SEMI-SUPERVISED CONSISTENCY REGULARIZATION FOR ACCURATE CELL TYPE FRACTION AND GENE EXPRESSION ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cell deconvolution is the estimation of cell type fractions and cell type-specific gene expression from mixed data with unknown composition. In biomedical research, cell deconvolution, which is a source separation task, is used to obtain mechanistic and diagnostic insights into human diseases. An unmet challenge in cell deconvolution, however, is the scarcity of realistic training data and the strong domain shift observed in synthetic training data that is used in contemporary methods. Here, we hypothesize that simultaneous consistency regularization of the target and training domains will improve deconvolution performance. By adding this biologically motivated consistency loss to two novel deep learning-based deconvolution algorithms, we achieve state-of-the-art performance on both cell fraction and gene expression estimation. Our method, DISSECT, outperforms competing algorithms across several gene expression datasets and can be easily adapted to deconvolve other biomedical data types, as exemplified by our spatial expression deconvolution experiments.

1 INTRODUCTION

A prominent approach to study tissue-specific gene expression changes in human development and disease is RNA sequencing (bulk RNA-seq). Tissues, however, usually consist of multiple cell types in different quantities, with different gene expression programs. As a consequence, bulk RNA-seq from tissues measures average gene expression across the constituent cells, disregarding cell type-specific changes. The quantification of the cellular composition and cell type-specific expression that underlies bulk RNA-seq data is therefore of pivotal importance to understand disease mechanisms and to identify potential therapeutic interventions (Li & Wang, 2021).

A recent technological advancement, single-cell RNA-seq, allows for the investigation of gene expression in single cells for thousands of individual cells of a given tissue sample in a single experiment. While it provides unprecedented insights into single cell biology, it suffers from severe technical limitations, most notably gene expression 'dropouts' (Lähnemann et al., 2020). In addition, the technology is still very costly, which largely prohibits its application in clinical and diagnostic settings. Bulk RNA-seq, on the other hand, can be performed for a fraction of the cost and is widely used in clinical oncology and drug discovery (Zhou et al., 2019; Roberts et al., 2014).

To computationally infer cell fraction and cell type-specific gene expression information from bulk RNA-seq data, recent computational methods utilize single cell sequencing data to create simulated references with known fraction and expression information for training (Avila Cobos et al., 2020; Menden et al., 2020; Newman et al., 2019; Wang et al., 2019). While this approach achieves good deconvolution results, its performance suffers from the strong domain gap between single-cell RNA-seq training (reference) data and the bulk RNA-seq target data. Among many possible sources of variation, two most obvious are the presence of batch effects which refers to technological differences between two sequencing experiments and gene expression differences of biological nature. In the next section we formally define the task of cell deconvolution, and present our hypothesis that semi-supervised consistency regularization minimizes the bulk RNA-seq deconvolution error while learning from single cell RNA-seq data.

2 CELL DECONVOLUTION

Given an $m \times n$ gene expression matrix $\mathbf{B}$ consisting of m bulk gene expression vectors measuring n genes, the goal of deconvlution is to find a $m \times c$ matrix $\mathbf{X}$ of cell type fractions, where c is the number of cell types present in bulk samples such that,

$$\mathbf{B} = \mathbf{X}\mathbf{S}, \tag{1}$$

where fractions and gene expression satisfy non-negativity $0 \leq \mathbf{X}_{ik}$, and $0 \leq \mathbf{S}_{kj}, \forall i \in [1, m], \forall j \in [1, n]$ and $\forall k \in [1, c]$ and sum-to-1 criterion *i.e.* $\sum_{k=1}^{c} \mathbf{X}_{ik} = 1, \forall i \in [1, m]$. Here, $\mathbf{S}$ is known as the signature matrix and is unobserved. Each row $\mathbf{S}_{k\cdot}$ is a gene expression profile (or signature) of cell type k. To utilize a reference based framework, $\mathbf{S}$ can be replaced with $\mathbf{S}_{ref}$ derived from a single-cell experiment.

The problem of reference-based cell deconvolution can alternatively be formulated as a learning problem, where a function f such that $f(\mathbf{B}) = \mathbf{X}$ is learnt. Since only $\mathbf{B}$ is available and $\mathbf{X}$ is generally unknown, simulations from single-cell reference can be used to learn f. Clearly, from the above formulation of the cell deconvolution task, it is reasonable to assume linearity of deconvolution, i.e., each bulk mixture is a linear combination of expression vectors of cells spanned with corresponding cell type fractions. Thus, as defined in (Menden et al., 2020), multiple single cells can be combined in random proportions to generate training examples $\mathbf{B}^{\text{sim}}$ and $\mathbf{X}^{\text{sim}}$, where each row of $\mathbf{B}^{\text{sim}}$ is defined as,

$$\mathbf{B}^{\text{sim}}_{i\cdot} = \sum_{k=1}^{c} \sum_{l=1}^{\alpha_{k,i}} \mathbf{e}^{k}_{l},$$

where $\mathbf{e}^{k}_{l}$ is expression vector of cell l belonging to cell type k, and $\alpha_{k,i}$ is the number of cells belonging to cell type k sampled to construct $\mathbf{B}^{\text{sim}}_{i\cdot}$. Correspondingly, each element of $\mathbf{X}^{\text{sim}}$ is proportion of a cell type k in that sample i and is defined as,

$$\mathbf{X}^{\text{sim}}_{ik} = \frac{\alpha_{k,i}}{\sum_{k=1}^{c} \alpha_{k,i}}, \text{ and}$$

In this case, since each simulated sample has a distinct signature (i.e. gene expression profile), $\mathbf{S}$ is a three dimensional matrix with each element $\mathbf{S}_{kji}$ denoting gene expression of gene j in cell type k for sample i. It is computed as following,

$$\mathbf{S}^{\text{sim}}_{k\cdot i} = \frac{\sum_{l=1}^{\alpha_{k,i}} \mathbf{e}^{k}_{l}}{\alpha_{k,i}}.$$

The predictor f, learned from a simulated dataset, can then be applied to $\mathbf{B}$ to estimate $\mathbf{X}$. Note that, the genes expressed may differ between vectors $\mathbf{e}_l$ and $\mathbf{B}$ and as such before learning function f, each $\mathbf{e}^{\mathbf{k}}_{\mathbf{l}}$ is subsetted to include genes common with $\mathbf{B}$. This is the reason why this learning problem is transductive and a separate model needs to be reconstructed for each $\mathbf{B}$.

2.1 EXPLOITING LINEARITY OF DECONVOLUTION

2.1.1 ASSUMPTION

From Section 2, it is evident that the relationship between $\mathbf{B}$ and $\mathbf{S}$ is linear. However, $\mathbf{S}$ is unobserved and learning is done using simulations. To address the inherent domain shift, we hypothesize that a consistency based regularization penalizing non-linearity of mixtures of real and simulated samples would result in a mapping f that is closer to true f. We define it in Section 2.1.2.

2.1.2 CONSISTENCY REGULARIZATION

To simplify the notation, we consider same index i for real bulk samples, simulations (sim) and their mixtures (mix, defined further). Given a true bulk RNA-seq $\mathbf{B}_{i\cdot}$ and a simulated sample with paired proportions $(\mathbf{B}_{i\cdot}^{\text{sim}}, \mathbf{X}_{i\cdot}^{\text{sim}})$, we can generate a mixture $\mathbf{B}_{i\cdot}^{\text{mix}}$ such that

$$\mathbf{B}_{i\cdot}^{\text{mix}} = \beta \mathbf{B}_{i\cdot} + (1-\beta)\mathbf{B}_{i\cdot}^{\text{sim}}, \tag{2}$$

Which gives us relation

$$\mathbf{X}_{i\cdot}^{\text{mix}} \mathbf{S}_{\cdot i\cdot}^{\text{mix}} = \beta \mathbf{S}_{\cdot i\cdot} \mathbf{X}_{i\cdot} + (1-\beta)\mathbf{S}_{\cdot i\cdot}^{\text{sim}} \mathbf{X}_{i\cdot}^{\text{sim}}. \tag{3}$$

Cell types are characterized by a few marker genes that are invariant across cell states and even across tissues (Domínguez Conde et al., 2022). A network that accurately predicts cell type fractions based on gene expression of simulated (or real) bulks would thus have to learn them. Thus, to estimate cell type fractions, we assume that the expression of these marker genes should be identical in signatures $\mathbf{S}_{\cdot i\cdot}^{\text{mix}}, \mathbf{S}_{\cdot i\cdot}$ and $\mathbf{S}_{\cdot i\cdot}^{\textbf{sim}}$. Hence,

$$\mathbf{X}_{i\cdot}^{\text{mix}} = \beta \mathbf{X}_{i\cdot} + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}, \tag{4}$$

where $\beta \in [0, 1]$. Equation 4 enables the use of consistency regularization without having to explicitly estimate signatures.

In an iterative learning process $\mathbf{X}_{i\cdot}$ can be replaced with predictions of the algorithm from the previous iteration. Naturally, it is also possible to only mix real samples with each other, however, the number of samples available from true bulk RNA-seq is considerably lesser (usually ranging from a couple to less than thousand) than the amount single-cells present in a single-cell experiment (usually in thousands). The equation 4 allows to generate pseudo ground truth proportions for mixtures $\mathbf{B}_{i\cdot}^{\text{mix}}$ at the each step of learning cell type fractions, while Equation 3 allows to generate pseudo ground truth signatures at each step of learning gene expression profiles. We define the network architecture and loss functions in 2.2.

2.2 NETWORK ARCHITECTURE AND LEARNING PARADIGM

We approach the two tasks, estimation of cell type fractions and estimation of gene expression profiles per cell type as two different tasks because of their differing assumptions. For estimation of cell type fractions, we assume that signatures are identical for each sample, both simulated and bulk, while to estimate gene expression, we relax this condition use full consistency regularization (Equation 3).

2.2.1 ESTIMATION OF CELL TYPE FRACTIONS

The underlying algorithm of the first part of our deconvolution method is an average ensemble of multilayered perceptrons (MLPs). Each MLP consists of the same architecture initialized with different weights. This is done to reduce the variance by averaging different runs (Ju et al., 2018). Each MLP has an architecture: Input (# genes) - ReLU6 (512) - ReLU6 (256) - ReLU6 (128) - ReLU6 (64) - Linear (# cell types) - Softmax. ReLU6 (Output of ReLU activation clipped by a maximum value of 6) (Hannun et al., 2014; Sandler et al., 2018) was chosen out of tested activations over grid search on [Linear, ReLU, ReLU6, Swish (Ramachandran et al., 2017)] The final application of Softmax activation allows to achieve non-negativity and sum to 1 criteria of deconvolution. we train the network with batch size 64 to minimize the loss function defined below with an Adam Optimizer with initial learning rate of $1e-5$.

$$\mathcal{L}_{\text{total}}(\mathbf{X}_{i\cdot}^{\text{sim}}, f(\mathbf{B}_{i\cdot}^{\text{sim}}), \mathbf{X}_{i\cdot}^{\text{mix}}, f(\mathbf{B}_{i\cdot}^{\text{mix}})) = \mathcal{L}_{\text{KLdivergence}}(\mathbf{X}_{i\cdot}^{\text{sim}}, f(\mathbf{B}_{i\cdot}^{\text{sim}})) + \lambda_1 * \mathcal{L}_{\text{cons}}(\mathbf{X}_{i\cdot}^{\text{mix}}, f(\mathbf{B}_{i\cdot}^{\text{mix}})), \tag{5}$$

where $\mathcal{L}_{\text{KLdivergence}}(\cdot,\cdot)$ is the Kullback-Leibler divergence and $\mathcal{L}_{\text{cons}}(\cdot,\cdot)$ is the consistency loss defined as:

$$\mathcal{L}_{\text{cons}}(\mathbf{X}_{i\cdot}^{\text{mix}}, f(\mathbf{B}_{i\cdot}^{\text{mix}})) = ||\mathbf{X}_{i\cdot}^{\text{mix}} - f(\mathbf{B}_{i\cdot}^{\text{mix}})||_2^2, \text{ and}$$

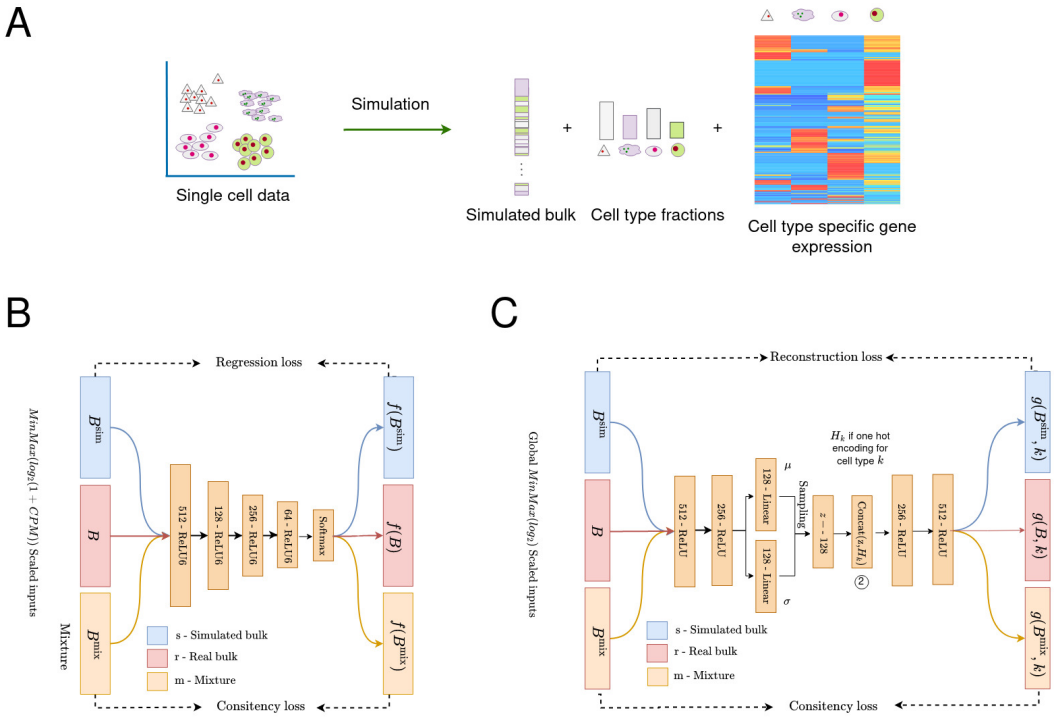

Figure 1: **A.** Illustration of simulation using reference single-cell data. The figure shows the simulation of one sample which consists of cell type fractions, simulated gene expression and cell type specific gene expression profiles (i.e. signature matrix). **B.** Detailed overview of an MLP used to estimate cell type fractions, and **C.** Overview of an autoencoder used to estimate cell type specific gene expression profiles.

$$\mathbf{X}_{i\cdot}^{\text{mix}} = \beta f(\mathbf{B}_{i\cdot}) + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}.$$

To generate mixtures, for each batch, we sample β and uniformly at random for Equation 4. The interval $[0.1, 0.9]$ was chosen for the uniform distribution to allow for at least some real and some simulated gene expression in the mixture. This loss is similar to the semi-supervised framework proposed in MixMatch (Berthelot et al., 2019). MixMatch uses unlabelled samples to MixUp and match sample predictions and generalizes semi-supervised framework, while the loss defined in Equation 5 addresses the limited samples available from true bulk RNA-seq, unavailability of sample fractions and is derived from the definition of task itself. In essence, Equation 5 integrates domain knowledge into the objective.

To avoid a scenario where the network doesn't learn and outputs predictions such that $f(\mathbf{B}_{i\cdot}^{\text{mix}}) = f(\mathbf{B}_{i\cdot}^{\text{sim}}) = f(\mathbf{B}_{i\cdot})$, which is a solution to Equation 4, we first let the model learn purely from simulated examples. This allows the model to learn meaningful expression profiles to achieve accurate results on simulated examples. We selected λ_1 based on a grid search over constant and step-wise functions. We adopt a step-wise function for λ_1, given as:

$$\lambda_1 = \begin{cases} 0 & \text{if step} \leq 2000, \\ 15 & \text{elif } 2000 \leq \text{step} \leq 4000, \\ 10 & \text{else.} \end{cases}$$

We train the network for a predefined number of steps as opposed to epochs, since it is possible to generate infinitely many simulated samples without increasing the intrinsic dimensionality of the data. In our experiments, we set their number to 5000 as found optimal in Menden et al. (2020).

2.2.2 ESTIMATION OF CELL TYPE SPECIFIC GENE EXPRESSION PROFILES

Estimation of cell type fractions from bulk RNA-seq requires an assumption that signatures of cell types are shared across single cell and bulk RNA-seq. However, cell type gene expression profiles (at least for genes that are not invariant across tissue states) may differ. Previously, works such as CSx (Newman et al., 2019) and TAPE (Chen et al., 2021) have explored utilizing cell type fractions to estimate gene expression. Here, we make use of a β-variational autoencoder with standard normal distribution as prior to estimate average gene expression of the different cell types from bulk RNA-seq expression levels. To jointly train the network on all cell types, we condition the decoder (at its input layer) with cell type labels. To make use of bulk RNA seq during the training, we regularize the reconstruction loss with a consistency loss defined over per cell type signature. Denoting f as before and $g(\cdot, k)$ as the output of the autoencoder with condition k (corresponding to cell type label) on the decoder input, this consistency loss is defined as:

$$\mathcal{L}_{\text{cons}}^{\text{VAE}}(f, g, \mathbf{B}_{i\cdot}^{\text{mix}}, \mathbf{B}_{i\cdot}, \mathbf{X}_{i\cdot}^{\text{sim}}, \mathbf{S}_{ki\cdot}^{\text{sim}}) = ||f(\mathbf{B}_{i\cdot}^{\text{mix}})_k g(\mathbf{B}_{i\cdot}^{\text{mix}}, k) - \beta f(\mathbf{B}_{i\cdot})_k g(\mathbf{B}_{i\cdot}, k) + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}\mathbf{S}_{ki\cdot}^{\text{sim}}||_2^2,$$

where $\mathbf{B}_i^{\text{mix}}$ is given by Equation 2, $f(\mathbf{B}_{i\cdot}^{\text{mix}})_k$ is the estimated proportion of cell type k in sample i. In implementation, we replace $f(\mathbf{B}_{i\cdot}^{\text{mix}})_k$ with $\beta f(\mathbf{B}_{i\cdot})_k + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}$. Thus, this loss forces the learned signature for cell type k, $g(\mathbf{B}_{i\cdot}^{\text{mix}}, k)$, to be closer to signatures for both real and simulated bulk samples. This loss function makes the assumption that mixing two bulk samples is similar to mixing individual cell type specific signatures that constitute those bulks. We added this loss function with a regularization parameter λ_2 (with default value 0.1) to the loss of the standard β-variational autoencoder (the weight on the KL divergence, denoted as β^{VAE}, is set to 0.1 by default). The total loss function sums up to:

$$\begin{aligned}\mathcal{L}_{\text{total}}^{\text{VAE}}(f, g, \mathbf{B}_{i\cdot}^{\text{sim}}, \mathbf{B}_{i\cdot}^{\text{mix}}, \mathbf{B}_{i\cdot}, \mathbf{X}_{i\cdot}^{\text{sim}}, \mathbf{S}_{ki\cdot}^{\text{sim}}) &= ||\mathbf{S}_{ki\cdot}^{\text{sim}} - g(\mathbf{B}_{i\cdot}^{\text{sim}}, k)||_2^2 \\ &+ \lambda_2 \mathcal{L}_{\text{cons}}^{\text{VAE}}(f, g, \mathbf{B}_{i\cdot}^{\text{mix}}, \mathbf{B}_{i\cdot}, \mathbf{X}_{i\cdot}^{\text{sim}}, \mathbf{S}_{ki\cdot}^{\text{sim}}) \\ &+ \beta^{VAE}\mathcal{L}_{\text{KLdivergence}}(\mathcal{N}(\mu, \sigma), \mathcal{N}(0, 1)),\end{aligned}$$

where $\mathcal{N}(0, 1)$ is standard normal distribution, μ and σ are the empirical mean and standard deviation estimated from the output of the encoder. Both the encoder and decoder consist of two hidden layers. We train the network to minimize the loss function with an Adam optimizer with initial learning rate of $1e-4$. By default, the network is trained with a batch size of 32 for $10000 \times c$ number of steps. The architecture of the network is summarized in Figure 1.

3 RELATED WORK AND COMPETING ALGORITHMS

Several methods for cell deconvolution have been developed. Avila Cobos et al. (2020) and Jin & Liu (2021) provided a benchmark and review of state of the art cell deconvolution algorithms. Here, we focus on MuSiC, CSx, Scaden and TAPE, because both MuSiC and CSx are single cell reference based methods that have performed well on simulation studies in aforementioned benchmarking studies. Scaden and TAPE are selected as both are deep learning based deconvolution approaches. We briefly detail these approaches below. Out of these methods, CSx and TAPE can also estimate cell type-specific gene expression signatures.

MuSiC (Wang et al., 2019) uses weighted non-negative least squares. MuSiC maintains cross-cell and cross-sample consistencies by appropriately weighting genes based on their informativity during an iterative procedure. MuSiC is provided as an R package. Deconvolution using MuSiC was performed according to the authors recommendations. Since MuSiC is a method that utilizes multi-subject scRNA-seq datasets, when available, we used cells from multiple subjects in deconvolution with MuSiC.

CibersortX (CSx) (Newman et al., 2019) is a deconvolution method that addresses domain gap problems with scRNA-seq and bulk samples by aiming to correct batch effects. It uses scRNA-seq to generate a cell type specific signature matrix and uses ν-support vector regression as the underlying algorithm. CSx comprises two modes, S- and B-modes, to address the domain gap. S-mode is used

when deconvolving with a signature matrix constructed using a scRNA-seq dataset, while B-mode is used when deconvolving with a signature matrix constructed using purified samples. CSx allows academic use of its implementation through docker containers and a web-based application. We followed the documentation provided by the authors to run CSx and used the S-mode. CSx can also predict gene expression signatures for each sample for which it uses a Non-negative matrix factorization based iterative algorithm.

Scaden (Menden et al., 2020) is an average ensemble of three deep neural networks with different architectures that was developed for cell fraction deconvolution. Each network is trained to minimize the mean squared error on simulated pseudo bulk data generated from an scRNA-seq reference similar to described above. Scaden is provided as a Python package. We used the official Scaden package with the instructions provided by the authors to train the networks. The middle network of Scaden has the same architecture as individual networks of DISSECT. Scaden uses dropout and does not feature a consistency regularization.

TAPE (Chen et al., 2021) is a fully-connected autoencoder where the bottleneck consists of cell type fractions. The architecture of the encoder is similar to the archictecture of Scaden but with CeLU activations. The decoder consists of linear activations and outputs gene expression of the input vector. TAPE has two modes - Overall and Adaptive. The adaptive mode aims at optimizing the network for bulk samples, while the overall mode trains for fractions with an added loss function that reconstructs input bulk expression from fractions. Since TAPE-A reconstructs gene expression from fractions (bottleneck), the signature matrix is visible in the (linear) decoder. To estimate gene expression signatures for each bulk sample, decoder weights are optimized per-sample using an iterative optimization strategy. Network weights are changed during the two modes , we compare with both and refer to TAPE in overall mode as TAPE-O and in adaptive mode as TAPE-A. TAPE-O is primarily a modification in network architecture and regularization of loss function with reconstruction and does not use real bulk samples in learning, while TAPE-A addressed domain shift.

Linear MLPs: The solution to the deconvolution problem could be, in principle, a linear function. For this reason we also compared to an MLP ensemble that is based on the architecture in Section 2.2, but in which we replaced all non-linear activations with an identity function and removed the consistency loss.

4 Experiments and results

4.1 Datasets

We evaluated the algorithms on six datasets consisting of peripheral blood mononuclear cells (PBMCs) and corresponding ground-truth cell type fractions experimentally quantified using flow cytometry. Details of these datasets are given in Table 1.

Table 1: Real bulk datasets with known ground truth cell fraction information for the evaluation of deconvolution performance. For each of these datasets, we focus on five cell types, namely CD4 T cells, CD8 T cells, NK cells, Monocytes and B cells, which are measured consistently across all datasets. We grouped the remaining cell type proportions in a single label "Unknown" following methodology in Menden et al. (2020) or calculate "Unknown" proportions as 1 - (Sum of fractions of aforementioned five cell types).

Dataset name	**# samples**	**Original Source**	**Type**
SDY67	12	Zimmermann et al. (2016)	RNA-seq
Monaco bulk	12	Monaco et al. (2019)	RNA-seq
Monaco microarray	164	Monaco et al. (2019)	Microarray
GSE65133	20	Newman et al. (2015)	Microarray
GSE107572	9	Finotello et al. (2019)	RNA-seq
GSE120502	250	Harrison et al. (2019)	RNA-seq

To deconvolve these datasets, we used the same reference single-cell datasets as used in previous deep learning-based deconvolution approaches (Menden et al., 2020). We use this single-cell data

from healthy donors to create simulated bulk data with known fractions as described in section 2. The parameter details for simulations are provided in Appendix C.2.

Table 2: Single-cell data used for the creation of simulated reference data for supervised training. We manually annotated cells with these five cell types based on expression of cell type marker genes. The cells that did not belong to any of these cell types were labelled as "Unknown". This unknown cluster allows comparison of fractions measured at relative scale.

Dataset name	**# cells**	**Title in 10X Genomics**[1]
PBMC8k	5419	8k PBMCs from a Healthy Donor
PBMC6k	8381	6k PBMCs from a Healthy Donor
DonorA	2900	Frozen PBMCs (Donor A)
DonorC	9519	Frozen PBMCs (Donor C)

To preprocess these single-cell datasets, we utilized the procedure described in C. For the *Monaco bulk* dataset (Table 1), more granular level fractions of cell type subsets are quantified using flow cytometry. We utilized this information to evaluate methods on discerning closely related or scarce cell type subsets. Since it is notoriously hard to identify cell types at such granularity in PBMC single-cell datasets from healthy donors, we considered 9852 RNA-seq samples of purified cells from Ota et al. (2021) as references. Purified RNA-seq samples are average profiles for thousands of cells of the same cell type. To match cell types between the purified reference and flow cytometry from *Monaco bulk*, we harmonize cell type labels. Resulting ground truth and reference had 18 cell subsets defined are given in Appendix D.

4.2 Evaluation metrics

We used Pearson correlation and root mean squared error (RMSE) for the evaluation of deconvolution results. Since some cell types are much more abundant than others, it is important to consider the overall and per cell type average correlation and RMSE (see Apppendix B).

4.3 Results

In this section, we present results on experiments detailed in Section 4.1. Additional experiments and results on datasets without corresponding flow cytometry fractions are presented in Appendix E. There we utilize validated biological hypotheses to evaluate DISSECT against competing methods.

4.3.1 Estimation of celltype fractions

To evaluate deconvolution performance, we deconvolved each of the datasets in Table 1 using the *PBMC8k* reference dataset (Table 2). Table 3 and Appendix Table 7 demonstrate that DISSECT shows significantly improved correlations in 9 out of 12 comparisons and the lowest RMSE in 11 out of 12 comparisons, across 6 different datasets.

Next, we evaluated the cell fraction deconvolution performance on the *Monaco bulk* (Section 4.1) dataset that contains several closely related and rare cell types and constitutes a relatively hard task. With a correlation of 0.6, DISSECT's average performance is 14 percentage points better than the second best performance by Scaden 4. For 8 out of 18 cell types it reaches the best correlation, while Scaden performs best for 3 out of 18 cell types. With an RMSE of 0.03, Scaden's performance is 1 percentage point better than DISSECT (Appendix Table 8). In summary, DISSECT displays the best correlation and a highly competitive RMSE in the cell fraction deconvolution task.

Finally, we performed an ablation study that validates our hypothesis that the consistency loss is primarily responsible for DISSECT's improved deconvolution performance (Appendix Table 12).

4.3.2 Estimation of cell type-specific gene expression

Next, we evaluated the performance of DISSECT's cell type-specific gene expression inference on simulated bulk RNA-seq data. We used simulated data as we could not obtain bulk RNA-seq and

[1]Calculated over real values.

Table 3: *Average (overall)* Pearson correlation between estimations and flow cytometry cell type fractions computed over all cell types. The reference dataset is *PBMC8k* for all methods.

Dataset	**MuSiC**	**CSx**	**Scaden**	**TAPE-O**	**TAPE-A**	**Linear MLPs**	**DISSECT**
SDY67	nan (-0.301)	0.59 (0.52)	0.507 (0.419)	0.53 (0.465)	0.513 (0.152)	0.51 (0.739)	**0.631** (**0.789**)
Monaco bulk	nan (0.006)	0.429 (0.634)	0.685 (0.649)	0.695 (0.569)	0.664 (0.662)	0.534 (0.73)	**0.727** (**0.783**)
Monaco Microarray	nan (-0.16)	0.66 (0.398)	**0.791** (0.189)	0.71 (0.317)	nan (0.03)	0.71 (0.233)	0.786 (**0.799**)
GSE65133	nan (-0.245)	0.642 (0.686)	0.771 (0.413)	0.729 (0.413)	0.719 (0.252)	0.77 (0.421)	**0.821** (**0.918**)
GSE107572	nan (0.155)	0.669 (0.527)	0.627 (0.723)	0.615 (0.658)	0.012 (0.617)	0.486 (0.71)	**0.705** (**0.79**)
GSE120502	0.103 (-0.302)	0.529 (0.776)	0.688 (0.718)	**0.692** (**0.883**)	0.632 (0.779)	0.63 (0.379)	0.662 (0.864)

Table 4: Pearson correlation between estimates from different methods and flow cytometry for granular cell type fractions in *Monaco bulk*.

Celltype	**MuSiC**	**CSx**	**Scaden**	**TAPE-O**	**TAPE-A**	**Linear MLPs**	**DISSECT**
B Ex	nan	**0.43**	0.18	0.22	0.030	0.10	0.32
B NSM	nan	-0.08	**0.12**	0.10	-0.15	-0.22	0.09
B Naive	nan	0.95	0.87	0.8	0.43	0.71	**0.96**
B SM	**0.85**	nan	0.57	0.45	0.15	0.26	0.63
Monocytes C	0.30	0.29	**0.63**	0.57	0.52	0.11	0.62
Monocytes I	0.41	0.36	0.90	0.87	0.81	0.54	**0.93**
Monocytes NC	0.25	0.09	0.31	0.35	0.48	0.19	**0.66**
NK	0.80	**0.82**	0.58	0.59	0.65	0.49	**0.82**
Neutrophils LD	0.2	nan	**0.89**	0.48	0.57	0.03	0.56
Plasmablasts	0.62	0.85	0.86	0.65	0.66	0.42	**0.92**
CD4 T Naive	0.66	0.47	0.68	0.70	0.34	0.14	**0.76**
CD4 T Memory	**0.47**	-0.15	0.27	0.27	0.12	0.08	0.24
CD8 T Naive	0.52	**0.7**	0.36	0.38	0.32	0.27	0.49
CD8 T CM	nan	-0.65	0.19	0.13	**0.21**	0.01	0.12
CD8 T EM	nan	nan	0.02	0.47	0.45	0.11	**0.62**
CD8 T TE	0.25	**0.9**	0.28	0.35	0.36	0.35	0.86
mDC	nan	0.46	0.47	0.39	0.40	0.05	**0.68**
pDC	0.55	**0.57**	0.19	0.42	0.31	0.3	0.55
Average	0.49 [1]	0.40 [1]	0.46	0.45	0.37	0.22	**0.60**

corresponding cell type-specific expression information. To maintain a domain shift between the training and test datasets, simulated data for training and testing were created using different single-cell datasets. Here, we compare our approach with the only two state-of-the-art methods that can infer cell type-specific gene expression per sample, TAPE and CSx (Section 3). We simulated bulk samples from one of the four reference single-cell PBMC datasets listed in Table 2, and created training simulations from the remaining three. To assess the performances, we computed sample- (Table 5) and gene-wise (Table 6) Pearson correlations.

DISSECT displays the best sample- and gene-wise correlations in 3 out of 4 experiments, each, taking first place in overall cell type-specific gene expression deconvolution performance.

5 Discussion

We detailed how the use of a linear consistency is suitable for the task of deconvolution, especially in the absence of real ground truth training information, as is often the case in biomedical settings. Our approach relies on the supervised learning on simulated data and an unsupervised domain adaptation to the target data of interest. This semi-supervised learning approach results in state-of-the-art

Table 5: Pearson correlation between ground truth and estimated gene expression profiles on simulated datasets averaged over samples. The column *Dataset* indicates the single-cell dataset used to create simulations for the test set.

Dataset	**CSx**	**TAPE-A**	**DISSECT**
PBMC6k	0.71 ± 0.12	$\mathbf{0.83 \pm 0.09}$	0.82 ± 0.08
PBMC8k	0.82 ± 0.13	0.79 ± 0.09	$\mathbf{0.84 \pm 0.11}$
DonorA	0.78 ± 0.12	0.85 ± 0.11	$\mathbf{0.89 \pm 0.10}$
DonorC	0.75 ± 0.09	0.81 ± 0.12	$\mathbf{0.83 \pm 0.08}$

Table 6: Pearson correlation between ground truth and estimated gene expression profiles on simulated datasets averaged over estimated genes.

Dataset	**CSx**	**TAPE-A**	**DISSECT**
PBMC6k	0.37 ± 0.22	0.42 ± 0.14	$\mathbf{0.46 \pm 0.15}$
PBMC8k	0.36 ± 0.25	$\mathbf{0.51 \pm 0.12}$	0.48 ± 0.14
DonorA	0.42 ± 0.21	0.48 ± 0.20	$\mathbf{0.48 \pm 0.18}$
DonorC	0.46 ± 0.18	0.45 ± 0.11	$\mathbf{0.49 \pm 0.12}$

deconvolution performance, for both cell fraction and gene expression estimation. While we only focused on MLPs for estimation of cell type fractions and autoencoders for gene expression estimation in this work, we surmise that consistency regularization might improve other deconvolution algorithms as well. We envision further work in this area.

The task of deconvolution plays an important role in spatial transcriptomics (ST) and cell-free DNA methylation (cfDNA). Recently, several algorithms have been developed for ST (Li et al., 2022) and cfDNA deconvolution (Jeong et al., 2022). We surmise that consistency regularization might also improve ST and cfDNA deconvolution by adjusting DISSECT's simulation procedure to mimic ST or cfDNA. We provide two proof-of-concept ST deconvolution results using consistency regularization in Appendix F. Similar arguments can be made for the deconvolution of several other biomedical data types, such as epigenetic, proteomic, and metabolomic data, for instance.

6 Limitations

While DISSECT displays favorable deconvolution performance compared to other methods the results are far from perfect. Especially for hard deconvolution tasks, such as samples with many similar cell types and very scarce cell populations, an increase in deconvolution performance is warranted. Future research into semi-supervised and contrastive algorithms as well as data augmentation and integration techniques should further enhance DISSECT's performance.

7 Code and data availability

All considered datasets are publicly available from respective sources. Code is available anonymously at `https://anonymous.4open.science/r/DISSECT-F0C4`.

REFERENCES

Cell Types Database: RNA-Seq Data - brain-map.org — portal.brain-map.org. https://portal.brain-map.org/atlases-and-data/rnaseq. [Accessed 27-June-2022].

Emilyn U Alejandro, Brigid Gregg, Manuel Blandino-Rosano, Corentin Cras-Méneur, and Ernesto Bernal-Mizrachi. Natural history of β-cell adaptation and failure in type 2 diabetes. *Molecular aspects of medicine*, 42:19–41, 2015.

Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11(1):1–14, 2020.

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

Pazit Beckerman, Jing Bi-Karchin, Ae Seo Deok Park, Chengxiang Qiu, Patrick D Dummer, Irfana Soomro, Carine M Boustany-Kari, Steven S Pullen, Jeffrey H Miner, Chien-An A Hu, et al. Transgenic expression of human apol1 risk variants in podocytes induces kidney disease in mice. *Nature medicine*, 23(4):429–438, 2017.

David A Bennett, Aron S Buchman, Patricia A Boyle, Lisa L Barnes, Robert S Wilson, and Julie A Schneider. Religious orders study and rush memory and aging project. *Journal of Alzheimer's disease*, 64(s1):S161–S189, 2018.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Xin Chen and Rhonda Dzakpasu. Observed network dynamics from altering the balance between excitatory and inhibitory neurons in cultured networks. *Physical Review E*, 82(3):031907, 2010.

Yanshuo Chen, Yixuan Wang, Yuelong Chen, Yuqi Cheng, Yumeng Wei, Yunxiang Li, Ting-Fung Chan, and Yu Li. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *bioRxiv*, 2021.

Diego Contreras. Electrophysiological classes of neocortical neurons. *Neural Networks*, 17(5-6): 633–646, 2004.

C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

João Fadista, Petter Vikman, Emilia Ottosson Laakso, Inês Guerra Mollet, Jonathan Lou Esguerra, Jalal Taneera, Petter Storm, Peter Osmark, Claes Ladenvall, Rashmi B Prasad, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences*, 111(38):13924–13929, 2014.

Francesca Finotello, Clemens Mayer, Christina Plattner, Gerhard Laschober, Dietmar Rieder, Hubert Hackl, Anne Krogsdam, Zuzana Loncova, Wilfried Posch, Doris Wilflingseder, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data. *Genome medicine*, 11(1):1–20, 2019.

Hongjun Fu, Andrea Possenti, Rosie Freer, Yoshikazu Nakano, Nancy C Hernandez Villegas, Maoping Tang, Paula VM Cauhy, Benjamin A Lassus, Shuo Chen, Stephanie L Fowler, et al. A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. *Nature neuroscience*, 22(1):47–56, 2019.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Genelle F Harrison, Joaquin Sanz, Jonathan Boulais, Michael J Mina, Jean-Christophe Grenier, Yumei Leng, Anne Dumaine, Vania Yotova, Christina M Bergey, Samuel L Nsobya, et al. Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. *Nature ecology & evolution*, 3(8):1253–1264, 2019.

Jean-Claude Henquin and Jacques Rahier. Pancreatic alpha cell mass in european subjects with type 2 diabetes. *Diabetologia*, 54(7):1720–1725, 2011.

John V Hindle. Ageing, neurodegeneration and parkinson's disease. *Age and ageing*, 39(2):156–161, 2010.

Yunhee Jeong, Lisa Barros de Andrade e Sousa, Dominik Thalmeier, Reka Toth, Marlene Ganslmeier, Kersten Breuer, Christoph Plass, and Pavlo Lutsik. Systematic evaluation of cell-type deconvolution pipelines for sequencing-based bulk dna methylomes. *Briefings in bioinformatics*, 23(4):bbac248, 2022.

Haijing Jin and Zhandong Liu. A benchmark for rna-seq deconvolution analysis under dynamic testing environments. *Genome biology*, 22(1):1–23, 2021.

Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671, 2022.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, pp. 1–9, 2022.

Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, 13(1):1–6, 2021.

Bi-Cheng Liu, Tao-Tao Tang, Lin-Li Lv, and Hui-Yao Lan. Renal tubule injury: a driving force toward chronic kidney disease. *Kidney international*, 93(3):568–579, 2018.

Rakesh Malhotra, Timothy Craven, Walter T Ambrosius, Anthony A Killeen, William E Haley, Alfred K Cheung, Michel Chonchol, Mark Sarnak, Chirag R Parikh, Michael G Shlipak, et al. Effects of intensive blood pressure lowering on kidney tubule injury in ckd: a longitudinal subgroup analysis in sprint. *American Journal of Kidney Diseases*, 73(1):21–30, 2019.

Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science advances*, 6(30):eaba2619, 2020.

Zhen Miao, Michael S Balzer, Ziyuan Ma, Hongbo Liu, Junnan Wu, Rojesh Shrestha, Tamas Aranyi, Amy Kwan, Ayano Kondo, Marco Pontoglio, et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nature communications*, 12(1):1–17, 2021.

Gianni Monaco, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carre, Nicolas Burdin, Lucian Visan, Michele Ceccarelli, Michael Poidinger, et al. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell reports*, 26(6):1627–1640, 2019.

Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

Mineto Ota, Yasuo Nagafuchi, Hiroaki Hatano, Kazuyoshi Ishigaki, Chikashi Terao, Yusuke Takeshima, Haruyuki Yanaoka, Satomi Kobayashi, Mai Okubo, Harumi Shirai, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell*, 184(11): 3006–3021, 2021.

Jihwan Park, Rojesh Shrestha, Chengxiang Qiu, Ayano Kondo, Shizheng Huang, Max Werth, Mingyao Li, Jonathan Barasch, and Katalin Suszták. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*, 360(6390):758–763, 2018.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Kathryn G Roberts, Yongjin Li, Debbie Payne-Turner, Richard C Harvey, Yung-Li Yang, Deqing Pei, Kelly McCastlain, Li Ding, Charles Lu, Guangchun Song, et al. Targetable kinase-activating lesions in ph-like acute lymphoblastic leukemia. *New England journal of medicine*, 371(11): 1005–1015, 2014.

Yoshifumi Saisho. β-cell dysfunction: Its critical role in prevention and management of type 2 diabetes. *World journal of diabetes*, 6(1):109, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4):593–607, 2016.

Wolfgang J Streit, Heiko Braak, Qing-Shan Xue, and Ingo Bechmann. Dystrophic (senescent) rather than activated microglial cells are associated with tau pathology and likely precede neurodegeneration in alzheimer's disease. *Acta neuropathologica*, 118(4):475–485, 2009.

Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

Manjeri A Venkatachalam, Joel M Weinberg, Wilhelm Kriz, and Anil K Bidani. Failed tubule recovery, aki-ckd transition, and kidney disease progression. *Journal of the American Society of Nephrology*, 26(8):1765–1776, 2015.

Xiaojun Wang, Ryosuke Misawa, Mark C Zielinski, Peter Cowen, Junghyo Jo, Vipul Periwal, Camillo Ricordi, Aisha Khan, Joel Szust, Junhui Shen, et al. Regional differences in islet distribution in the human pancreas-preferential beta-cell loss in the head region in patients with type 2 diabetes. *PloS one*, 8(6):e67454, 2013.

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10 (1):1–9, 2019.

Yurong Xin, Jinrang Kim, Haruka Okamoto, Min Ni, Yi Wei, Christina Adler, Andrew J Murphy, George D Yancopoulos, Calvin Lin, and Jesper Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell metabolism*, 24(4):608–615, 2016.

Jian-Guo Zhou, Bo Liang, Su-Han Jin, Hui-Ling Liao, Guo-Bo Du, Long Cheng, Hu Ma, and Udo S Gaipl. Development and validation of an rna-seq-based prognostic signature in neuroblastoma. *Frontiers in oncology*, 9:1361, 2019.

Michael T Zimmermann, Ann L Oberg, Diane E Grill, Inna G Ovsyannikova, Iana H Haralambieva, Richard B Kennedy, and Gregory A Poland. System-wide associations between dna-methylation, gene expression, and humoral immune response to influenza vaccination. *PloS one*, 11(3):e0152034, 2016.

A

Table 7: *Average (overall)* RMSE between estimations and flow cytometry cell type fractions computed over all cell types. The reference dataset is *PBMC8k* for all methods.

Dataset	**MuSiC**	**CSx**	**Scaden**	**TAPE-O**	**TAPE-A**	**Linear MLPs**	**DISSECT**
SDY67	0.3 (0.348)	0.147 (0.15)	0.134 (0.13)	0.11 (**0.10**)	0.136 (0.12)	0.13 (0.128)	**0.089** (**0.10**)
Monaco bulk	0.247 (0.256)	0.125 (0.094)	0.081 (0.07)	0.086 (0.097)	0.065 (**0.068**)	0.08 (0.077)	**0.06** (0.069)
Monaco Microarray	0.27 (0.296)	0.151 (0.134)	0.139 (0.162)	0.15 (0.15)	0.17 (0.183)	0.285 (0.307)	**0.07** (**0.069**)
GSE65133	0.308 (0.381)	0.11 (0.121)	0.13 (0.149)	0.122 (0.119)	0.14 (0.136)	0.21 (0.253)	**0.06** (**0.058**)
GSE107572	0.245 (0.267)	0.12 (0.137)	0.105 (0.128)	0.086 (0.099)	0.116 (0.128)	0.11 (0.11)	**0.068** (**0.076**)
GSE120502	0.358 (0.458)	0.12 (0.12)	0.104 (0.121)	0.087 (0.093)	0.093 (0.102)	0.143 (0.15)	**0.086** (**0.087**)

Table 8: RMSE between estimates from different methods and flow cytometry for granular cell type fractions in *Monaco bulk*.

Celltype	**MuSiC**	**CSx**	**Scaden**	**TAPE-O**	**TAPE-A**	**Linear MLPs**	**DISSECT**
B Ex	**0.01**	0.05	0.04	0.04	**0.01**	0.02	0.02
B NSM	0.02	**0.01**	0.02	0.03	0.05	0.04	0.02
B Naive	**0.01**	0.03	0.03	0.03	0.04	0.06	0.03
B SM	0.05	**0.01**	**0.01**	0.02	0.02	0.03	**0.01**
Monocytes C	0.05	**0.02**	0.04	**0.02**	**0.02**	0.06	0.06
Monocytes I	0.12	0.15	**0.03**	0.06	0.10	0.12	0.04
Monocytes NC	0.20	0.09	**0.02**	0.07	0.05	0.10	**0.02**
NK	**0.05**	0.08	0.08	0.11	0.11	0.05	0.08
Neutrophils LD	0.02	0.03	**0.01**	**0.01**	**0.01**	0.01	0.02
Plasmablasts	**0.01**	**0.01**	0.02	**0.01**	0.04	0.01	0.04
CD4 T Naive	**0.02**	0.03	0.05	0.05	**0.02**	0.02	0.05
CD4 T Memory	0.10	0.07	**0.03**	0.12	0.15	0.21	**0.03**
CD8 T Naive	0.21	0.07	**0.04**	0.05	0.05	0.01	**0.04**
CD8 T CM	**0.01**	0.08	0.02	0.05	0.12	0.01	0.03
CD8 T EM	0.02	**0.01**	0.02	0.02	**0.01**	0.03	0.02
CD8 T TE	**0.01**	0.07	0.09	0.08	0.11	0.16	0.08
mDC	**0.01**	0.04	0.04	0.08	0.09	0.03	0.03
pDC	0.02	**0.00**	0.02	0.01	0.01	0.01	0.02
Average	0.06	0.05	**0.03**	0.05	0.06	0.05	0.04

B NOTE ON AVERAGE AND OVERALL CORRELATIONS FOR PERFORMANCE EVALUATION.

Overall metrics can be deceiving when the model performs comparatively very well or very poor on a largely abundant cell type as illustrated by results in Figures 2 and 3. Overall metrics are provided for completeness as it has been extensively used in evaluation of other deconvolution works and benchmarks (Avila Cobos et al., 2020; Newman et al., 2019; Wang et al., 2019; Chen et al., 2021).

C QUALITY CONTROL AND PREPROCESSING

C.1 QUALITY CONTROL:

Before simulating from reference datasets, we remove cells with less than 200 expressed genes and genes which are expressed in less than 3 cells. Further, we also remove cells expressing more

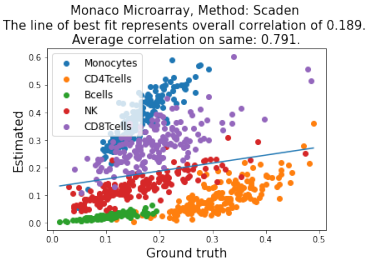

Figure 2: Correlations between Ground truth fractions from *Monaco Microarray* (Table 1) and the estimations using Scaden.

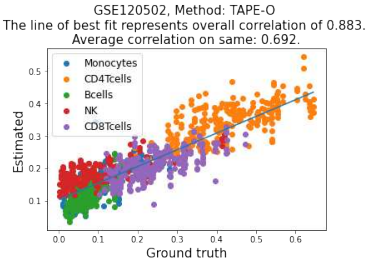

Figure 3: Correlations between the ground truth fractions from *GSE120502* (Table 1) and the estimations using TAPE-O.

than 4% mitochondrial genes. Thereafter, before each deconvolution, we subset reference and bulk datasets to include only the common genes between the two.

C.2 Simulations

For deep learning methods, we sampled $\alpha_{k,i}$ uniformly to simulate based on procedure described in Section 2 with $\sum_{k=1}^{c} \alpha_{k,i} = 100, \forall i$ if the dataset is single-cell. For experiments on granular level cell types where simulations are done from purified cell samples, we modify simulation procedure to reflect this. In this case, a simulated sample is given by $\mathbf{B}_{i\cdot}^{\text{sim}} = \sum_{k=1}^{c} \mathbf{X}_{ik}^{\text{sim}} \mathbf{b}_{\mathbf{l}}^{\mathbf{k}}$, where $\mathbf{b}_{\mathbf{l}}^{\mathbf{k}}$ is the expression vector of purified sample l belonging to cell type k. All other notations are same as in Section 2.

For all experiments, we simulated total $1000 \times c$ simulations where c is number of cell types in the reference dataset.

C.3 Preprocessing:

Estimation of cell type fractions:

Before passing simulated and real bulk samples to the network, we normalize samples to sum to a million counts (CPM: Counts per million) and log scale them with base 2 after adding 1. CPM normalization was performed to maintain total mRNA expressed per gene to be out of a fixed total gene expression, and CPM is widely used in computational genomics. During training, for each batch, we normalize each sample by $MinMax$ scaling. These are standard preprocessing steps (Menden et al., 2020).

Estimation of cell type specific gene expression profiles: To estimate cell type specific gene expression profiles, we need to maintain relationship between gene expression of individual cell types and simulated bulks, which would be lost if we perform CPM normalization of both simulated samples and corresponding cell type specific gene expression profiles. Hence, instead of performing CPM normalization of simulated bulks, we normalize each test bulk sample to sum to the mean of sums of simulated samples. Further, for estimating cell type specific gene expression, we want to maintain gene level information across samples. To achieve this, instead of normalizing each sample using $MinMax$ scaling, we perform $MinMax$ scaling globally over all samples.

D Cell subsets for estimations of granular subsets on Monaco bulk.

B cells (4 subsets): B Naive (Naive B cells), B Ex (Exhausted B cells), B NSM (Non-switched memory B cells), B SM (Switched memory B cells)

CD4 T cells (2 subsets): CD4 T Naive (Naive CD4 T cells), CD4 T Memory (Memory CD4 T cells),

CD8 T cells (4 subsets): T CD8 Naive (Naive CD8 T cells), CD8 T CM (Central Memory CD8 T cells), CD8 T TE (Terminally effector CD8 T cells), and CD8 T EM (Effector Memory CD8 T cells),

Monocytes (3 subsets): Monocytes C (Classical monocytes), Monocyte NC (Non classical monocytes) and Monocytes I (Intermediate monocytes),

Dendritic cells (2 subsets): mDC (myeloid dendritic cells), pDCs (Plasmacytoid dendritic cells)

Plasmablasts

Neutrophils LD (Low density neutrophils)

NK cells.

Table 9: Harmonization of cell subset labels between *Monaco bulk* and reference from Ota et al. (2021)

Cell subset	**Subset(s) in flow cytometry of *Monaco bulk***	**Merged subset(s) in Ota et al. (2021)**
B Ex	B Ex	DN_B
B NSM	B NSM	USM_B
B Naive	B Naive	Naive_B
B SM	B SM	SM_B
Monocytes C	Monocytes C	CL_Mono
Monocytes I	Monocytes I	Int_Mono
Monocytes NC	Monocytes NC	NC_Mono
NK	NK	NK
Neutrophils LD	Neutrophils LD	LDG
Plasmablasts	Plasmablasts	Plasmablast
CD4 T Naive	T CD4 Naive	Naive_CD4
CD4 T Memory	Th1, Th2,Th17,Th1/Th17, Tregs	Th1, Th2, Th17, Tfh, Fe_II_eTreg
CD8 T Naive	T CD8 Naive	Naive_CD8
CD8 T CM	T CD8 CM	CM_CD8
CD8 T EM	T CD8 EM	EM_CD8
CD8 T TE	T CD8 TE	TEMRA_CD8, Mem_CD8
mDC	mDCs	mDC
pDC	pDCs	pDC

E RELATIONSHIP BETWEEN CELL TYPE FRACTIONS AND BIOLOGICAL PHENOTYPES.

To assess the performance of DISSECT and other algorithms on further bulk RNA-seq datsaets, we rely on established biological findings. Table 10 lists these datasets and corresponding hypothesis based on literature. To deconvolve these datasets, we used single-cell datasets listed in table 11. The cell types present in each single cell dataset is provided in the description of each corresponding figure (Figures 4 - 7.)

For *GSE50244*, we performed deconvolution using three reference single-cell datasets. These references differ in technologies with which the cells were sequenced. These technologies are *Baron*: inDrop, *Segerstolpe*: Smart-seq2, and *Xin*: SMARTer. Further, these reference datasets contain cells belonging to different states. We used cells belonging to only healthy individuals in *Baron* and *Segerstolpe* while both healthy and T2D individuals are used in *Xin*. Figure 4 shows the distribution of alpha, beta, gamma and delta cells for each deconvolution method. There is a lack of concordance between three distributions of cell types across all methods. Beta and alpha cells are generally the two most abundant of these four cell types in pancreatic islets (Henquin & Rahier, 2011). This is correctly observed with estimations from the considered deep learning based methods. However, for Scaden, the relative proportions of alpha and beta cells are inverted between *Baron* and *Xin*. While for DISSECT, they are predicted almost at the same level for the three datasets, which more varying

Table 10: Details on bulk RNA-seq datasets used to evaluate deconvolution methods on biological phenotypes. Biological hypotheses based on literature serve as proxy ground truths.

Tissue	**Dataset**	**# samples**	**Biological hypothesis based on literature**	**original Source**
Pancreas	GSE50244	89 (77 with information on hemoglobic 1C levels)	Fraction of beta cells are negatively associated with severity of type 2 diabetes indicated by hemoglobin A1c (hba1C) level (Alejandro et al., 2015; Saisho, 2015; Wang et al., 2013).	Fadista et al. (2014)
Kidney	GSE81492	10	Tubule cells diminish with chronic kidney disease (CKD) (Venkatachalam et al., 2015; Liu et al., 2018; Malhotra et al., 2019).	Beckerman et al. (2017)
Brain	ROSMAP	508 (463 with corresponding annotation of Braak stages)	1. Neurodegeneration with advanced Braak stage (Streit et al., 2009; Hindle, 2010; Fu et al., 2019), and 2. Approximately 70-30 ratio of excitatory and inhibitory neurons (Contreras, 2004; Chen & Dzakpasu, 2010).	Bennett et al. (2018)

Table 11: Details on single-cell datasets used to deconvolve corresponding tissue samples.

Tissue	**Dataset name**	**# cells**	**Original Source**
Pancreas	Baron	8569	Baron et al. (2016)
Pancreas	Segerstolpe	3514	Segerstolpe et al. (2016)
Pancreas	Xin	1492	Xin et al. (2016)
Kidney	Park	43745	Park et al. (2018)
Kidney	Miao	16887	Miao et al. (2021)
Brain	Allen Brain Atlas	49418	bra.

beta cell fractions when *Baron* is the reference. TAPE-O also achieves this trend, however, TAPE-O incorrectly predicts Delta cells as being at the same level as alpha cells for reference *Baron*. Linear MLPs show the most variance and predict almost 80% alpha cells for *Segerstolpe* and *Xin*. Further, with Linear MLPs, the least abundant gamma and delta cells are predicted to be negative for almost all samples. Next, we looked at the association between beta cell fractions and T2D severity (Table 10). Across all three datasets, DISSECT estimations are significantly negatively correlated with hb1ac. As in other experiments, we observe a wide discrepancy between TAPE-O and TAPE-A results.

Next, we deconvolved *ROSMAP* using reference from *Allen Brain Atlas*. The results are presented in Figure 6. Nearly all methods capture a negative association between the median fractions of excitatory neurons and Braak stages. Scaden, TAPE-O, DISSECT maintain a higher proportion of excitatory neurons compared to inhibitory neurons. However, DISSECT estimates show the excitatory-inhibitory neurons ratio to be almost 70-30. TAPE-A and MuSiC on the other hand show opposite of what is expected.

Lastly, we deconvolved *GSE81492* using two single cell reference datasets *Park* and *Miao*. We present the results per method per reference in Figure 7. Both of these datasets come from same technologies (10x Genomics). Despite this, almost all methods show variation in their estimates when changing the reference, wit DISSECT showing the least variance and giving similar associations between cell type fractions and tissue state.

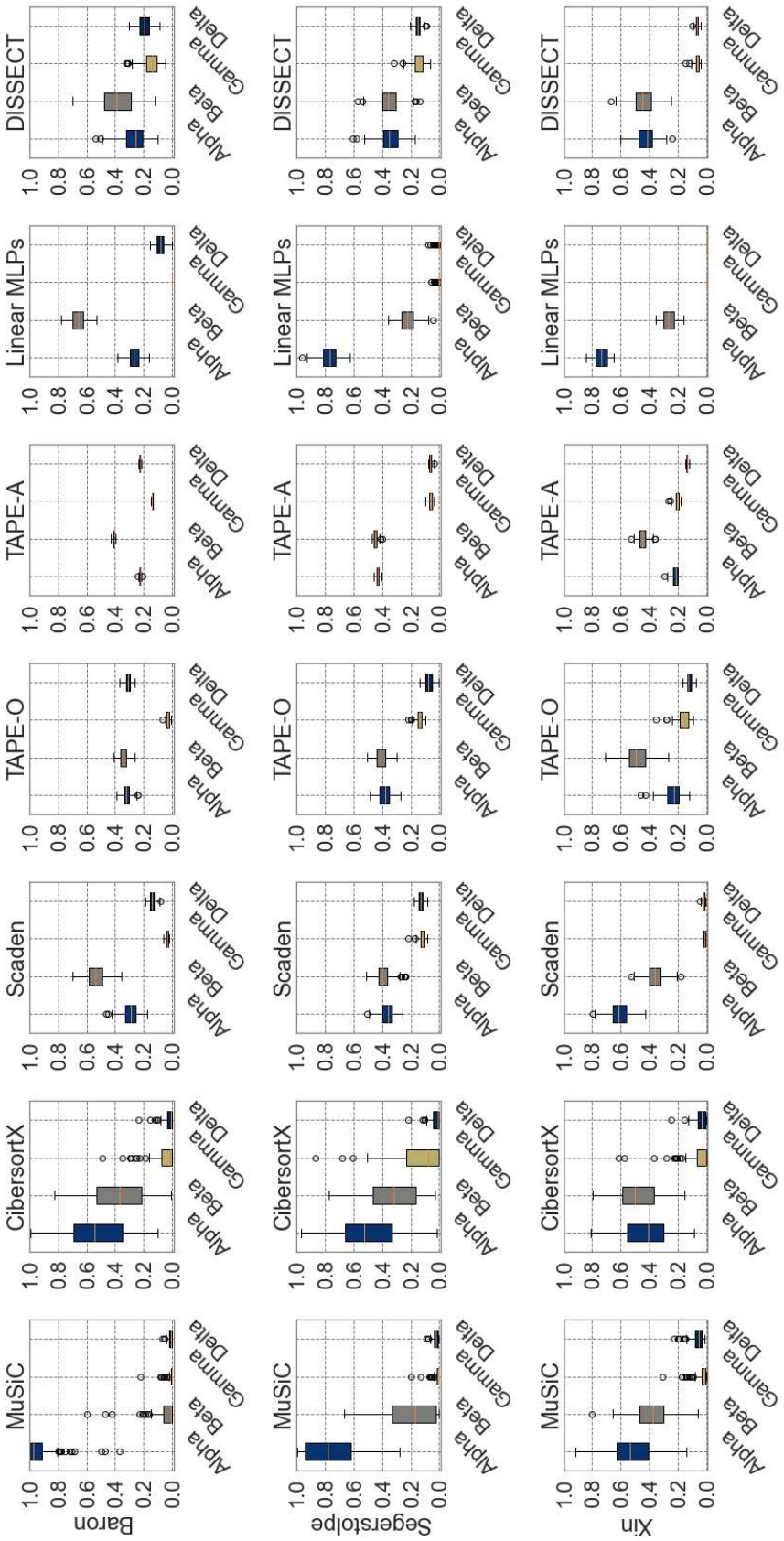

Figure 4: Distribution of relative fractions of alpha, beta, gamma and delta cells (present in all three single cell datasets) estimated on 77 bulk RNA-seq samples from *GSE52044* using three different single-cell reference datasets. Each row corresponds to the single-cell reference dataset.

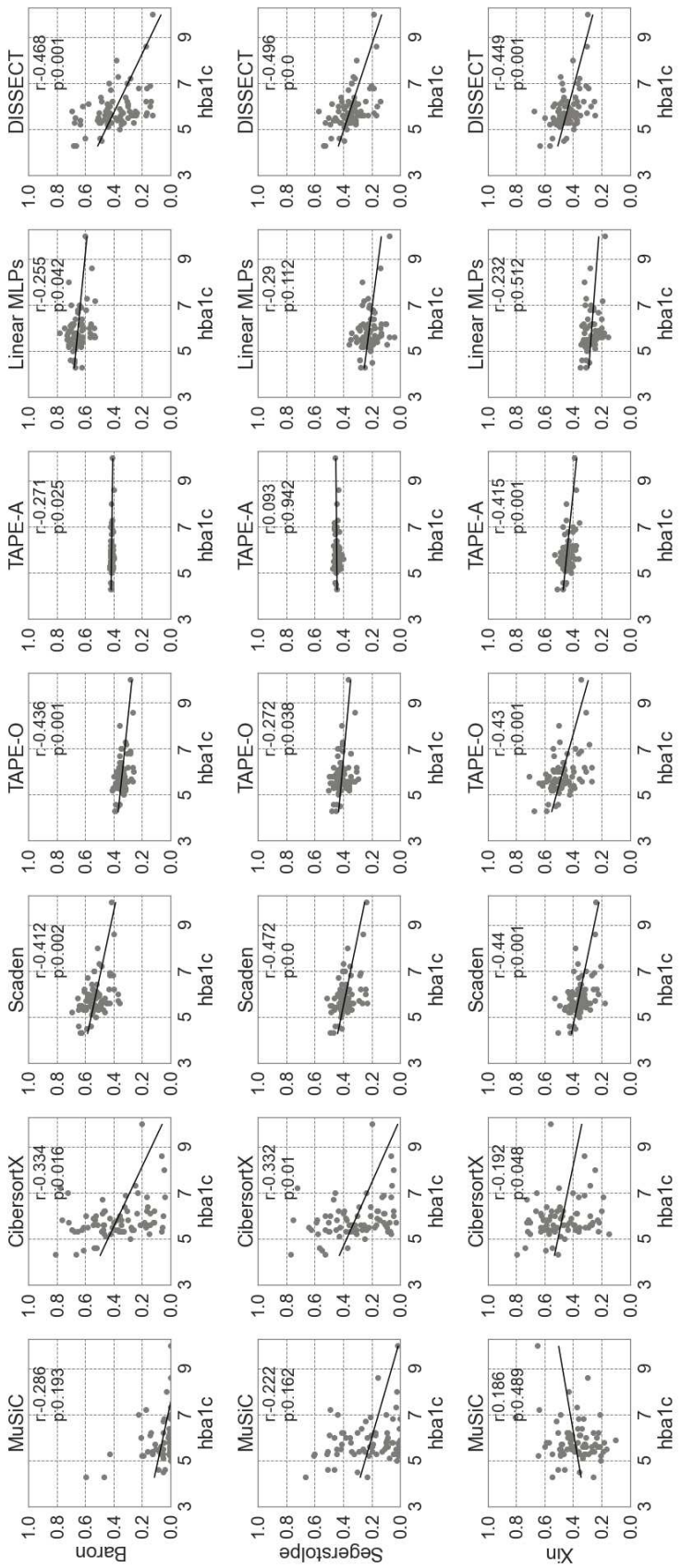

Figure 5: Relationship between Hemoglobin A1C (hba1c) levels and estimated fractions of beta cells in 77 bulk RNA-seq samples from *GSE52044* using three different single-cell reference datasets. Corresponding Method for each plot is indicated in the title. Each row corresponds to the single-cell reference dataset. Pearson correlation of the relationship and associated p-value is indicated in the plot. p-value shown is obtained for beta cells from a multiple linear regression model considering age, sex and bod mass index (BMI) as covariates, *i.e.* model hb1ac $\sim$ Constant + fractions of beta cells + Age + Sex + BMI.

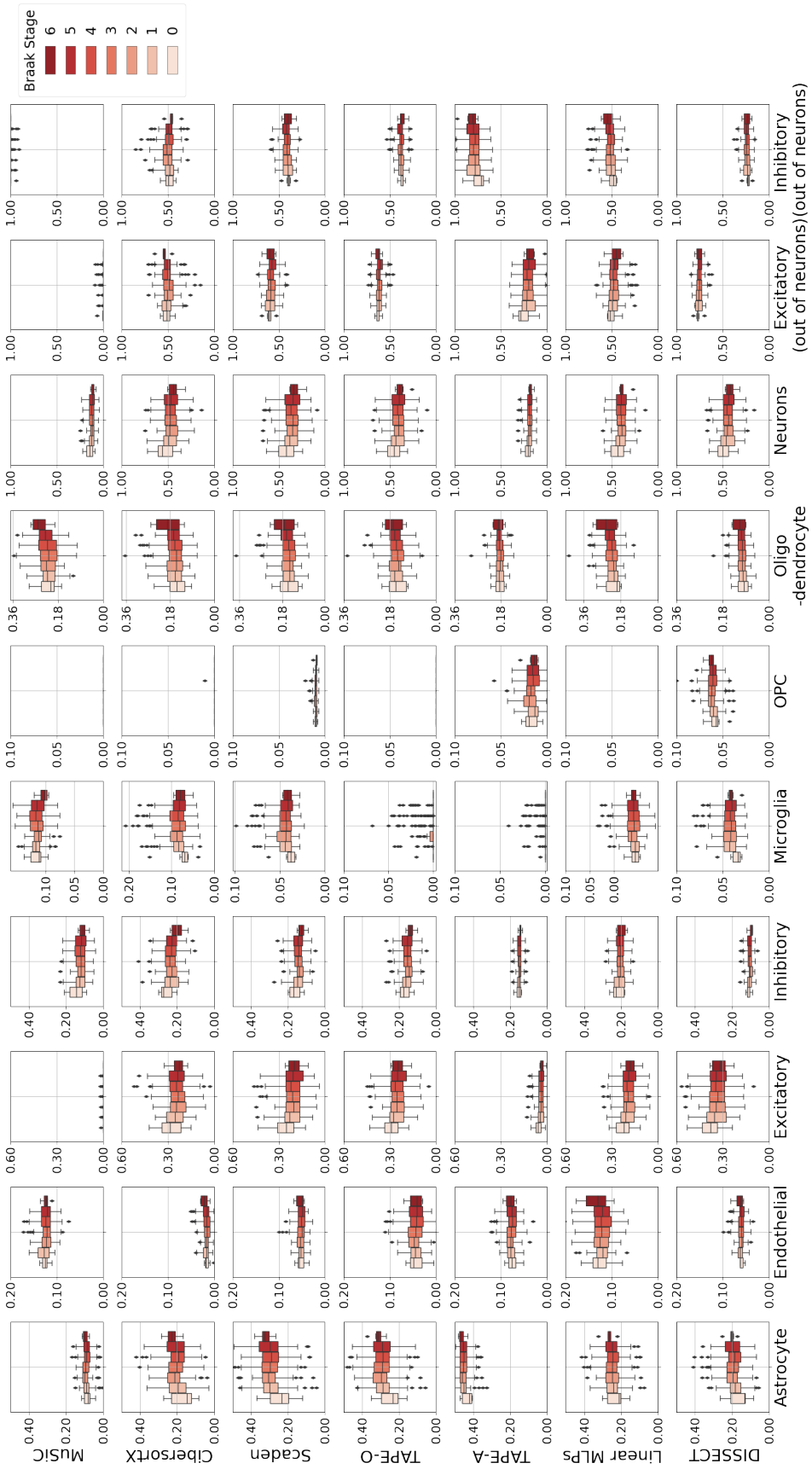

Figure 6: Fractions of cell types in 463 bulk RNA-seq samples from ROSMAP Alzheimer's Disease cohort for whom corresponding Braak stages are available. Allen Brain Atlas is used as reference single-cell data (Table 11). Rows indicate methods and each column is a cell type. *OPC:* Oligodendrocyte Precursor Cells. Excitatory and Inhibitory are two neuron subsets. Last two columns show fractions of excitatory and inhibitory neurons out of total neuronal content.

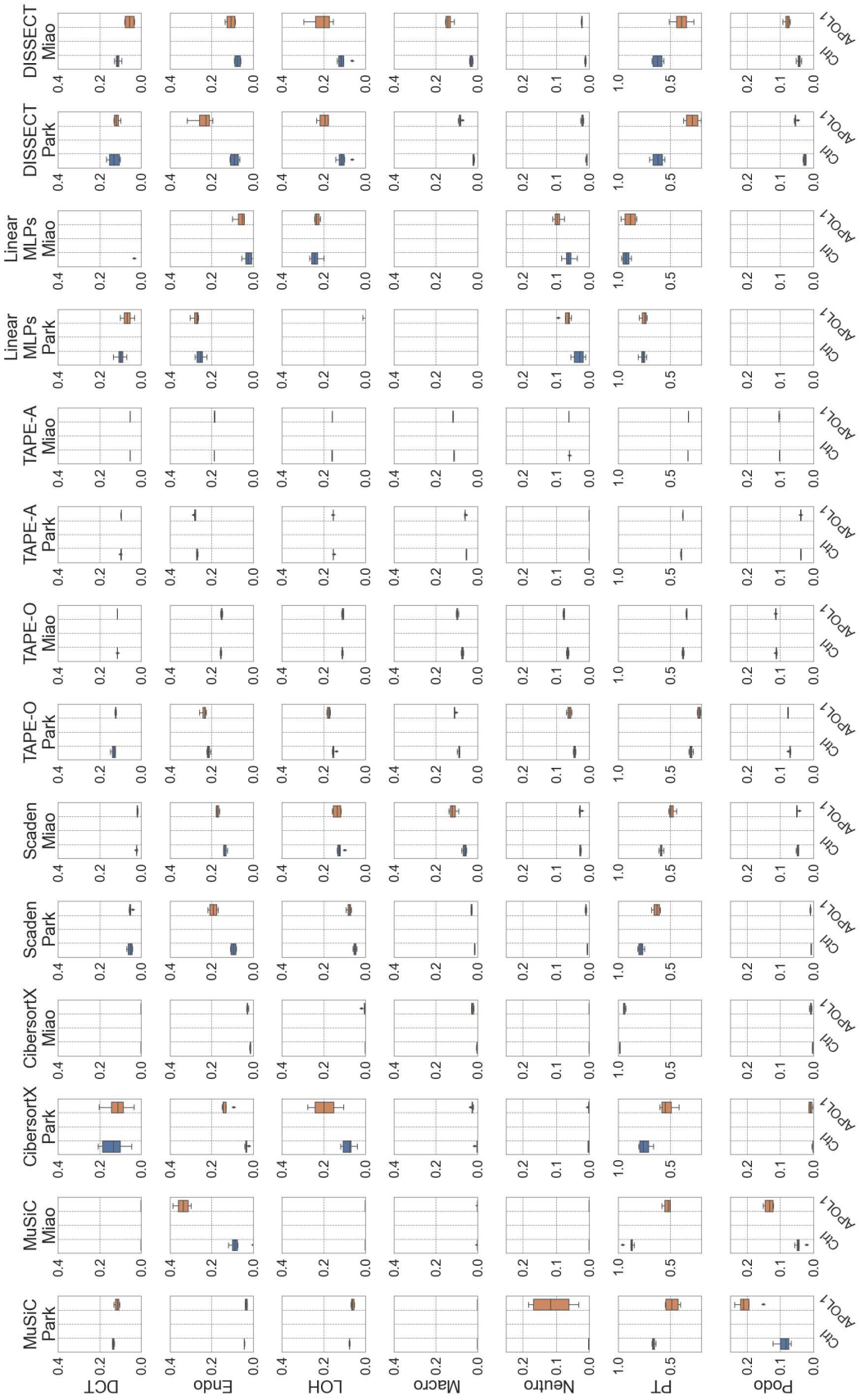

Figure 7: Estimated fractions of different cell types in 10 bulk RNA-seq samples from *GSE81492* (Ctrl: Control mice, n=6 and APOL1: Apolipoprotein L1 transgenic mice, n=4) mice using two separate single-cell reference datasets (*Park* and *Miao*). Corresponding Methods and reference datasets for each plot are indicated in the title. Each row corresponds to a cell type. DCT: Distal convoluted tubule, Endo: Endothelial cells, LOH: Loope of Henle, Macro: Macrophages, Neutro: Neutrophils, PT: Proximal Tubule, Podo: Podocytes

F APPLICATION TO SPATIAL TRANSCRIPTOMICS

Here we illustrate applicability of DISSECT on spatial transcriptomics (ST). We focused on two publicly available tissue slides (Mouse brain and human lymph node) on which ST has been performed.

In brain, cortical neuronal layers are structured spatially. To verify whether DISSECT estimates may be valid in ST, we deconvolved a sagittal mouse brain slice available as part of Seurat [3]. As reference, we used a mouse brain single-cell dataset from Allen Brain Institute consisting of approximately 14000 cells sequenced using Smart-seqv2 protocol (Tasic et al., 2016). We adjusted simulation procedure to mimic ST datasets. 10x Visium (one of the technologies to generate ST samples) consists of around 10 cells per spot [4] [5].To reflect this, we simulated between 5-12 cells to generate one spot (*i.e.* $\sum_{k=1}^{c} \alpha_{ki} \sim [5, 12]$). Since ST is much sparser, to generate one spot, we kept between 2-6 cell types. Figure 8 shows fractions of cell types overlaid on the hematoxylin and eosin (H&E) stained images of tissue slide, and Figure 9 shows jointly cortical neuronal proportions which shows a spatially structured arrangement of neurons.

To evaluate how DISSECT behaves on ST deconvolution on granular level subsets, we deconvolved a human lymph node slide using corresponding integrated single-cell datasets on which granular level cell types are annotated. Both of these datasets are obtained from (Kleshchevnikov et al., 2022). Remarkably, DISSECT is able to identify spatial patterns of cell type fractions, along with correctly predicting co-localization of cycling B cells and germinal center B cells. This is illustrated in Figures 11 and 12. Several germinal center zones are correspondigly visible in h&E stained image (Figure 11).

These results demonstrate the usability of DISSECT on spatial deconvolution and warrant evaluation of consistency loss further in data modalities other than bulk RNA-seq.

[3] https://satijalab.org/seurat/articles/spatial_vignette.html

[4] Each spot is a location which is sequenced in tissue slide. Thus, each spot is analogous to a bulk RNA-seq, albeit much sparser due to less number of cells per spot.

[5] https://kb.10xgenomics.com/hc/en-us/articles/360035487952-How-many-cells-are-captured-in-a-single-spot-

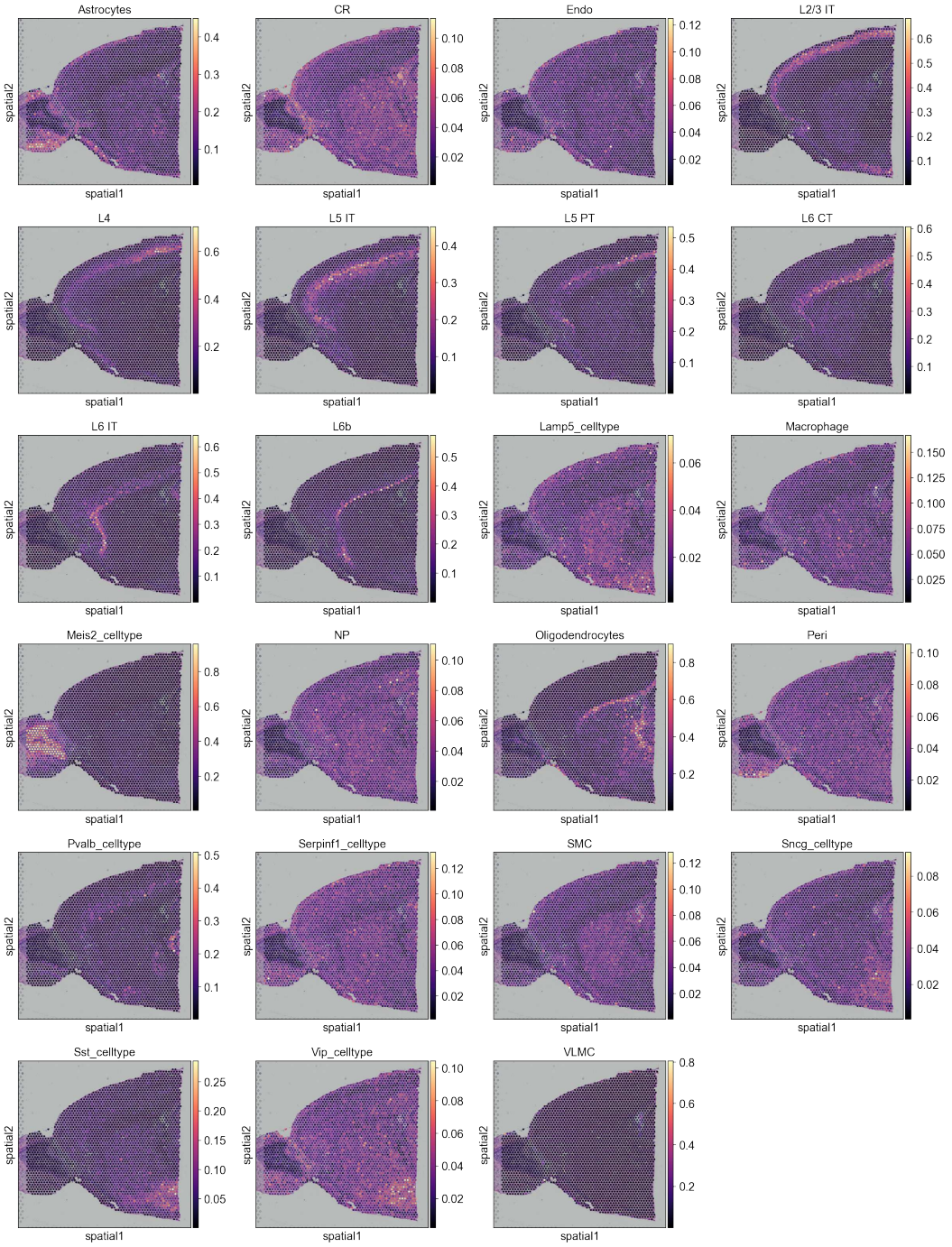

Figure 8: Fractions of all cell types shown on H&E image. Each dot corresponds to a spot sequenced in the tissue slice.

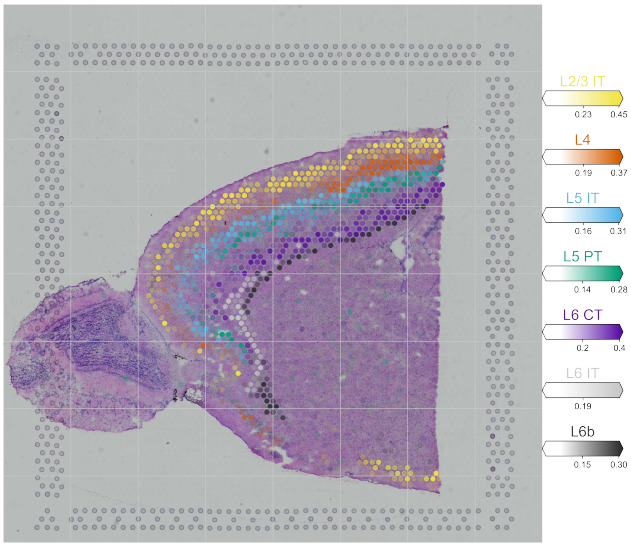

Figure 9: Cortical neurons overlaid on H&E image. The legend shows fractions.

F.1

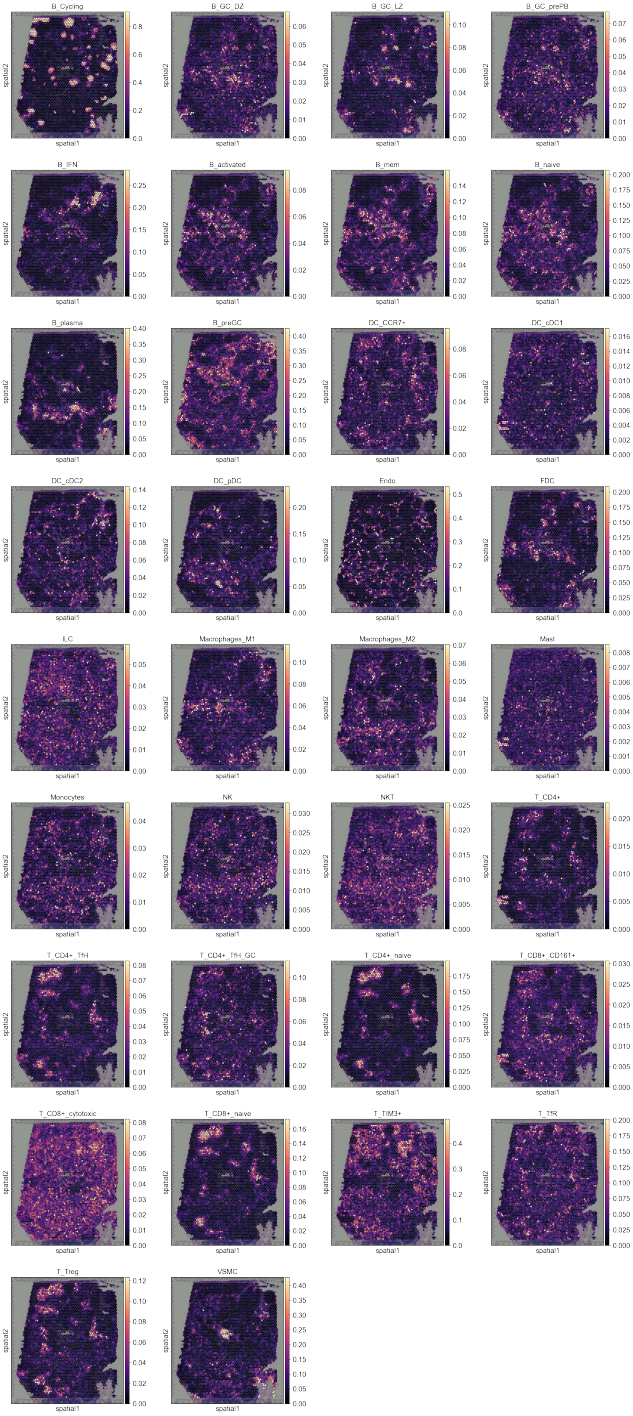

Figure 10: Estimated fractions of 34 granular level cell types overlaid on H&E image of a lymph node tissue slice.

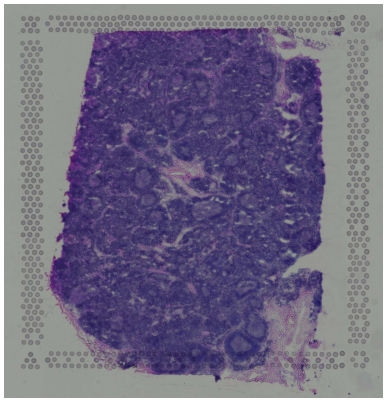

Figure 11: H&E image of lymph node tissue slice.

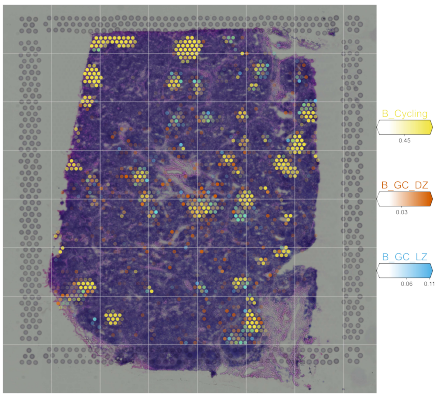

Figure 12: Fractions of Cycling and light zone (LZ) and dark zone (DZ) Germinal center B cells expected to be present in germinal centers.

G

G.1 Ablation

Table 12: Average performance over five random experiments for SDY67 (Table 1.) Each column indicates the additional part.

Metric	Linear MLP	Activations	KL Divergence	KL Divergence + Consistency
r	0.51 ± 0.018	0.55 ± 0.016	0.54 ± 0.006	0.63 ± 0.005
rmse	0.13 ± 0.008	0.13 ± 0.006	0.11 ± 0.004	0.09 ± 0.002