Bioinformatics, 2024, **40(2)**, btae067
https://doi.org/10.1093/bioinformatics/btae067
Advance Access Publication Date: 23 February 2024
Original Paper

OXFORD

Gene expression

Phenotype prediction from single-cell RNA-seq data using attention-based neural networks

Yuzhen Mao[1,‡], Yen-Yi Lin [2,3,‡], Nelson K.Y. Wong[4], Stanislav Volik[3], Funda Sar[2,3], Colin Collins[2,3,*], Martin Ester[1,3,*]

[1]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[2]Department of Urologic Sciences, University of British Columbia, Vancouver BC V5Z 1M9, Canada
[3]Vancouver Prostate Centre, Vancouver, BC V6H 3Z6, Canada
[4]Department of Experimental Therapeutics, BC Cancer, Vancouver BC V5Z 1L3, Canada

*Corresponding authors. School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada. E-mail: ester@sfu.ca (M.E.); Department of Urologic Sciences, University of British Columbia, 2775 Laurel Street, Vancouver, BC V5Z 1M9, Canada. E-mail: colin.collins@ubc.ca (C.C.)

‡These authors contributed equally to this work.

Associate Editor: Christina Kendziorski

Abstract

Motivation: A patient's disease phenotype can be driven and determined by specific groups of cells whose marker genes are either unknown or can only be detected at late-stage using conventional bulk assays such as RNA-Seq technology. Recent advances in single-cell RNA sequencing (scRNA-seq) enable gene expression profiling in cell-level resolution, and therefore have the potential to identify those cells driving the disease phenotype even while the number of these cells is small. However, most existing methods rely heavily on accurate cell type detection, and the number of available annotated samples is usually too small for training deep learning predictive models.

Results: Here, we propose the method ScRAT for phenotype prediction using scRNA-seq data. To train ScRAT with a limited number of samples of different phenotypes, such as coronavirus disease (COVID) and non-COVID, ScRAT first applies a mixup module to increase the number of training samples. A multi-head attention mechanism is employed to learn the most informative cells for each phenotype without relying on a given cell type annotation. Using three public COVID datasets, we show that ScRAT outperforms other phenotype prediction methods. The performance edge of ScRAT over its competitors increases as the number of training samples decreases, indicating the efficacy of our sample mixup. Critical cell types detected based on high-attention cells also support novel findings in the original papers and the recent literature. This suggests that ScRAT overcomes the challenge of missing marker genes and limited sample number with great potential revealing novel molecular mechanisms and/or therapies.

Availability and implementation: The code of our proposed method ScRAT is published at https://github.com/yuzhenmao/ScRAT.

1 Introduction

Accurate prediction of phenotypes for patients in given cohorts is critical in advancing diagnosis, prognosis, and therapy (Ching *et al.* 2018). Heterogeneous symptoms may lead to ambiguous predictions (Morley *et al.* 2021), and therefore the analyses based on high-throughput omics data have started to enter clinical routine in the last decade (Chen *et al.* 2021). A challenging step in these analyses is to dissect cellular content from patients' genomic profiles, including the detection of cell types defined by gene expression profiles and their proportions in different patients (Newman *et al.* 2019). While phenotype information, such as tumor metastasis, disease stage, and treatment response for bulk tissue samples are widely collected from various consortia, their gene expression profiles are measured by averaging cells across the whole tissue, which often do not reveal the full complexity of diverse cell types within patients.

Recent advances of single-cell and single-nuclei RNA-Sequencing (sc/snRNA-Seq) enable gene expression profiling at the unprecedented single-cell resolution. While this technology improves our understanding of cell-type markers and disease-specific signatures, analysis of large-scale cohorts is not clinically practical, especially for cancer research, for the following reasons. (i) Dependence of accurate cell type identification which might be biased or unavailable. Most single-cell RNA sequencing (scRNA-seq) analysis starts with detecting cell types using unsupervised clustering, followed by cell-type annotations based on marker genes. Patients' phenotypes are then predicted by distributions of cell types or identifying specific cell types. However, accurate cell type identifications are affected by the proper clustering resolution for a given sample, and the marker gene information that might be suboptimal or missing. Therefore, many existing scRNA-seq analysis methods even require users to provide the number of cell types, which is unknown before analysis, or set up a universal value to provide the corresponding analysis results. (ii) Limited number of samples. Most well-annotated scRNA-seq datasets involve few than 20 samples, whose statistical power is too weak to support the phenotype prediction and the findings of phenotype-specific cell

Received: 5 June 2023; **Revised:** 15 December 2023; **Editorial Decision:** 19 January 2024

types. Such small size of samples can also lead to serious overfitting for most machine learning models and significantly affect their prediction performance. (iii) Lack of interpretability. Many computational methods try to resolve the above issues but rarely provide users much insight into cell types and molecular mechanisms driving or related to phenotypes. Due to the above reasons, scRNA-seq often needs to be integrated with bulk assays in the analysis, and people mainly apply these methods to study compositions of single tissues rather than their phenotypes for diagnosis and prognosis applications.

Here, we present ScRAT, a phenotype prediction framework that can learn from limited numbers of scRNA-seq samples with minimal dependence on cell-type annotations. Compared to most available scRNA-seq analysis algorithms that model gene expression profiles of different cell clusters by separate Gaussian distributions (He *et al.* 2021, Zeng *et al.* 2022), the first contribution in ScRAT is that we utilize the attention mechanism to measure interactions between cells as their correlations, or attention weights. For each cell, we incorporate all of its interaction patterns and attention weights to establish its connections with the corresponding phenotypes. Secondly, we introduce a mixup module in our framework as a data augmentation approach to mitigate the potential overfitting issue caused by the high model complexity together with the very limited number of labeled samples. Lastly, ScRAT establishes the connection between the input (cells) and the output (phenotypes) of the Transformer model simply using the attention weights. This is cost-effective compared to existing approaches from the literature that tend to be computationally expensive, such as gradients propagation or training probing classifiers (Clark *et al.*, 2019, Chefer *et al.* 2020, Chefer *et al.* 2021). ScRAT hence selects cells containing the most discriminative information to specific phenotypes, or *critical cells*, using their attention weights. It provides a natural way to construct phenotype-specific subpopulations in clinical cohorts that suggest prognostic markers and potential therapeutic information.

We evaluate ScRAT on three public coronavirus disease (COVID) datasets compared to five baseline frameworks. In each dataset, we would split the samples into two phenotypes based on the given annotation: COVID versus non-COVID, mild/moderate versus severe/critical, or convalescence versus progression. We also reduce the number of training samples to investigate the predictive power of each framework. ScRAT achieves the best area under the receiver of characteristic curve (AUC) in all comparisons and provides leading precision and recall in most scenarios. The performance edge of ScRAT over its competitors increases as the number of training samples decreases, indicating the efficacy of our sample mixup module. Since these public datasets come with cell type annotations in various resolutions, we also examine the connections between phenotypes and subpopulations enriched with high-attention cells. Our experiment shows that ScRAT can detect disease-critical and phenotypic-driver subpopulations using high-attention cells that can potentially help to identify novel conditions of druggable populations.

In short, ScRAT is the first deep neural network based method to predict phenotypes from scRNA-Seq, and among the first attention-based framework for scRNA-seq analysis. Our integration of attention mechanism and mixup allows ScRAT to be independent of cell-type annotations, capable of learning from a limited number of training samples. Lastly, we propose a simple method to explain the prediction of Transformer that is more cost-effective than the existing methods. This indicates ScRAT can provide interpretable information to guide biologists.

2 Related work

2.1 Deep learning in single-cell RNA-seq analysis

Single-cell RNA-seq has become a popular tool for gene expression analysis at a single-cell resolution, and deep learning techniques have been shown promising results on many related tasks (Lopez *et al.* 2018). For example, Yin *et al.* (2022) propose an autoencoder-based classification framework to obtain compressed representations of scRNA-seq data. These representations are then fed into subsequent classifiers to predict the cell types. Ravindra *et al.* (2020) use graph attention networks to construct a graph representation of the scRNA-seq data, where each node represents a cell and each edge represents the similarity between two cells. The disease state for each cell is then predicted based on the learned graph representations.

2.2 Phenotype prediction using bulk RNA-seq

Gene expression profiling has been used to predicting phenotypes in many clinical settings (Lonsdale *et al.* 2013, Uhlen *et al.* 2017). PAM50 classifies breast tumor based on expression profiles of 50 genes (Perou *et al.* 2000). Molecular phenotypes of prostate cancer also relie on gene expression profiling (Cancer Genome Atlas Research Network 2015), and multiple expression-based diagnosis tests have also been developed. For example, The Prolaris cell-cycle progression predicts aggressiveness for prostate cancer using expressions of 31 genes from the cell cycle proliferation pathway (Cuzick *et al.* 2012). A signature of 157 genes was developed to predict lethal prostate cancer (Penney *et al.* 2011). Oncotype Dx genomic prostate score (Cullen *et al.* 2015) and Decipher Biopsy score (Erho *et al.* 2013) also identify gene signatures to predict the risk of metastasis as the tumor outcome. These methods are mainly developed using bulk assays, and can not benefit from cell-level resolution information in scRNA-seq to improve diagnosis and prognosis.

2.3 Analysis of phenotype-driving (sub-)cell type(s) in RNA-seq

To better utilize information from scRNA-seq datasets, cell-type deconvolution methods (Newman *et al.* 2019, Fan *et al.* 2022) have been developed to dissect compositions of cell populations in bulk RNA-seq. LRCell (Ma *et al.* 2022) further extends the deconvolution methods to estimate the impacts of each cell type on differentially expressed genes (DEGs) between bulk RNA-seq datasets. LRcell takes a pre-defined list of (sub-)cell type(s) and their marker genes to analyze the contribution of these cell types to the DEGs in RNA-Seq datasets. While LRCell was designed to incorporate marker genes from scRNA-seq datasets, the problem it tries to solve is different from that of scRAT in the following ways. First of all, LRCell can not predict the phenotypes of individual patients. The final output of LRCell is the rank of predefined (sub-)cell types based on their impacts driving DEGs in RNA-Seq data using linear regression models. The rank of cell (sub)types can not directly predict the phenotypes of a patient. In addition, LRCell requires sets of (pre-embedded) marker genes from all (sub-)cell types of the bulk tissue acquired either from scRNA-seq datasets a priori or from MSigDB C8, cell-type signature gene sets.

2.4 Phenotype prediction using single-cell RNA-seq

CloudPred (He *et al.* 2021) models the individual points as samples from a mixture of Gaussians, probabilistically assigns points to clusters, then estimates prevalence of the subpopulations and uses it to predict the phenotype of that patient. scPheno (Zeng *et al.* 2022) constructs gene expression profiles by a joint distribution of cell states and disease phenotypes based on a deep generative probabilistic model, and feeds the distribution as the predictive features to support vector machine for the phenotype prediction. Their assumption of modeling single cell populations as Gaussian is limiting. In addition, while these methods can work under a small number of labeled training data, their limited model capacity might not be able to fully capture latent information hidden in high-dimensional scRNA-seq datasets.

3 Problem definition

A *cell* is the most basic unit in scRNA-seq experiments and will be denoted by a vector c over m genes including the measure of gene expression level, e.g. the count of Unique Molecular Identifiers (UMIs). The ultimate prediction unit is denoted as a *sample*, which is extracted from a single patient. A sample consists of n cells and is represented as an $n \times m$ matrix S, where S_{ij} corresponds to the raw or normalized UMIs of the j-th gene in the i-th cell. Each sample is associated with a specific one-hot encoded phenotype label from a pre-defined set $\mathcal{P} = \{P^1, P^2, \ldots P^o\}$. Based on that, we formally define our problem as follows:

Problem: Phenotype prediction for scRNA-seq samples.
Input: A set of labeled samples represented by scRNA-seq matrices $\mathcal{D} = \{S_1, S_2, \ldots, S_C\}$, and their corresponding labels $\mathcal{Y} = \{P_1, P_2, \ldots, P_C\}$; a set of unlabeled samples $\mathcal{D}' = \{S_1', S_2', \ldots, S_L'\}$.
Output: A prediction model which maps $S_i' \in \mathcal{D}'$ to $P^i \in \mathcal{P}$.

4 Materials and methods

In this section, we present a neural network-based method called ScRAT to predict the phenotype of an scRNA-seq sample. An overview of ScRAT is presented in Fig. 1, which consists of three major modules: Sample Mixup, Attention Layer, and Phenotype Classifier. Our method takes an scRNA-seq sample from a single patient as input. Note that the order of cells within each sample does not matter and the size of each sample is variable. To alleviate the possible overfitting issue, we employ a data augmentation technique called "sample mixup" during the training time to increase the amount and diversity of training samples. The backbone of ScRAT is a multi-head attention layer (Vaswani *et al.* 2017), which aims to learn a task-orientated embedding for each cell within the sample. Considering its poor scalability (Tay *et al.* 2020), a cropping strategy is applied to the input sample before passing it to the attention layer. As the last step, a one-layer Multi-Layer Perceptron (MLP) takes the output of the attention layer and predicts the phenotype as a probability distribution over the different values of the phenotype. In the following subsections, we delve into these three modules of ScRAT in detail.

4.1 Sample mixup

The size of currently available scRNA-seq datasets is very small, and it is expected to remain relatively small in the near future, which will likely result in overfitting when training a deep learning model. Mixup and its variants (Zhang *et al.* 2017, Verma *et al.* 2019) are interpolation-based and widely-adopted data augmentation techniques for regularizing neural networks and improving model generalizability (Carratino *et al.* 2022). For instance, in computer vision setting, mixup convexly combines random pairs of images and their associated labels to generate new training data. Inspired by this, for scRNA-seq analysis, we introduce a simple but efficient data augmentation method, sample mixup, to generate new samples during training process. Specifically, given two scRNA-seq samples S and S' together with a fixed $\lambda \in [0, 1]$, sample mixup is defined as follows (Zhang *et al.* 2017):

$$\begin{aligned} \{\tilde{x} | \tilde{x} &= \lambda x_i + (1-\lambda) x_i'\}, \\ \tilde{y} &= \lambda y + (1-\lambda) y', \end{aligned} \tag{1}$$

where x_i and x_i' are gene expression profile of cells drawn from S and S', and y and y' are corresponding one-hot phenotype label encodings.

Compared to the computer vision setup, samples here correspond to images, cells in each sample correspond to pixels in each image, and phenotypes of samples correspond to labels of images. The main differences between these two scenarios are that pixels in one image can only be mixed with pixels in the same spatial location of another image, and mixup can only be applied to images having the same size. scRNA-seq data are not limited by these two constraints.

The proposed scRNA-seq sample mixup aims to increase the number and diversity of samples. Specifically, given a pair of samples S_1 and S_2 with the same or different phenotypes, we first randomly sample a batch S_{11} containing N cells only from S_1, and sample another batch S_{21} with the same amount of cells only from S_2. Each batch is allowed to include duplicate cells during sampling. Then mixup is applied to S_{11} and S_{21} based on Equation (1), where $\lambda \sim \text{Beta}(\alpha, \alpha)$, for $\alpha \in (0, \infty)$ (Zhang *et al.* 2017, Carratino *et al.* 2022), to generate N augmented cells forming a new sample S_3 called pseudo-sample, with the phenotype label equals to the linear combination of phenotype labels of S_1 and S_2.

Notably, since cells of different populations are biologically very different, it does not make much sense to directly apply mixup to them. Therefore, although our model does not require cell type information, during the sample mixup, we only mix cells of the same cell population, assuming that this information has either been annotated by a human expert or been determined automatically by state-of-the-art annotation methods such as MARS (Brbić *et al.* 2020). For cell populations that appear only in one of the samples, we add Gaussian noise to the gene expression profile of cells that belong to those unique cell populations during the mixup.

Sample mixup also ensures that the proportion of each cell population in the pseudo-sample is the linear combination of the proportions of that cell population in two original samples. For example, given $\lambda = 0.2$ and the proportions of cell population A in two original samples are 30% and 20%, respectively, then the proportion of A in the pseudo-sample is calculated as: $0.2 \times 30\% + 0.8 \times 20\% = 22\%$.

The effectiveness of sample mixup has been evaluated in our ablation study (Supplementary Section S4).

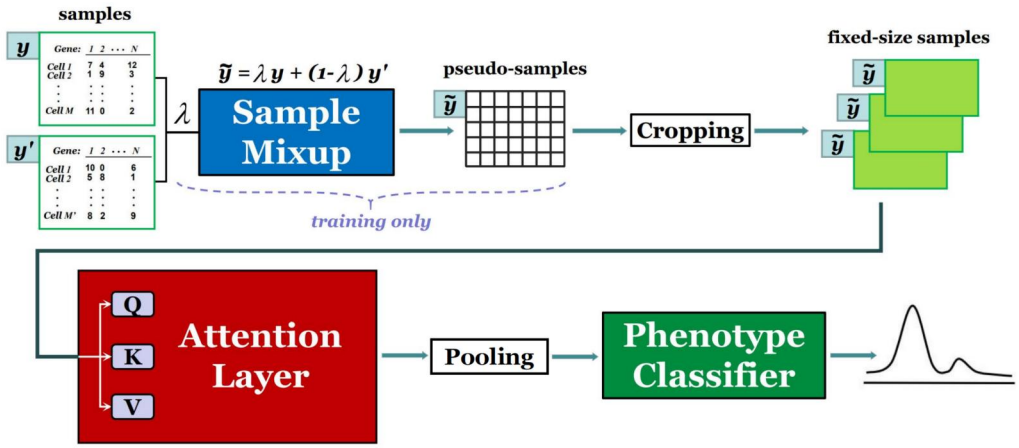

Figure 1. An overview of ScRAT, which consists of three main modules: Sample Mixup, Attention Layer, and Phenotype Classifier. It takes a scRNA-seq sample (a set of cells) as input, and outputs the predicted phenotype for the input sample.

4.2 Attention layer

Attention mechanisms (Bahdanau *et al.* 2014) have achieved state-of-the-art performance in a wide range of machine learning tasks which take a set of elements as input, such as words (Devlin *et al.* 2018) and pixels (Dosovitskiy *et al.* 2020). An attention mechanism pays more attention to the relatively important elements by assigning high weights to them during the forward pass. Multi-head Attention is one of the most popular versions of this mechanism which was first proposed in (Vaswani *et al.* 2017), and we use attention as a synonym for this version here. Compared with classical neural nets such as MLP and convolutional neural network (CNN) (Krizhevsky *et al.* 2017), attention can not only deal with variable-sized inputs but also assign weights to different elements dynamically, which is necessary for unordered inputs.

Specifically, the input of the attention layer is a set of cell embeddings $\mathbf{c} = \{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_N\}, \vec{c}_i \in \mathbb{R}^{d_{\text{in}}}$, where N is the number of cells, and d_{in} is the number of features in each embedding. Following the previous work (Vaswani *et al.* 2017) closely, our attention layer maps the input embeddings to three different kinds of vectors: key, query and value using three weight matrices with the same shape respectively: $\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v \in \mathbb{R}^{d_{\text{kqv}} \times d_{\text{in}}}$, where d_{kqv} is the dimensionality of key, query and value. Afterwards, a self-attention with scaled dot-product is applied to each pair of cells to compute their attention weights based on their key and query vectors:

$$s_{ij} = \frac{\text{dot-product}(\mathbf{W}_q \vec{c}_i, \mathbf{W}_k \vec{c}_j)}{\sqrt{d_{\text{kqv}}}}, \tag{2}$$

which denote the importance of cell j to cell i and are normalized using the softmax function:

$$a_{ij} = \text{softmax}_j(s_{ij}) = \frac{\exp(s_{ij})}{\sum_{k=1}^{N} \exp(s_{ik})}. \tag{3}$$

These attention weights are then treated as the weights in the following linear combination process which outputs a new embedding for each cell based on the value vectors of all cells:

$$\vec{h}_i = \sum_{j=1}^{N} a_{ij} \mathbf{W}_v \vec{c}_j. \tag{4}$$

To extract information at different positions as well as make the training process more stable (Liu *et al.* 2021), multi-head attention (Vaswani *et al.* 2017) is applied in our attention layer. Specifically, instead of utilizing only one attention head with one group of $\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v$, we utilize K attention heads with K different groups of mapping matrices and run them in parallel. Afterwards, we concatenate the outputs from each head and apply an additional linear layer to it at the end.

In a nutshell, our attention layer is formulated as:

$$\text{Attention}(\vec{c}_i) = \text{Concat}(\vec{h}_i^1, \ldots, \vec{h}_i^K)\mathbf{W}_o, \tag{5}$$

where $\mathbf{W}_o \in \mathbb{R}^{Kd_{\text{kqv}} \times d_{\text{out}}}$ is a weight matrix, $\vec{h}_i^k$ is the output of the k-th head based on Equation (4).

One limitation of existing attention-based models is that they cannot handle very long sequences as input since the self-attention operation has quadratic run-time and memory complexity (Beltagy *et al.* 2020, Zhou *et al.* 2021). Therefore, after augmenting the whole dataset using mixup, we introduce a cropping strategy to both training and test data, which randomly selects several subsets from each sample and only uses these subsets to train the model. We call these subsets of cells "fixed-size samples" in this article.

More specifically, for each sample, we randomly select NC cells as one fixed-size sample, and generate NS fixed-size samples for each sample. During the training process, each fixed-size sample is calculated a loss which is added up in the final loss computation used to update the model parameters; while during the testing process, we assign a (categorical) predicted label to each fixed-size sample by setting a threshold, and use majority vote to assign the predicted label to each sample based on their fixed-size samples. Here, NC and NS are both hyper-parameters which can be tuned by the users. Since NC could be relatively small, this cropping strategy improves the model scalability. Moreover, this strategy is an analogy to the cropping in the computer vision setting, and therefore can also be treated as a useful data augmentation

approach. In the next section, comprehensive experiments demonstrate its effectiveness.

4.3 Phenotype classifier

The outputs of the attention layer are the embeddings of all cells within the input sample. Similar to the way the average pooling function operates in image classification, we aggregate the cell embeddings for each sample by computing the average value along each dimension. While this method may cause some loss of information, it is a commonly used and effective technique to simplify the feature map representation and improve the model's generalization performance. Moreover, it ensures that the cell order does not affect the final results. Finally, the aggregated embedding is passed to the phenotype classifier, a one-layer MLP, which outputs the predicted phenotype for the input sample, i.e. a probability distribution over the different values of the phenotype.

5 Experiments

We evaluate the performance of ScRAT on three large-scale public COVID scRNA-seq datasets, and compare it with five state-of-the-art methods. We perform an ablation study to determine the impact of the different ScRAT components. Finally, we design a cost-effective method to convert cell attention weights in ScRAT as the relevance score which determines the relevance of a given cell population with respect to the phenotype. Our biological analysis demonstrates the potential of revealing disease mechanisms based on the critical cell types identified using attention weights in ScRAT.

5.1 Experimental setup

5.1.1 Datasets

Our experiments include four tasks based on the following three scRNA-seq COVID19 cell datasets. For COMBAT (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 2022) and Haniffa (Stephenson *et al.* 2021) datasets, we perform the task of disease diagnosis (i.e. COVID versus Non-COVID). For SC4 (Ren *et al.* 2021) which includes mostly COVID samples, we perform two separate tasks of predicting severity (i.e. mild/moderate versus severe/critical) and stage (i.e. convalescence versus progression). Please refer to Supplementary Table S1 for more information.

5.1.2 Design of experiments

To reflect the limited number of labeled scRNA-seq samples in real applications, we first define the *Training Ratio* for each task as the number of samples included in the training data divided by total number of samples in the dataset, and split the original dataset into the training and test datasets accordingly. For each given training ratio ranging from 9% to 50% in our design, we run the experiments for 100 random splits to better evaluate the performance of different methods.

The limited number of patients in the datasets makes it impractical to control for different values of clinical variables, such as age and sex, in subsampling. Therefore, we randomly split patients into training and testing in each iteration. The sampling results should reflect the overall trends of clinical variables in the original datasets (Ren *et al.* 2021, Stephenson *et al.* 2021, COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 2022), and this is the best available strategy to minimize the potential impacts driven by other confounding factors.

Considering the high dimensionality of scRNA-seq data which is likely to result in serious over-fitting, we map the original input to a low dimensional latent space and keep only 50 principal components by principal components analysis (PCA) (Halko *et al.* 2011). The AUC is used as the evaluation metric in the following discussions.

5.1.3 Baselines

We compare ScRAT with five popular phenotype prediction methods, including two pseudo-bulk methods: (i) "Linear" and (ii) "Feedforward (bulk)," and three single-cell methods: (iii) "Feedforward (single)," (iv) "Attention," and (v) "CloudPred." Please refer to Supplementary Section S2 for more details.

5.1.4 Configuration of ScRAT

Throughout our experiments, we observed that the performance of the model remained largely unaffected when the number of pseudo-samples exceeded 250 (Supplementary Fig. S1). Therefore, for each experiment of ScRAT, we apply mixup to generate 300 pseudo-samples with 10 000 cells in each from the original training samples, and only use these 300 pseudo-samples to train the model. α in the beta-distribution of mixup is set to 0.5. For the cropping strategy, we set the number of cells in each fixed-size sample (NC) to 500, and set the number of the fixed-size samples (NS) to 20 and 50 for training and testing, respectively. We only use one attention layer, and set the number of attention heads K to 8 and the dimension of each head d_{kqv} to 16. We use Adam optimizer with learning rate as 1e−2. All the hyper-parameters are decided using the 5-fold cross validation technique. See Supplementary Section S2 for more details.

5.2 Prediction results

We compare ScRAT with five baseline methods on four tasks, and provide the AUC of all methods in Fig. 2. In general, we have the following observations: (i) ScRAT consistently outperforms all baseline methods on four tasks, which demonstrates the effectiveness and generalizability of ScRAT. More specifically, the performance edge of ScRAT over the second best method (usually the vanilla attention) increases as the training dataset size (number of samples) decreases, verifying the usefulness of our proposed sample mixup as a data augmentation approach. For example, at training ratio = 9%, the P-value of the t-test between AUCs of ScRAT and vanilla attention is much smaller than 0.01 in all but the SC4-Severity tasks. (ii) Vanilla attention layer is the second best model for all four tasks, which indicates the strengths of the attention mechanism in phenotype prediction task using scRNA-seq data. (iii) Feed-forward (single) has the worst AUCs in the Stage and Severity tasks, and also has low precision in both COMBAT and Haniffa datasets compared to ScRAT, the attention model, and even Feed-forward (bulk) (Supplementary Figs S3 and S4). While Feed-forward (single) can use the information at the cell-level resolution, it processes each cell separately without considering their connections as mentioned in Supplementary Section S2. On the other hand, Feedforward (bulk) takes the averaged gene expression profiles as input, which automatically aggregates the whole sample's information in predictions. This suggests that processing each cell separately and naively averaging their embeddings like Feed-forward (single) will not be able to utilize the advantage of single-cell datasets and also

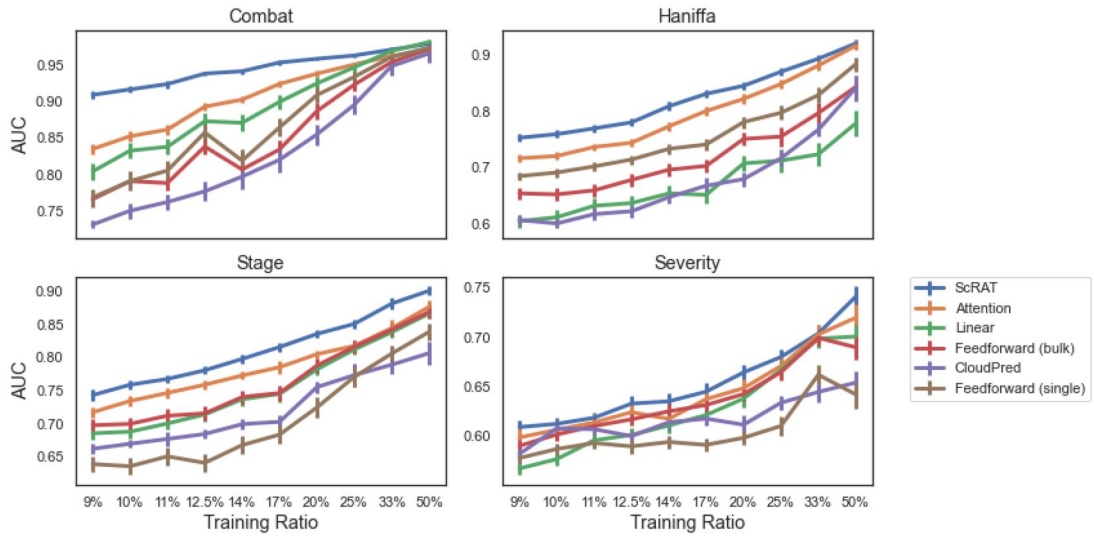

Figure 2. Comparison of different methods on four different tasks. For each task, we report the prediction results of all methods using AUC ± 95% confidence intervals for 10 different training ratios. ScRAT outperforms other methods in all settings, followed by vanilla attention (the *P*-value of *t*-test between ScRAT and vanilla attention $\ll$0.01 in all but the SC4-Severity tasks at Training Ratio = 9%). The performance edge of ScRAT over vanilla attention increases as the training ratio decreases, especially for the Combat datasets. See Supplementary Figs S3 and S4 for more information.

demonstrates the necessity of processing all cells from one sample integratively as shown in ScRAT.

5.3 Biological interpretation of cell attention

The accuracy of predicting phenotypes using ScRAT relies heavily on high-attention cells, indicating a strong connection between these cells and the critical cell types that drive clinical phenotypes. ScRAT calculates attention weights for each cell without assuming an annotation of cells with cell types. Therefore, it allows biologists to analyze any given group of cells using different clustering and annotation methods and infer the relevance of any given cell population with respect to the phenotype using attention weights. In addition, the use of ScRAT may identify cell types whose involvement in the pathogenesis of the disease was previously under-appreciated. This will help biologists to generate new hypothesis and further examine the roles of previously under-appreciated cell types in preclinical models and clinical specimens.

We first define the relevance of a cell with respect to the phenotype prediction as follows. Given a trained model with H attention heads and an input sample S_j with N cells, ScRAT generates one attention matrix per attention head. For a cell c_i in S_j, its *High-attention Occurrence Value* (HOV) is defined as the total number of times its attention weight ranked top k in a row across all rows and all H attention matrices, or

$$HOV_i^{S_j} = \sum_{h=1}^{H}\sum_{n=1}^{N} \mathbb{I}(a_{hni}^{S_j} \geq k_{hn}^{S_j}), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $a_{hni}^{S_j}$ is the attention weight at the n-th row and i-th column of the h-th head's attention matrix, and $k_{hn}^{S_j}$ denotes the k-th highest attention weight in the same row of the same attention matrix.

Once we have cell annotations for all cells in S_j, we extend the cell-level HOV to derive the *Relevance score* (R-score) for any given cell type $\mathcal{T}$ with respect to sample S_j by adding together the HOVs of all the cells in $\mathcal{T} \cap S_j$, and normalize it:

$$\text{R-scores}_{\mathcal{T}}^{S_j} = \frac{\sum_{i=1}^{N} \mathbb{I}(c_i \in C_{\mathcal{T}}^{S_j}) HOV_i^{S_j}}{|C_{\mathcal{T}}^{S_j}|}, \tag{7}$$

where $C_{\mathcal{T}}^{S_j}$ is the set of all the cells of cell type $\mathcal{T}$ in S_j.

For every phenotype, we then average the R-scores of the same cell type across all samples of this phenotype. The top k' cell types with the highest averaged R-scores are then selected as the critical cell types of this phenotype.

Here, we use the Haniffa dataset, the most comprehensively analyzed one among the three datasets used in the experiments, to demonstrate the clinical relevance of high-attention cells and critical cell types reported by ScRAT. The top 10 critical cell types (among 51 cell groups defined in the original paper) ranked by R-score are p_DC, RBC, Plasmablast, Platelets, HSC_CD38pos, Plasma_cell_IgG, CD83_CD14_mono, CD14_mono, Plasma_cell_IgA, and DC3 as shown in Table 1. Assuming critical cell types would better separate patients of different phenotypes, we test how well a phenotype can be predicted using only cells from that cell type and a simple feed-forward network. The AUC reported in Table 1 corresponds to a 50% training ratio. Most of our critical cell types have AUC >0.85. We repeat the experiment for all 51 cell types and the AUC of critical cell types selected by ScRAT are among top 10 AUC except for RBC and pDC, which demonstrates the relevance of cell types with a high R-score for phenotype prediction.

Next we compare these critical cell types to the corresponding analysis in the original paper (Stephenson *et al.* 2021), and discover that their major findings are related to the critical cell types listed in Table 1, as discussed in the following. (i) "Humoral immune response": ScRAT detected multiple subtypes of Plasma cells as critical, i.e. Plasmablasts, Plasma_cell_IgG, and Plasma_cell_IgA, which are the key effectors of the humoral immunity that produce antibodies. Consistent with our finding, the authors of the original paper also reported a larger population of Plasmablasts, Plasma_cell_IgA, and Plasma_cell_IgG in COVID-19 patients with severe symptoms. Notably, one characteristic of the humoral response against severe acute respiratory syndrome coronavirus 2 is the short-lived

Table 1. Top 10 critical cell types with their R-score and AUC of phenotype classification when using only the corresponding critical cell type.[a]

Critical cell types	R-score	AUC
pDC	2.58	0.58
RBC	2.20	0.68
Plasmablast	2.13	0.94
Platelets	2.13	0.86
HSC_CD38pos	1.69	1.00
Plasma_cell_IgG	1.62	0.95
CD83_CD14_mono	1.56	0.89
CD14_mono	1.50	0.94
Plasma_cell_IgA	1.24	0.92
DC3	1.18	0.77

[a] We rank these critical cell types based on their R-score of COVID phenotype. The larger R-score indicates the higher relevance to the phenotype. We use the Feedforward (single) model to predict the phenotype using only cells from single-cell type. The AUC is based on 50% training ratio, using half of patients as the training data and the other half of patients for testing. Most critical cell types selected by ScRAT also achieve high AUC except for RBC and pDC. A cell type of higher AUC is more discriminative in predicting different phenotypes, and hence more likely a real critical cell type. The concordance between cell types with high R-scores and high AUCs shows that high-attention cells detected by ScRAT are phenotype-specific. See Supplementary Table S7 for detailed information.

neutralizing antibodies, both IgG and IgA, manifested in the different humoral responses during COVID-19 infection and of other inflammatory conditions (Nguyen *et al.* 2022). More than 21% of cells from these three subtypes have high-attention weight, suggesting that ScRAT can detect the significance of humoral immune response during COVID-19 infection. (ii) "Impacts of monocytes": ScRAT also identified CD14_mono as critical cell types. The data in Stephenson *et al.* (2021) implies that CD14+ monocytes preferentially replenish the bronchoalveolar macrophages in health, while a much smaller and specific subset of monocytes, namely the C1QA/B/C+/CD16+ monocytes, replenish the bronchoalveolar macrophages of the COVID-19 patients. The latter denoted as C1_C16_mono ranks 13^{th} based on the R-score, and its population expansion is also more often observed in patients admitted to intensive care units (ICUs). The differential behaviors of these monocytes constitute a distinguishing feature between COVID-19 and non-COVID-19 patients. (iii) "Monocytes and platelet aggregates": Pathological monocyte-platelet interactions have been associated with aberrant coagulation and thrombosis formation in COVID-19 patients (Levi *et al.* 2020, Hottz *et al.* 2020). Since such interaction requires receptor–ligand interactions, the original authors suggested several receptor-ligand pairs between monocytes and platelets that may contribute to the aberrant interactions in COVID patients. This finding supports the selection of more than 40% of platelet cells as critical by ScRAT. (iv) "Hematopoietic stem cells": HSC_CD38pos are early hematopoietic progenitors and are rarely observed in peripheral blood mononuclear cells (PBMC) samples. The authors hypothesized that their presence in the PBMC samples of COVID-19 patients reflected perturbations of the bone marrow homeostasis during COVID-19 infection. Since HSC_CD38pos only constitutes less than 0.27% of cells in the dataset, this demonstrates that ScRAT can detect important phenotype-specific cell types of very small size.

We also want to highlight the detection of DC3 at the bottom of Table 1. This newly identified dendritic cell type has been shown to promote inflammatory functions of CD4+ and CD8+ T cells (Villar and Segura 2020), but their specific functions are yet to be deciphered. There are recent reports about their association with COVID, including an increased of CD163+ CD14+ cells within the DC3 cell type in COVID patients with severe symptoms (Winheim *et al.* 2021). Although the exact roles of these DC3 cells in COVID infection are yet to be uncovered, their high R-scores show the ability of ScRAT to detect cells of interest in a specific biological context.

We have used the manually annotated cell types assigned by the authors of the COVID datasets to validate the interpretation of scRAT predictions. The most relevant cell types among the high-attention cells, such as specific monocytes and platelets, are supported by the recent literature. The consistency between experts' analysis and critical cell types inferred using attention weights from ScRAT confirms the relevance of high-attention cells to phenotypes.

5.4 Prediction of multi-level and continuous traits

While the experiments in this work are binary classifications for a given patient, we would like to emphasize that ScRAT was not designed only for binary traits. Our implementation can handle multi-level discrete traits (e.g. more detailed severity level) and continuous traits (e.g. time to recover). We did not conduct such experiments simply because the number of patients in each multi-level trait from three public datasets was too small, not because of the limitation of scRAT. More descriptions of the corresponding formulations for multi-level and continuous traits can be found in Supplementary Section S3.

6 Conclusion

In this article, we introduce the problem of phenotype prediction using scRNA-seq data. We present ScRAT, an attention-based method that is designed to learn from limited samples without prior knowledge of marker genes or critical cell types, and provides accurate phenotype predictions. ScRAT consists of three module: Sample Mixup, Attention Layer, and Phenotype Classifier. Sample Mixup increases the size of training data to avoid overfitting. The Attention Layer models interactions between cells without any given cell-type annotations and provides a way to extract critical cells important in phenotype predictions. The Phenotype Classifier takes the latent representation of the input data produced by the attention layer and predicts the phenotype. We perform experiments on four tasks from three benchmarks and demonstrate that ScRAT consistently outperforms five baselines. We also show the biological meaningfulness of the cell types which ScRAT determines to be critical for phenotype prediction, through an analysis of the papers of the consortia which create the benchmarks and of several more recent studies. These findings suggest that ScRAT has the potential to discover phenotypic-driver cell types that suggest novel molecular mechanisms and/or targeted therapies.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the NSERC Discovery Grant "Transfer learning and causal models for biomedical data" and by grants from the Canadian Institutes of Health Research (PJT-175238) and Cancer Research Society (840847).

References

Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. 2014 Sep 1.

Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. 2020 Apr 10.

Brbić M, Zitnik M, Wang S *et al.* MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat Methods* 2020; **17**:1200–6.

Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* 2015;**163**:1011–25.

Carratino L, Cissé M, Jenatton R, Vert JP. On mixup regularization. *J Mach Learn Res* 2022;**23**:14632–62.

Chefer H, Gur S, Wolf L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2021 (pp. 397–406).

Chefer H, Gur S, Wolf L. Transformer Interpretability Beyond Attention Visualization. In: *2021 IEEE. InCVF Conference on Computer Vision and Pattern Recognition (CVPR)(2020)* 2020 (pp. 782–791).

Chen F, Wendl MC, Wyczalkowski MA *et al.* Moving pan-cancer studies from basic research toward the clinic. *Nat Cancer* 2021; **2**:879–90.

Ching T, Himmelstein DS, Beaulieu-Jones BK et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;**15**:20170387.

Clark K, Khandelwal U, Levy O, Manning CD. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*. 2019 Jun 11.

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* 2022;**185**:916–38.e58.

Cullen J, Rosner IL, Brand TC *et al.* A biopsy-based 17-gene genomic prostate score predicts recurrence after radical prostatectomy and adverse surgical pathology in a racially diverse population of men with clinically low- and intermediate-risk prostate cancer. *Eur Urol* 2015;**68**:123–31.

Cuzick J, Berney DM, Fisher G *et al.*; Transatlantic Prostate Group. Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. *Br J Cancer* 2012;**106**:1095–9.

Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 2018 Oct 11.

Dosovitskiy A, Beyer L, Kolesnikov A et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 2020 Oct 22.

Erho N, Crisan A, Vergara IA *et al.* Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* 2013;**8**:e66855.

Fan J, Lyu Y, Zhang Q et al. MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief Bioinform* 2022; **23**:bbac430.

Halko N, Martinsson PG, Tropp JA *et al.* Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev* 2011;**53**:217–88.

He B, Thomson M, Subramaniam M et al. Cloudpred: Predicting patient phenotypes from single-cell rna-seq. In: *Pacific Symposium On Biocomputing2022*. 2021 (pp. 337–348).

Hottz ED, Azevedo-Quintanilha IG, Palhinha L *et al.* Platelet activation and platelet-monocyte aggregate formation trigger tissue factor expression in patients with severe COVID-19. *Blood* 2020; **136**:1330–41.

Krizhevsky A, Sutskever I, Hinton GE *et al.* ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; **60**:84–90.

Levi M, Thachil J, Iba T *et al.* Coagulation abnormalities and thrombosis in patients with COVID-19. *Lancet Haematol* 2020; **7**:e438–40.

Liu L, Liu J, Han J. Multi-head or single-head? an empirical comparison for transformer training. *arXiv preprint arXiv:2106.09650*. 2021 Jun 17.

Lonsdale J, Thomas J, Salvatore M *et al.* The genotype-tissue expression (GTEx) project. *Nat Genet* 2013;**45**:580–5.

Lopez R, Regier J, Cole MB *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.

Ma W, Sharma S, Jin P *et al.* LRcell: detecting the source of differential expression at the sub–cell-type level from bulk RNA-seq data. *Brief Bioinform* 2022;**23**:bbac063.

Morley TJ, Han L, Castro VM *et al.* Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat Med* 2021;**27**:1097–104.

Newman AM, Steen CB, Liu CL *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**:773–82.

Nguyen DC, Lamothe PA, Woodruff MC *et al.* COVID-19 and plasma cells: is there long-lived protection? *Immunol Rev* 2022;**309**:40–63.

Penney KL, Sinnott JA, Fall K *et al.* mRNA expression signature of gleason grade predicts lethal prostate cancer. *J Clin Oncol* 2011; **29**:2391–6.

Perou CM, Sørlie T, Eisen MB *et al.* Molecular portraits of human breast tumours. *Nature* 2000;**406**:747–52.

Ravindra N, Sehanobish A, Pappalardo JL et al. Disease state prediction from single-cell data using graph attention networks. In: *Proceedings of the ACM conference on health, inference, and learning* 2020 Apr 2 (pp. 121–130).

Ren X, Wen W, Fan X *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 2021;**184**: 1895–913.e19.

Stephenson E, Reynolds G, Botting RA *et al.*; Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* 2021;**27**:904–16.

Tay Y, Dehghani M, Bahri D et al. Efficient transformers: A survey. *arXiv preprint cs.LG/2009.06732*. 2020.

Uhlen M, Zhang C, Lee S *et al.* A pathology atlas of the human cancer transcriptome. *Science* 2017;**357**:eaan2507.

Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. *Adv Neural Inf*. 2017;**30**.

Verma V, Lamb A, Beckham C et al. *Manifold mixup: learning better representations by interpolating hidden states*. 2019.

Villar J, Segura E. The more, the merrier: DC3s join the human dendritic cell family. *Immunity* 2020;**53**:233–5.

Winheim E, Rinke L, Lutz K *et al.* Impaired function and delayed regeneration of dendritic cells in COVID-19. *PLoS Pathog* 2021; **17**:e1009742.

Yin Q, Wang Y, Guan J *et al.* sciae: an integrative autoencoder-based ensemble classification framework for single-cell RNA-seq data. *Brief Bioinform* 2022;**23**:bbab508.

Zeng F, Kong X, Yang F et al. scPheno: A deep generative model to integrate scRNA-seq with disease phenotypes and its application on prediction of COVID-19 pneumonia and severe assessment. *bioRxiv. 2022* 2022–06.

Zhang H, Cisse M, Dauphin YN et al. mixup: beyond empirical risk minimization. *arXiv preprintarXiv:1710.09412*. 2017 Oct 25.

Zhou H, Zhang S, Peng J et al. Informer: beyond efficient transformer for long sequence time-series forecasting. In: *Proceedings of the AAAI conference on artificial intelligence* 2021 May 18.