Khatri *et al. Genome Biology* (2024) 25:112
https://doi.org/10.1186/s13059-024-03251-5

Genome Biology

METHOD

Open Access

DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation

Robin Khatri[1], Pierre Machart[1] and Stefan Bonn[1*]

*Correspondence:
sbonn@uke.de

[1] Institute of Medical Systems Biology, Center for Molecular Neurobiology, Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract

Cell deconvolution is the estimation of cell type fractions and cell type-specific gene expression from mixed data. An unmet challenge in cell deconvolution is the scarcity of realistic training data and the domain shift often observed in synthetic training data. Here, we show that two novel deep neural networks with simultaneous consistency regularization of the target and training domains significantly improve deconvolution performance. Our algorithm, DISSECT, outperforms competing algorithms in cell fraction and gene expression estimation by up to 14 percentage points. DISSECT can be easily adapted to other biomedical data types, as exemplified by our proteomic deconvolution experiments.

Keywords: Cell deconvolution, Semi-supervised learning, Deep learning

Background

A prominent approach to studying 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 and with different gene expression programs. Consequently, 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 understanding disease mechanisms and identifying potential therapeutic interventions [1].

A recent technological advancement, single-cell RNA-seq, allows for investigating gene expression in single cells for thousands of individual cells of a given tissue sample in a single experiment. However, while it provides unprecedented insights into single-cell biology, it suffers from severe technical limitations, most notably the presence of zero values in gene expression due to methodological noise, termed as "dropouts" [2]. In addition, the technology is still very costly, which essentially 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 [3, 4].

Computational inference of cell type fraction and cell type-specific gene expression is a source-separation task, termed as "cell deconvolution" within the context of cell biology. The estimation of cell type-specific gene expression is a well established and challenging problem in the field. Prior work includes but is not limited to TAPE [5], bMIND [6], BayesPrism [7], and CibersortX (CSx) [8]. The basic aim is to provide cell type-specific gene expression information at a group or sample level. The resultant information allows deep biological insights into cell type-specific gene expression and pathway changes from bulk data. For cell deconvolution, recent computational methods utilize single-cell sequencing data to create simulated references with known fraction and expression for training [9]. While this approach achieves good deconvolution results, its performance suffers from the substantial domain shift between single-cell RNA-seq training (reference) data and the bulk RNA-seq target data. Domain refers to the statistical distribution of the source of a dataset [10]. Domain shift refers to a change in the statistical distribution of samples, which can be due to covariate shift, the presence of open sets, or both. In gene expression datasets, the covariate shift between real data and simulated datasets occurs due to changes in cell type-specific gene expression and can arise from different dropout rates and tissue conditions, for instance. When domain shifts have purely technical reasons, they are often termed batch effects. CSx [8] has previously approached the problem of batch effect removal between single cell gene expression datasets [11], using Combat [12] to remove changes in cell type-specific gene expression between a single-cell reference signature matrix and bulk RNAseq data. Open sets may occur when new cell types are encountered during test time, such as the presence of differing cell lineages [13]. Since cells go through different differentiation states, domain shift between real data and simulations may be a combination of both, the covariate shift and presence of open sets. Among many possible sources of domain variation, the most prevalent might be the presence of batch effects that refer to technological differences between two sequencing experiments and gene expression differences of biological nature.

In this work, we first formally define the task of cell deconvolution and outline the hypothesis that semi-supervised consistency regularization should improve bulk RNA-seq deconvolution when learning from single cell RNA-seq data. We then provide evidence that two novel deep learning algorithms with semi-supervised consistency regularization outperform competing state-of-the-art algorithms in deconvolution, both on a cellular and gene expression level, across a wide range of datasets. On the datasets with ground truth flow cytometry cell type proportions, DISSECT achieves consistently better Jensen-Shannon distance (*JSD*): 0.063 ± 0.015 and root mean squared error (*rmse*): 0.021 ± 0.019. In addition, DISSECT shows state-of-the-art gene expression deconvolution performance, achieving the best sample- and gene-wise correlations. Our algorithm can easily be adapted to other biomedical data types, as exemplified by our bulk proteomics and spatial expression deconvolution experiments.

Results

In this section, we first formally define the cell deconvolution task, then present our hypothesis and DISSECT deep learning models, and compare DISSECT's performance to other state-of-the-art deconvolution algorithms.

Task of 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 deconvolution is to find an $m \times c$ matrix $\mathbf{X}$ of cell type fractions, where c is the number of cell types present in bulk samples such that,

$$\mathbf{B} = \mathbf{XS}, \tag{1}$$

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 by identifying the most representative cell type specific gene expression [8].

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 a 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 previously in Scaden [9], 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 the 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 the proportion of a cell type k in that sample i and is defined as,

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

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 $\mathbf{e}_\mathbf{l}^\mathbf{k}$ 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 $\mathbf{B}$.

Exploiting the linearity of deconvolution

The deconvolution task is to learn a cell type-specific gene-expression matrix (or signature matrix) $\mathbf{S}$, which serves to accurately predict cell fractions and their corresponding gene expression from a bulk gene expression matrix $\mathbf{B}$. The actual mixing process of cells to form a tissue is assumed to be linear and, as such, the relationship between $\mathbf{B}$ and $\mathbf{S}$ is linear. However, $\mathbf{S}$ is unobserved, and the deconvolution algorithm is learned using simulations. This learning process involving simulations is highly dependent on the reference being the single-cell dataset used to generate simulations, and is subjected to an inherent strong domain shift [14]. To address this, we hypothesize that a consistency-based regularization penalizing the non-linearity of mixtures of real and simulated samples would result in a mapping $\hat{f}$ that is closer to true mapping f. Non-linearity of mixtures of real and simulated samples refers to the violation of Eq. 4, defined later, for estimated $\mathbf{X}_{i\cdot}$, $\mathbf{X}_{i\cdot}^{\text{sim}}$ and $\mathbf{X}_{i\cdot}^{\text{mix}}$ using mapping f.

Consistency regularization

Consider that $\mathbf{B}$ represents gene expression matrices of real (test) bulk RNA-seq that we want to deconvolve and and $\mathbf{B}^{\text{sim}}$ represents gene expression matrix of simulated bulk samples. The number of rows (representing samples) in these two matrices may differ. To simplify the notation, we use the same index i to denote indices for real bulk samples, simulations (sim) and their mixtures (mix, defined further). Given a true bulk RNA-seq sample $\mathbf{B}_{i\cdot}$, and a simulated sample $\mathbf{B}_{i\cdot}^{\text{sim}}$ with paired proportions $\mathbf{X}_{i\cdot}^{\text{sim}}$ defined over a common set of genes, 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 the relation

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

where $\mathbf{X}_{i\cdot}$ represents cell fractions of sample i and where $\beta \in [0, 1]$. Cell types are characterized by a few marker genes that are invariant across cell states and even across tissues [15]. A network that accurately predicts cell type fractions based on gene expression of simulated or real bulk RNA-seq data would thus have to learn them. In the estimation of cell type fractions, we therefore 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 $\mathbf{S}_{\cdot i\cdot}^{\mathbf{sim}}$. Hence,

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

Equation 4 serves as the formulation to generate pseudo ground-truths for these mixtures during learning, and it 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. The number of bulk RNA-seq samples is, however, considerably lower (tens to hundreds) than the amount of single-cells present in a single-cell experiment (thousands or more). Equation 4 allows to generate pseudo ground truth proportions for mixtures $\mathbf{B}^{\text{mix}}_{i\cdot}$ at each step of learning cell type fractions, while Eq. 3 allows to generate pseudo ground truth signatures at each step of learning gene expression profiles.

Network architecture and learning procedure

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 the 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 and involve complete consistency regularization (Eq. 3). An illustration of the method is presented in Fig. 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). The ensembling is performed to reduce the variance by averaging different runs [16]. Each MLP consists of the same architecture initialized with different weights. 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) [17, 18] was chosen out of tested activations over grid search on (Linear, ReLU, ReLU6, Swish [19]). The final

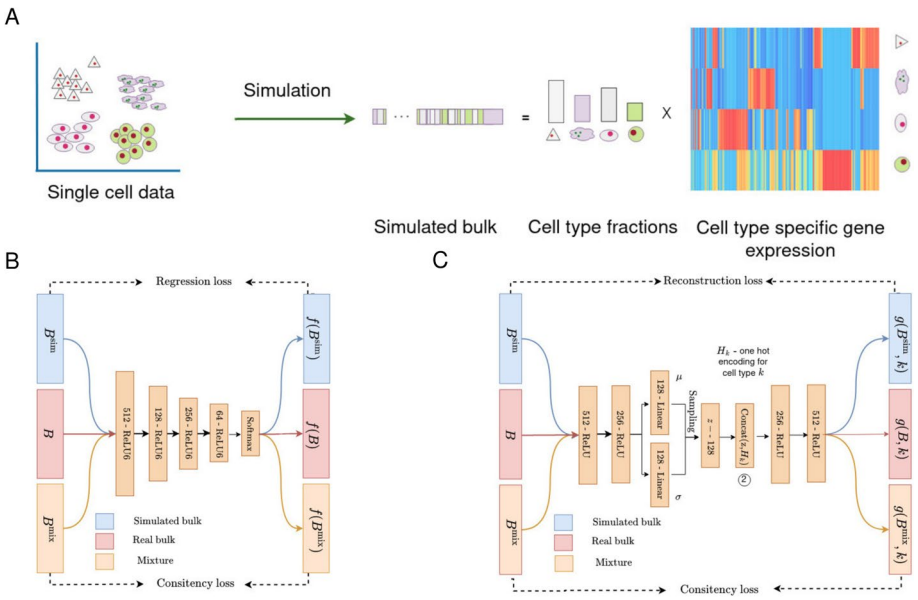

Fig. 1 **A** Illustration of the simulation procedure using reference single-cell data. The figure shows the simulation of one sample which consists of cell type fractions, simulated gene expression and cell type specific gene expression profiles (i.e., signature matrix). **B** Detailed overview of an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

application of a softmax activation function allows to achieve the non-negativity and sum to 1 criteria of deconvolution. We train the network with batch size 64 to minimize the loss function per batch defined below with an Adam Optimizer with initial learning rate of $1e-5$.

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

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:

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

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

To generate mixtures, for each batch, we sample β uniformly at random for Eq. 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. Since the number of simulations is generally larger (in our experiments, set to 1,000 times the number of cell types) than that of real data, we sample real data to create additional bulk samples, $\mathbf{B}_{i\cdot}$, until the size equals that of the simulated data, $\mathbf{B}_{i\cdot}^{\text{sim}}$. This pair of data together with simulated proportions, $\mathbf{X}_{i\cdot}^{\text{sim}}$, is then used to create training batches of size 64. For every batch, we generate mixtures according to Eq. 2.

Our loss is inspired by MixMatch [20], which uses unlabelled samples to mix up and match sample predictions. Our adaptation in Eq. 5 addresses the limited samples available from true bulk RNA-seq, unavailability of sample fractions and is derived from the definition of the task itself. In essence, Eq. 5 integrates domain knowledge into the objective.

To avoid a scenario where the network does not 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 Eq. 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 limit the number of steps to 5000 as found optimal in Scaden [9].

Estimation of per sample 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 between samples. Previously, works such as CSx [8] and TAPE [5] have explored utilizing cell type fractions to estimate gene expression per sample. 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. This allows for training a single model to estimate gene expression of each cell type for a sample. 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:

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

where $\mathbf{B}_{i}^{\text{mix}}$ is given by Eq. 2, and $f(\mathbf{B}_{i\cdot}^{\text{mix}})_k$ is the proportion of cell type k in sample i as estimated during cell type fraction estimation and is fixed during training. 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:

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

where $\mathcal{N}(0,1)$ is standard normal distribution, and 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. Under default settings used throughout this work, we train the network to minimize the loss function with an Adam optimizer with initial learning rate of $1e-3$, and the values for hyperparameters λ_2 and β^{VAE} are respectively 0.1 and $1e-2$. The network is trained for $5000 \times k$, k being the number of cell types.

Estimation of cell type fractions and comparison with flow cytometry

To quantitatively assess the deconvolution algorithm, we first deconvolve six different peripheral blood mononuclear cells (PBMC) bulk datasets for which cell type proportions have already been quantified using flow cytometry (Additional file 1: Table S1). To evaluate deconvolution performance, we utilize root-mean-squared error (*rmse*) and Pearson correlation (r) for cell type-wise comparisons and Jensen-Shannon distance

(*JSD*) for sample-wise comparisons between estimated fractions and ground truth proportions. The evaluation metrics are defined in the "Evaluation metrics" section. To evaluate our approach, we compared it to state-of-the-art deconvolution methods, MuSiC [21], CSx [8], Scaden [9] and TAPE (TAPE-O and TAPE-A) [5], BayesPrism and BayesPrism-M [7], and bMIND [6]. MuSiC and CSx were chosen for their best performances in benchmarking studies [22, 23]. Scaden and TAPE are selected as both are deep learning-based deconvolution approaches, the latter of which, TAPE-A, performs an adaptation of the network weights for test samples. Since deconvolution is linear, we also considered linear MLPs as a deconvolution algorithm. Further details can be found under the "State of the art" section.

We utilize the *PBMC8k* single cell RNA-seq dataset as reference (Additional file 1: Table S2) for all methods. The first two principal components of combined simulated and real PBMC datasets are visualized in Additional file 2: Fig. S1A, illustrating a domain shift between datasets.

For each dataset, DISSECT always obtained the best *JSD* across all datasets (Fig. 2A), leading to an average improvement over the second-placed algorithms of 6 percentage points. On the *GSE65133* dataset, for instance, DISSECT outperforms second-paced Scaden by 8 percentage points (DISSECT: $JSD = 0.145$, Scaden: $JSD = 0.222$). Similarly, DISSECT always obtains the best *rmse* across all datasets and improves over

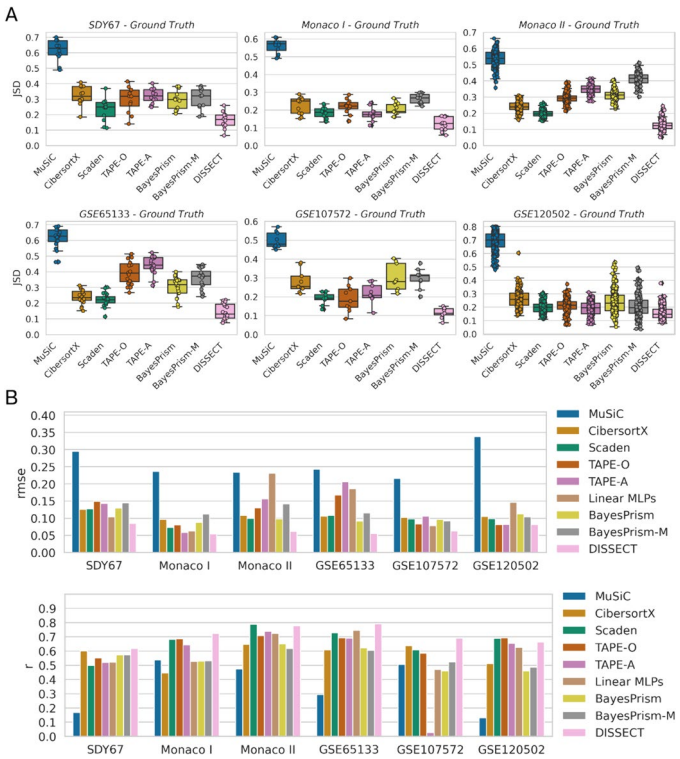

Fig. 2 Evaluation of deconvolution algorithm on six datasets with ground truth information. **A** Per-sample Jensen-Shannon divergence (*JSD*). Each plot corresponds to a dataset. From left to right and top to bottom: *SDY67*, *Monaco I*, *Monaco II*, *GSE65133*, *GSE107572*, and *GSE120502*. **B** Root mean-squared-error (*rmse*, top) averaged over cell types for each of the dataset. Datasets are listed on *x*-axis. Pearson's correlation (*r*, bottom) averaged over cell types

second-placed algorithms by 2 percentage points, on average (Fig. 2B). In addition, it achieved the best *r* on 4 out of 6 datasets (Fig. 2B).

Furthermore, we computed *macro-* level *r* and *rmse* by computing the metrics without making a distinction of cell types as performed previously in [9]. Note that in this setting, *JSD* remains unaffected as it is a sample-level metric and is therefore excluded. We observe that DISSECT achieves consistently best *rmse* across all datasets while achieving best *r* on 5 out of the 6 datasets (Additional file 2: Fig. S1).

Since MuSiC can take advantage of multi-sample references, we also evaluated MuSiC using blood data from the Immune Cell Atlas (ICA) (Additional file 1: Table S2). We also evaluated MuSiC with pre-selected marker genes (MuSiC-M) that were selected by CSx. MuSiC-M showed increased performance in 4 out of 6 datasets (Additional file 2: Fig. S2A-B). MuSiC also shows improved performance in the multi-sample setting in both rmse (Additional file 2: Fig. S2A) and r (Additional file 2: Fig. S2B). DISSECT still reaches best performance in *rmse* (on average 8 percentage points better) and *r* (on average 13 percentage points better) across all datasets.

Next, we evaluated the cell fraction deconvolution performance on the *Monaco I* (Additional file 1: Table S1) dataset, which contains several closely related and rare cell types and constitutes a relatively hard cell deconvolution task, using *Ota* dataset (Additional file 1: Table S1). With a correlation of 0.6, DISSECT's average performance is 14 percentage points better than the second placed Scaden (Additional file 1: Table S3), while Scaden's average RMSE was marginally (1 percentage point) better than second placed DISSECT (Additional file 1: Table S4). To validate that the performance improvement in DISSECT is due to the semi-supervised learning and consistency loss, we performed an ablation study on data *SDY67* by successively and cumulatively removing components of the algorithm and testing it again. The following components were removed successively: consistency regularization, KL Divergence loss (mean squared error instead), and the nonlinear activation function (identity function instead). The ablation results are shown in Additional file 1: Table S5.

In summary, these results provide strong evidence that DISSECT consistently outperforms current state-of-the-art cell type deconvolution algorithms across six different datasets with ground truth information.

Consistency of predictions and relationship between cell type fractions and biological phenotypes

To further corroborate the above results, we evaluate DISSECT's performance on three datasets that do not have paired flow cytometry data. In this section, we compare to other established biological facts as well as divergences over different reference single-cell datasets. The bulk datasets together with literature-based expected biological relationships of cell types are listed in Additional file 1: Table S1.

Brain

The *ROSMAP* dataset consists of 508 bulk RNA-seq samples from the dorsolateral prefrontal cortex (DLPFC) of patients with Alzheimer's disease (AD) as well as non-AD samples (Additional file 1: Table S1). For 463 of these samples, Braak stages of disease severity have been quantified. Correspondingly, single-nuclei RNA-seq (snRNA-seq)

for 48 individuals from the same cohort is available [24]. For 41 of these samples, cell type fractions based on immunohistochemistry (IHC) from a previous work exist [25]. It should be noted that IHC was performed for all neurons and as a result, comparison with respect to excitatory vs inhibitory neurons was not possible. Here, we consider two biological ground truths: first is the ratio of excitatory neurons to inhibitory neurons (Additional file 1: Table S1), and second is the neurodegeneration, or the loss of neurons with increasing Braak Stages [26]. We deconvolved *ROSMAP* using the *Allen Brain Atlas* reference (Additional file 1: Table S2).

We computed the *JSD* between the estimated fractions and IHC cell type proportions. DISSECT estimated fractions had the best average *JSD*s and provides the expected excitatory-inhibitory neuron ratio of (3:1–9:1), while other methods generally underestimated this ratio (Fig. 3A). All methods recover a negative correlation between increasing Braak stages and the fraction of neurons (Additional file 2: Fig. S3).

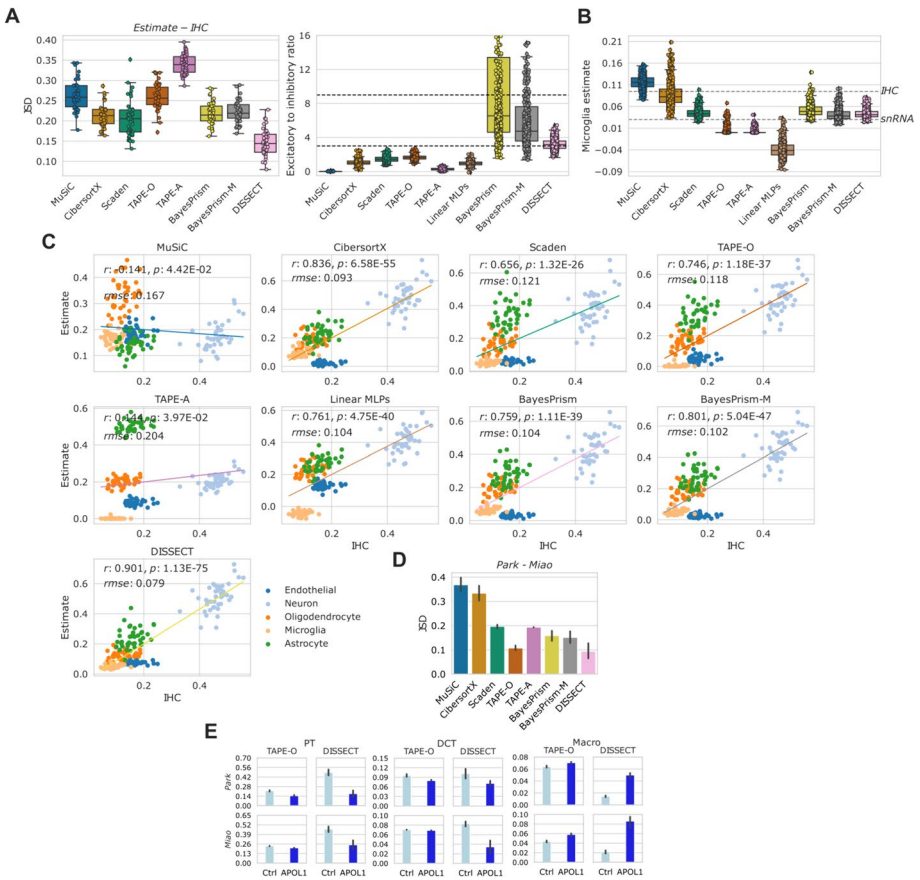

Fig. 3 **A** Left: Box-plots showing *JSD* between estimated fractions and IHC based cell type proportions from 41 individuals from *ROSMAP*. Right: Ratio of excitatory to inhibitory neurons computed from *ROSMAP*. Expected ratios lie between 3:1 and 9:1 as indicated by dashed lines. **B** Boxplots showing microglia proportion as estimated by different methods. Median proportions of microglia estimated using snRNA-seq and IHC are labeled. **C** Correlations between estimates (*y-axis*) and IHC cell type proportions (*x-axis*). **D** *JSD* between predicted proportions from *Kidney* between experiments with *Miao* and *Park* as references. **E** Predictions from TAPE-O and DISSECT from *Kidney*. From left to right: Proximal tubule (PT), ductal convoluted tubules (DCT), and macrophages (Macro). Each row indicates a reference. Error bars show standard deviations, while height of the bars shown mean prediction

Previously, it has been noted that snRNA-seq and IHC data provide different estimates for some cell types, notably microglia and endothelial cells [25]. It is interesting to observe that DISSECT and Scaden were the only methods where the estimates of microglia resembled closely those obtained from snRNA-seq and IHC data (Fig. 3B). We also computed *r* and *rmse* between the IHC cell type proportions and estimated fractions (Fig. 3C). With a correlation *r* of 0.901 DISSECT proved to be 14 percentage points better than the second-placed linear MLP. DISSECT also displayed the best *rmse* at 0.079.

Overall, the comparison to IHC and snRNA-seq ground truth information for the ROSMAP data further strengthens our claim that consistency regularization with DISSECT robustly improves cell deconvolution.

Pancreas

The *GSE50244* bulk RNAseq dataset consists of 89 pancreas samples from healthy and type 2 diabetes (T2D) individuals (Additional file 1: Table S1). For 77 of these samples, hemoglobic 1C levels are available as ground truth information. We performed the deconvolution using three single-cell reference datasets *Baron*, *Segerstolpe*, and *Xin* (Additional file 1: Table S2). Both *Baron* and *Segerstolpe* datasets contain alpha, beta, gamma, delta, acinar, and ductal cell types. While only alpha, beta, gamma, and delta cell types were present in the *Segerstolpe* dataset. To measure the consistency of deconvolution algorithms, we measured *JSD*s between estimated fractions using each of the three references (Additional file 2: Fig. S4A). While several methods showed considerable divergences, indicating reference-dependent deconvolution results, DISSECT displayed the most consistent results with a *JSD* of ~0.1–0.2 across the three pairs. In terms of recovery of significant negative correlations between the estimated fractions of beta cells and hemoglobin 1C (hba1c) levels, DISSECT provided highly significant correlations of between − 0.45 and − 0.47 across the three references (Additional file 2: Fig. S4B). These results further suggest that DISSECT is both precise and robust in cell type deconvolution on real data and is comparatively less affected by the choice of single-cell reference.

Kidney

The *GSE81492* dataset consists of 10 kidney samples of APOL1 mutant mice, which is a mouse model of chronic kidney disease (CKD) (Additional file 1: Table S1). We deconvolved the dataset using two single cell references: *Miao* and *Park* (Additional file 1: Table S2). Similar to our experiments on the pancreas tissue, we computed *JSD* between the estimated cell type fractions from the two references. DISSECT provided the best average *JSD* (0.09) out of all considered methods (Fig. 3D). We further compare the methods on the recovery of expected relation of cell type fractions with the biological phenotype (Additional file 1: Table S1). Figure 3E compares two best methods on *JSD*, DISSECT, and TAPE-O, while Additional file 2: Fig. S5 presents these results on all cell types for all methods. It is known that CKD results in the decrease in proximal tubule cells (PT) and distal convoluted tubules (DCT). Cell type fractions estimated with DISSECT showed a significant loss of PTs and DCTs and a corresponding increase in macrophages, while TAPE-O provided much smaller differences between the control and CKD model (Fig. 3E). PTs are the most abundant cell type in kidney making up around

50% of a mouse kidney [27]. DISSECT correctly estimates the high abundance of PTs in healthy kidney, while TAPE-O underestimates them (Fig. 3E).

In summary, it is noteworthy that DISSECT shows state-of-the-art precision and robustness in cell type deconvolution across various ground truth information and 9 datasets, including PBMC, brain, pancreas, and kidney bulk RNA-seq samples. DISSECT also shows superior robustness to the choice of single cell reference.

Application to proteomics and spatial transcriptomics

It is conceivable that DISSECT's consistency regularization for bulk RNA-seq cell type deconvolution should also lend itself to other biomedical datatypes in which domain shifts might be a problem. Applications might include, for example, the deconvolution of spatial transcritomic (ST) and bulk proteomic data with supra-cellular resolution. In order to evaluate these potential use-cases, we performed deconvolution of spatial transcriptomics and proteomics samples. Here, our aim is to test the hypothesis of applicability of DISSECT on these data modalities and we do not intend to perform an exhaustive comparison to multiple methods developed for these modalities. For comparisons on spatial transcriptomics, we consider four state-of-the-art spatial deconvolution methods, RCTD [28], Cell2location (C2L) [29] as shown to perform among the best in the benchmarking study [30]. We also include SONAR [31] and CARD [32], both of which can utilize spatial information. For comparisons on proteomic deconvolution, we consider the tested bulk deconvolution methods.

Spatial transcriptomics

We evaluated DISSECT on the task of spatial deconvolution using mouse brain and human lymph node samples (Additional file 1: Table S1). As a ground truth, we considered relationships with biological phenotypes in line with our application of kidney and pancreas datasets (Additional file 1: Table S1). Due to the spatial nature of the ST, we could verify the recovery of neuronal layers in brain (Additional file 2: Fig. S6) and discernment of germinal centers in lymph node (Additional file 2: Fig. S7). DISSECT performs on par with C2L and RCTD on both datasets. The results are provided and discussed in detail in the Additional file 2: Supplementary Note.

Proteomics

To compare the ability of the tested deconvolution methods to recover cell type proportions from proteomics mixtures, we utilized 50 human brain samples (Additional file 1: Table S1). We applied each deconvolution method on these samples using the Allen Brain Atlas reference (Additional file 1: Table S2). Compared to other methods, DISSECT recovered excitatory neurons to be the expected majority population in both datasets while maintaining the excitatory to inhibitory neuron ratio to be around expected range of (3:1–9:1) (Additional file 2: Fig. S8). These results strongly suggest that DISSECT reaches state-of-the-art performance on proteomic cell type deconvolution and might be applicable to other biomedical data types.

Evaluation of DISSECT under domain shifts

To assess the impact of consistency regularization on the performance of DISSECT and other algorithms, we used *Ota* dataset (Additional file 1: Table S1). Using this dataset in a dynamic domain shift setup (see the "Domain shift experimental setup" section), we evaluated the performance of deconvolution methods. We also included DISSECT without consistency (DISSECT w/o consistency) to asses the impact of semi-supervised learning under varying shifts. The performance of all methods dropped significantly for test sets with domain shifts (Additional file 2: Fig. S9). However, the drop in performance was much lower for DISSECT than other methods. Furthermore, a clear advantage of semi-supervised learning with consistency regularization is observed in comparison to DISSECT without consistency, especially in terms of *rmse*.

Estimation of cell type-specific gene expression

So far, we have shown that DISSECT can reliably deconvolve cell fractions. In this section, we focus on the deconvolution and inference of cell type-specific gene expression from bulk RNA-seq mixtures using our novel conditional autoencoder based algorithm (Fig. 1). While we were able to use ground truth flow cytometry data for the evaluation of cell fractions, no such gold-standard is available for cell type-specific gene expression information. In consequence, we measure DISSECT's gene expression inference performance on simulated bulk RNA-seq data. To maintain a domain shift between the training and test datasets, we simulated data for training and testing using different single-cell datasets. We compared the performance of DISSECT with that of TAPE-A, bMIND, and BayesPrism, all of which can infer cell type-specific gene expression per sample. We simulated bulk samples from one of the four reference single-cell PBMC datasets listed in Additional file 1: Table S2 and created training simulations from the remaining three. Simulations from each single-cell dataset consisted of 6000 samples. To evaluate the performance of DISSECT and other methods, we compared the true and estimated gene expression profiles of each cell type for each simulated sample (sample-wise) and for each gene (gene-wise) using Spearman correlation. These sets of results were aggregated across cell types and averaged. DISSECT displays the best sample- and gene-wise correlations in 6 out of 8 experiments, outperforming TAPE-A by 0.025 ± 0.023 in the sample-wise comparisons and by 0.012 ± 0.029 in the gene-wise comparisons (Table 1). Moreover, DISSECT exhibited an improvement in both sample and gene-wise metrics, exemplifying its advantage.

These results indicate that DISSECT's consistency regularization robustly performs state-of-the-art cell type-specific gene expression deconvolution.

Discussion

In this work, we first formally define the task of cell deconvolution and outline the hypothesis that semi-supervised consistency regularization should improve bulk RNA-seq deconvolution when learning from single cell RNA-seq data. We then provide evidence that our novel deep learning-based algorithm, DISSECT, outperforms competing state-of-the-art algorithms in deconvolution, both on a cellular and gene expression level, across many different datasets. This included 6 PBMC datasets with

Table 1 Spearman 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	TAPE-A	bMIND	BayesPrism	DISSECT
sample-wise r				
PBMC6k	**0.83±0.09**	0.80 ± 0.07	**0.83 ± 0.11**	0.82 ± 0.08
PBMC8k	0.79 ± 0.09	0.80 ± 0.08	0.81 ± 0.09	**0.84±0.11**
DonorA	0.85 ± 0.11	0.84 ± 0.09	0.80 ± 0.09	**0.89±0.10**
DonorC	0.81 ± 0.12	**0.83±0.11**	0.80 ± 0.08	**0.83±0.08**
gene-wise r				
PBMC6k	0.42 ± 0.14	**0.46±0.14**	0.41 ± 0.14	**0.46±0.15**
PBMC8k	**0.51±0.12**	0.44 ± 0.18	0.45 ± 0.12	0.48 ± 0.14
DonorA	**0.48 ± 0.20**	0.45 ± 0.16	0.46 ± 0.18	**0.48±0.18**
DonorC	0.45 ± 0.11	0.43 ± 0.15	0.45 ± 0.12	**0.49±0.12**

For each dataset, values with the highest mean correlation are displayed in bold font

ground truth flow cytometry information and 3 datasets (brain, pancreas, and kidney) with other established biological facts as ground truth information. Across the board, DISSECT provided the best cell type deconvolution results when compared to four state-of-the-art methods, while also being comparatively robust to the choice of single-cell reference. We follow a two-step procedure because the assumptions for each of the algorithms differ, and we do not foresee any significant benefit from iteratively deconvolving cell type fractions and gene expression. In a case study, we also show how our algorithm can easily be adapted to deconvolve cell types of proteomic and spatial expression data. For the spatial transcriptomics data, DISSECT estimates cell type fractions per spot, which are constrained to sum to 1. To be able to estimate the number of cells per cell type for each spot, and to map single cells, DISSECT estimates can be used as a prior for algorithms such as CytoSpace [33]. CytoSpace infers both the number of cells in each spot and solves an optimization problem to map single cells to their spatial locations. To estimate only the number of cells per cell type for each spot, the total number of cells as estimated by CytoSpace can be multiplied with the output of DISSECT. While these results are not exhaustive, they nevertheless show the applicability of DISSECT on other biomedical data types, a research avenue we might pursue in more depth in the future. In addition to DISSECT's state-of-the-art cell type fraction deconvolution (an average improvement of 0.063 in *JSD* and 0.021 in *rmse* over the state of the art on the datasets with ground truth cell type fractions), it achieved best cell type-specific gene expression deconvolution results in 6 out of 8 comparisons across four simulated datasets with an average improvement of 0.025 in the sample-wise and 0.012 in the gene-wise comparisons.

While we focused on MLPs for the estimation of cell type fractions and an autoencoder for gene expression estimation in this work, consistency regularization might also improve other deconvolution algorithms.

No gold standard ground truth exists for quantitative assessment of estimated cell type-specific gene expression between two conditions for real bulk RNA-seq datasets. This is a limitation of the experimental setup presented for cell type-specific gene

expression estimation. A potential solution will be to develop biologically valid benchmark datasets that can be evaluated at scale.

While DISSECT outperforms competing algorithms in cell type fraction and cell type-specific gene expression deconvolution, some results leave room for further improvement. DISSECT accurately distinguishes cell types where the transcriptional difference reflects cell subtypes, for instance PBMCs (CD4 T cells and CD8 T cells), pancreas (pancreatic islets), kidney (tubular epithelial cells), and brain (OPC and oligodendrocytes). However, when estimating granular cell type proportions in the Monaco I dataset, error rates exceeded the ground truth proportions (*rmse*>0.01 for cell subsets present at less than 1%). Therefore, for cell types that make up less than 1% of all cells and cells with very similar gene expression, for instance CD4 T and activated CD4 T cells, deconvolution algorithms should be used with caution. Future research into semi-supervised and contrastive algorithms as well as data augmentation and integration techniques should further enhance DISSECT's performance on hard deconvolution tasks.

Conclusions

In conclusion, DISSECT provides a semi-supervised deep learning framework to estimate cell type proportions and per-sample cell type-specific gene expression, is robust across datasets and tissues, and is easily applicable to other data modalities. DISSECT delivers state-of-the-art deconvolution performance, as long as cell types are not too closely related and make up more than 1% of all cells.

Methods

Evaluation metrics

To quantitively evaluate estimated cell type fractions across samples, we used two metrics, namely Pearson's correlation (*r*) and root-mean-squared error (*rmse*). Given x and y as estimated fractions and ground truth respectively,

$$r = \frac{cov(x, y)}{\sigma_x, \sigma_y} \tag{6}$$

$$rmse = \sqrt{Avg(x - y)^2} \tag{7}$$

To compute sample-wise divergences two list of fractions x_i and y_i for the same sample i, we used Jensen-Shannon distance (JSD) which is the square root of Jensen-Shannon divergence. *JSD* is given as

$$JSD(x \| y) = \sqrt{\frac{D(x_i \| m_i) + D(y_i \| m_i)}{2}}, \tag{8}$$

where $m_i = \frac{(x_i + y_i)}{2}$ and D is the Kullback-Leibler divergence.

State of the art

Here, we briefly detail the state-of-the-art deconvolution approaches. Out of these methods, CSx, TAPE, BayesPrism, and bMIND can also estimate per sample cell type-specific gene expression signatures.

MuSiC

MuSiC [21] 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. We used MuSiC R package (version 1.0.0). 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. We used the default hyperparameters to execute MuSiC. For single-cell datasets with multiple donors (Additional file 1: Table S2), we ran MuSiC with single-cell data from all available donors.

CSx

CSx [8] 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. To construct the signature matrix, we used the following hyperparameters for CSx as recommended by the authors: kappa $= 999$, q-value $=$ 0.01 and number of genes within a range of 300 and 500. The quantile normalization was also disabled. 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. 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. However, CSx only estimates genes likely to be differentially expressed in one of the bulk samples and as such the evaluations for simulations from healthy PBMC single-cells are not possible. We ran CSx through docker container obtained from [34].

Scaden

Scaden [9] is an average ensemble of three deep neural networks with different architectures that was developed for cell fraction deconvolution. Each network is trained only 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 (version 1.1.2) with the instructions provided by the authors to train the networks.

TAPE

TAPE [5] 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. The adaptive mode of TAPE (TAPE-A) 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. We used the official scTAPE package (version 1.1.2) implemented in Python.

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 has similar architecture to DISSECT, but in which we replaced all non-linear activations with an identity function and removed the consistency loss.

BayesPrism and BayesPrism-M

Primarily a method developed for oncology bulk datasets, BayesPrism [7] is a Bayesian framework to infer cell type fraction and cell type specific per-sample gene expression. It models gene expression as multinonmial distribution and calculates the cumulative posterior across cell states to derive the statistics for individual cell types. To evaluate BayesPrism with preselected marker genes using *select.marker* function. We utilize official implementation of BayesPrism in R (version 2.1.2).

bMIND

bMIND [6] is a Bayesian method to infer cell type specific gene expression per sample based on single-cell gene expression for given cell types. Using the prior from single-cell gene expression, bMIND models bulk gene expression as the product of gene expression and cell type fractions as a Bayesian mixed-effects model. bMIND uses cell type fractions as estimated by other deconvolution methods as its input. We used default settings of bMIND in our experiemnts with its R implementation (version 0.3.3).

Pre-processing and simulations

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. Furthermore, we also remove cells expressing more than 4% mitochondrial genes. Thereafter, before each deconvolution, we subset reference and bulk datasets to include only the common genes between the two. This quality control step was identical for all methods.

Simulations for deconvolution of bulk RNA-seq samples and proteomics

For deep learning methods, we sampled $\alpha_{k,i}$ uniformly to generate simulations *s.t.* $\sum_{k=1}^{c} \alpha_{k,i} = 100, \forall i$ if the dataset is single-cell. For experiments on granular level cell types where simulations are done from purified cell samples, we modified the 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. For all experiments, we simulated total $1000 \times c$ simulations where c is number of cell types in the reference dataset.

Simulations for deconvolution of 10x Visium ST samples

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. To reflect this, we simulated between 5 and 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 and 6 cell types. Due to sparsity of spots, not all cell types are present in a given spot. To account for this and to make comparison across spots possible, we utilized the outputs of the last layer (before performing softmax operation) and set negative predictions to zero. Thereafter, we re-normalized these absolute scores by such that each prediction sum to one. For all experiments, we simulated total $1000 \times c$ simulations where c is number of cell types in the reference dataset.

Deconvolution of proteomics data

For deconvolution of proteomics data, it is not valid to mix protein intensities and gene expression due to different normalizations. Instead of mixing simulated samples with real samples, proteomics samples were mixed with each other, i.e., at each training step, $\mathbf{B}_{i\cdot}^{\text{mix}} = \beta \mathbf{B}_{r_1\cdot} + (1-\beta)\mathbf{B}_{r_2\cdot}$, where r_1 and r_2 are two randomly selected proteomics samples at the training step.

Pre-processing for estimation of cell type fractions

For Scaden, TAPE, linear MLPs, and DISSECT, before passing simulated and real bulk samples to the network, we normalize samples to sum to a million counts (counts per million (CPM)) 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 [9].

For MuSiC and CSx (under S-mode), data was supplied on a linear scale as suggested in their respective publications and no change was made to the default normalization methods of both [8, 21].

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. Furthermore, 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.

For TAPE, since the signature matrix is observed in decoder (see the "State of the art" section), preprocessing step is similar to the preprocessing done in estimating cell type fractions.

Hyperparameters and fine-tuning

We fine tuned the network for activation functions, learning rate, and batch size using randomized search with hyperopt [35] with the root mean squared error as the objective function. The following grids were used for the optimization: activations = [linear, ReLU, ReLU6, Swish], learning rate = [5e−3, 1e−3, 5e−4, 1e−4, 5e−5, 1e−5], λ_1 = [0,1,5,10,15] with or without scheduled change at every 2000 steps and batch sizes = [32, 64, 128, 256] with 50 iterations on Ascites bulk dataset as used in Scaden [9]. Other hyperparameters were fixed to the default hyperparameters of Scaden. The optimal hyperparameters were fixed for all experiments, with batch size = 64, learning rate = 1e−5, activation function = ReLU6, λ_1 according to schedule [0,15,10] at steps [0,2000,4000], and number of steps = 5000.

Domain shift experimental setup

Using the *Ota* dataset (Additional file 1: Table S1) that contains 9852 purified samples belonging to immune cell subsets including several B cell and T cell subsets as shown in Additional file 1: Table S3, we created an experimental setup with domain shifts involving the following 4 scenarios. *20% split*: We randomly split the dataset into training (80%) and test sets (20%). *Activated 1*: We used the same split as in *20% split*. We removed certain CD4 and CD8 T cell subsets, namely, CD4 T memory, CD8 TEM, and CD8 TE from the training split while they were kept in the test set. In the test set, on the other hand, other subsets (CD4 T naive, CD8 T naive, and CD8 TCM) were removed. *Activated 2*: We followed the same procedure as in *Activated 1* except we removed certain B cell subsets, namely, B NSM, BEx, and BSM from the training set while they were kept in the test set. B naive subset was removed from the test set. Finally, for a model-based domain shift, we used DISCERN [36] to project the test set of *20% split* to the dataset simulated from *pbmc8k* and used in deconvolving PBMC bulk RNAseq. The CD4 T cell, CD8 T cell, and B cell subsets, regardless of their subtype identity, were labeled as CD4Tcells, CD8Tcells, and Bcells to allow comparisons. In each scenario, 6000 samples were simulated.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03251-5.

Additional file 1. Supplementary tables. The file contains supplementary tables [47-74].

Additional file 2. Supplementary figures. The file contains supplementary figures and supplementary note [75, 76].

Additional file 3. Review history. The file contains the peer review history.

Acknowledgements
We would like to thank Fabian Hausmann for helpful discussions and the MAXOMOD consortium for providing the proteomic data.

Review history
The review history is available as Additional file 3.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
SB initiated, and SB, PM, and RK conceptualized the study. RK wrote the code and implemented DISSECT. RK analyzed the data with help from SB. SB and RK interpreted the results and wrote the manuscript. PM provided ideas and reviewed the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. R.K. was supported by FOR5068 P9 and the 3R initiative of the UKE, P. M. by SFB1286 SP02, and S.B. by EU E-rare MAXOMOD, the M3I excellence initiative of the UKE, and SFB1192 B8 and C3.

Availability of data and materials
The datasets analyzed in this work are publicly available. Summary of the datasets are provided in Additional file 1: Tables S1 (bulk) and S2 (single-cell). PBMC single-cell RNA seq was obtained from 10x Genomics [37]. Single-cell RNA-seq dataset of the brain was obtained from Allen Brain Map [38]. Single-cell RNA-seq datasets from kidneys and pancreas can be accessed using Gene Expression Omnibus using corresponding accession codes: GSE157079 (***Miao***) and GSE107585 (***Park***), GSE81608 (***Xin***), GSE84133 (***Baron***). The raw single-cell data for ***Segertolpe*** is available at ArrayExpress (EBI) with accession code E-MTAB-5061. Cross-tissue Immune Cell Atlas (ICA) is available from [39]. Bulk RNA-seq datasets titled ***Monaco I***, ***Monaco II***, ***GSE120502***, ***GSE107572***, ***GSE50244***, and ***GSE81492*** are available from Gene Expression Omnibus [40] with following accession codes: GSE107011, GSE106898, GSE120502, GSE107572, GSE50244, and GSE81492. Bulk RNA-seq dataset ***SDY67*** was obtained from data resource provided in [9]. The original source for ***SDY67*** is ImmPort with accession code SDY67. ***ROSMAP*** cohort dataset is available from Synapse [41] with accession code syn3219045. The pre-processed data was obtained from [42]. The bulk proteomics data of post-mortem human brain samples was obtained from MAXOMOD consortium [43]. Allen brain reference with cortex annotations was obtained from [44].

Code availability
DISSECT is implemented in Python using tensorflow and keras (both versions 2.7.0) frameworks. The code is available at GitHub [45] with installation and usage instructions and has been deposited to Zenodo [46] under MIT license.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 July 2023 Accepted: 17 April 2024
Published online: 30 April 2024

References

1. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021;13(1):1–6.
2. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):1–35.
3. Zhou JG, Liang B, Jin SH, Liao HL, Du GB, Cheng L, et al. Development and validation of an RNA-seq-based prognostic signature in neuroblastoma. Front Oncol. 2019;9:1361.
4. Roberts KG, Li Y, Payne-Turner D, Harvey RC, Yang YL, Pei D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N Engl J Med. 2014;371(11):1005–15.

5. Chen Y, Wang Y, Chen Y, Cheng Y, Wei Y, Li Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nat Commun. 2022;13(1):6735.
6. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. Genome Res. 2021;31(10):1807–18.
7. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer. 2022;3(4):505–17.
8. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82.
9. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, et al. Deep learning–based cell composition analysis from tissue expression profiles. Sci Adv. 2020;6(30):eaba2619.
10. Long M, Cao Y, Wang J, Jordan M. Learning transferable features with deep adaptation networks. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. vol. 37 of Proceedings of Machine Learning Research. Lille: PMLR; 2015. pp. 97–105. https://proceedings.mlr.press/v37/long15.html.
11. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9.
12. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
13. Wang S, Pisco AO, McGeever A, Brbic M, Zitnik M, Darmanis S, et al. Leveraging the Cell Ontology to classify unseen cell types. Nat Commun. 2021;12(1):5556.
14. Maden SK, Kwon SH, Huuki-Myers LA, Collado-Torres L, Hicks SC, Maynard KR. Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets. Genome Biol. 2023;24(1):288.
15. Domínguez Conde C, Xu C, Jarvis L, Rainbow D, Wells S, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376(6594):eabl5197.
16. Ju C, Bibaut A, van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. J Appl Stat. 2018;45(15):2800–18.
17. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep speech: scaling up end-to-end speech recognition. 2014. arXiv preprint arXiv:14125567. https://arxiv.org/abs/1412.5567.
18. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA. 2018. pp. 4510–20. https://ieeexplore.ieee.org/document/8578572.
19. Ramachandran P, Zoph B, Le QV. Searching for activation functions. 2017. arXiv preprint arXiv:171005941. https://arxiv.org/abs/1710.05941.
20. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA. Mixmatch: a holistic approach to semi-supervised learning. Adv Neural Inf Process Syst. 2019;32:5050–60.
21. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):1–9.
22. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11(1):1–14.
23. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. Genome Biol. 2021;22(1):1–23.
24. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332–7.
25. Patrick E, Taga M, Ergun A, Ng B, Casazza W, Cimpean M, et al. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. PLoS Comput Biol. 2020;16(8):e1008120.
26. Braak H, Del Tredici K, Rüb U, De Vos RA, Steur ENJ, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging. 2003;24(2):197–211.
27. Clark JZ, Chen L, Chou CL, Jung HJ, Lee JW, Knepper MA. Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. Kidney Int. 2019;95(4):787–96.
28. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40(4):517–26.
29. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell 2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40(5):661–71.
30. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, Hu Y, Zhang X, Yao X, Tang M, Liu K. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods. 2022;19(6):662–70.
31. Liu Z, Wu D, Zhai W, Ma L. SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics. Nat Commun. 2023;14(1). https://doi.org/10.1038/s41467-023-40458-9.
32. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. Nat Biotechnol. 2022;40(9):1349–59. https://doi.org/10.1038/s41587-022-01273-7.
33. Vahid MR, Brown EL, Steen CB, Zhang W, Jeon HS, Kang M, Gentles AJ, Newman AM. High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. Nat Biotechnol. 2023;41(11):1543–8.
34. CIBERSORTx. https://cibersortx.stanford.edu/. Accessed 30 Jan 2024.
35. Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: International conference on machine learning. PMLR; 2013. pp. 115–23.
36. Hausmann F, Ergen C, Khatri R, Marouf M, Hänzelmann S, Gagliani N, et al. DISCERN: deep single-cell expression reconstruction for improved cell clustering and cell subtype and state detection. Genome Biol. 2023;24(1):212.
37. 10x Genomics. https://www.10xgenomics.com. Accessed 30 Jan 2024.
38. Allen Brain Map. https://portal.brain-map.org. Accessed 30 Jan 2024.

39. Cross-tissue Immune Cell Atlas. https://www.tissueimmunecellatlas.org. Accessed 30 Jan 2024.
40. Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/. Accessed 30 Jan 2024.
41. Synapse. https://www.synapse.org. Accessed 30 Jan 2024.
42. Deconvolution of cellular heterogeneity in brain transcriptomes. https://github.com/ellispatrick/CortexCellDeconv. Accessed 30 Jan 2024.
43. Caldi Gomes L, Hänzelmann S, Hausmann F, Khatri R, Oller S, Parvaz M, et al. Multiomic ALS signatures highlight sex differences and molecular subclusters and identify the MAPK pathway as therapeutic target. bioRxiv. 2023;2023–08.
44. Reference Atlas :: Allen Brain Atlas: Mouse Brain. https://mouse.brain-map.org/static/atlas. Accessed 30 Jan 2024.
45. Khatri R, Machart P, Bonn S. Deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. https://github.com/imsb-uke/DISSECT. Accessed 30 Jan 2024.
46. Khatri R, Machart P, Bonn S. Deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. Zenodo. 2024. https://doi.org/10.5281/zenodo.10570404.
47. Zimmermann MT, Oberg AL, Grill DE, Ovsyannikova IG, Haralambieva IH, Kennedy RB, et al. System-wide associations between DNA-methylation, gene expression, and humoral immune response to influenza vaccination. PloS ONE. 2016;11(3):e0152034.
48. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carre C, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep. 2019;26(6):1627–40.
49. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
50. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):1–20.
51. Harrison GF, Sanz J, Boulais J, Mina MJ, Grenier JC, Leng Y, et al. Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nat Ecol Evol. 2019;3(8):1253–64.
52. Ota M, Nagafuchi Y, Hatano H, Ishigaki K, Terao C, Takeshima Y, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell. 2021;184(11):3006–21.
53. Alejandro EU, Gregg B, Blandino-Rosano M, Cras-Méneur C, Bernal-Mizrachi E. Natural history of β-cell adaptation and failure in type 2 diabetes. Mol Asp Med. 2015;42:19–41.
54. Saisho Y. β-cell dysfunction: its critical role in prevention and management of type 2 diabetes. World J Diabetes. 2015;6(1):109.
55. Wang X, Misawa R, Zielinski MC, Cowen P, Jo J, Periwal V, 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. 2013;8(6):e67454.
56. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci. 2014;111(38):13924–9.
57. Venkatachalam MA, Weinberg JM, Kriz W, Bidani AK. Failed tubule recovery, AKI-CKD transition, and kidney disease progression. J Am Soc Nephrol. 2015;26(8):1765–76.
58. Liu BC, Tang TT, Lv LL, Lan HY. Renal tubule injury: a driving force toward chronic kidney disease. Kidney Int. 2018;93(3):568–79.
59. Malhotra R, Craven T, Ambrosius WT, Killeen AA, Haley WE, Cheung AK, et al. Effects of intensive blood pressure lowering on kidney tubule injury in CKD: a longitudinal subgroup analysis in SPRINT. Am J Kidney Dis. 2019;73(1):21–30.
60. Beckerman P, Bi-Karchin J, Park ASD, Qiu C, Dummer PD, Soomro I, et al. Transgenic expression of human APOL1 risk variants in podocytes induces kidney disease in mice. Nat Med. 2017;23(4):429–38.
61. Streit WJ, Braak H, Xue QS, Bechmann I. Dystrophic (senescent) rather than activated microglial cells are associated with tau pathology and likely precede neurodegeneration in Alzheimer's disease. Acta Neuropathol. 2009;118(4):475–85.
62. Hindle JV. Ageing, neurodegeneration and Parkinson's disease. Age Ageing. 2010;39(2):156–61.
63. Fu H, Possenti A, Freer R, Nakano Y, Hernandez Villegas NC, Tang M, et al. A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. Nat Neurosci. 2019;22(1):47–56.
64. Alreja A, Nemenman I, Rozell CJ. Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices. PLoS Comput Biol. 2022;18(1):e1009642.
65. Winer J, Larue D. Populations of GABAergic neurons and axons in layer I of rat auditory cortex. Neuroscience. 1989;33(3):499–515.
66. Ouellet L, de Villers-Sidani E. Trajectory of the main GABAergic interneuron populations from early development to old age in the rat primary auditory cortex. Front Neuroanat. 2014;8:40.
67. Braitenberg V, Schüz A. Cortex: Statistics and geometry of neuronal connectivity. 2nd thoroughly revised edition of: Anatomy of the cortex. Statistics and geometry (1991), 249. Springer Verlag Tiergarten. 1998;17:69121.
68. Beaulieu C. Numerical data on neocortical neurons in adult rat, with special reference to the GABA population. Brain Res. 1993;609(1–2):284–92.
69. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci. 2018;21(6):811–9.
70. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Syst. 2016;3(4):346–60.
71. Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24(4):593–607.
72. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, et al. RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metab. 2016;24(4):608–15.
73. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science. 2018;360(6390):758–63.

74. Miao Z, Balzer MS, Ma Z, Liu H, Wu J, Shrestha R, et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nat Commun. 2021;12(1):1–17.
75. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. IEEE; 2010. pp. 3121–4.
76. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

Publisher's Note