Interdisciplinary Sciences: Computational Life Sciences (2025) 17:12–26
https://doi.org/10.1007/s12539-024-00655-6

ORIGINAL RESEARCH ARTICLE

scCrab: A Reference-Guided Cancer Cell Identification Method based on Bayesian Neural Networks

Heyang Hua[1] · Wenxin Long[1] · Yan Pan[2] · Siyu Li[1] · Jianyu Zhou[3] · Haixin Wang[4] · Shengquan Chen[1]

Received: 30 January 2024 / Revised: 20 August 2024 / Accepted: 21 August 2024 / Published online: 30 September 2024

Abstract

Cancer is a significant global public health concern, where early detection can greatly enhance curative outcomes. Therefore, the identification of cancer cells holds significant importance as the primary method for cancer diagnosis. The advancement of single-cell RNA sequencing (scRNA-seq) technology has made it possible to address the problem of cancer cell identification at the single-cell level more efficiently with computational methods, as opposed to the time-consuming and less reproducible manual identification methods. However, existing computational methods have shown suboptimal identification performance and a lack of capability to incorporate external reference data as prior information. Here, we propose scCrab, a reference-guided automatic cancer cell identification method, which performs ensemble learning based on a Bayesian neural network (BNN) with multi-head self-attention mechanisms and a linear regression model. Through a series of experiments on various datasets, we systematically validated the superior performance of scCrab in both intra- and inter-dataset predictions. Besides, we demonstrated the robustness of scCrab to dropout rate and sample size, and conducted ablation experiments to investigate the contributions of each component in scCrab. Furthermore, as a dedicated model for cancer cell identification, scCrab effectively captures cancer-related biological significance during the identification process.

Heyang Hua, Wenxin Long and Yan Pan contributed equally.

✉ Jianyu Zhou
jyzhou@nankai.edu.cn

✉ Haixin Wang
wanghaixin@301hospital.com.cn

✉ Shengquan Chen
chenshengquan@nankai.edu.cn

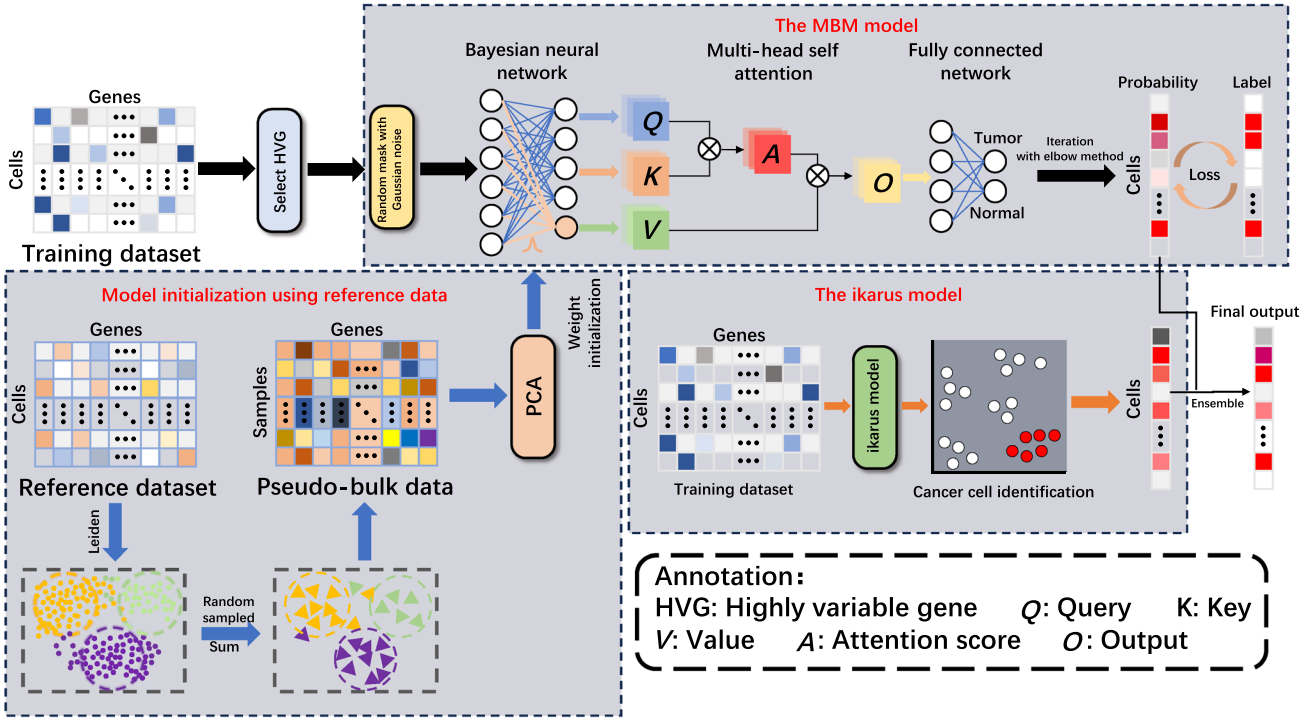
[1] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China

[2] Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

[3] College of Software, Nankai University, Tianjin 300071, China

[4] Cadre Medical Department, The 1St Clinical Center, Chinese PLA General Hospital, Beijing 100853, China

Graphical Abstract



Keywords Cancer cell identification · Reference-guided method · Bayesian neural network · Self-attention mechanism

1 Introduction

Cancer, as a leading cause of morbidity and mortality among all diseases, remains to be a major public health concern worldwide. Several studies have shown that early intervention can greatly improve cure rates if the disease is still manageable [1–3]. Therefore, as a prime method for diagnosing cancer, the identification of cancer cells has underscored its importance. Traditional approaches to cancer cell annotation largely rely on manual annotation and other biochemical methods using molecular markers, which tend to be relatively time-consuming, unstable, and irreproducible [4]. Hence, the need for an automatic, efficient, and precise methodology for cancer cell identification is of utmost urgency. Single-cell RNA sequencing (scRNA-seq) technology has evolved to become a profoundly impactful tool in transcriptome sequencing [5]. Compared to bulk RNA-seq that primarily concentrates on average gene expression among cells, scRNA-seq aims to obtain expression profiles at single-cell level, which permits the elucidation of cell heterogeneity overlooked by bulk RNA-seq [6, 7]. Therefore, the advent of scRNA-seq data has opened up the potential for identifying cancer cells at the single-cell level and thus revolutionizing the landscape of cancer cell detection.

As the problem of cancer cell identification can be regarded as a specific instance of cell type identification, it is necessary to understand existing state-of-the-art cell type identification methods. Numerous tools for automated cell type identification based on scRNA-seq data have been developed, falling into three main categories: marker gene-based methods, correlation-based methods, and supervised classification-based methods [4, 8]. Marker gene-based methods prioritize highly expressed marker genes identified in the literature that are likely to reveal cell heterogeneity [9]. Correlation-based methods measure the correlations between the reference and the query datasets to perform annotation [10]. Supervised classification-based methods basically consider cells as targets to be classified and genes as features, and generally show promising results by utilizing and integrating various machine learning methods [4, 11]. Abdelaal et al. have benchmarked the widely-used methods and concluded that the traditional machine learning methods of support vector machine (SVM), random forest (RF), and K-nearest neighbors (KNN) have great performance in cell type annotation for scRNA-seq data [12]. However, these methods are not designed for cancer cell identification, which may lead to the neglect of some specific characteristics of cancer cells. Furthermore, the incorporation of

reference data into the analysis of scRNA-seq data proves to be more effective in addressing the noise and technical variations [13–15], but these methods tend to overlook the utilization of information within the reference dataset.

Recently, Dohmen et al. proposed a workflow named ikarus for cancer cell identification, which demonstrates remarkable performance [16]. By utilizing differential expression analysis, ikarus first selects a list of tumor and normal marker genes from expert-labeled datasets, and uses AUCell [17] to score every cell with the selected marker gene list. Then, ikarus constructs a logistic classifier using the AUCell scores as input and the cell type labels as target variables. Finally, ikarus implements the cell annotation task as an iterative two-step process of cell type assignment and label propagation. However, ikarus selects the marker genes from two fixed datasets, regardless of the training and testing datasets. When utilizing other datasets to select markers and train, ikarus often presents a much worse performance. Besides, AUCell outputs only two scores for each cell, representing the activity of the tumor marker gene set and the normal marker gene set, respectively. Specifically, when training or testing datasets originate from cancer subtypes different from the two reference datasets, certain genes that characterize cellular heterogeneity may be absent from the predefined set of marker genes. This omission could potentially result in a significant loss of important information. As a result, ikarus suffers from a lack of robustness and loss of information, restricting its application scope.

To address the shortcomings of existing models and the information loss problem of ikarus, we proposed scCrab, a reference-guided **s**ingle-**c**ell **c**ancer cell identification model based on flexible **r**eference **a**nd **B**ayesian neural network (BNN). As an ensemble model, scCrab utilizes the ikarus model in addition to a network model. The network model within scCrab (referred to as MBM) encompasses random **m**asking with Gaussian noise, **B**ayesian neural networks, **m**ulti-head self-attention mechanisms, and fully connected networks. To take full advantage of a wide range of existing datasets, scCrab utilizes a projection weight matrix obtained by principal component analysis (PCA) to extract information from external reference data and incorporates it as prior information into the BNN, which enables scCrab to flexibly integrate information from the reference data as prior information into its network model. Through comprehensive intra- and inter-dataset experiments on seven datasets, we demonstrated that scCrab consistently exhibits outstanding performance and remarkable stability for cancer cell identification. Additionally, along with robustness tests on different cancer types, we demonstrated the strong robustness and noise resistance of scCrab. Moreover, by conducting Gene Ontology (GO) enrichment analysis to identify pathways enriched with differentially expressed genes obtained from annotated cell labels, we demonstrated that scCrab is capable of revealing cancer cellular heterogeneity. On the contrary, existing methods suffer from failing to utilize information from reference datasets and GO enrichment analysis. In conclusion, the innovation of scCrab lies in its clever integration of flexible references and BNN, which forms an ensemble model that significantly enhances performance in cancer cell identification. We believe that the successful implementation of scCrab not only guides the future development of cancer cell identification but also lays a solid foundation for deeper exploration in complex diseases.

2 Materials and Methods

2.1 The Framework of scCrab

scCrab is an ensemble model that includes a neural network called the MBM model, as well as the ikarus framework. The MBM model is a neural network that includes a BNN module and a multi-head self-attention mechanism module. The cancer cell recognition model, ikarus, is based on linear regression. As shown in Fig. 1, the workflow of scCrab can be delineated into the following steps: (1) preprocessing the training and testing dataset; (2) generating pseudo-bulk data from a reference scRNA-seq dataset which is not used for training scCrab; (3) extracting prior information using the projection weight matrix generated from PCA. The projection weight matrix is then used to set the initial weights of the BNN; (4) training the network model of scCrab with preprocessed training data; (5) applying the ikarus [16] method to training data; (6) integrating the outcomes obtained from the trained network and the ikarus model. We note that gene expressions are used as input data for training models and PCA is utilized to generate the projection weight matrix from the reference dataset.

2.2 The Construction of MBM

The network model in scCrab is named as MBM, of which composition can be divided into four main modules: (1) the random mask with Gaussian noise, (2) the BNN module, (3) the multi-head self-attention mechanism module, and (4) the fully connected network module. First, to circumvent the limitations of existing methods in handling high-noise data, scCrab masks the raw data by adding Gaussian noise with zero as the mean and unit variance to each value, which can be considered as a regularization mechanism that bolsters the resistance to noise and generalization ability of the model [18, 19]. Second, scCrab feeds the features of each cell in the preprocessed training set into the BNN module and obtains the corresponding latent representations. Third, scCrab utilizes a multi-head self-attention module following the BNN module to bolster the capacity of the network

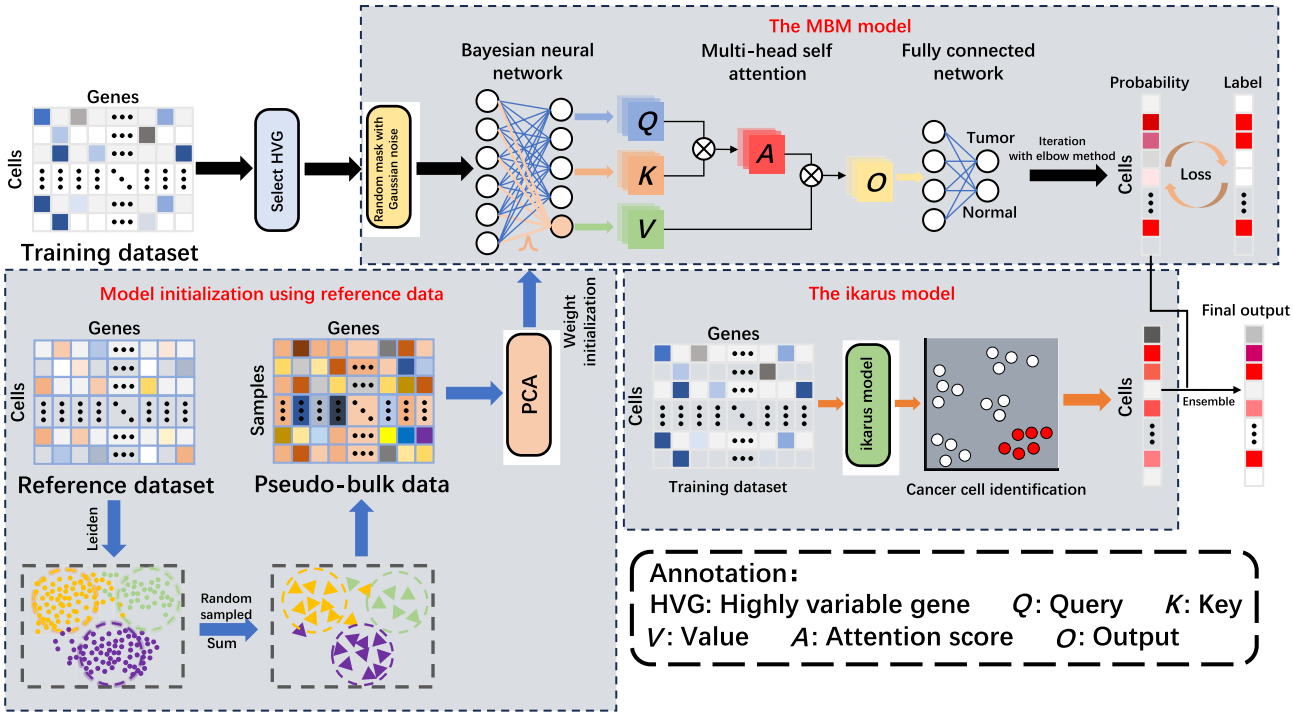


Fig. 1 A graphical representation of scCrab. Initially, scCrab performs preprocessing on the training dataset and combines it with prior information extracted from a reference scRNA-seq dataset. Following that, the data is inputted into the MBM, which integrates and analyzes the combined information. The same training dataset is independently trained by ikarus, and the two components of the ensemble model only intersect at the final probability output stage

for extracting pertinent features. Finally, the fully connected network consists of multiple linear layers with dimensions transitioning from the latent dimension to 64, followed by an output linear layer with dimensions 64 to 2.

The BNN module is a neural network layer formed by connecting Bayesian layers and linear layers. The Bayesian layer outputs 1/5 of the latent representations, while the linear layer outputs the remaining 4/5. BNN encapsulates the uncertainty inherent in the neural network weights [20]. By viewing the network weights in the Bayesian layer as values sampled from multiple conditioned distributions rather than as fixed values in conventional neural networks, we introduce prior information into the network by imposing a prior distribution of weights. Parameters of the Bayesian layer represent the elements in the matrix $\boldsymbol{X}_{d_{\text{in}} \times d_{\text{out}}}$, which is subject to Gaussian distribution,

$$\boldsymbol{X}_{ij} \sim \mathrm{N}\left(\mu_b^{(ij)}, \sigma_b^{(ij)^2}\right), i = 1, 2, \ldots, d_{\text{in}}, j = 1, 2, \ldots, d_{\text{out}},$$

where $\mu_b^{(ij)}$ and $\sigma_b^{(ij)^2}$ are elements of the mean matrix μ_b and the variance matrix σ_b^2 respectively. BNN is trained in batches, with a batch size of 256 set in our experiment. For the output of BNN, the output matrix of shape $b \times d_{\text{out}}$ is

$$f_{\text{BNN}}(\boldsymbol{I}) = \boldsymbol{I}\boldsymbol{X} = \boldsymbol{I}\left(\mu_b + \delta \odot \sigma_b\right),$$

where b represents batch size and $\boldsymbol{I}$ represents input matrix with a shape of $b \times d_{\text{in}}$ in a minibatch. Besides, $f_{\text{BNN}}(\cdot)$ represents BNN function, δ is the Gaussian noise with zero mean and unit variance and $\odot$ indicates Hadamard product which takes in two matrices and returns a matrix of the multiplied corresponding elements.

To introduce prior information into the Bayesian layer from an external reference dataset that can be different from the training and test datasets, we assume that the elements in $\boldsymbol{X}$ are subject to Gaussian distribution parameterized by μ_x and σ_x^2. The activation function for the BNN module and the fully connected network module is ReLU [21],

$$f_{\text{ReLU}}(x) = \max(0, x),$$

where x is the input of ReLU activation function. The optimization objective of MBM is based on a combination of Kullback–Leibler (KL) divergence loss and cross-entropy (CE) loss. KL divergence is used to better restrict the differences between BNN parameter distribution and prior distribution,

$$L_{\mathrm{KL}} = \mathrm{KL}(\mathrm{N}(\mu_b, \sigma_b^2), \mathrm{N}(\mu_x, \sigma_x^2))$$
$$= \frac{1}{d_{\mathrm{in}} d_{\mathrm{out}}} \sum_{i=1}^{d_{\mathrm{in}}} \sum_{j=1}^{d_{\mathrm{out}}} (\log\left(\frac{\sigma_k^{(ij)}}{\sigma_b^{(ij)}}\right)$$
$$+ \frac{\sigma_b^{(ij)^2} + \left(\mu_b^{(ij)} - \mu_k^{(ij)}\right)^2}{2\sigma_x^{(ij)^2}} - \frac{1}{2}),$$

where μ_x represents the mean of prior information from external data, as reflected in the values of the projection weight matrix generated from PCA of the gene expression matrix from the reference dataset, serving as the prior gene expression matrix. Additionally, $\sigma_x^{(ij)}$ takes the logarithm of one-tenth of the $\mu_x^{(ij)}$, signifying the standard deviation of the prior information.

The CE loss not only serves as a metric for assessing the predictive accuracy of the model but also provides insights into the capability of MBM to handle parameter uncertainty. This property enhances the capacity of MBM to address uncertainty-related challenges effectively. The formula for calculating CE loss can be expressed as follows:

$$L_{\mathrm{CE}} = -\int_{\theta} P(\theta) \int_{X} Q(\boldsymbol{Y}|\boldsymbol{X}, \theta) \log P(\boldsymbol{Y}|\boldsymbol{X}, \theta) d\boldsymbol{X} d\theta,$$

$$\theta \sim \mathcal{N}(\mu_b, \sigma_b^2),$$

where $P(\theta)$ denotes the prior distribution of parameters, indicating parameter uncertainty. $Q(\boldsymbol{Y}|\boldsymbol{X}, \theta)$ is the conditional distribution of the predictions of MBM for the input $\boldsymbol{X}$ given the parameters θ. $P(\boldsymbol{Y}|\boldsymbol{X}, \theta)$ is the conditional distribution of true labels given the input $\boldsymbol{X}$ and parameters θ.

The loss function of the MBM can be expressed as follows:

$$L_{\mathrm{BMF}} = L_{\mathrm{CE}} + \alpha L_{\mathrm{KL}},$$

where α is a hyperparameter used to adjust the relative importance between the two loss functions, which is set as 1 in our experiments. To manage the instability caused by BNN, we employ ensemble iteration, an ensemble learning technique that aims to iteratively train models for improved accuracy and robustness in predictions, which is detailed in Supplementary Text 1.

In the output linear layer, where the dimensions are reduced from 64 to 2, the outputs of the two neurons are passed through a softmax function:

$$\mathrm{Softmax}\left(\left[z_1, z_2\right]\right) = \left[\frac{e^{z_1}}{e^{z_1} + e^{z_2}}, \frac{e^{z_2}}{e^{z_1} + e^{z_2}}\right],$$

where z_1, z_2 are the outputs of the two neurons in the final layer, respectively. The softmax function calculates the probability of a cell belonging to cancer cells, from which MBM determines the probability of a cell being a cancer cell.

2.3 Multi-head Self-attention

In the part of MBM, we employ a multi-head self-attention mechanism to process the input data. This mechanism enables our model to have a deeper understanding of cell-to-cell interactions and intra-cellular structures [22]. Initially, the vectors $\boldsymbol{x}$ processed by the BNN module are linearly transformed into three separate tensors: keys, queries, and values.

$$\boldsymbol{Q} = \boldsymbol{W_q x}, \boldsymbol{K} = \boldsymbol{W_k x}, \boldsymbol{V} = \boldsymbol{W_v x},$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ represent the vector matrices of queries, keys, and values, respectively. The learnable $\boldsymbol{W_q}$, $\boldsymbol{W_k}$ and $\boldsymbol{W_v}$ are the corresponding linear transformation matrices of these vector matrices. All the $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are then divided into h heads and undergo a scaling operation to ensure stable attention weights by making the result of the dot product be not too large or too small,

$$\boldsymbol{Q}^h = \mathrm{head}(\boldsymbol{Q}), \boldsymbol{K}^h = \mathrm{head}(\boldsymbol{K}), \boldsymbol{V}^h = \mathrm{head}(\boldsymbol{V}),$$

$$\boldsymbol{Q}_i^h = \frac{\boldsymbol{Q}_i^h}{\sqrt{\frac{d}{h}}}, \boldsymbol{K}_i^h = \frac{\mathbf{K}_i^h}{\sqrt{\frac{d}{h}}}, \boldsymbol{V}_i^h = \frac{\boldsymbol{V}_i^h}{\sqrt{\frac{d}{h}}},$$

where $\frac{d}{h}$ is the dimension of each head. In our experiments, vectors are divided into 8 heads. Next, attention scores are computed by taking the dot product between queries and keys. Moreover, softmax is applied to convert the scores into attention weights $\boldsymbol{A}$,

$$\mathrm{score}\left(\boldsymbol{Q}^h, \boldsymbol{K}^h\right) = \boldsymbol{Q}^h \boldsymbol{K}^{h\intercal},$$

$$\boldsymbol{A} = \mathrm{softmax}\left(\mathrm{score}\left(\boldsymbol{Q}^h, \boldsymbol{K}^h\right)\right).$$

These attention weights are applied to the values tensor, resulting in a weighted fusion of information from different heads,

$$\boldsymbol{z}^h = \boldsymbol{A}\boldsymbol{V}^h.$$

Finally, the weighted tensors are recombined and transformed using a linear operation to produce the new representations to obtain the latent space after feature extraction,

$$\boldsymbol{z} = \mathrm{concat}(\boldsymbol{z}^1, \boldsymbol{z}^2, \ldots, \boldsymbol{z}^h),$$

$$\mathrm{Output} = \boldsymbol{W}_{\mathrm{o}} \boldsymbol{z},$$

where W_0 represents a learnable linear transformation matrix that maps z back to the original dimension.

2.4 Model Initialization Using Reference Data

Compared to the limitations of other existing identification models that cannot leverage external reference data, scCrab utilizes either labeled or unlabeled data from open databases as prior information for the model. For labeled data, scCrab directly uses its expert annotation, while for unlabeled data, scCrab applies the Leiden algorithm with default parameters in scanpy to perform clustering on the reference dataset, with the labels of cell clusters assigned as cell labels for the individual cells [23, 24]. Cells with the same label are randomly sampled 100 times, and the values of their respective features are summed to obtain a piece of pseudo-bulk data, which corresponds to a row in a pseudo-bulk data matrix. This generation process is repeated 100 times to generate individual pseudo-bulk matrices, each containing 100 rows for every cluster and then these matrices are subsequently concatenated for further analysis. After that, the pseudo-bulk data $\boldsymbol{D}$ subsequently experiences dimensional reduction through PCA, from which the prior information of the BNN is gleaned from the projection weight matrix $\boldsymbol{M}$ generated from PCA,

$$\boldsymbol{P}_{\mathrm{PCA}} = \boldsymbol{MD}$$

$$\boldsymbol{M} = \boldsymbol{P}_{\mathrm{PCA}}\boldsymbol{D}^{-1},$$

where $\boldsymbol{P}_{\mathrm{PCA}}$ is the matrix of PCs generated by PCA. BNN assimilates prior information by employing the projection weight matrix generated from PCA of the pseudo-bulk data derived from the reference data as the mean of the weight of the first BNN layer and taking the logarithm of one-tenth of the mean as the variance of the prior distribution.

Due to the versatile approaches for incorporating reference data, scCrab has the capability to utilize scRNA-seq data from public databases. The approaches for utilizing reference data are as follows: (1) If users choose not to introduce prior information into the BNN, the weights of the network can be set as a Gaussian distribution with zero mean and unit variance. This implies that the weight initialization methods mentioned above will not be used. (2) If external data is selected as the reference dataset, the aforementioned weights initialization method can be employed to extract the prior information from the external dataset. (3) In cases where no external data is available, scCrab still uses the training dataset itself as the reference dataset and applies the aforementioned method to extract the prior information from the training dataset.

2.5 The Ensemble Strategy of scCrab

Ensemble learning aims to boost performance by integrating multiple base learners. This approach has several benefits, including enhanced classification accuracy, improved model robustness, and reduced likelihood of overfitting. ikarus, the state-of-the-art cancer cell identification method, exhibits outstanding performance through the selection of gene signatures and the training of a logistic regression classifier using AUCell scores [17], culminating in cell-label propagation predicated on a custom cell–cell network. Therefore, to improve identification performance through ensemble techniques, we utilize ikarus to complement our MBM model. Specifically, the same training dataset is independently trained in two separate models and the details of model training are presented in Supplementary Text 2. During the predicting process, both the MBM and the ikarus models are capable of predicting the probability of cells manifested as tumor cells. We compute the average of the probabilities yielded by both methods as the final output probability of scCrab. If the output probability of a cell being classified as a cancer cell surpasses a threshold (i.e., 0.5), we categorize the cell as a cancer cell; otherwise, we classify the cell as a normal cell.

2.6 Data Collection and Preprocessing

We evaluated the predictive performance of scCrab on seven human scRNA-seq datasets with distinct sources, tissues,

Table 1 Summary of the datasets used in this study

Dataset	No. of cells	No. of tumor cells	No. of normal cells	Cancer type	Protocol	References
Lambrechts (LA)	52,698	7447	45,251	Lung cancer	10x	[25]
Ma (MA)	56,721	17,164	39,557	Hepatocellular carcinoma	10x	[26]
Bischoff (BI)	63,327	8097	55,230	Lung carcinoid	10x	[27]
Tirosh (TIS)	5578	2215	3363	Head/Neck carcinoma	10x + FACS sorted	[28]
Tirier (TIR)	177,880	74,181	103,699	Myeloma	10x	[29]
Kildisiute_10x (KLA)	6442	1766	4676	Neuroblastoma	10x	[30]
Kildisiute_celseq2 (KLB)	13,281	1630	11,651	Neuroblastoma	CELseq2	[30]

sizes and single-cell sequencing technologies. As shown in Table 1, these datasets include: (1) Lung cancer, a leading cause of cancer-related mortality globally. In this study, we utilized a non-small-cell lung cancer dataset (LA) [25] and a lung carcinoid dataset (BI) [27]. (2) Hepatocellular carcinoma, the most prevalent form of primary liver cancer. We utilized a hepatocellular carcinoma dataset (MA) [26] in our study. (3) Head and neck squamous cell carcinoma (HNSCC), a varied group of cancers originating in the upper aerodigestive tract. We utilized a head and neck cancer dataset (TIS) [28] in our study. (4) Multiple myeloma, a hematological cancer marked by the uncontrolled growth of plasma cells in the bone marrow. We utilized a myeloma dataset (TIR) [29] in our study. (5) Neuroblastoma, a solid tumor predominantly affecting children and deriving from the developing sympathetic nervous system. We utilized two neuroblastoma datasets (KLA, KLB) [30] in our study.

In the data preprocessing phase, we followed the standard preprocessing pipeline from scanpy, in which we normalized the original gene expression matrix, applied a logarithmic transformation, and selected HVGs from each dataset using sc.pp.highly_variable_genes with default parameters [24].

2.7 Baseline Methods

We compared scCrab with four supervised machine learning models, including SVM, RF, KNN, and ikarus. Among these methods, ikarus stands as a state-of-the-art cancer cell identification method, while the others are conventional machine learning methods recommended by recent benchmark studies (SVM with a linear kernel, RF with 50 estimators, and KNN with 9 neighbors) [12]. We benchmarked the performance of baseline methods following their default tutorials. Specifically, SVM, KNN, and RF use gene signatures selected via the sc.pp.highly_variable_genes function from scanpy as input [24]. In contrast, ikarus employs its built-in gene selection method to determine the gene signatures input. For inter-dataset experiments, we selected genes from the training dataset. For intra-dataset experiments, we divided the dataset into five folds and selected genes from four training folds. All experiments were conducted on a machine equipped with two Intel Xeon Platinum 8375C CPUs, two NVIDIA RTX A6000 GPUs, and 256 GB of RAM.

2.8 Evaluation Metrics

In cancer cell identification tasks, the number of cancer cells is typically much smaller than that of normal cells. In our dataset, the highest ratio of normal to cancer cells is 7.14 (i.e., the KLB dataset), which indicates an extremely imbalanced dataset. When evaluating classification performance in imbalanced datasets, the normal accuracy score can often be uninformative. To address this issue, we used three widely-used metrics: balanced accuracy score [31], area under the precision-recall curve (AUPRC) [32], and Cohen's Kappa (kappa) [33]. In both intra-dataset and inter-dataset experiments, we employed all of the three metrics as benchmarks. The balanced accuracy score independently evaluates the classification performance for both positive and negative cases instead of considering global accuracy. AUPRC is calculated using the precision-recall curve which illustrates the trade-off between precision and recall at various decision thresholds. The unique aspect of this metric is that it remains unaffected by the choice of decision threshold. AUPRC demonstrates robust stability in its results, ensuring reliability irrespective of the selected threshold, including the 0.5 thresholds employed by scCrab. Kappa is a metric for classification accuracy based on the confusion matrix. The calculation formulas for all metrics are detailed in Supplementary Text 3.

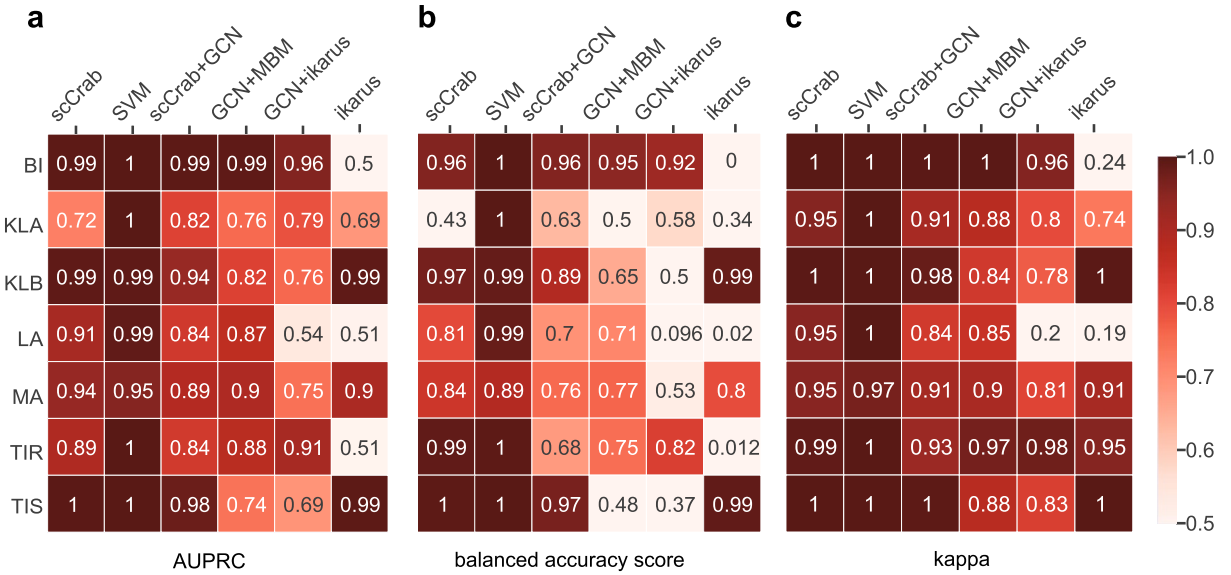
2.9 Model Ablation

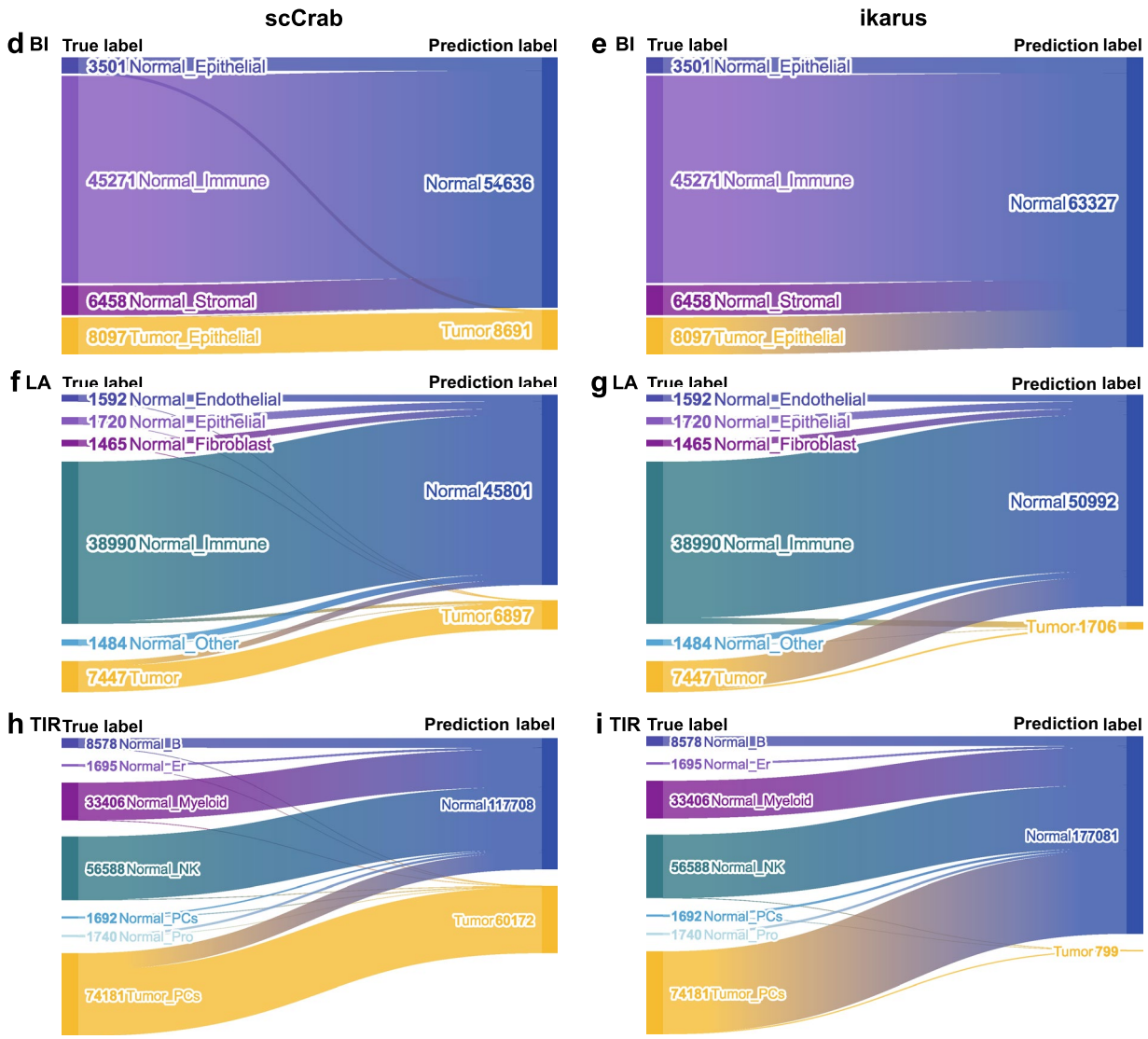
To investigate the benefit of each component within scCrab, we evaluated the performance of alternative models where certain components were ablated from scCrab, including (1) the variant that did not incorporate the randomly masked Gaussian noise during preprocessing (referred as scCrab_w/o_masking), (2) the variant that replaced the BNN with a multilayer perceptron (MLP), where the number of layers and neurons per layer in the MLP is identical to that of the BNN (referred as scCrab_w/o_BNN), (3) the variant that operated without the multi-head self-attention mechanism (referred as scCrab_w/o_Attn), and (4) the variant that excluded the integration of the GCN approach (referred as scCrab_w/o_GCN).

3 Results

3.1 scCrab Outperforms Other Methods in Intra-dataset Cross-validation

To assess the performance of the model across heterogeneous datasets, we first conducted a five-fold cross-validation experiment using datasets of LA, MA, BI, TIS, TIR, KLA, and KLB. In this experiment, we randomly divided all cells into five folds and predicted cell labels as Tumor or Normal for each fold using models trained on the remaining four folds. We utilized AUPRC (Fig. 2a), balanced accuracy score (Fig. 2b), and kappa (Fig. 2c) as the measures for the performance of the model. In our experiments, we discerned that SVM outperformed in cancer cell identification in intra-dataset tasks, while its performance in inter-dataset tasks is only at a moderate level. Therefore, during intra-dataset cross-validation trials, we considered SVM as a performance


a
b
c
scCrab
SVM
scCrab+GCN
GCN+MBM
GCN+ikarus
ikarus
BI 0.99 1 0.99 0.99 0.96 0.5
KLA 0.72 1 0.82 0.76 0.79 0.69
KLB 0.99 0.99 0.94 0.82 0.76 0.99
LA 0.91 0.99 0.84 0.87 0.54 0.51
MA 0.94 0.95 0.89 0.9 0.75 0.9
TIR 0.89 1 0.84 0.88 0.91 0.51
TIS 1 1 0.98 0.74 0.69 0.99
AUPRC
BI 0.96 1 0.96 0.95 0.92 0
KLA 0.43 1 0.63 0.5 0.58 0.34
KLB 0.97 0.99 0.89 0.65 0.5 0.99
LA 0.81 0.99 0.7 0.71 0.096 0.02
MA 0.84 0.89 0.76 0.77 0.53 0.8
TIR 0.99 1 0.68 0.75 0.82 0.012
TIS 1 1 0.97 0.48 0.37 0.99
balanced accuracy score
BI 1 1 1 1 0.96 0.24
KLA 0.95 1 0.91 0.88 0.8 0.74
KLB 1 1 0.98 0.84 0.78 1
LA 0.95 1 0.84 0.85 0.2 0.19
MA 0.95 0.97 0.91 0.9 0.81 0.91
TIR 0.99 1 0.93 0.97 0.98 0.95
TIS 1 1 1 0.88 0.83 1
kappa
1.0
0.9
0.8
0.7
0.6
0.5


scCrab
ikarus
d BI
True label
Prediction label
3501 Normal_Epithelial
45271 Normal_Immune
6458 Normal_Stromal
8097 Tumor_Epithelial
Normal 54636
Tumor 8691
e BI
True label
Prediction label
3501 Normal_Epithelial
45271 Normal_Immune
6458 Normal_Stromal
8097 Tumor_Epithelial
Normal 63327
f LA
True label
Prediction label
1592 Normal_Endothelial
1720 Normal_Epithelial
1465 Normal_Fibroblast
38990 Normal_Immune
1484 Normal_Other
7447 Tumor
Normal 45801
Tumor 6897
g LA
True label
Prediction label
1592 Normal_Endothelial
1720 Normal_Epithelial
1465 Normal_Fibroblast
38990 Normal_Immune
1484 Normal_Other
7447 Tumor
Normal 50992
Tumor 1706
h TIR
True label
Prediction label
8578 Normal_B
1695 Normal_Er
33406 Normal_Myeloid
56588 Normal_NK
1692 Normal_PCs
1740 Normal_Pro
74181 Tumor_PCs
Normal 117708
Tumor 60172
i TIR
True label
Prediction label
8578 Normal_B
1695 Normal_Er
33406 Normal_Myeloid
56588 Normal_NK
1692 Normal_PCs
1740 Normal_Pro
74181 Tumor_PCs
Normal 177081
Tumor 799

◂**Fig. 2** Performance of intra-dataset cross-validation. (**a**) A comparison of five-fold cross-validation mean AUPRC between scCrab and other methods. (**b**) A comparison of five-fold cross-validation mean balanced accuracy score between scCrab and other methods. (**c**) A comparison of five-fold cross-validation mean kappa between scCrab and other methods. (**d**)–(**i**) Sankey plots of scCrab and ikarus, revealing the specific numbers of cells annotated as Tumor or Normal in intra-dataset experiments. The left column is prediction results of scCrab and right column is that of ikarus

standard for classification tasks. If the performance of our model aligns closely with this upper limit, it implies that the model is proficient in intra-dataset tasks.

We further noticed that traditional classification models tend to ignore the prior information of gene interaction. To address this problem, we tried to introduce Graph Convolutional Network (GCN) into our ensemble model [34, 35]. Graph Convolutional Networks (GCN) are intricate deep learning models that harness the structural data of graphs, enabling them to generalize effectively and deliver remarkable performance across diverse tasks [34, 36]. In our implementation, each cell was regarded as a graph and its expression profile was regarded as the graph's feature. We considered genes as nodes and gene interactions as edges (Supplementary Text 4). With incorporating the GCN into the ensemble model, the probability of a cell being classified as a cancer cell is determined by averaging the probabilities outputted by the MBM component, ikarus component, and GCN component. This exploration led to the following combination strategies: (1) the integration of MBM, ikarus, and GCN (scCrab + GCN), (2) the integration of GCN and MBM (GCN + MBM), (3) the integration of GCN and ikarus (GCN + ikarus). In addition to these three combinations, we compared scCrab with SVM and ikarus for comprehensive analysis in the intra-dataset experiment. As shown in Fig. 2a–c, SVM, functioning as a superior boundary model, displays exemplary performance in intra-dataset cross-validation. scCrab exhibits comparable performance to SVM on many datasets, even surpassing SVM in some cases. Furthermore, scCrab outperforms the other three combination strategies (i.e., scCrab + GCN, GCN + MBM and GCN + ikarus) and ikarus across all datasets. In conclusion, the integration of MBM and ikarus into scCrab represents an optimal ensemble approach.

Throughout our experiments, we noted that ikarus exhibited inferior performance on certain datasets. To scrutinize the specific misclassifications, we generated Sankey diagrams for both scCrab and ikarus across all datasets. The results for the BIS, LA, and TIR datasets are presented in Fig. 2d–i, while the results for the remaining datasets are available in Supplementary Figure 1. The Sankey diagrams revealed that scCrab outperformed ikarus in intra-dataset cancer cell identification tasks. Specifically, as shown in Fig. 2d, f, h, the majority of cells were accurately classified by scCrab, while many cells with true normal labels were incorrectly classified as tumor cells by ikarus (Fig. 2e, g, i). Moreover, different cell types showed different probabilities of misclassification, with immune cells particularly difficult to be misclassified. This observation suggests a substantial distinction between immune cells and cancer cells. In summary, scCrab exhibits excellent performance in intra-dataset predictions, with a clear advantage over ikarus, which is the state-of-the-art method specifically designed for cancer cell identification.

3.2 scCrab Shows Remarkable Performance in Inter-dataset Prediction Experiments

In real-world application scenarios, newly acquired or yet-to-be-predicted data often exhibit significant differences compared to the dataset used during model training. Therefore, the capability to make predictions across datasets is more practical and holds greater significance compared to intra-dataset predictions. Through experimentation across multiple datasets, we can give a more comprehensive benchmark of machine learning methods' performance in cancer cell identification.

In our investigation, we conducted inter-dataset prediction experiments using the seven previously mentioned datasets. In these experiments, one dataset was designated as the training dataset, while the remaining six datasets functioned as testing datasets. Based on divergent comparison targets, we divided the experiments into two groups. Figure 3a compared the inter-dataset classification performance between scCrab and machine learning methods including SVM, KNN, RF, and ikarus. In Fig. 3b, we compared scCrab with different ensemble approaches to identify the optimal one. The results of performance trained on the dataet LA are presented in Fig. 3, while the remaining results are available in Supplementary Figure 2.

As shown in the radar charts, scCrab demonstrated superior performance in inter-dataset prediction tasks. While SVM performed as the "upper-bound" model in intra-dataset cross-validation experiments, the performance of SVM was inferior to that of scCrab in inter-dataset prediction experiments (Fig. 3a), indicating the limitations of classical machine learning models when predicting newly acquired data. In contrast, the outstanding performance of scCrab in inter-dataset prediction underscores its robustness and superior capacity for predicting new data. Through calculating the difference of all of our evaluation metrics between scCrab and ikarus, we discovered that scCrab increased the balanced accuracy score by 17.22% and increased AUPRC by 14.35%, averaging all training and testing datasets. Concerning various ensemble approaches, we discovered that scCrab, which ensembles MBM and ikarus, provided the

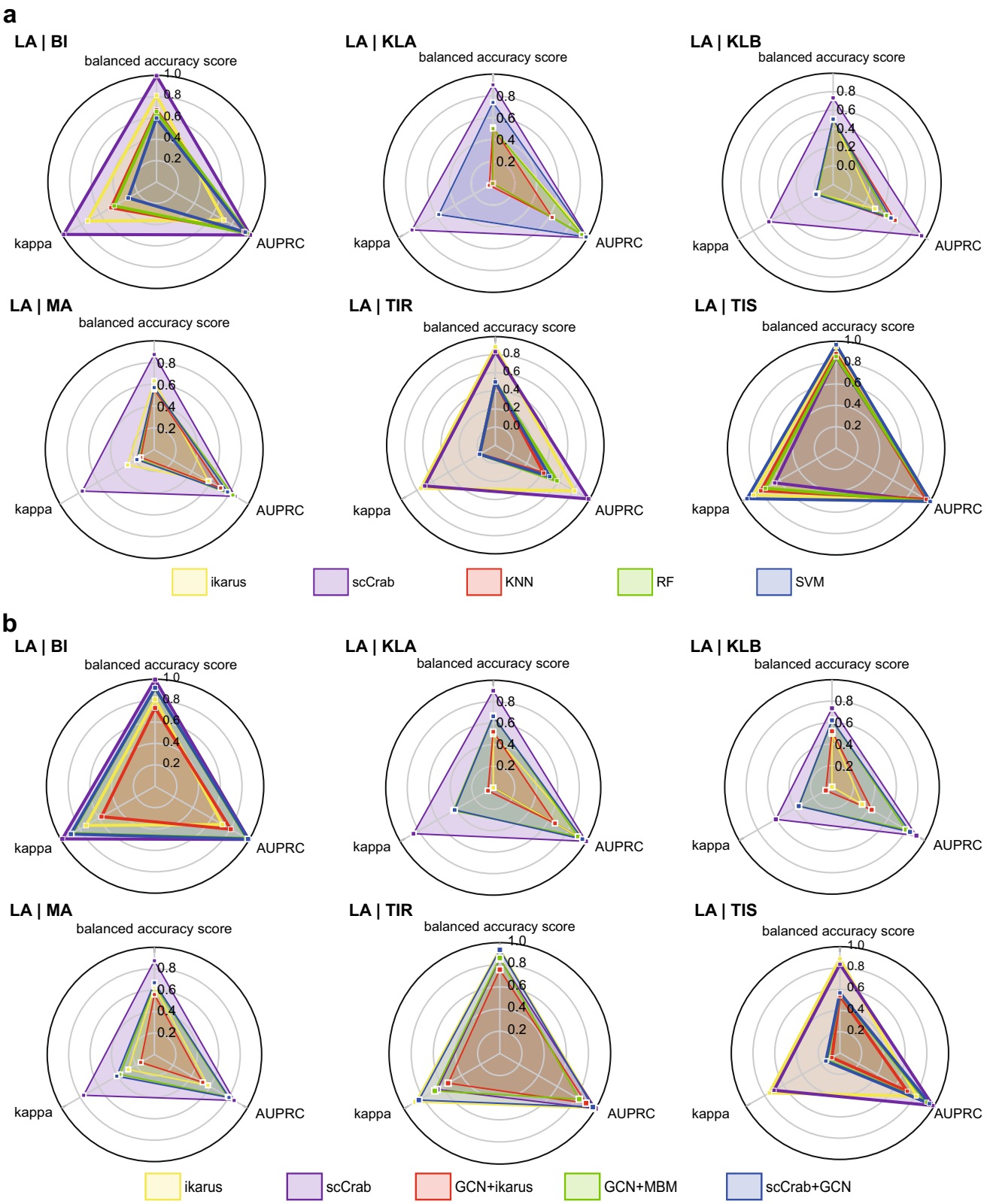


Fig. 3 Radar plots of scCrab and other models. Radar plots of scCrab and other models trained on LA. We utilized the vertical line symbol "|" to connect the training set and test set, for example, "training_set | test_set". In the legend, "+" means ensemble. (**a**) The comparisons of scCrab with other baseline machine learning methods. (**b**) The comparisons of scCrab with other ensembling methods

best performance among others (Fig. 3b). As a result, scCrab was the optimal ensemble approach and outperformed ikarus and other machine learning models in inter-dataset prediction.

3.3 Bayesian Neural Networks Significantly Improve the Performance of scCrab

In order to identify the importance of each component within scCrab, we evaluated the performance of scCrab excluding the self-attention mechanism (wo_Attn), excluding the BNN (wo_BNN), and excluding the random mask with Gaussian noise (wo_Noise). Considering GCN's unique ability to incorporate prior information on gene interactions, we also integrated it into our ensemble model (w_GCN). To guarantee equitable experiments, we conducted inter-dataset prediction experiments employing the seven datasets. The results evaluated by AUPRC are presented in Fig. 4a, while the results of the evaluation metrics kappa and balanced accuracy score are shown in Supplementary Figure 3. With each column representing the model and each row representing the training set, each value in the heatmap of Fig. 4a signifies the average score of testing on the other six datasets. As shown in Fig. 4a, the introduction of GCN did not lead to an improvement in model performance, which suggests that scCrab has opted for the appropriate ensemble approach. Furthermore, the incorporation of the BNN markedly improved the predictive performance of the model. Through calculating the difference in AUPRC results between the ablated models and scCrab, we inferred that the exceptional performance of scCrab is primarily attributed to the BNN, resulting in a 13.10% decrease of AUPRC, compared to wo_Attn (1.36%), wo_Noise (4.75%), w_GCN (6.92%).

3.4 scCrab Exhibits Outstanding Robustness and Stability

Considering the typical presence of substantial noise and varying levels of sparsity in scRNA-seq data is of great importance to identifying the robustness of computational methods to noise and data sparsity. Accordingly, we performed artificial dropout and subsampling experiments to evaluate the robustness and stability of scCrab. In the dropout experiment, we deliberately set 20%, 40%, 60%, and 80% of the non-zero values in the expression matrix of the training set to zero and performed inter-dataset prediction experiments with scCrab. As illustrated in Fig. 4b,

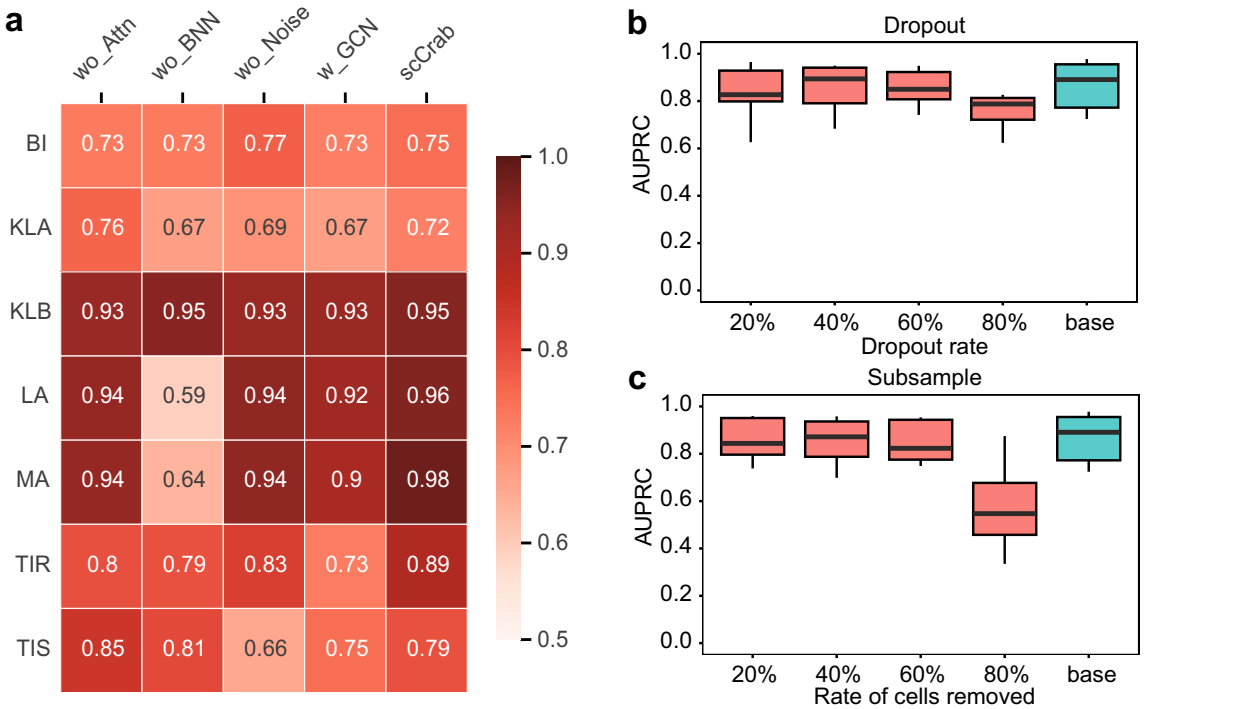


Fig. 4 (**a**) Model ablation experiment of scCrab in 7 scRNA-seq datasets. AUPRC is used as the metric. (**b**) The AUPRC score of classification results with random manual dropout rate of gene expression values. Each box contains the mean scores of training with one dataset and testing with other 6 datasets. Base (the blue box) is the performance of scCrab with the whole scRNA-seq dataset as input. (**c**) The AUPRC score of classification results with random and stratified input subsampling rate of cells. Each box contains the mean scores of training with one dataset and testing with other 6 datasets. Base (the blue box) is the performance of scCrab with the whole scRNA-seq dataset as input

the performance of scCrab without intentional dropout is denoted as the "Base". The results indicate that even with an 80% dropout on the data, the impact on the performance of scCrab is relatively minimal. This highlights scCrab's robustness and stability of prediction, regardless of the random absence of features.

In the subsampling experiment, we randomly removed 20%, 40%, 60%, and 80% of the cells in the training set and carried out inter-dataset prediction experiments with scCrab. We contrasted these results with those trained using the dataset without subsampling (Fig. 4c). Notably, results showed that only when subsampling 80% of the datasets does the performance of scCrab significantly decline. This indicates that scCrab sustains reliable performance despite data subset variations, demonstrating remarkable resilience. This strong robustness and stability prove scCrab an optimal choice for handling complex real-world environments and uncertainties.

3.5 scCrab Effectively Captures the Unique Biological Functions of Cancer Cells

To demonstrate that scCrab can capture underlying biological significance in the process of cancer cell identification, we utilized tumor and normal labels predicted by scCrab to conduct differential expression analysis, and then performed Gene Ontology (GO) enrichment analyses to identify pathways enriched with these differentially expressed genes [37, 38]. Then we can explore the biological functions captured by scCrab. For example, we trained scCrab using the KLB dataset, made predictions on the BI dataset and computed differentially expressed genes between tumor cells and normal cells, and then utilized them for GO enrichment analysis. In GO enrichment analysis, the false discovery rate (FDR) serves to control the proportion of pathways that are incorrectly marked as enriched, and the formula of FDR is detailed in Supplementary Text 5 [39]. Figure 5 illustrated the top 10 significantly enriched annotations in biological processes, cellular components, and molecular functions.

Our analysis revealed numerous GO terms associated with cellular carcinogenesis. For example, protein phosphorylation is a biological processes GO term, playing a common role in cellular signal transduction, controlling various biological processes, including cell growth and proliferation [40]. Aberrant protein phosphorylation is often observed in cancer cells, leading to irregular cell signal transduction. As the inhibition of cell apoptosis results in the uncontrolled proliferation of cancer cells, the apoptotic process represents a self-destructive cell death process that plays a crucial role in normal cell biology, which is another biological processes GO term [41]. Dysfunction of the endoplasmic reticulum membrane, a cellular component GO term critical

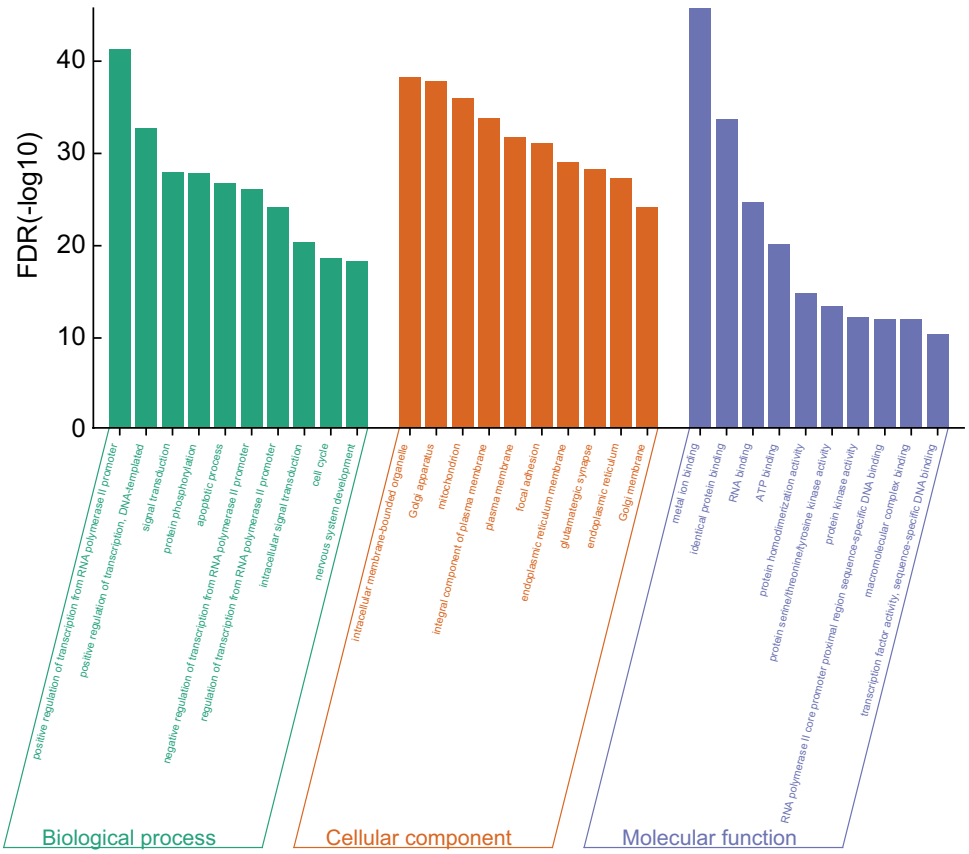


Fig. 5 Enrichment results. The enrichment results of differentially expressed genes between tumor cells and normal cells predicted by scCrab. GO terms are divided as biological processes, cellular components, and molecular functions. There are ten GO terms under each enrichment function

for protein synthesis and folding, can lead to abnormal expression and function of cancer-related proteins when its function is compromised [42]. The GO terms obtained from the predictions of scCrab encompass several key biological processes and molecular functions related to cancer cell genesis and carcinogenesis, with supporting literature linking them to cancer. Moreover, through inter-dataset prediction experiments, we demonstrated that scCrab effectively learns cancer-related information during the training process to identify cancer cells. Taken together, scCrab not only accurately identifies cancer cells, but also effectively captures the unique biological functions of cancer cells during the process of cancer cell identification.

4 Discussion

With the rapid development of scRNA-seq technology, it has become possible to identify cancer cells in a certain tissue through efficient computational methods. However, existing methods fail to incorporate external reference data as prior information, and their identification performance is suboptimal. Here, we propose a novel method called scCrab, a reference-guided cancer cell identification model based on ensemble learning. Since PCA dimensions represent the directions of greatest variance in the source data, scCrab flexibly integrates information from the reference dataset as prior information into a network model by using PCA to generate the projection weight matrix and initialize the weights of the BNN. This allows scCrab to learn intricate patterns and non-linear relationships, leading to higher accuracy and precision in classification. Furthermore, the network model in scCrab combines BNN with a random mask with Gaussian noise and a multi-head self-attention mechanism to enhance its predictive capabilities. Finally, by integrating the network model with ikarus, scCrab showcases enhanced performance. Through comprehensive experiments on seven scRNA-seq datasets, we have shown that scCrab outperforms other state-of-the-art methods, not only in terms of predictive accuracy of inter- and intra-datasets cancer cell identification experiments but also in its utilization of external reference data. Furthermore, we have demonstrated the robustness of scCrab when confronted with dropout and subsampling challenges and highlighted the significance of BNN as a crucial component of scCrab. Moreover, as an identification model specialized in cancer cells, scCrab has the capability to uncover cell heterogeneity related to cancer cells during the identification process. Therefore, scCrab can provide essential guidance to the automatic identification of cancer cells in scRNA-seq data. We have also benchmarked the time and memory resource requirements of scCrab and alternative methods, explained more detailedly in Supplementary Text 6. We anticipate that scCrab will assist in future cancer research and facilitate the development of personalized cancer therapy.

We also provide several avenues for enhancing scCrab. Firstly, as appropriate external datasets can offer valuable insights and information to enhance performance, we can explore the utilization of different datasets as external reference data. Secondly, with the rapid development of large language models, we can further learn and extract the underlying information from huge amounts of single-cell data based on pre-trained models. Finally, with the development and integration of single-cell multi-omics techniques such as cellular epigenomics, we can further explore the gene regulatory mechanisms of cancer cells by integrating multi-omics data [43].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12539-024-00655-6.

Funding This work was supported by the National Natural Science Foundation of China [62203236], the Fundamental Research Funds for the Central Universities, Nankai University [63231137], and the Young Elite Scientists Sponsorship Program by CAST [2023QNRC001].

Data Availability The TIR dataset was collected from NCBI Gene Expression Omnibus (GEO) with the accession number GSE161801. The LA dataset is available at GSE123904. The MA dataset is available at GSE151530. The BI dataset is available at GSE196303. The TIS dataset is available at GSE103322. The KLA and KLB datasets are available at https://neuroblastoma-cell-atlas.cog.sanger.ac.uk/nb_GOSH_cellxgene.h5ad and https://neuroblastoma-cell-atlas.cog.sanger.ac.uk/nb_PMC_cellxgene.h5ad.

Code Availability The MIT-licensed scCrab software, including detailed documents and tutorials, is freely available at https://github.com/BioX-NKU/scCrab.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Schiffman JD, Fisher PG, Gibbs P (2015) Early detection of cancer: past, present, and future. Am Soc Clin Oncol Educ Book 35(1):57–65. https://doi.org/10.14694/EdBook_AM.2015.35.57
2. Gonzalez-Silva L, Quevedo L, Varela I (2020) Tumor functional heterogeneity unraveled by scRNA-seq technologies. Trends Cancer 6(1):13–19. https://doi.org/10.1016/j.trecan.2019.11.010
3. Smith RA, Cokkinides V, von Eschenbach AC et al (2002) American cancer society guidelines for the early detection of cancer. Ca-Cancer J Clin. https://doi.org/10.3322/canjclin.52.1.8
4. Pasquini G, Rojo Arias JE, Schafer P et al (2021) Automated methods for cell type annotation on scRNA-seq data. Comput Struct Biotechnol J 19:961–969. https://doi.org/10.1016/j.csbj.2021.01.015

5. Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. Nat Rev Genet 17(3):175–188. https://doi.org/10.1038/nrg.2015.16
6. Kanter I, Kalisky T (2015) Single cell transcriptomics: methods and applications. Front Oncol 5:53. https://doi.org/10.3389/fonc.2015.00053
7. Liu S, Trapnell C (2016) Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research 1:2. https://doi.org/10.12688/f1000research.7223.1
8. Chen X, Chen S, Song S et al (2022) Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. Nat Mach Intell 4(2):116–126. https://doi.org/10.1038/s42256-021-00432-w
9. Pliner HA, Shendure J, Trapnell C (2019) Supervised classification enables rapid annotation of cell atlases. Nat Methods 16(10):983–986. https://doi.org/10.1038/s41592-019-0535-3
10. Ranjan B, Sun W, Park J et al (2021) DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. Nat Commun 12(1):5849. https://doi.org/10.1038/s41467-021-26085-2
11. Zhang Z, Chen S, Lin Z (2023) RefTM: reference-guided topic modeling of single-cell chromatin accessibility data. Briefings Bioinf 24(1):1–11. https://doi.org/10.1093/bib/bbac540
12. Abdelaal T, Michielsen L, Cats D et al (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20(1):194. https://doi.org/10.1186/s13059-019-1795-z
13. Li H, Courtois ET, Sengupta D et al (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet 49(5):708–718. https://doi.org/10.1038/ng.3818
14. Miao Z, Moreno P, Huang N et al (2020) Putative cell type discovery from single-cell gene expression data. Nat Methods 17(6):621–628. https://doi.org/10.1038/s41592-020-0825-9
15. De Kanter JK, Lijnzaad P, Candelli T et al (2019) CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res 47(16):e95–e95. https://doi.org/10.1093/nar/gkz543
16. Dohmen J, Baranovskii A, Ronen J et al (2022) Identifying tumor cells at the single-cell level using machine learning. Genome Biol 23:123. https://doi.org/10.1186/s13059-022-02683-1
17. Aibar S, González-Blas CB, Moerman T et al (2017) SCENIC: single-cell regulatory network inference and clustering. Nat Methods 14(11):1083–1086. https://doi.org/10.1038/nmeth.4463
18. Bishop CM (1995) Training with noise is equivalent to Tikhonov regularization. Neural Comput 7(1):108–116. https://doi.org/10.1162/neco.1995.7.1.108
19. Vincent P, Larochelle H, Bengio Y et al (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp 1096–1103. https://doi.org/10.1145/1390156.1390294
20. MacKay DJC (1995) Bayesian neural networks and density networks. Nucl Instrum Methods Phys Res Sect A 354(1):73–80. https://doi.org/10.1016/0168-9002(94)00931-7
21. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp 807–814. https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf
22. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA. pp 6000–6010. https://user.phil.hhu.de/~cwurm/wp-content/uploads/2020/01/7181-attention-is-all-you-need.pdf
23. Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. Rep Sci 9:5233. https://doi.org/10.1038/s41598-019-41695-z
24. Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19:15. https://doi.org/10.1186/s13059-017-1382-0
25. Lambrechts D, Wauters E, Boeckx B et al (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 24(8):1277. https://doi.org/10.1038/s41591-018-0096-5
26. Ma L, Wang L, Khatib SA et al (2021) Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. J Hepatol 75(6):1397–1408. https://doi.org/10.1016/j.jhep.2021.06.028
27. Bischoff P, Trinks A, Wiederspahn J et al (2022) The single-cell transcriptional landscape of lung carcinoid tumors. Int J Cancer 150(12):2058–2071. https://doi.org/10.1002/ijc.33995
28. Puram SV, Tirosh I, Parikh AS et al (2017) Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 171(7):1611–1624.e24. https://doi.org/10.1016/j.cell.2017.10.044
29. Tirier SM, Mallm JP, Steiger S et al (2021) Subclone-specific microenvironmental impact and drug response in refractory multiple myeloma revealed by single-cell transcriptomics. Commun Nat 12:6960. https://doi.org/10.1038/s41467-021-26951-z
30. Kildisiute G, Kholosy WM, Young MD et al (2021) Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma cancer cell. Sci Adv 7(6):1–13. https://doi.org/10.1126/sciadv.abd3311
31. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
32. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, pp 233–240. https://doi.org/10.1145/1143844.1143874
33. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104
34. Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Inf Process Syst 29:29. https://proceedings.neurips.cc/paper_files/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf
35. Wang T, Bai J, Nabavi S (2021) Single-cell classification using graph convolutional networks. BMC Bioinf 22(1):364. https://doi.org/10.1186/s12859-021-04278-2
36. Wu Z, Pan S, Chen F et al (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24. https://doi.org/10.1109/TNNLS.2020.2978386
37. Sherman BT, Hao M, Qiu J et al (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res 50(W1):W216–W221. https://doi.org/10.1093/nar/gkac194
38. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. https://doi.org/10.1038/nprot.2008.211
39. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57(1):289–300. http://www.jstor.org/stable/2346101
40. Cohen P (2000) The regulation of protein function by multisite phosphorylation—a 25 year update. Trends Biochem Sci 25(12):596–601. https://doi.org/10.1016/S0968-0004(00)01712-6

41. Elmore S (2007) Apoptosis: a review of programmed cell death. Toxicol Pathol 35(4):495–516. https://doi.org/10.1080/01926230701320337
42. Hetz C, Papa FR (2018) The unfolded protein response and cell fate control. Mol Cell 69(2):169–181. https://doi.org/10.1016/j.molcel.2017.06.017
43. Gao Z, Chen X, Li Z et al (2023) scEpiTools: a database to comprehensively interrogate analytic tools for single-cell epigenomic data. J Genet Genomics 51(4):462–465. https://doi.org/10.1016/j.jgg.2023.09.011