



Multimodal Single-Cell Translation and Alignment with Semi-Supervised Learning

RAN ZHANG,¹ LAETITIA MENG-PAPAXANTHOS,²
JEAN-PHILIPPE VERT,³ and WILLIAM STAFFORD NOBLE^{1,4}

ABSTRACT

Single-cell multi-omics technologies enable comprehensive interrogation of cellular regulation, yet most single-cell assays measure only one type of activity—such as transcription, chromatin accessibility, DNA methylation, or 3D chromatin architecture—for each cell. To enable a multimodal view for individual cells, we propose **Polarbear**, a semi-supervised machine learning framework that facilitates missing modality profile prediction and single-cell cross-modality alignment. Polarbear learns to translate between modalities by using data from co-assay measurements coupled with the large quantity of single-assay data available in public databases. This semi-supervised scheme mitigates issues related to **low cell quantities** and **high sparsity** in co-assay data. Polarbear first pre-trains a **beta-variational autoencoder** for each modality using both co-assay and single-assay profiles to learn robust representations of individual cells, and it then uses the co-assay labels to train a translator between these cell representations. This semi-supervised framework enables us to predict missing modality profiles and match single cells across modalities with improved accuracy compared with fully supervised methods, thus facilitating multimodal data integration.

Keywords: cross-modality translation and multi-omics alignment, single cell multi-omics.

1. INTRODUCTION

SINGLE CELL OMICS, including epigenomics, transcriptomics, proteomics, etc., are extremely valuable for studying cell-to-cell variation, because each type of assay provides a unique perspective on cellular regulation. However, in most cases, each type of measurement is carried out on different sets of cells, with only one type of activity captured for any single cell.

¹Department of Genome Sciences, University of Washington, Seattle, Washington, USA.

²Google Research, Brain Team, Zurich, Switzerland.

³Google Research Brain Team, Paris, France.

⁴Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington, USA.

An early version of this paper was published as part of the 2022 Annual International Conference on Research in Computational Molecular Biology (RECOMB) and was deposited to bioRxiv.

The recent emergence of single-cell co-assays, in which multiple types of measurements are conducted on the same cell, enables us to directly measure multiple forms of molecular activities within each cell. However, co-assay measurements are often lower throughput than standard single-cell measurements, and they are in general more challenging to produce than single-assay data.

Multiple machine-learning methods have been proposed to translate between different single-cell omics measurements by using co-assay data (Ashuach et al, 2021; Gayoso et al, 2021; Hao et al, 2021; Lotfollahi et al, 2022; Minoura et al, 2021; Wu et al, 2021). In most of these studies, a fully supervised model is learned based on the cross-modality cell matching provided by the co-assay data. Thus, the performance of these models is necessarily limited by the sparsity and limited amount of co-assay data.

Meanwhile, public databases house orders of magnitude more single-cell data. We hypothesized that training a translation model using both co-assay data (labeled data) and single-assay data (unlabeled data) from independent studies will improve cross-modality translation performance compared with using only co-assay data.

Accordingly, we propose a semi-supervised framework, Polarbear, that learns to translate between single-cell measurements by employing both co-assay and single-assay data. In this article, we focus on translation between scRNA-seq and scATAC-seq; that is, given a scRNA-seq profile of a cell, Polarbear will generate the scATAC-seq profile of the same cell, and vice-versa. Polarbear can be applied to several types of co-assays that measure gene expression and chromatin accessibility within single cells, including CAR-seq (Cao et al, 2018), SNARE-seq (Chen et al, 2019), Paired-seq (Zhu et al, 2019), and SHARE-seq (Ma et al, 2020).

Polarbear operates in two main stages (Fig. 1). In the first stage, we train two deep beta variational autoencoder (beta-VAE) neural networks that learn, in an unsupervised fashion, to reduce each given type of data to a latent representation (the encoder) and then expand that representation to recover the original data (the decoder). (V)AEs have already been successfully applied to scRNA-seq and scATAC-seq data, primarily for the purpose of de-noising (Ashuach et al, 2022; Eraslan et al, 2019; Lopez et al, 2018; Talwar et al, 2018; Trong et al, 2020; Wang and Gu, 2018; Xiong et al, 2019). Here, we train one beta-VAE for each type of data and learn latent cell representations that are independent of sequencing depth and batch effect (Ashuach et al, 2022; Lopez et al, 2018).

In stage two, we stitch together the encoder for one data type with the decoder of a second data type, interposing between them a single, fully connected “translator” layer. During this phase, the parameters of the encoder and decoder are frozen, and the translator parameters are trained in a supervised fashion using co-assay data. Repeating this procedure in reverse, Polarbear allows for bidirectional translation between scRNA-seq and scATAC-seq data. In principle, our method can also be applied to co-assays operating on other data modalities, since there is no requirement for feature correspondence between different types of measurements.

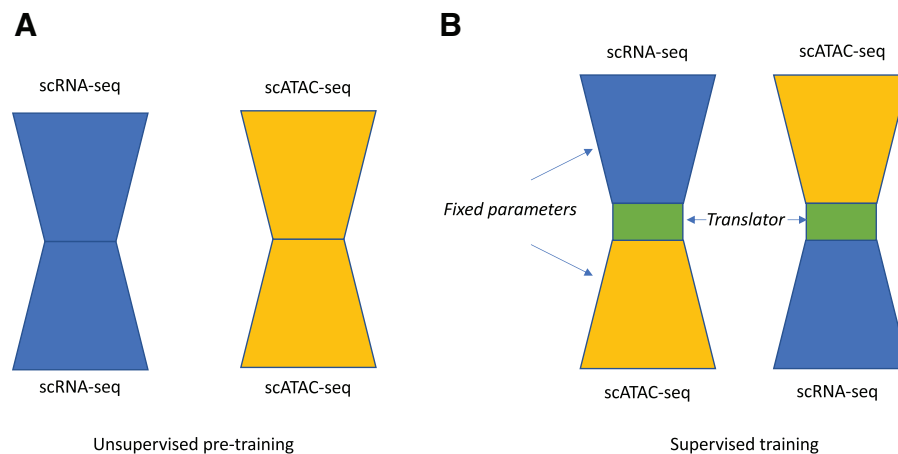


FIG. 1. Polarbear’s semi-supervised framework. (A) In stage 1, Polarbear trains an autoencoder for each data modality, using both single-assay and co-assay data. (B) In stage 2, the encoder from one modality is stitched together with a decoder from the other modality (and vice versa), and the translation layers are trained in a supervised fashion using co-assay data.

To evaluate the performance of Polarbear, we propose a set of evaluation metrics for single-cell translation and alignment tasks, with the aim of teasing out the individual-cell level differences. A drawback of current methods lies in the choice of evaluation metrics used for cross-modality profile prediction. Many previous methods report the correlation (for scRNA-seq) or area under the receiver operating characteristic curve (AUROC, for scATAC-seq) between the overall observed and predicted profiles. However, these performance measures can be strongly driven by the average profiles across cells, failing to reflect whether the prediction method accurately captures cell-to-cell variation. Although MultiVI (Ashuach et al, 2021) systematically demonstrates that the proposed method can predict differential expression between cell clusters or cell types, it does not address whether the model accurately captures differences among single cells.

Using an extensive set of performance measures, we demonstrate that Polarbear’s translation performance improves when we add single-assay data to the training procedure in the first phase. We also show that Polarbear outperforms BABEL (Wu et al, 2021), a state-of-the-art translation method, using several different performance measures. Finally, we demonstrate that Polarbear can be used to accurately match cells between modalities. Overall, our work illustrates the utility of exploiting single-assay data to aid in the prediction and alignment of cross-modality profiles.

1.1. RELATED WORK

Several previous methods have been developed for single-cell multi-omics cross-modality prediction (Table 1). **TotalVI** builds a VAE that takes as input the concatenation of thousands of gene and hundreds of protein expression profiles from the CITE-seq co-assay. The autoencoder then learns to impute missing protein expression profiles based on scRNA-seq profiles (Gayoso et al, 2021). **BABEL** joins two autoencoders, one from each data domain, to translate between single-cell modalities; the model forces the corresponding cell embeddings to be shared between the autoencoders and optimizes the autoencoder model and cross-modality translation simultaneously (Wu et al, 2021).

The **scMM** method joins two VAEs with a mixture-of-experts multimodal deep generative model, aiming at learning a joint embedding between modalities and at predicting missing modalities (Minoura et al, 2021). **Seurat** computes nearest neighbor graphs in each data modality and predicts the missing domain profile of a cell by identifying neighboring cells in co-assay data and then computing the average profile of those neighbor cells in the second modality (Hao et al, 2021). **Multigrade** jointly embeds data from two or more modalities and uses the joint embedding to infer profiles in each domain (Lotfollahi et al, 2022).

MultiVI also embeds scRNA-seq and scATAC-seq profiles into a shared space by joining two VAEs. The model is able to take single-assay data as input and predict missing modality profiles (Ashuach et al, 2021).

Besides single-cell cross-modality translation, a similar approach has been used in a related area. Gala et al (2019) build a **k-coupled autoencoder** on partially paired scRNA-seq and electrophysiological profiles of neurons. They then impute latent representations and missing modality profiles for cells only measured in one modality.

TABLE 1. METHOD COMPARISON

	<i>totalVI</i>	<i>BABEL</i>	<i>scMM</i>	<i>Seurat</i>	<i>Multigrade</i>	<i>MultiVI</i>	<i>Polarbear</i>
scRNA → scATAC		✓	✓			✓	✓
scATAC → scRNA		✓	✓			✓	✓
Batch correction	✓		✓	✓	✓	✓	✓
Uses single-assay data						✓	✓
Neural network model	✓	✓	✓		✓	✓	✓
Uses a joint or shared embedding	✓	✓	✓		✓	✓	
Translates between embeddings							✓
Peak-wise evaluation						✓ ^a	✓
Gene-wise evaluation						✓ ^a	✓
Cluster matching evaluation			✓	✓	✓	✓	
Cell matching evaluation						✓	✓

^aEvaluation is done across cell types per-peak or per-gene, rather than across individual cells.

Polarbear improves upon previous models in several ways. First, previous methods tend to optimize the model on the multimodal data jointly, by optimizing different tasks of the model (e.g., autoencoder model within each modality, cross-modality translation, and cell embedding matching) at the same time. This makes the model performance vulnerable to the choice of hyperparameters. In contrast, Polarbear takes a **stepwise optimization** approach. It first learns embeddings with both single and co-assay data by separately optimizing each autoencoder, and it then learns to translate between embeddings across modalities based on co-assays by minimizing translation loss. Thus, Polarbear is less likely to be biased toward optimization of a specific task and requires less hyperparameter tuning.

Second, previous models generate predictions based on a joint or shared embedding of both modalities. Polarbear does not require a shared or joint embedding; instead, it adds a **translation layer** between the embeddings across modalities based on co-assay data. Thus, Polarbear is more flexible at leveraging single-assay data and incorporating pre-trained models from new data modalities.

More importantly, most previous methods learn the translation model using only co-assay data, which are available only in limited quantities. Although MultiVI can learn from **single-assay data**, the question of whether adding single-assay data from unrelated datasets improves translation performance has not been addressed. Polarbear is able to use single-assay data collected from public datasets to improve its translation performance.

In this study, we choose to compare our method with BABEL, for several reasons. First, BABEL does not require subsetting features and can be trained on the original gene and peak space. This is achieved by only allowing for within-chromosome connections in certain layers of the AE model. In contrast, most models require significant computational resources and consequently operate on a subset of features for the sake of memory and run time. Indeed, multiVI, which was published on bioRxiv very recently, ran out of memory when trained on the SNARE-seq data used in our study.

Second, BABEL directly addresses the task of translating between scRNA-seq and scATAC-seq, and it has been applied to the SNARE-seq co-assay data we use in this study, so it is most likely that we can make a fair comparison with BABEL.

Most importantly, since the focus and novelty of Polarbear is the semi-supervised framework that leverages single-assay data from unrelated studies, instead of comparing extensively with current methods that are not specifically designed for this task, we demonstrate the power of our semi-supervised framework leveraging single-assay data (“Polarbear”) by comparing it with a Polarbear model that is only trained on co-assay data (“Polarbear co-assay”). We envision that the semi-supervised training framework proposed by Polarbear may be adapted to other existing architectures to boost their performance.

2. METHODS

2.1. Polarbear model

Polarbear first constructs separate beta-VAEs for the scRNA-seq and scATAC-seq domains. Compared with traditional VAEs, we **correct for batch effects and the bias of sequencing depth across cells**, adapting ideas from scVI and peakVI (Ashuach et al, 2022; Lopez et al, 2018) (Fig. 2). Specifically, Polarbear learns the latent cell representations independent of the bias of sequencing depth and batch by explicitly taking those factors into account during model training. For batch correction, we one-hot encode **batch factors (b)** and concatenate them to the input and embedding layers. The sequencing-depth factor is calculated as the sum of counts per cell, which is reintroduced to calculate the reconstruction losses of the VAEs jointly with the sequencing-depth corrected estimations.

The scRNA-seq VAE takes **gene expression raw counts** as input (x) and encodes each cellular profile in the embedding layer (z_x). The scRNA-seq VAE has **two hidden layers** in the encoder the decoder. The dimension of the bottleneck layer is chosen as a hyperparameter, and the dimension of each hidden layer is calculated as half of the geometric mean between the input gene space and the latent dimensions. The inference step assumes that the gene expression counts follow **zero-inflated negative binomial (ZINB) distributions**, and it estimates the mean, variance, and dropout likelihood based on the decoder (Lopez et al, 2018).

For each gene, we assume that the variance term is shared across cells within each batch. The **sequencing-depth term (d_x)**, together with the three output predictions, are used to fit a ZINB distribution. We use the maximum log likelihood of the ZINB distribution as the loss function, and “**expected frequency**” as the normalized scRNA prediction.

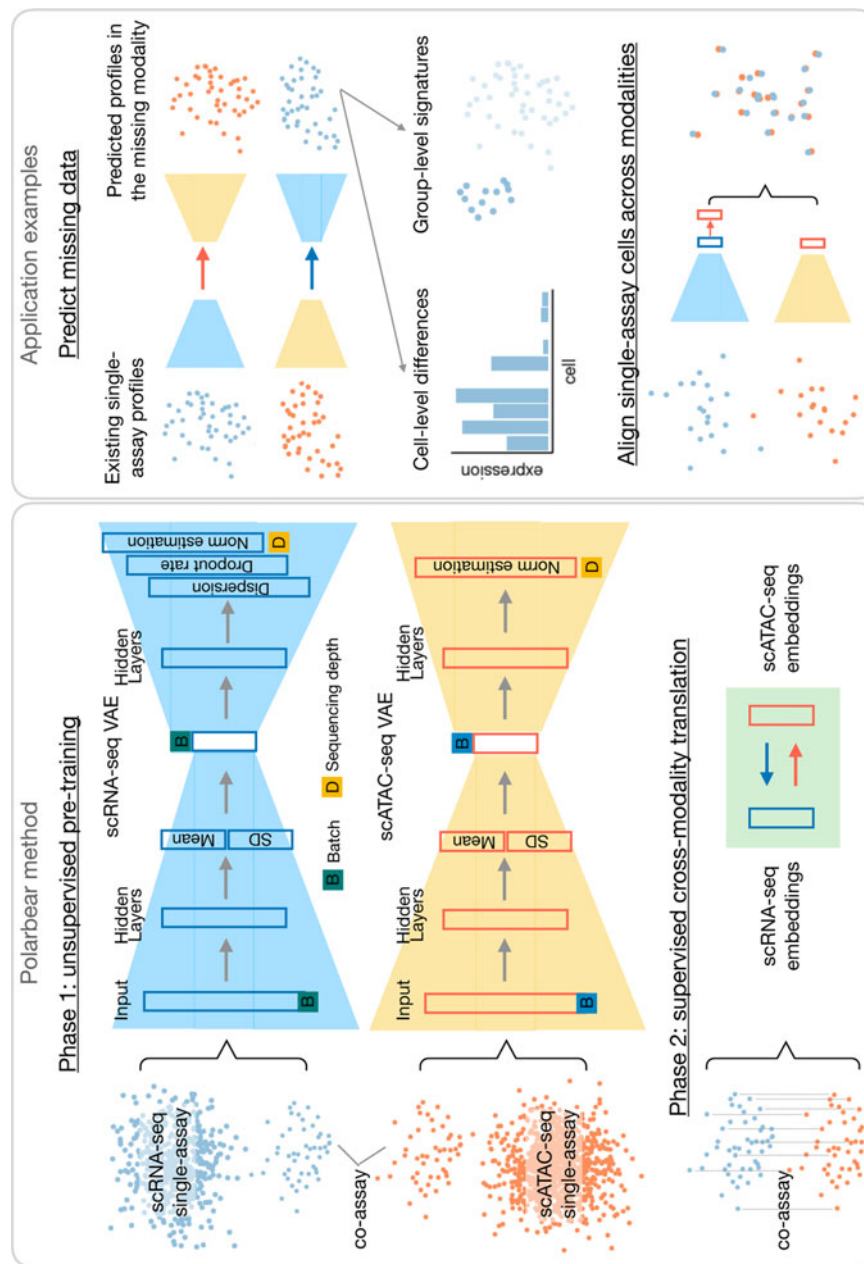


FIG. 2. Polarbear's semi-supervised framework and applications. Left: Polarbear's semi-supervised framework. In Phase 1, Polarbear trains a VAE for each data modality, using both single-assay and co-assay data. In Phase 2, the encoder from one modality is stitched together with a decoder from the other modality (and vice versa), and the translation layers are trained in a supervised fashion using co-assay data. Specifically, the VAEs take into account sequencing depth (D) and batch effects (B). The scRNA-seq VAE assumes that counts are drawn from a zero-inflated negative binomial distribution, and the scATAC-seq VAE assumes a Bernoulli distribution. Right: Sampled applications of Polarbear. Polarbear can predict missing domain profiles based on the known domain, capturing individual cell-level differences and group-level signatures in the missing domain. Further, Polarbear can match single-cell profiles across modalities.

The scATAC-seq autoencoder takes in binarized peak counts (y). It has **three hidden layers** in both the encoder and decoder. It projects the scATAC-seq profiles to the embedding layer with the same number of latent dimensions as the scRNA-seq VAE (with latent representation as z_y), and the dimension of each hidden layer is defined as half the **geometric mean** between the input peak dimension and the latent dimensions.

We assume that each peak follows a **Bernoulli distribution** and learn a sequencing-depth normalized probability for each peak as the output of the decoder. This probability is combined with the sequencing depth (d_y) to maximize for the **binary cross-entropy of the estimation**. To be able to incorporate the large number of peaks and save memory in the training process, we only allow for within-chromosome connections in the first two encoder layers and the last two decoder layers.

After the autoencoders in both domains are optimized, we learn **a single linear translation layer** between the embedding layers of the scRNA-seq and scATAC-seq, supervised by co-assay data, and minimize the translation loss on each modality.

In both translation directions, since the distributional variables are independent of sequencing depth, the size-normalized expectations of the distributions (i.e., the “**norm estimation**” in Fig. 2) can be directly used for subsequent tasks such as differential expression analysis.

More specifically, the model is optimized in four separate steps.

- First, we optimize the scATAC-seq VAE [encoder $q_y(z_y|y, b_y, d_y)$ and decoder $p_y(y|z_y, b_y, d_y)$], with the loss as a weighted sum of the reconstruction loss and Kullback–Leibler (KL) divergence term. The

$$\text{loss}_{\text{ATACreconst}} = -E_{z_y \sim q_y(z_y|y, b_y, d_y)} \log p_y(y|z_y, b_y, d_y) + \beta D_{\text{KL}}(q_y(z_y|y, b_y, d_y) \parallel p_y(z_y))$$

- Second, we optimize the scRNA-seq VAE [encoder $q_x(z_x|x, b_x, d_x)$ and decoder $p_x(x|z_x, b_x, d_x)$] by minimizing $\text{loss}_{\text{RNAreconst}}$.

$$\text{loss}_{\text{RNAreconst}} = -E_{z_x \sim q_x(z_x|x, b_x, d_x)} \log p_x(x|z_x, b_x, d_x) + \beta D_{\text{KL}}(q_x(z_x|x, b_x, d_x) \parallel p_x(z_x))$$

- Third, we freeze the VAEs and optimize the linear translation layer from scATAC-seq to scRNA-seq embeddings, supervised by co-assay labels. In this step, we first project the scATAC-seq profiles (y) to the embedding space (z_y) using the trained scATAC-seq encoder, then translate it to the scRNA-seq embedding space with t_{yx} , and finally project the translated embeddings to the original scRNA-seq space using the scRNA-seq decoder. In this step, the negative log likelihood of the estimated scRNA-seq profile (x) from the input (y) is minimized.

$$\text{loss}_{\text{RNAtranslat}} = -E_{z_y \sim q_y(z_y|y, b_y, d_y)} \log p_x(x|t_{yx}(z_y), b_x, d_x)$$

- Fourth, we optimize the linear translation layer from scRNA-seq to scATAC-seq embeddings, supervised by co-assay labels. Similarly, the translation loss of estimating the scATAC-seq profile (y) based on scRNA-seq input (x) is minimized to learn an optimized t_{xy} transformation.

$$\text{loss}_{\text{ATACtranslat}} = -E_{z_x \sim q_x(z_x|x, b_x, d_x)} \log p_y(y|t_{xy}(z_x), b_y, d_y)$$

2.2. Hyperparameter tuning

The Polarbear neural network architecture has two primary hyperparameters: the latent dimensions of the autoencoders, and the weight of the KL divergence term in each VAE: β . In this study, we use the validation set to choose hyperparameters, selecting the number of latent dimensions ($\{10, 25, 50\}$) and β ($\{0, 0.5, 1, 2\}$). In the random test set setup, we randomly split the SNARE-seq dataset, assigning 60% of cells to the training set, 20% to the validation set, and 20% to the test set. In the unseen cell type scenario, we use the same validation and test set as BABEL, where the validation and test set are the largest two cell clusters based on the SNARE-seq scRNA-seq dataset. The rest of cells are used as the training set.

We downloaded BABEL’s scripts and followed the instructions to generate predictions on SNARE-seq (Wu et al, 2021). We verified that we are able to reproduce the performance reported in the article. **In all scenarios, we make sure that BABEL’s train/validation/test splits are the same as Polarbear’s**. BABEL has proposed a set of default parameters (latent dimension: 16, weight factor: 1.33); however, for a fair

comparison we tune the following 2D grid of hyperparameters: number of latent dimensions in {10, 16, 25, 50} and weight factor to balance scATAC loss in {0.2, 1.33, 5}. We then select BABEL's best performing model based on each task's performance on the validation set.

2.3. Performance measures of cross-modality translation

In designing performance measures for cross-modality translation, we tried to place ourselves in the shoes of a prospective end user of our predictive model. Imagine a scenario in which we are interested in leveraging an existing scRNA-seq dataset to predict chromatin accessibility in a particular biological system, applying our trained model to the scRNA-seq matrix to yield a predicted matrix of scATAC values. Given the predicted peak activations, we can imagine trying to solve two different problems.

In the first setting, we begin by identifying cell types using the original scRNA-seq data or identifying cell groups based on the experimental design (e.g., disease and control groups). We may then be interested in the pattern of predicted chromatin peak activations within each cell type or group. In this setting, a classifier-based measure such as the area under the precision-recall (AUPR) or area under the receiver operating characteristic (AUROC), computed separately for each peak, would accurately capture the per-peak predictive behavior across single cells and thus be indicative of per-peak predictive power across clusters or groups.

Each of these two measures has advantages. The AUPR emphasizes enriching the top of the ranked list of predictions with positives. On the other hand, AUROC explicitly corrects for differences in “skew” (i.e., differences in the number of non-zero values) for each peak. To correct for the skew in AUPR measurement toward peaks with large positive proportions, we calculate AUPRnorm as follows, where AUPRnorm of 0 represents the behavior of a random predictor and 1 indicates a perfect predictor.

$$\text{AUPRnorm} = \frac{\text{AUPR} - \text{PP}}{1 - \text{PP}}$$

where

$$\text{PP} = \frac{\text{\#cells with peak expressed}}{\text{\#cells}}$$

In the second setting, we can use the profile of predicted peak activations across each cell to match scRNA-seq profiles to corresponding scATAC-seq profiles. In this setting, we identify these matches based on Euclidean distance. We want to ensure that each predicted profile's nearest neighbor is the correct match; hence, we can use the **fraction of samples closer than the true match** (FOSCTTM) as a performance measure (Liu et al, 2019).

Based on these scenarios, we report here the average per-peak AUROC and AUPRnorm, as well as the FOSCTTM. Similarly, when predicting gene expression from scATAC-seq, we report the average per-gene Pearson correlation (on log-scaled expression) and the FOSCTTM. We do not foresee a scenario in which the overall “flattened” performance of the model, in which we treat all values in the matrix as a single list and compute a single score (Pearson correlation, mean-squared error (MSE), AUPR, or AUROC), will be of primary interest to an end user.

2.4. Correcting for the sequencing-depth bias in the evaluations

To assess the predictions with awareness of sequencing-depth bias, we further post-process the data only for evaluation purposes.

For scRNA-seq prediction, since the Pearson correlation coefficient for each gene across cells is used, we would like to correct for sequencing depth to calculate relative signal across cells. To do that, we use the “norm estimation” as predicted value, and we compare it with normalized scRNA-seq profiles (Lun et al., 2016).

For scATAC-seq prediction, since we are using the binarized scATAC-seq profiles as the true labels, the labels can be highly biased toward cells with high sequencing depth. To mitigate this effect, we generate an “unnormalized” scATAC-seq prediction so that we can evaluate on the true profiles and make a fair comparison to methods that do not take into account sequencing depth.

To estimate sequencing depth information for the missing scATAC-seq profile, we take advantage of the positive correlation between the scATAC-seq and scRNA-seq depth factors when both modalities are

observed, and we predict the scATAC-seq depth factor in the translation task based on the known scRNA-seq profile. Specifically, we use ridge regression for this prediction task, with a penalty term selected from $\{0.01, 0.1, 1, 10, 100, 1000, 10000, 100000\}$ determined by cross-validation within the training set. Finally, we multiply the learned sequencing depth with the normalized scATAC-seq predictions to generate unnormalized predictions in the test set, and we use that for the evaluation purpose.

2.5. Single-cell data pre-processing

For co-assay data, we use SNARE-seq data from mouse adult brains ($\sim 10k$ cells) (Chen et al, 2019). We filter out peaks that occur in fewer than 5 cells or $>10\%$ of cells, as in the original SNARE-seq paper. To learn robust representations of each domain, we collect publicly available scRNA-seq and scATAC-seq profiles from adult mouse brains (Fang et al, 2021; Li et al, 2021; Zeisel et al, 2018). The single-assay data contains $\sim 160k$ cells sequenced by scRNA-seq, and scATAC-seq profiles from $\sim 855k$ cells. We randomly downsampled the latter dataset to $\sim 170k$ cells for use in training (Table 2).

We then set up a mapping between the single-assay features and co-assay features. For scRNA profiles, we use SNARE-seq genes as a reference, map genes in the other datasets to the gene symbols in the SNARE-seq data, and **further filter out non-protein-coding genes based on Gencode annotations** (Harrow et al, 2006). We remove sex chromosome genes for consistency across datasets. In this way, 17,271 genes are maintained for input to Polarbear. For the scATAC profile, we first lift all peaks to the mm10 reference assembly (Hinrichs et al, 2006).

Because the ATAC-seq peak locations vary across datasets, we use the SNARE-seq peaks as a reference and map features from other datasets to SNARE-seq peaks if there is an overlap of 1 bp or more. In the end, $>92\%$ SNARE-seq peaks are successfully mapped to at least one other dataset, and around 82% SNARE-seq peaks are mapped across all data sets and technologies. Peaks from sex chromosomes are again filtered out. In the end, 220,526 peaks are input to the Polarbear model. Finally, to ensure the quality of the single-assay cells, we further filter out cells with fewer than 50 genes or peaks expressed in the mapped dataset.

To test the performance using a subset of features, we generated a filtered dataset in which genes and peaks expressed in $<1\%$ of SNARE-seq cells removed (“Polarbear-exp”). This filter produces a dataset with 9570 genes and 52,974 peaks. We also tested another way of feature subsetting, where we retain the top 25% most variable genes and peaks (identified through *scanpy*). This approach ends up with 4317 genes and 56,644 peaks (“Polarbear-var”).

2.6. Cell-type specific marker prediction and differential expression analysis

An important application of single-cell analysis is to cluster the cells according to the similarity of their scRNA-seq or scATAC-seq profiles; accordingly, a good translator should be able to produce predicted profiles that yield clusters and cluster-level signature genes similar to the ones produced by the observed data.

To predict cell-type specific markers in the missing modality based on measurements in the known domain, we calculate whether a gene or a peak is specifically expressed in a specific cell type compared with the rest of the population. For each cell type, we label its corresponding cells as positives and cells in other cell types as negatives, and we calculate the AUROC of the predicted gene/peak-wise profile relative to these labels. A high AUROC score suggests that the gene or peak is specifically expressed in the corresponding cell type and thus likely to be a cell type specific marker.

We use the cell types derived in the SNARE-seq study as ground truth, and we validate the marker prediction on an expert-curated list of marker genes for each cell cluster (Chen et al, 2019).

We can also calculate differential expression patterns based on the predicted profiles. To do that, we perform a one-sided Wilcoxon rank-sum test between the predicted expression pattern (“norm estimation”) in one cell type and that in all other cell types, and we control for the false discovery rate (FDR) using the

TABLE 2. DATA SETS

<i>Data set</i>	<i>Cells</i>	<i>Assay</i>	<i>Platform</i>
SNARE-seq	$\sim 10k$	Co-assay	Illumina HiSeq 2500/4000
Li <i>et al.</i>	$\sim 800k$	snATAC-seq	Illumina HiSeq 2500
Fang <i>et al.</i>	$\sim 55k$	scATAC-seq	Illumina HiSeq 2500
Zeisel <i>et al.</i>	$\sim 160k$	scRNA-seq	10x Genomics

Benjamini–Hochberg procedure. This yields a genome-wise list of differential expression p -value and FDRs. To validate differential expression predictions, we label differentially expressed genes derived from the normalized true profile ($\text{FDR} \leq 0.01$) as positive and other genes as negative, and we calculate a precision-recall curve using the predicted differential expression p -value.

3. RESULTS

3.1. Polarbear accurately translates between single-cell data domains

We begin by testing Polarbear’s ability to translate between scRNA and scATAC profiles in a SNARE-seq adult mouse brain co-assay dataset. To learn robust representations of each domain, we train the autoencoders with large-scale, publicly available scRNA and scATAC single-assay profiles, also derived from adult mouse brains (Section 2.5). We then train Polarbear’s translator layers in a supervised fashion using a training set of 80% of the cells from the SNARE-seq dataset, evaluating translation performance on the test set comprising the remaining 20%.

To evaluate the performance of our model, we measure how well the predicted scRNA-seq profiles based on scATAC-seq measurements allow us to **recapitulate gene expression differences across cells**. To do this, we calculate the gene-wise correlation between the predicted profile and the true normalized profile (Lun et al., 2016). Since we are mostly interested in capturing meaningful gene expression patterns across cell clusters, we specifically evaluate our prediction on the **differentially expressed genes** in each cell type identified in the SNARE-seq study (Chen et al, 2019).

In this analysis, Polarbear outperforms BABEL, yielding an improved gene-wise correlation for 1064 out of 1205 genes (Wilcoxon rank-sum test p -value 1.20×10^{-33} , Fig. 3A). We also observe that Polarbear strongly outperforms a Polarbear variant (“Polarbear co-assay”) that is trained only with co-assay data, where 1052 out of 1205 genes have improved correlation (Fig. 3B), demonstrating that the added single-assay data are important for improved translation performance.

We also evaluate the scATAC-seq predictions when only scRNA-seq measurements in those cells are available. We calculate the peak-wise AUROC and AUPRnorm of the predicted profile relative to the observed, binarized scATAC-seq profile. Similarly, we focus on the peaks that are known to be differentially expressed in SNARE-seq scATAC-seq profiles, identified based on the SNARE-seq study (Chen et al, 2019). Our analysis demonstrates that Polarbear outperforms both competing methods in recapitulating the cell-to-cell variation (Fig. 3C–F and Table 3).

3.2. Polarbear can recapitulate and predict cell-type-specific signatures

Using the translated profiles, Polarbear can be used to derive interesting biological insights. In one scenario, Polarbear can be used to infer cell type labels of cells utilizing the cell-type specific marker gene expression prediction, based on a population of scATAC-seq profiles and prior knowledge of cell type marker genes. To validate this, we use as a gold standard a predefined sets of cell-type marker genes that are annotated in the SNARE-seq study (Chen et al, 2019), and we ask whether Polarbear’s predictions on the marker genes allow us to assign these cells to the corresponding cell types.

Specifically, we calculate the AUROC for a signature gene’s predicted expression in a one-vs-all fashion for one corresponding cell type versus all others. We find that Polarbear is able to predict the gold standard cell type markers correctly with a median AUROC of 0.933. This performance is significantly better than both BABEL (median AUROC = 0.869; Fig. 4A) and the co-assay variant of Polarbear (median AUROC = 0.914; Fig. 4B). Interestingly, Polarbear is especially better at labeling **rare cell types** compared with Polarbear co-assay, suggesting that the rare cell types might benefit more from incorporating single-assay data. This offers a great opportunity for Polarbear to infer cell types based on scATAC-seq profiles.

Besides recapitulating the knowledge that can be derived from experimental profiles, Polarbear’s cross-modality prediction is able to reveal insights that even cannot be captured by experimentally generated profiles in the “missing” domain. Here, we focus on a relatively rare cell type, **microglia**, which consists of only 91 cells in the SNARE-seq dataset. Based on scATAC-seq profiles in the test set, we predict microglia-specific genes by calculating AUROC for each gene’s predicted expression in microglia cells against all other cells.

Based on Polarbear’s prediction, the *Sal11* gene is specifically expressed in microglia (AUROC = 0.800), but this gene is not highly expressed in microglia based on the observed scRNA-seq profiles (AUROC =

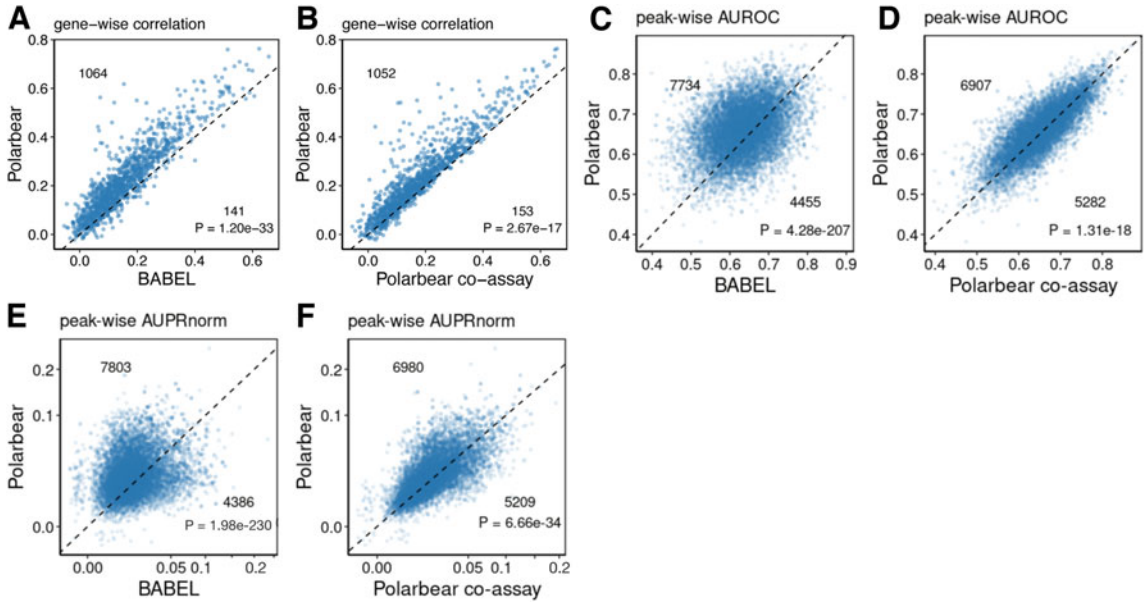


FIG. 3. Cross-modality prediction on the random test set. (A, B) Gene-wise correlation between the true and predicted profile, comparing Polarbear with BABEL (A) or with Polarbear co-assay (B), only showing genes that are differentially expressed across cell types. BABEL performance is reported based on the best performing model in each task after a hyperparameter grid search. Each dot is a gene, and numbers indicate the number of dots above and below the diagonal line. p -Values are calculated by a one-sided Wilcoxon rank-sum test. “Polarbear co-assay” only uses co-assay data to train the model. (C, D) Peak-wise AUROC, comparing Polarbear with BABEL (C) or with Polarbear co-assay (D). Each dot represents a peak, and only peaks differentially expressed across cell types are shown. (E, F) Peak-wise AUPRnorm, comparing Polarbear with BABEL (E) or with Polarbear co-assay (F). Each dot represents a peak, and only peaks differentially expressed across cell types are shown.

0.498). Interestingly, *Sal11* has been previously found to be a microglia signature gene, and it encodes the transcription factor, *Sal11*, that maintains microglia identity (Buttgereit et al, 2016).

3.3. Polarbear can predict inter- and intra-cell type variations in new cell types

Because the amount of available co-assay data is limited relative to single-assay data, a common challenge for models such as Polarbear is to translate between modalities in cell types for which no training data are available. The authors of the BABEL model simulated this scenario by creating a train/test split in which an entire SNARE-seq cell cluster is held out for testing. Accordingly, we also investigate this setting, using BABEL’s train/test split.

First, we investigate whether Polarbear predictions can capture variations within this unseen population. Because the scRNA-seq and scATAC-seq profiles within a cell type are expected to be relatively homogeneous, successfully translating across modalities in this scenario requires the model to capture

TABLE 3. TRANSLATION PERFORMANCE REPRESENTED BY MEAN AND STANDARD DEVIATION

Test set	Evaluation metric	BABEL	Polarbear co-assay	Polarbear
		Mean (SD)	Mean (SD)	Mean (SD)
Random 20%	Gene-wise correlation	0.155 (0.117)	0.173 (0.122)	0.219 (0.141)
	Peak-wise AUROC	0.636 (0.0872)	0.654 (0.0623)	0.661 (0.0617)
	Peak-wise AUPRnorm	0.0185 (0.0173)	0.0219 (0.0178)	0.0247 (0.0200)
	Gene-wise correlation	0.0712 (0.0504)	0.0713 (0.0763)	0.137 (0.0834)
Unseen cell type	Peak-wise AUROC	0.560 (0.0661)	0.567 (0.0625)	0.604 (0.0682)
	Peak-wise AUPRnorm	0.0120 (0.0143)	0.0127 (0.0128)	0.0207 (0.0183)

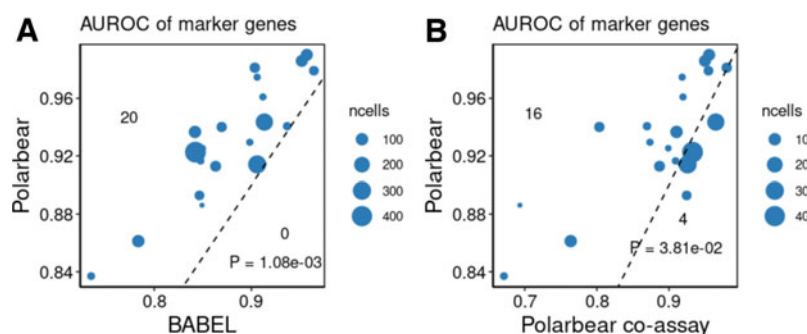


FIG. 4. Cell-type specific marker gene prediction. (A, B) In the random 20% test set, for each cell-type specific marker gene, we calculate the AUROC of its prediction to be higher in cells in corresponding cell type compared with unrelated ones. Each dot is a marker gene, with size indicating the number of positive labels (i.e., number of cells in the corresponding cell type). Polarbear is compared against BABEL (A) and Polarbear co-assay (B).

differences between individual cells, not just differences across cell types. We observe that Polarbear consistently outperforms BABEL and Polarbear co-assay in translation in both directions, suggesting that Polarbear predictions are able to recapitulate meaningful variations across cells within a cell type (Fig. 5A–F).

Besides capturing the within-cell type variations, a successful translation should also recapitulate the group-level signatures of the unseen cell type, so that the model can infer biological knowledge from new cell types. Accordingly, based on the predicted gene expression profile based on scATAC-seq in the unseen cell type, we calculate differentially gene expression pattern of the unseen cell type compared with the rest of the cells in the SNARE-seq population (Section 2.6).

To evaluate how well our prediction can recapitulate the true signatures, we define the true labels as genes significantly highly expressed in the unseen cell type based on the true scRNA-seq profiles. We then

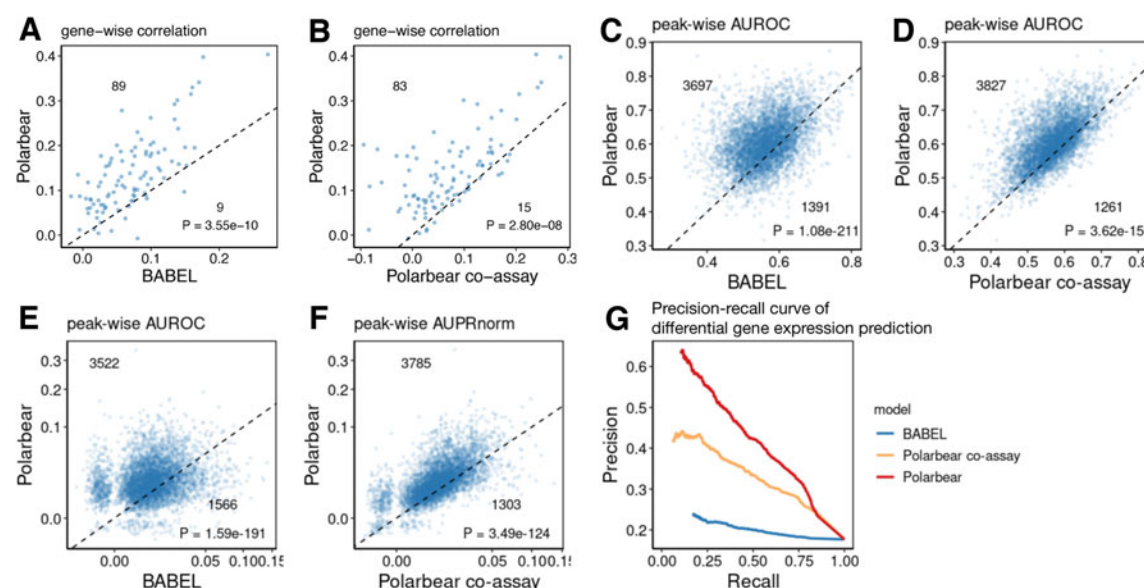


FIG. 5. Cross-modality prediction on an unseen cell type. (A, B) Gene-wise correlation between the true and predicted profile, comparing Polarbear with BABEL (A) or with Polarbear co-assay (B). Each dot is a gene, and numbers indicate the number of dots above and below the diagonal line. p -Values are calculated by one-sided Wilcoxon rank-sum test. “Polarbear co-assay” only uses co-assay data to train the model. (C, D) Peak-wise AUROC, comparing Polarbear with BABEL (C) or with Polarbear co-assay (D). (E, F) Peak-wise AUPRnorm, comparing Polarbear with BABEL (E) or with Polarbear co-assay (F). (G) Precision-recall curve on prioritizing the true set of differentially expressed genes based on differential expression pattern on the predicted profiles.

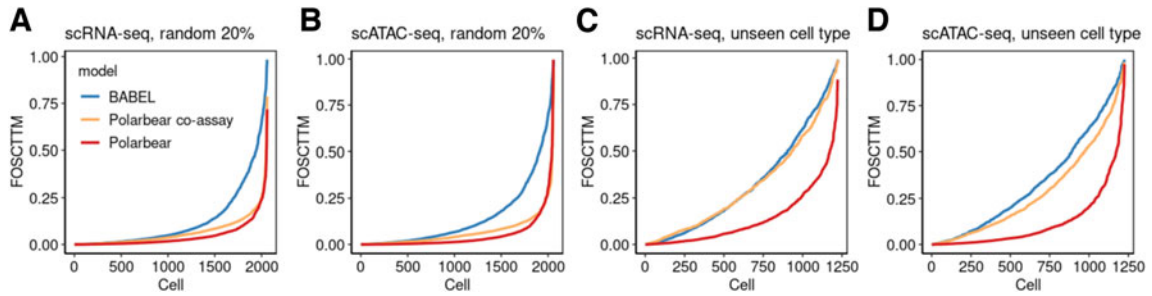


FIG. 6. Evaluation of cross-modality matching on different models. The FOSCTTM score for Polarbear (red), Polarbear trained only with co-assay data (orange), as well as BABEL (blue). Cells are sorted based on FOSCTTM score for each method. (A, B) Matching performance on the random 20% test set, using either scRNA-seq (A) or scATAC-seq (B) as queries. (C, D) Matching performance on the unseen cell type, using either scRNA-seq (C) or scATAC-seq (D) as queries. FOSCTTM, fraction of samples closer than the true match.

calculate a precision-recall curve relative to these cell labels, ranking genes by their differential expression from the predicted profiles. We find that Polarbear's predictions are able to accurately recapitulate differentially expressed genes, significantly outperforming other methods (Fig. 5G). These results suggest that Polarbear can correctly predict intra- and inter-cell type variations, even for cell types for which no co-assay data are available to train the model.

3.4. Polarbear can match corresponding cells across modalities

Polarbear can also be used to match corresponding cells from different modalities. Given unpaired single-assay profiles in each modality, we can use Polarbear to match those cells between modalities, supervised by the co-assay data. To simulate this setting, we project the scRNA-seq and scATAC-seq profiles in the held-out test set to the bottleneck layer of the Polarbear model, and we match cells from different modalities in a greedy fashion based on Euclidean distance in the latent space. To assess the matching performance, we calculate for each cell the **FOSCTTM** (Liu et al, 2019).

Although BABEL does not claim to perform cross-modality alignment, the model learns to translate by forcing corresponding cells from the two modalities to be located in the same embedding space. Thus, to be consistent in this article, we still include BABEL as a baseline model to compare against, and we expand a grid search of hyperparameters for BABEL to select the best alignment model (Section 2.2). Polarbear is able to achieve a lower FOSCTTM score than BABEL and Polarbear co-assay in matching cells in the random test set (Fig. 6A, B), as well as matching cells within the unseen cell type (Fig. 6C, D), suggesting that adding single-assay data improves cross-modality matching.

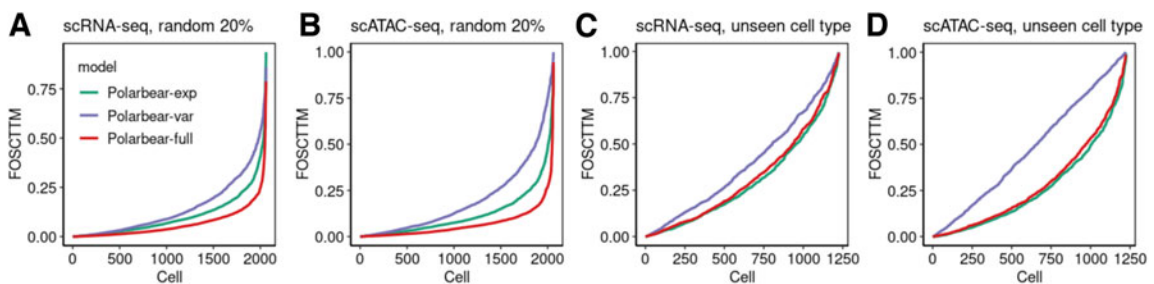


FIG. 7. Evaluation of cross-modality matching on different feature pre-filtering methods. The FOSCTTM score for Polarbear-full (red), Polarbear-var trained with top quantile highly variable features (purple), as well as Polarbear-exp that is trained on highly expressed features (green). Cells are sorted based on FOSCTTM score for each method. Only co-assays are used to train the model. (A, B) Matching performance on the random 20% test set, using either scRNA-seq (A) or scATAC-seq (B) as queries. (C, D) Matching performance on the unseen cell type, using either scRNA-seq (C) or scATAC-seq (D) as queries.

We noticed that several methods choose to pre-filter genes and peaks to exclude either sparse features or less variable features (Ashuach et al, 2021; Minoura et al, 2021). Besides the obvious drawback of failing to predict on potentially biologically interesting peaks and genes, we hypothesize that this type of feature subsetting will decrease the performance of cross-modality cell alignment. To test this hypothesis, we further filtered the SNARE-seq dataset used in our model (**Polarbear-full**) to exclude lowly expressed genes and peaks (**Polarbear-exp**) or only retain highly variable genes and peaks (**Polarbear-var**; Section 2.5), and we use the same model and training procedures to align cells in the test set.

Only SNARE-seq co-assay data are used here to train the model. Our experiment shows that Polarbear-full is the best at aligning cells in the random test set, and it performs similarly to Polarbear-exp at aligning cells within the unseen cell type. Meanwhile, Polarbear-var is not able to perform well in either scenario (Fig. 7). Our analysis demonstrates the importance of including features that are relatively sparse or have low variability to achieve optimal cross-modality alignment performance.

4. DISCUSSION

We propose Polarbear, a semi-supervised framework that leverages both co-assay and publicly available single-assay data to translate between scRNA-seq and scATAC-seq profiles. We demonstrate that Polarbear improves upon methods that only train on co-assay data. Polarbear predictions are able to capture cell-type and individual cell-level differences, and they can predict missing domain knowledge for cell types without any co-assay data available. Polarbear code and data used in this study can be found on <http://github.com/Noble-Lab/Polarbear>.

Polarbear can be used to generate biological hypotheses in the missing domain, such as inferring differentially expressed genes/peaks between cell types or experimental groups. We expect Polarbear to be used to facilitate biological discoveries on uncharacterized domains at the single-cell level, such as identifying individual cell or subclone-specific regulatory elements based on scRNA-seq profiles in tumor samples.

Currently, Polarbear predictions do not improve scATAC-seq predictions as much as scRNA-seq predictions. Possible reasons for this difference are that scATAC-seq profiles are sparse and noisy, and scATAC-seq data potentially contain more information than scRNA-seq because a single gene can be regulated by multiple scATAC-seq peaks. We foresee that models taking into account prior knowledge (e.g., **DNA sequence features** or **regulatory region annotations**) may further improve scATAC-seq predictions.

Polarbear can also match single cells across modalities with high accuracy. We envision that our semi-supervised matching framework could be adapted for aligning the large compendium of publicly available single-assay profiles, so that we can generate new hypothesis (e.g., gene-peak relationships and cell clustering based on joint features) based on the predicted paired scRNA-seq and scATAC-seq profiles.

Thanks to Polarbear's flexible training framework, the current Polarbear model could, in the future, be combined with pre-trained models from other data domains by learning the translation layer based on a limited number of co-assay data, and thus be generalized to translate among multiple modalities.

ACKNOWLEDGMENTS

The authors would like to thank Noble Lab members, especially Yang Lu, Gang Li, Ayse Dincer, Anupama Jha, and Dejun Lin, for valuable discussions.

AUTHORS' CONTRIBUTIONS

R.Z.: Conceptualization, Methodology, Formal analysis, Investigation, Validation, Visualization, Writing—Original draft preparation, and Software. L.M.-P.: Methodology, Writing—Review and Editing, and Software. J.-P.V.: Supervision, Methodology, and Writing—Review and Editing. W.S.N.: Conceptualization, Supervision, Methodology, Writing—Review and Editing, and Funding acquisition.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This work was funded in part by National Institutes of Health award UM1 HG011531.

REFERENCES

- Ashuach T, Gabitto MI, Jordan MI, et al. Multivi: Deep generative model for the integration of multi-modal data. *bioRxiv* 2021; doi: 10.1101/2021.08.20.457057.
- Ashuach T, Reidenbach DA, Gayoso A, et al. Peakvi: A deep generative model for single-cell chromatin accessibility analysis. *Cell Rep Methods* 2022;2(3):100182; doi: 10.1016/j.crmeth.2022.100182.
- Buttgereit A, Lelios I, Yu X, et al. Sall1 is a transcriptional regulator defining microglia identity and function. *Nat Immunol* 2016;17(12):1397–1406; doi: 10.1038/ni.3585.
- Cao J, Cusanovich DA, Ramani V, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 2018;361(6409):1380–1385; doi: 10.1126/science.aau0730.
- Chen S, Lake BB, Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;37(12):1452–1457; doi: 10.1038/s41587-019-0290-0.
- Eraslan G, Simon LM, Mircea M, et al. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;10(1):390; doi: 10.1038/s41587-019-0290-0.
- Fang R, Preissl S, Li Y, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* 2021;12(1):1–15; doi: 10.1038/s41467-021-21583-9.
- Gala R, Gouwens N, Yao Z, et al. A coupled autoencoder approach for multi-modal analysis of cell types. *Adv Neural Inf Process Syst* 2019;32.
- Gayoso A, Steier Z, Lopez R, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* 2021;18(3):272–282; doi: 10.1038/s41592-020-01050-x.
- Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell* 2021;184(3):3573–3587; doi: 10.1016/j.cell.2021.04.048.
- Harrow J, Denoeud F, Frankish A, et al. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* 2006;7(Suppl 1):S4; doi: 10.1186/gb-2006-7-s1-s4.
- Hinrichs AS, Karolchik D, Baertsch R, et al. The ucsc genome browser database: Update 2006. *Nucleic Acids Res* 2006;34(suppl_1):D590–D598; doi: 10.1093/nar/gkj144.
- Li YE, Preissl S, Hou X, et al. An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* 2021;598(7879):129–136; doi: 10.1038/s41586-021-03604-1.
- Liu J, Huang Y, Singh R, et al. Jointly embedding multiple single-cell omics measurements. In: 19th International Workshop on Algorithms in Bioinformatics (WABI 2019), volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*. (Huber KT, Gusfield D, eds.) Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany; 2019; pp. 10:1–10:13; doi: 10.4230/LIPIcs.WABI.2019.10.
- Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;15(12):1053–1058; doi: 10.1038/s41592-018-0229-2.
- Lotfollahi M, Litinetskaya A, Theis FJ. Multigrade: Single-cell multi-omic data integration. *bioRxiv* 2022; doi: 10.1101/2022.03.16.484643.
- Lun ATL, Bach K, and Marioni JC. Pooling across cells to normalize single-cell ma sequencing data with many zero counts. *Genome Biology* 2016;17(1):75; doi: 10.1186/s13059-016-0947-7.
- Ma S, Zhang B, LaFave LM, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020;183(4):1103–1116; doi: 10.1016/j.cell.2020.09.056.
- Minoura K, Abe K, Nam H, et al. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods* 2021;1(5):100071; doi: 10.1016/j.crmeth.2021.100071.
- Talwar D, Mongia A, Sengupta D, et al. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci Rep* 2018;8(1):16329; doi: 10.1016/j.crmeth.2021.100071.
- Trong TN, Kramer R, Mehtonen J, et al. Semisupervised generative autoencoder for single-cell data. *J Computat Biol* 2020;27(8):1190–1203; doi: 10.1089/cmb.2019.0337.
- Wang D, Gu, J. VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genomics Proteomics Bioinformatics* 2018;16(5):320–331; doi: 10.1016/j.gpb.2018.08.003.

- Wu KE, Yost KE, Chang HY, et al. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci U S A* 2021;118(15); doi: 10.1073/pnas.2023070118.
- Xiong , Xu K, Tian K, et al. Scale method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* 2019;10(1):1–10; doi: 10.1038/s41467-019-12630-7.
- Zeisel A, Hochgerner H, Lönnerberg P, et al. Molecular architecture of the mouse nervous system. *Cell* 2018; 174(4):999–1014; doi: 10.1016/j.cell.2018.06.021.
- Zhu C, Yu M, Huang H, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* 2019;26(11):1063–1070; doi: 10.1038/s41594-019-0323-x.

Address correspondence to:
Dr. William Stafford Noble
Department of Genome Sciences
University of Washington
Seattle, WA 98195-5065
USA

E-mail: william-noble@uw.edu