annoation than the original raw gene expression of
225 the cell based on the clustering performance (Adjusted Rand Index: 0.9514 vs 0.8731). It indicates the
226 efficiency of scBERT on learning cell-level representation, which can also be used for downstream
227 analysis (Fig. 5d).

228
Discussion

229 In order to improve the generalization ability of the cell type annotation algorithm and the interpretability
230 of individual gene importance, we developed scBERT, a deep learning model with a multi-head attention
231 mechanism and self-supervised strategy, to learn domain-irrelevant gene expression pattern and
232 interaction from the whole genome expression of large-scale unlabeled scRNA-seq data, and made it
233 possible to trace back to the importance of each individual gene for model interpretability.

234 scBERT surpasses the existing advanced methods on diverse benchmarks collectively involving 6 single-
235 cell datasets, 5 organs/tissues, more than 50 cell types, about 100,000 cells, and 4 single-cell omics
236 technologies. Among the current mainstream methods that use marker genes for classification, some
237 methods that utilize the gene expression pattern for cell type annotation. Both kinds of methods were
238 reported to achieve good performance on variable cell type annotation tasks, indicating that both types of
239 data imply discriminative information for different cell types. To investigate the effect of marker genes
240 and the discriminant ability of remaining expression patterns consisting of only the non-marker genes, we
241 conducted experiments where marker genes were eliminated gradually, leaving the remaining expression
242 profiles for cell type annotation(Fig. S6). The results prove that the marker genes are important for the
243 cell type annotation. However, in addition to the marker genes, there are still informative gene patterns
244 that have a good distinguishing power on cell type classification.

245 To the best of our knowledge, there is currently no research on applying Transformer architectures to
246 gene expression data analysis. After extensive literature review on the advances in the NLP field, we
247 designed the gene expression profile as rational gene embedding and expression embedding for model
248 inputs that work well with Transformer architectures. In addition, the existing deep learning methods

requires HVG selection or various ways of dimensionality reduction to reduce the data dimension. HVG selection reduces its generalization ability due to the introduction of bias in the parameter setting and dimensionality reduction directly impairs the gene-level interpretability of the deep learning model. In this regard, we employed an effective Transformer architecture, Performer, to process all genes in parallel, fusing the interaction information between genes and optimize the gene expression representation. Finally, the gene-level importance can be naturally traced back through the attention mechanism.

The originally designed end-to-end scBERT framework with gene expression embedding and self-learning strategy has superior performance, interpretability and generalization potential on cell type annotation task. Beyond that, scBERT can also be migrated to other tasks by simply modifying output and supervision signals. scBERT, as an effective cell type annotation tool, has been released on the platform for public usage. We hope that our model could stimulate an improved understanding of cell type associated gene-gene interactions and facilitate the development of cell type annotation and downstream analysis. For future work, we would like to train an advanced scBERT model to infer various tissue types and cell types simultaneously through larger-scale pre-training and incremental learning. Furthermore, since gene-gene interactions have been captured and described in signaling pathways, gene regulation networks, etc., it can be foreseen that Graph Transformer constructed in conjunction with this kind of prior information may also contribute to the scRNA-seq data analysis[39].

Methods

The scBERT model

Gene embedding. In natural language processing (NLP), the inputs of the BERT model are word embeddings, a set of real-valued vectors in a predefined vector space that represent individual words. The word embedding technology helps to better represent the text by assuring the words with similar meanings have a similar representation[40]. While from the aspect of scRNA-seq, the inputs are constituted

by individual genes and a predefined vector space is in need for representing the similarity between genes. Hence we employed gene2vec[26] to specifically encode gene embeddings. In this way, the difficulty of model training is reduced, with the help of the inter-gene relationship provided by prior knowledge.

Expression embedding. In spite of the gene embedding, there is also a challenge on how to utilize the transcription level of each gene, which is actually a single continuous variable. It's worth noting that the frequency of the word occurrence in a text is valuable information for text analysis and is often transformed as bag-of-words by term frequency statistic analysis for downstream tasks in the area of NLP[41]. The gene expression can also be considered as the occurrence of each gene that has already been well-documented in a biological system. From this insight, we applied the conventionally-used term frequency analysis method that discretizes the continuous expression variables by binning and converts them into 200-dimensional vectors, which are used as token embedding for the scBERT model.

Model Building. The quadratic computational complexity of the BERT model with Transformer as the basic unit does not scale very well to long sequence while the gene number of scRNA-seq can be up to more than 20, 000. To this end, a matrix decomposition version of the Transformer, Performer was employed to enlarge the sequence length. The regular dot-product attention in Transformer is a mapping of Q, K, V, $which$ are encoded representations of the input *queries*, *keys*, and *values* created for each unit respectively. The bidirectional attention matrix is formulated as:

$$Att(Q,K,V) = D^{-1}(QK^T)\,V, \qquad D = diag(QK^T 1_L) \qquad (1)$$

where $Q = W_q X$, $K = W_K X$, $V = W_V X$ are linear transformations of the input X, and W_Q, W_K, W_V are the weight matrices as parameters. 1_L is the all-ones vector of length L, and $diag(\cdot)$ is a diagonal matrix with the input vector as the diagonal.

The attention matrix in Performer is described as follows:

$$\widehat{Att}(Q,K,V) = \widehat{D}^{-1}\big(Q'((K')^T V)\big), \qquad \widehat{D} = diag\big(Q'((K')^T 1_L)\big) \qquad (2)$$

where $Q' = \emptyset(Q)$, $K' = \emptyset(K)$, and the function $\emptyset(x)$ is defined as:

297 $$\emptyset(X) = \frac{c}{\sqrt{m}} f(\omega^T X) \qquad (3)$$

298 where c is a positive constant, ω is a random feature matrix, and m is the dimesionality of the matrix.
299 Here, we constructed our model with 6 Performer enocoder layers and 10 heads for each layer.

300 The model training process contains two stages: self-supervised learning on unlabeled data to get a pre-
301 trained model and supervised learning on the specific cell type annotation tasks to get the fine-tuned
302 model.

303 *Self-supervised learning on unlabeled data.* In this study, we followed the conventional self-learning
304 strategy of the BERT model in NLP tasks by randomly masking the input data value and making a
305 prediction based on the remaining inputs. Considering the dropout zeros phenominan[42], we randomly
306 masked the non-zero gene expression and then reconstructed the original inputs by model predictions
307 using the remaining genes. We ultilized cross-entropy loss as the reconstruction loss , formulated as:

308 $$L_{Rec} = -\sum_{i=1}^{M}\sum_{j=1}^{N} y_{i,j} \log(p_{i,j}) \qquad (4)$$

309 where M is the number of cells, N is the number of masked gene expression values, while $y_{i,j}$ and $p_{i,j}$ are
310 the true expression and predicted expression of gene j in cell i respectively. Through this self-supervised
311 strategy, the model can learn general deep representations of gene expression patterns on the large amount
312 of unlabeled data which might alleviate the efforts of the downstream fine-tuning process.

313 *Supervised learning on specific tasks.* The output of scBERT was a 200-dimension feature corresponding
314 to each gene and a one-dimensional convolution was applied for abstract information extraction for each
315 gene feature. Then a 3-layer neural network was applied as the classification head and transformed the
316 gene features into the probability for each cell type. Cross-entropy loss was also employed as the cell type
317 label prediction loss, calculated as:

318 $$L_{Pred} = -\sum_{i=1}^{M} z_i \log(q_i) \qquad (5)$$

319 where z_i and q_i indicates the ground truth cell type label and predicted label of cell i.

320 Datasets

321 Since the model training consists of two stages: self-supervised learning on unlabeled data and fine-tuning
322 on task-specific data, the dataset used in the two stages were collected from different sources to avoid
323 data leakage. At the first stage, large amounts of data without annotations were used for general pattern
324 learning. At the second stage, task-specific data with well-annotated cell labels were required for the
325 subsequential systematic benchmarking of scBERT and SOTA methods. To this end, we only included
326 scRNA-seq datasets that provided highly credible cell type annotation and had been cited by the majority
327 of the cell type annotation methods for performance evaluation.

328 *The Panglao dataset.* The Panglao dataset[43] was downloaded from the PanglaoDB website
329 (https://panglaodb.se/). In brief, PanglaoDB integrated 209 human single-cell datasets consisting of 74
330 tissues, 1,126,580 cells that originated from different experiment sources via various platforms. In this
331 study, we used scRNA-seq data from PanglaoDB for first stage pre-training. No annotations or cell labels
332 were used at the first stage since the self-learning strategy was employed and only the genes and their
333 expression level were needed as inputs for the scBERT model.

334 *Zheng68k dataset.* The Zheng68k is a classic PBMC dataset that are widely used for cell type annotation
335 performance acessment[30]. It contains about 68, 000 cells within eleven subtypes of cells: CD8+ Cytotoxic
336 T cells (30.3%), CD8+/CD45RA+ Naive Cytotoxic cells (24.3%), CD56+ NK cells (12.8%), CD4+/CD25
337 T Reg cells (9.0%), CD19+ B cells (8.6%), CD4+/CD45RO+ Memory cells (4.5%), CD14+ Monocyte
338 cells (4.2%), Dendritic cells (3.1%), CD4+/CD45RA+/CD25- Naive T cells (2.7%), CD34+ cells (0.4%),
339 CD4+ T Helper2 cells (0.1%). Zheng68k dataset contains rare cell types and the distribution of cell types

340 in this dataset is imbalanced. Meanwhile, strong correlations between cell types make cell types difficult
341 to differentiate.

342 *Pancreas datasets.* The pancreas datasets consist of Baron, Muraro, Segerstolpe, and Xin. The cell type
343 labels were aligned and four cell types were included. The Baron dataset was downloaded from the Gene
344 Expression Omnibus (GEO) (accession number: GSE84133), the protocol was inDrop[31]. The Muraro
345 dataset was downloaded from GEO (accession number: GSE85241), the protocol was CEL-Seq2[36]. The
346 Segerstolpe dataset was accessed from ArrayExpress (accession number: E-MTAB-5061), the protocol
347 was Smart-Seq2[37]. The Xin dataset was downloaded from GEO (accession number: GSE81608), the
348 protocol was SMARTer[34]. The above pancreas datasets were generated from different experiment
349 platforms(Table S1).

350 *McParland dataset.* McParland dataset[35] from human liver tissue contains 20 hepatic cell populations
351 from the transcriptional profiling of 8444 cells. We downloaded the data from GEO (accession number:
352 GSE115469) and generated the cell type annotation following the authors' reported procedure.

353 Data preprocessing

354 As for the data provided in gene expression matrix format, log-normalization using a size factor of 10,
355 000 and quality control by filtering cell outliers with less than 200 genes expressed were performed on the
356 data. As for the input of scBERT, no dimension reduction or HVG selection was processed since scBERT
357 has a capacity of more than 20 000 genes as input and retains full gene-level interpretability.

358 Benchmarking

359 For benchmarking, we implemented state-of-the-art methods from the three categories. Among them,
360 SCINA, Garnett, and scSorter stand for the annotation using marker gene databases; Seurat, SingleR,
361 CellID, and scmap are correlation-based methods; scNym and Scibet are the SOTA methods that conduct

362 annotation by supervised/semi-supervised classification. Notably, this categorization depends on how the
363 most important process is conducted. As for marker gene-based annotation, CellMarker database with
364 manually curated cell-type markers using a literature search of over 100 000 papers was applied for
365 the marker database[29]. No mannual selection of the marker genes was included for an unbiased and
366 fair comparison of all the methods.

367 *scNym.* scNym is a recently proposed semi-supervised learning annotation method that leverages the
368 unlabeled target data through training a domain adversary[44]. It requires no prior manual specification of
369 marker genes. It makes use of the target data by domain adaptation and achieved the SOTA performance
370 on several tasks. However, users have to endure the inconvenience that they must re-train the model on
371 each batch of the new coming data.

372 *SciBet.* Scibet is a supervised classification method that selects genes using E-test for multinomial model
373 building and annotates cell types for a new cell in the test set[45]. We adopted SciBet in R package for
374 benchmarking.

375 *Seurat.* As a popular single-cell data analysis pipeline, Seurat is widely used by biologists and clinical
376 experts. Seurat maps the query samples to the reference dataset in a reference-based annotation manner[46].
377 In this study, we adopted the implementation of the cell type annotation of Seurat V4.0 and followed the
378 cell type annotation tutorial provided by Seurat for benchmarking.

379 *SingleR.* SingleR is a reference-based analysis method that calculates the Spearman coefficient on
380 variable genes and aggregates the coefficients to score the cell for each cell type[47]. It iterates the above
381 process by subsampling top genes until the most closely related cell types are distinguished. The SingleR
382 package was applied for benchmarking.

383 *CellID.* CellID is a clustering-free multivariant statistical method for cell type annotation that performs
384 dimensionality reduction, evaluates gene-to-cell distance and extracts gene signatures for cells (cell-to-

385 cell strategy) and groups (group-to-cell strategy)[48]. In this study, both strategies from the R package were
386 used for benchmarking.

387 *scmap.* scmap is a reference-based annotation method including two strategies: scmap-cluster and scmap-
388 cell. Scmap-cluster maps individual cell from query samples to certain cell types in the reference dataset
389 while scmap-cell maps individual cell from query samples to individual cells in a reference dataset[49].
390 Both scmap-cluster and scmap-cell perform feature selection and calculate the distance (cosin and
391 euclidean distance). The reference is searched for the nearest neighbors to a query cell. We used the R
392 package of scmap for the scmap-cluster and scmap-cell tools.

393 *SCINA.* SCINA is a typical marker gene-based annotation method that requires a list of marker genes for
394 different cell types and identifies the cell types based on the assumption that there exists a bimodal
395 distribution for each marker gene and the higher modes belong to the relevant cell type[9]. We used the
396 Scina package for benchmarking.

397 *Garnett.* Garnett requires a user-defined cell hierarchy of cell types and marker genes as input. Garnett
398 aggregates marker gene scores using term frequency-inverse document frequency transformation and uses
399 an elastic-net regression-based model for annotation[10]. We adopted the original R package to use the
400 garnet model for benchmarking.

401 *scSorter.* Scsorter employs marker genes and the HVGs for clustering and cell type annotation based on
402 the observation that most marker genes do not consistently preserve high expression levels in all the cells
403 that belong to the related cell types[50]. Here we adopted the R implement of Scsorter.

404 Benchmarking

405 To assess the performance of the annotation methods under different scenarios, five pairs of reference and
406 test datasets were generated and the performance was evaluated using scBERT and all the above methods.
407 The details were listed below.

408 *Performance on the intra-dataset data using cross-validation.* The PBMC data from Zheng68k with high
409 inter-class similarity were employed for testing the performance on the intra-dataset in a 5-fold cross-
410 validation manner. Notably, the reference dataset in this section also refers to the training dataset for the
411 supervised methods, including scBERT.

412 *Performance on the inter-dataset data.* To evaluate the robustness of the methods on cross-cohort data
413 with batch effects from different single-cell sequencing platforms, we tested the methods on four pancreas
414 datasets (Baron, Muraro, Segerstolpe, and Xin), taking three datasets as the training set and the remaining
415 one as the test set each time. In real application scenarios, it's a common way to include all available data
416 for training a model with better generalization ability and test the model on a set of new coming data.

417 *Estimating the influence of reference cell amount on the performance.* The number of reference cells is
418 prone to influence the model performance. In this study, 10%, 30%, 50%, 70%, and 90% of the PBMC
419 cells from the Zheng 68K dataset were randomly selected as the reference for finetuning while the
420 remaining as the query samples for testing.

421 *Class-imbalanced data tests.* Following the construction method for class-imbalanced data444, we
422 collected four PBMC cell types, CD19+B, CD8+Cytotoxic T, CD34+, and CD8+/CD45RA Naïve
423 Cytoxic cells that contain various levels of similarity across cell types from Zheng68K data. The cells of
424 the four types were randomly selected with the cell numbers 10000, 100, 10000, 100 for finetuning. As
425 for model testing, 100 cells were randomly selected per cell type.

426 *Novel cell type detection.* Human liver tissue was reported to assess the unknown cell type identification.
427 Here we adopted this dataset with 8434 cells belonging to 14 cell types. Following the schema proposed
428 in the previous study[7], we performed leave-out one cell type evaluation by removing one cell type from
429 the reference dataset while keeping the cell type groups in the query dataset. The evaluation process was
430 iterated on each cell type.

Assessment on the necessity of self-learning on unlabeled data. To illustrate the necessity of the self-learning process of the scBERT method, the performance gain was evaluated on the model after self-learning and finetuning compared to the model training from scratch.

Evaluation metrics. Accuracy, macro-F1 score, and confusion matrix were applied in all the benchmarking experiments to evaluate the performance of each classifier.

Model interpretability.

Most supervised methods perform HVG selection and dimensionality reduction before inputting the data into the model due to their simplified network architecture and low model capacity hence destroying the full gene-level resolution and interpretability. In contrast, the attention mechanism employed in scBERT can naturally provide hints for the decision-making of the model using every individual gene. The scBERT model extracts the useful attention pattern on gene expressions and interactions from a large scale of various scRNA-seq data. It alleviates the efforts of the fine-tuning process on the specific downstream tasks. The attention weights reflect the contribution of each gene and the interaction of gene pairs. The attention weights can be obtained from equation (1) modified by replacing V with V^0, where V^0 contains one-hot indicators for each position index. The attention weights were visualized and the top genes were sent to Enrichr[33] for enrichment analysis.

It's well known that marker genes play a key role in cell type annotation for marker gene-based annotation and most of the reference-based annotation. Even some of the supervised-based methods are heavily dependent on the prior marker gene knowledge. In this study, we remove the marker genes from the reference dataset and query dataset to explore whether the genes that are not in the marker gene lists can constitute an informative pattern for distinguishing different cell types.

Statistical analysis

The Wilcoxon test was applied for the significance test. Cross-validation was employed in all the benchmarking experiments and standard deviations were drawn in the figures. The significance was calculated by wilconxin test on the paired groups.

Data Availability

All data used in this study are publicly available and the usages are fully illustrated in the Method section.

Code Availability

The codes are implemented in Python. The preprocessing, scBERT modeling, and benchmarking processes are available on https://github.com/yfzon/scBERT with detailed instructions.

References

1. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, (2018).
2. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature 2019 566:7745* **566**, 496–502 (2019).
3. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature 2018 562:7727* **562**, 367–372 (2018).
4. Zhao, X., Wu, S., Fang, N., Sun, X. & Fan, J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Briefings in Bioinformatics* **21**, 1581–1595 (2020).
5. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal* **19**, 961–969 (2021).
6. Cao, Y., Wang, X. & Peng, G. SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. *Frontiers in Genetics* **0**, 490 (2020).
7. Huang, Q., Liu, Y., Du, Y. & Garmire, L. X. Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics, Proteomics & Bioinformatics* (2020) doi:10.1016/J.GPB.2020.07.004.
8. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).
9. Zhang, Z. *et al.* SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes 2019, Vol. 10, Page 531* **10**, 531 (2019).

484 10. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell
485 atlases. *Nature Methods 2019 16:10* **16**, 983–986 (2019).

486 11. Grabski, I. N. & Irizarry, R. A. A PROBABILISTIC GENE EXPRESSION BARCODE FOR ANNOTATION OF
487 CELL-TYPES FROM SINGLE CELL RNA-SEQ DATA A PREPRINT. (2020)
488 doi:10.1101/2020.01.05.895441.

489 12. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-
490 sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*
491 *2018 36:5* **36**, 421–427 (2018).

492 13. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA
493 sequencing data. *Genome Biology 2020 21:1* **21**, 1–32 (2020).

494 14. Serra, A., Coretto, P., Fratello, M. & Tagliaferri, R. Robust and sparse correlation matrix
495 estimation for the analysis of high-dimensional genomics data. *Bioinformatics* **34**, 625–634
496 (2018).

497 15. Ma, F. & Pellegrini, M. ACTINN: automated identification of cell types in single cell RNA
498 sequencing. *Bioinformatics* **36**, 533–538 (2020).

499 16. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: accurate supervised
500 method for cell-type classification from single-cell RNA-seq data. *Genome Biology 2019 20:1* **20**,
501 1–17 (2019).

502 17. Cao, Z.-J., Wei, L., Lu, S., Yang, D.-C. & Gao, G. Searching large-scale scRNA-seq databases via
503 unbiased cell embedding with Cell BLAST. *Nature Communications 2020 11:1* **11**, 1–13 (2020).

504 18. Xie, P. *et al.* SuperCT: a supervised-learning framework for enhanced characterization of single-
505 cell transcriptomic profiles. *Nucleic Acids Research* **47**, e48–e48 (2019).

506 19. Li, C. *et al.* SciBet as a portable and fast single cell type identifier. *Nature Communications 2020*
507 *11:1* **11**, 1–8 (2020).

508 20. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications 2020*
509 *11:1* **11**, 1–9 (2020).

510 21. Wang, T. *et al.* MOGONET integrates multi-omics data using graph convolutional networks
511 allowing patient classification and biomarker identification. *Nature Communications 2021 12:1*
512 **12**, 1–13 (2021).

513 22. Wang, T. *et al.* BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing
514 batch correction reveals hidden high-resolution cellular subtypes. *Genome Biology 2019 20:1* **20**,
515 1–15 (2019).

516 23. Menden, K. *et al.* Deep learning–based cell composition analysis from tissue expression profiles.
517 *Science Advances* **6**, (2020).

518 24. Parmar, N. *et al.* Image Transformer. 4055–4064 (2018).

519 25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional
520 Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North*
521 *American Chapter of the Association for Computational Linguistics: Human Language*
522 *Technologies - Proceedings of the Conference* **1**, 4171–4186 (2018).

523 26. Du, J. *et al.* Gene2vec: distributed representation of genes based on co-expression. *BMC*
524 *Genomics 2019 20:1* **20**, 7–15 (2019).

525 27. Choromanski, K. *et al.* Rethinking Attention with Performers. (2020).

526 28. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA
527 sequencing data. *Genome Biology 2019 20:1* **20**, 1–19 (2019).

528 29. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse.
529 *Nucleic Acids Research* **47**, D721–D728 (2019).

530 30. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature*
531 *Communications 2017 8:1* **8**, 1–12 (2017).

532 31. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals
533 Inter- and Intra-cell Population Structure. *Cell Systems* **3**, 346-360.e4 (2016).

534 32. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems* **3**, 385-
535 394.e3 (2016).

536 33. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and
537 Type 2 Diabetes. *Cell Metabolism* **24**, 593–607 (2016).

538 34. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell*
539 *Metabolism* **24**, 608–615 (2016).

540 35. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic
541 macrophage populations. *Nature Communications 2018 9:1* **9**, 1–21 (2018).

542 36. Nica, A. C. *et al.* Cell-type, allelic, and genetic signatures in the human pancreatic beta cell
543 transcriptome. *Genome Research* **23**, 1554–1562 (2013).

544 37. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nature*
545 *Precedings 2010* 1–1 (2010) doi:10.1038/npre.2010.4282.1.

546 38. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-
547 density imputation and islet-specific epigenome maps. *Nature Genetics 2018 50:11* **50**, 1505–
548 1513 (2018).

549 39. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. Graph Transformer Networks. *Advances in Neural*
550 *Information Processing Systems* **32**, (2019).

551 40. Goldberg, Y. Neural Network Methods for Natural Language Processing. *http://dx.doi.org/*
552 *10.2200/S00762ED1V01Y201703HLT037* **10**, 1–311 (2017).

553 41. Zhang, Y., Jin, R. & Zhou, Z.-H. Understanding bag-of-words model: a statistical framework.
554 *International Journal of Machine Learning and Cybernetics 2010 1:1* **1**, 43–52 (2010).

555 42. Kharchenko, P. v, Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential
556 expression analysis. *Nature Methods 2014 11:7* **11**, 740–742 (2014).

557 43. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse
558 and human single-cell RNA sequencing data. *Database* **2019**, 46 (2019).

559 44. Kimmel, J. C. & Kelley, D. R. Semi-supervised adversarial neural networks for single-cell
560 classification. *Genome Research* **31**, gr.268581.120 (2021).

561 45. Li, C. *et al.* SciBet as a portable and fast single cell type identifier. *Nature Communications 2020*
562 *11:1* **11**, 1–8 (2020).

563 46. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

564 47. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional
565 profibrotic macrophage. *Nature Immunology 2019 20:2* **20**, 163–172 (2019).

566 48. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity
567 recognition at the single-cell level with Cell-ID. *Nature Biotechnology 2021 39:9* **39**, 1095–1102
568 (2021).

569 49. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data
570 sets. *Nature Methods 2018 15:5* **15**, 359–362 (2018).

571 50. Guo, H. & Li, J. scSorter: assigning cells to known cell types according to marker genes. *Genome*
572 *Biology 2021 22:1* **22**, 1–18 (2021).

573

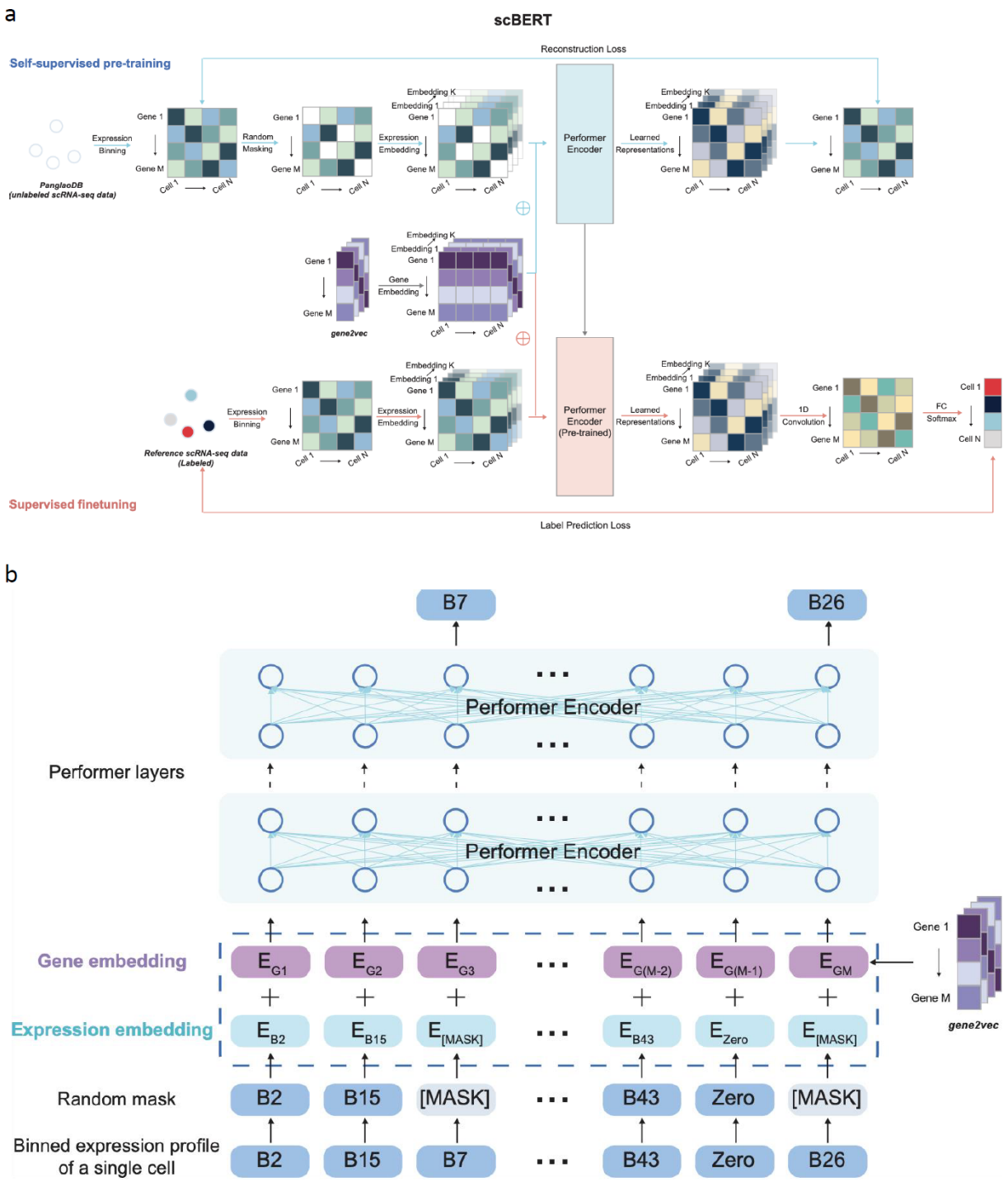

Fig. 1: Overview of the scBERT model.

a, Self-supervised learning on unlabeled data and finetuning on task-specific data. The data collected from PanglaoDB were standardized and preprocessed. We binned the expression of each gene into discrete variable as fixed-length expression embedding by adopting the bag-of-word method in NLP, considering the gene expression acquired from the count matrix is naturally a kind of frequency statistic of the gene

transcription in a cell. By masking the expression of certain genes and then using the expression pattern of remaining genes to fill in the missing positions, scBERT can recognize and utilize the whole expression pattern lying in the scRNA-seq data. The gene identity is converted into gene embedding using gene2vec, ensuring that the genes with similar expression profiles have closer embeddings, which is favorable for language model training. The encoder used in this study is the Performer, which has an enlarged receptive field that accommodates up to 20,000 genes to retain gene-level backtracking capabilities and can capture gene-gene interactions with self-attention mechanism for better gene-level representation. A reconstruction loss function is applied to estimate the discrepancy between the model prediction and the original unmasked data. By this means, we generated a pre-trained model that is robust to batch effects from various sequencing platforms and experiments and tolerant to the dropout by self-supervised learning on the large-scale scRNA-seq data. To deal with a specific cell type annotation task, the original scRNAseq data are input to the pre-trained model followed with 1-dimensional convolution layer and fully connected layers to generate the final cell type prediction, under the supervision of the cell type label for fine-tuning. b, Illustration of the embeddings for scBERT. The preprocessed scRNA-seq data are converted into discretized expression, and non-zero expressions are randomly masked. Taking the first gene as an example, the gene embedding E_{G1} (represents the gene identity from gene2vec) and the expression embedding E_{B2} (represents that the gene expression ranks in the second bin and transformed to the same dimension as the E_{G1}) are summed and inputted to the scBERT to generate gene representation for subsequential self-supervised learning and finetuning.

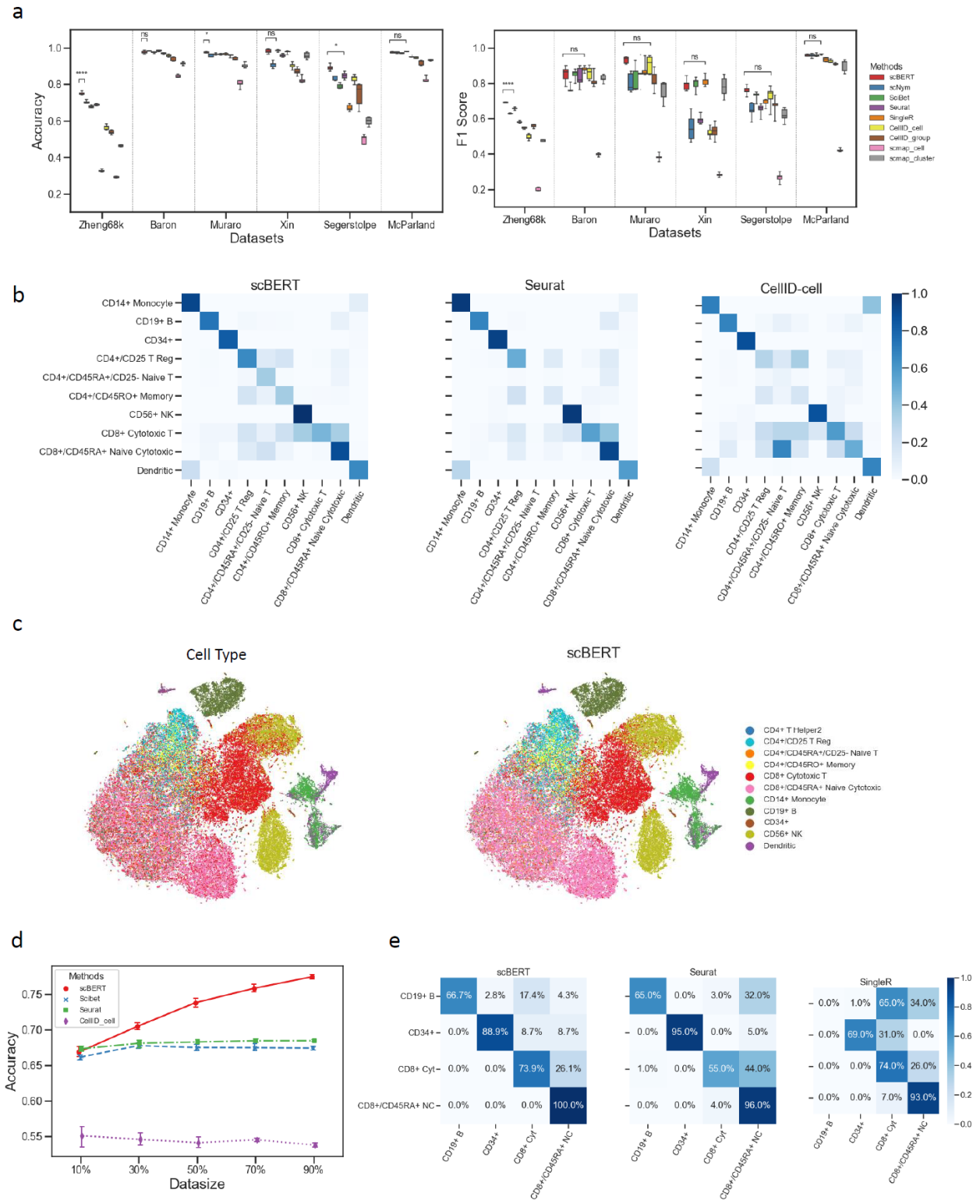

Fig. 2: Benchmarking and robustness evaluation by intra-dataset cross-validation.

a, Performance of cell type annotation methods measured by accuracy and F1 score on n=6 datasets using 5-fold cross-validation across 5 random repeats. Box plots showing the median (center lines), interquartile

range (hinges), and 1.5 times the interquartile range (whiskers). Performance of marker gene-based methods on the Zheng68K dataset are shown in Figure S2. b, Heatmaps for the confusion matrices of the cross-validation results on the Zheng68K dataset for scBERT, Seurat and CellD-Cell. c, t-SNE plot of the whole Zheng68K dataset (n= 68450 cells). Left panel is colored by expert annotated type from the original research, right panel is colored by scBERT prediction results. d, The influence on the cell type annotation performance by spliting different proportions of the Zheng68K dataset as the reference set for finetuning. e, Heatmaps for the confusion matrices of cross-validation on the imbalanced dataset constructed from Zheng 68K dataset. The detailed construction process is introduced in the method section. The marker genes used for the marker gene-based methods are listed in Table S2.

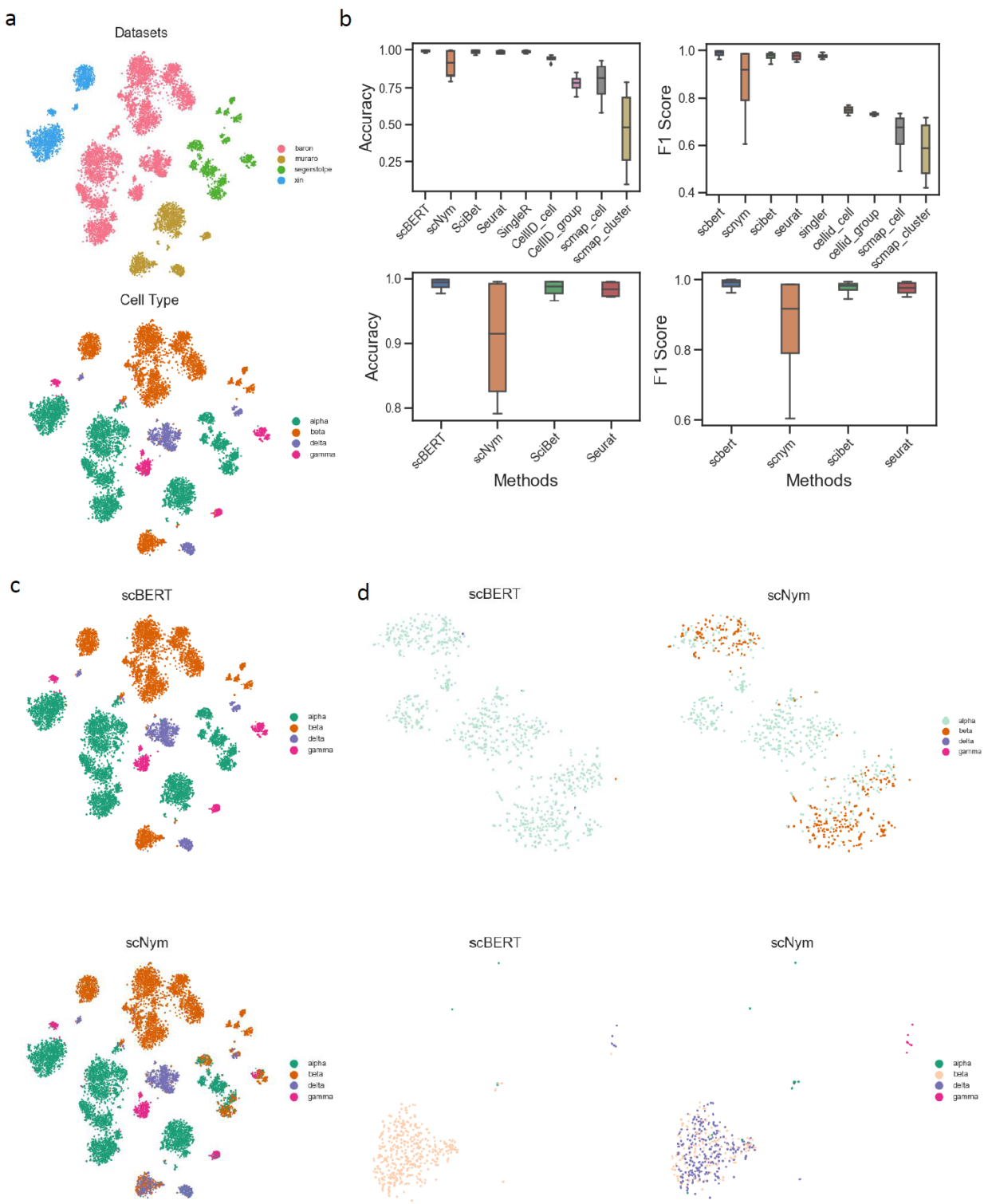

Fig. 3: Performance of scBERT on cell type annotation task across independent datasets generated using different single-cell sequencing technologies.

a, t-SNE representation of 34806 cells from four independent datasets (Baron, Muraro, Segerstolpe, Xin) by different sequencing platforms (inDrop, CEL-Seq2, SMART-Seq2, SMARTer). Cells are colored by the source dataset and annotated cell types provided by the atlas from the original researches. b, Comparison between different methods using accuracy and F1 score (top panel) and a zoomed-in plot of accuracy and F1 score of the top tier methods (bottom panel). The lower and upper hinges denote the first and third quartiles, with the whiskers in the range of 1.5 times the interquartile. c, t-SNE representation of the cells colored by scBERT prediction and scNym prediction, resepectively. d, t-SNE representation of the alpha cells from Segerstolpe dataset (top panel) and beta cells from Muraro dataset (bottom panel) colored by scBERT prediction and scNym prediction, respectively.

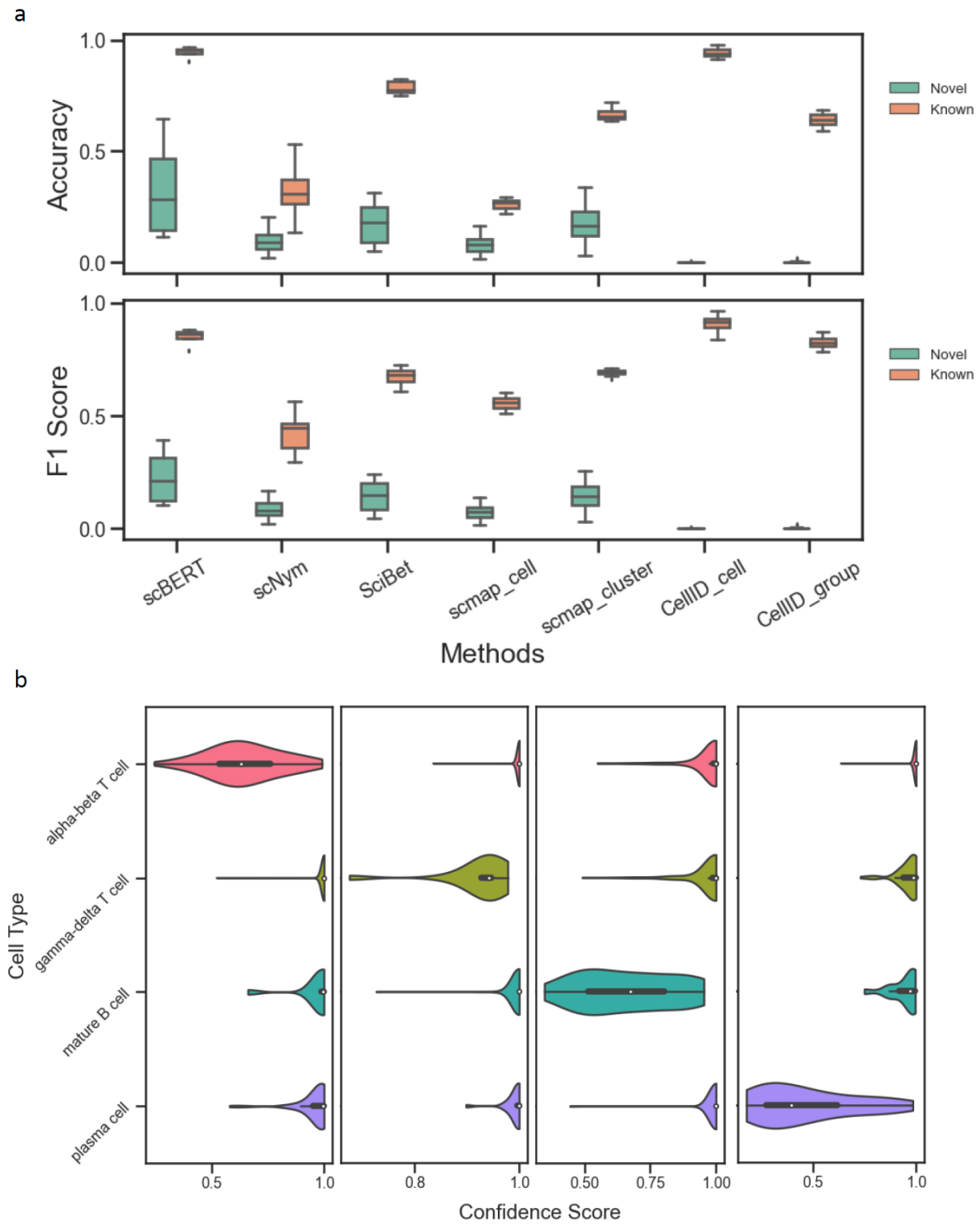

Fig. 4: Identification of novel cell types.

a, Performance of scBERT on the MacParland dataset from human liver tissue by removing alpha-beta T cell, gamma-delta T cell, mature B cell, plasma cell populations during the scBERT training process. The

accuracy and F1 score of both novel cell types and known cell types are shown in the boxplot. b, The confidence scores provided by scBERT for the four cell type are drawn from left to right.

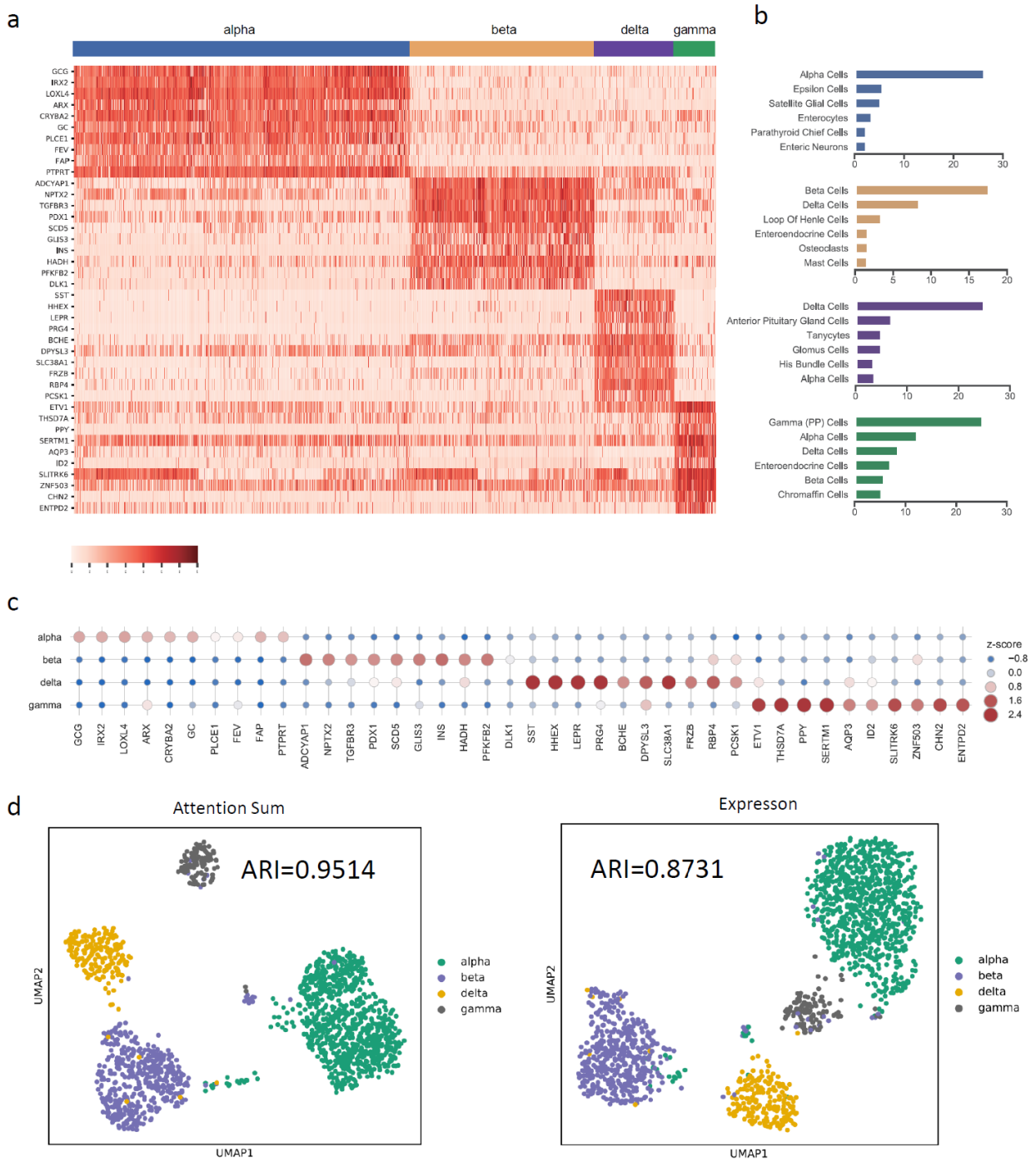

Fig. 5: Model interpretability. a, Heatmaps for the attention weights provided by scBERT on Pancreas cell type annotation task. The detailed attention estimation process is described in Methods. Top 10 genes with highest attention weights are listed for each cell type. b, The results of enrichment analysis of the top genes from scBERT using PanglaoDB cell type gene sets, with the complete information in Table S3-S7. c, Dotplot showing z-scores between the 10 genes receiving highest attention and the cell types. As shown

in the legend, the size of each plot reflects the z-score. d, UMAP for the cells colored by alpha, beta, delta, and gamma cell types. Left UMAP is based on the scBERT representation of each cell, while right UMAP is based on the raw expression of each cell. The ARI score is calculated and shown in the plot.

Acknowledgements

We thank Biaobin Jiang and Yan Ji for their valuable suggestions on model buiding and experimental design. We thank Tao Shen for advice on the large-scale model pre-training. This work was supported by the National Key R&D Program of China [2018YFC0910500]; SJTU-Yale Collaborative Research Seed Fund; and the Neil Shen's SJTU Medical Research.

Ethics declarations

Competing interests

No competing interests.

Supplementary Information

Supplementary tables are in the attached files.

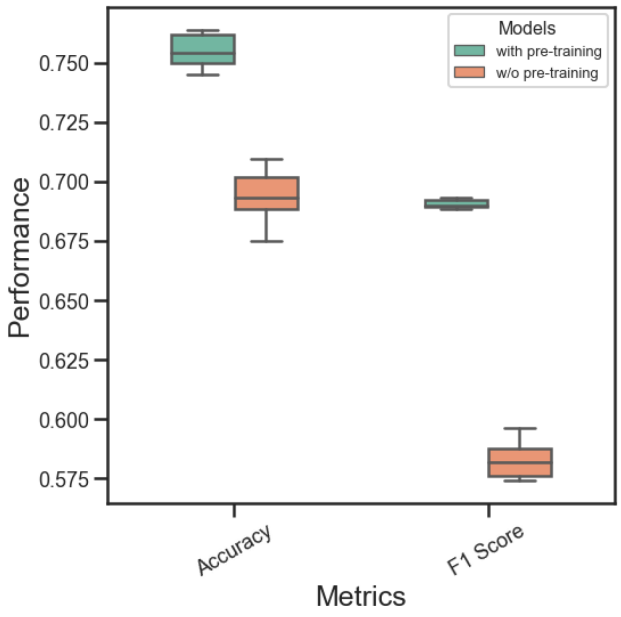

Figure S1. The impact of pre-training on model performance.
Performance of scBERT (with/without pre-training) measured by accuracy and F1 score on Zheng68K dataset using 5-fold cross-validation across 5 random repeats. The scBERT with pre-training is trained on more than one million public scRNA-seq data from PanglaoDB. In the contrast, the model weights of scBERT without pre-training are initiated randomly. Box plot shows the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers).

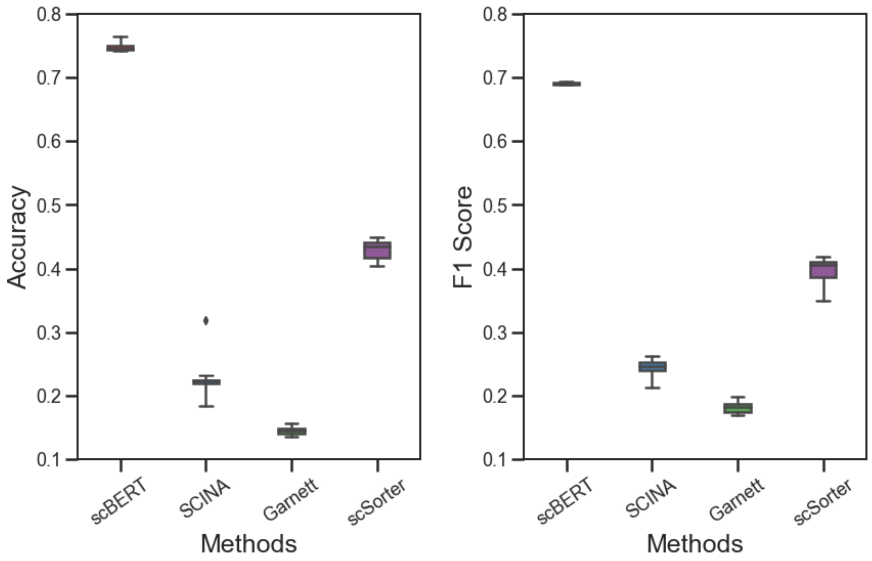

Figure S2. Performance comparison between scBERT and marker-based cell type annotation methods.
Performance of scBERT and marker-based methods (SCINA, Garnett, scSorter) measured by accuracy (left) and f1 score (right) on Zheng68K dataset using 5-fold cross-validation across 5 random repeats. The marker genes are adopted from CellMarker database, a manually curated resource of cell markers. Box plot shows the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers).

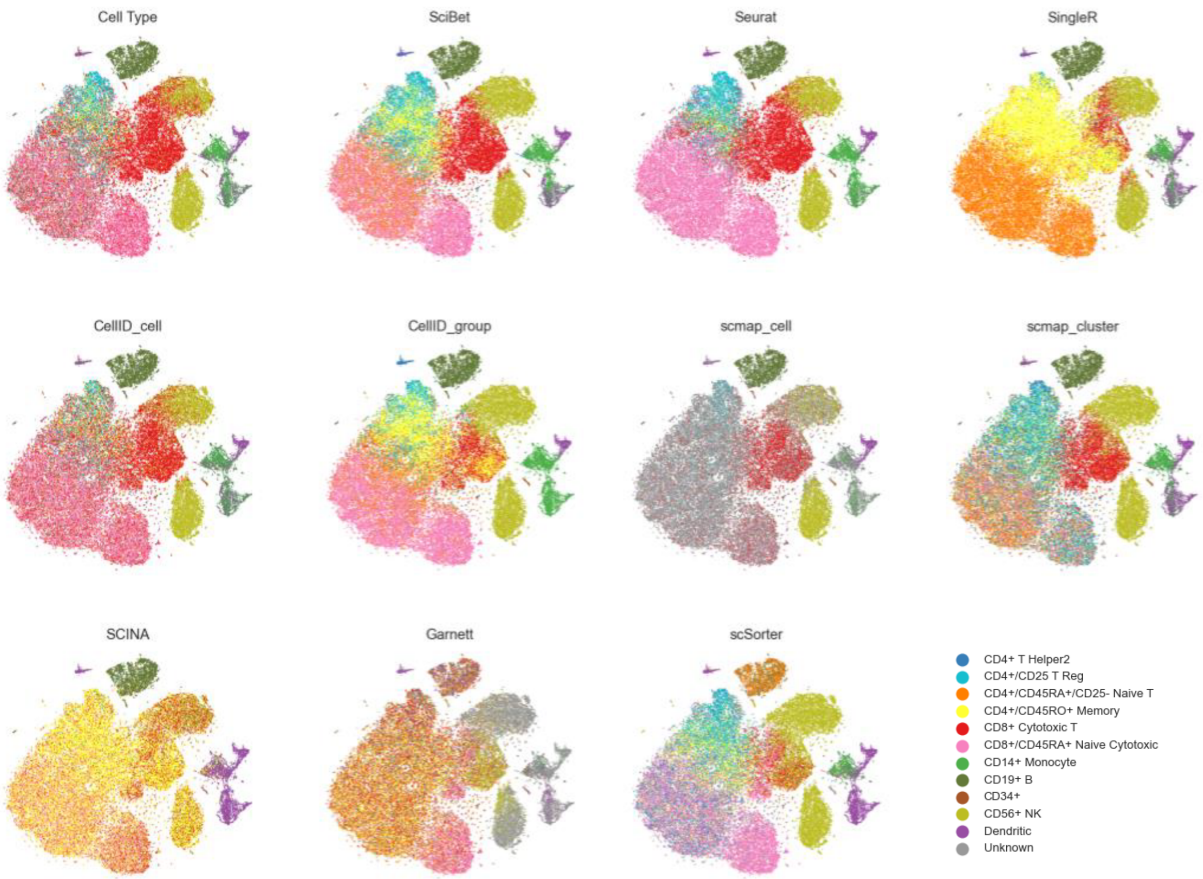

Figure S3. t-SNE plot of the cell type annotation results on Zheng68K dataset (n= 68450 cells). The top left plot is colored by expert annotated type from the original research, other plots show the cell type annotation results of comparison methods (SciBet, Seurat, SingleR， CellID_cell, CellID_group, scmap_cell, scmap_cluster, SCINA, Garnett, scSorter). The colors indicate the cell types annotation result from each individual method.

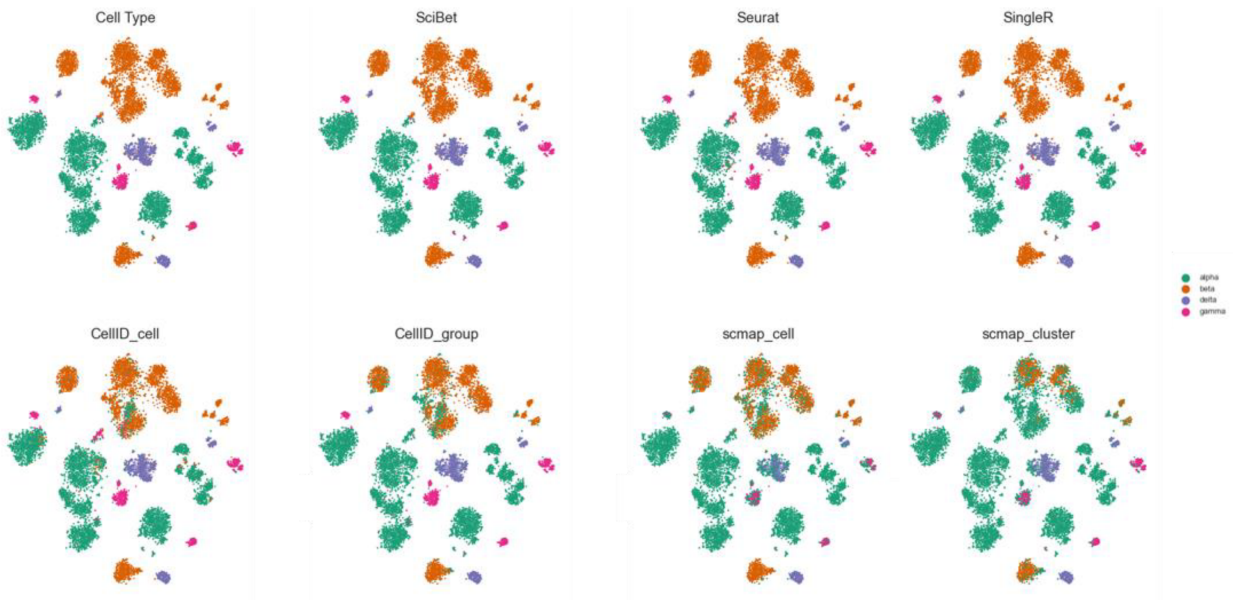

Figure S4. t-SNE plot of the cell type annotation results on Pancreas dataset (n= 34806 cells). The top left plot is colored by expert annotated types (alpha, beta, delta, gamma) from the original research, other plots show the cell type annotation results of comparison methods (SciBet, Seurat, SingleR, CellID_cell, CellID_group, scmap_cell, scmap_cluster). The colors indicate the cell types annotation result from each individual method.

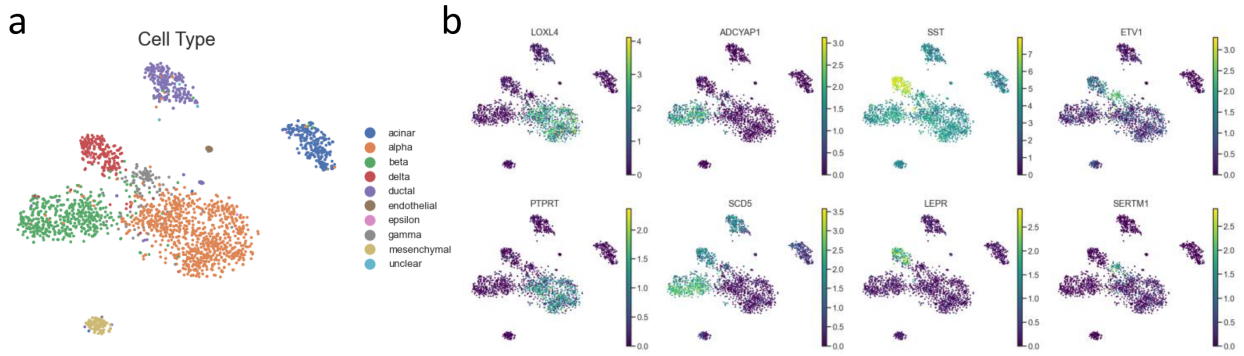

Figure S5. The distribution of the top attention sum genes across the four cell types of the Muraro dataset.
a, The cell type distribution of the Muraro dataset. b, The examples of top attention sum genes for alpha, beta, delta and gamma cells are visualized on the whole dataset using t-SNE from left to right. The top panel includes the top attention sum of genes that are consistent with reported markers for alpha, beta, delta and gamma cells, respectively. The bottom panel includes the top attention sum of genes that have distinguishing patterns on corresponding cell types but have not been reported as markers yet.

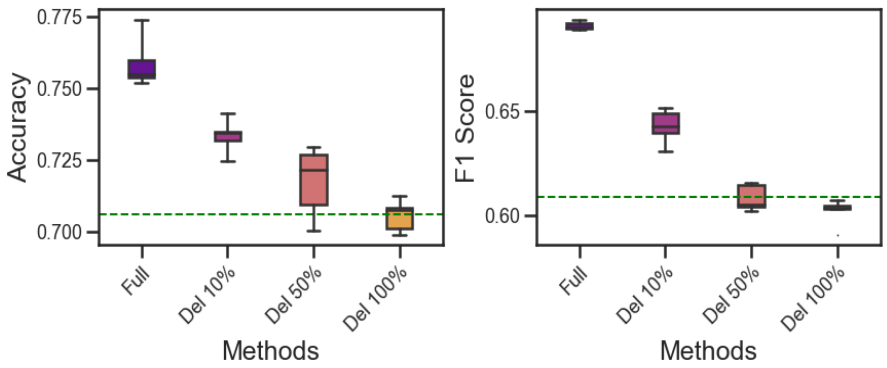

Figure S6. Explore the impact of marker genes on the scBERT model performance on Zheng68K dataset (n= 68450 cells).
The performance evaluation shows the effect of gradually removing marker genes (full, delete 10% delete 50% and delete 100% markers) on accuracy and f1 score, respectively. Box plot shows the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers). The green dashed line represents the best performance achieved by other cell type annotation methods with all marker genes.