BuDDI: *Bulk Deconvolution with Domain Invariance* to predict cell-type-specific perturbations from bulk

Natalie R. Davidson 0000-0002-1745-8072 · nrosed · n_rose_d
Department of Biomedical Informatics, University of Colorado Anschutz School of Medicine, Aurora, Colorado, United States of America · Funded by the Gordon and Betty Moore Foundation (GBMF 4552), NHGRI of the National Institutes of Health (1K99HG012945-01), NCI of the National Institutes of Health (5R01CA237170-05, 5R01CA243188-04, 5R01CA200854-06)

Casey S. Greene 0000-0001-8713-9213 · cgreene · GreeneScientist
Department of Biomedical Informatics, University of Colorado Anschutz School of Medicine, Aurora, Colorado, United States of America · Funded by the Gordon and Betty Moore Foundation (GBMF 4552), NCI of the National Institutes of Health (5R01CA237170-05, 5R01CA243188-04, 5R01CA200854-06)

✉ — Correspondence possible via GitHub Issues or email to Casey S. Greene

Abstract

While single-cell experiments provide deep cellular resolution within a single sample, some single-cell experiments are inherently more challenging than bulk experiments due to dissociation difficulties, cost, or limited tissue availability. This creates a situation where we have deep cellular profiles of one sample or condition, and bulk profiles across multiple samples and conditions. To bridge this gap, we propose BuDDI (BUlk Deconvolution with Domain Invariance). BuDDI utilizes domain adaptation techniques to effectively integrate available corpora of case-control bulk and reference scRNA-seq observations to infer cell-type-specific perturbation effects. BuDDI achieves this by learning independent latent spaces within a single variational autoencoder (VAE) encompassing at least four sources of variability: 1) cell-type proportion, 2) perturbation effect, 3) structured experimental variability, and 4) remaining variability. Since each latent space is encouraged to be independent, we simulate perturbation responses by independently composing each latent space to simulate cell-type-specific perturbation responses.

We evaluated BuDDI's performance on simulated and real data with experimental designs of increasing complexity. We first validated that BuDDI could learn domain invariant latent spaces on data with matched samples across each source of variability. Then we validated that BuDDI could accurately predict cell-type-specific perturbation response when no single-cell perturbed profiles were used during training; instead, only bulk samples had both perturbed and non-perturbed observations. Finally, we validated BuDDI on predicting sex-specific differences, an experimental design where it is not possible to have matched samples. In each experiment, BuDDI outperformed all other comparative methods and baselines. As more reference atlases are completed, BuDDI provides a path to combine these resources with bulk-profiled treatment or disease signatures to study perturbations, sex differences, or other factors at single-cell resolution.

Introduction

Single-cell RNA sequencing (scRNA-Seq) technologies have provided methods to interrogate how cell-type proportions and cell-type-specific expression profiles vary within biological systems. In contrast, bulk RNA-Seq sequencing technologies average cell-type-specific differences, but are easier and cheaper to perform. Due to these inherent differences, larger single-cell experiments typically provide more cell types and numbers of cells but are still lacking in the breadth of individuals, diseases, and perturbations of existing bulk RNA-Seq data. However, understanding cell-type-specific responses is key to understanding treatment response and disease etiology. For example, the method of action of traditional disease-modifying antirheumatic drugs (tDMARDs) is not well understood but is believed to target T-cells[1]. Unfortunately, there is very limited single-cell data with tDMARDs treatments. However, there are large single-cell studies measuring the arthritic synovial fluid[2,3] without tDMARDs and bulk studies that track patients before and after taking tDMARDs[1]. This pattern of missing data is not particular to arthritis and tDMARDs; it is also present in cohorts of rare diseases where the recruitment of new patients to perform single-cell sequencing is infeasible. To effectively utilize the existing large bulk studies and growing single-cell references, we need methodological advances that combine multi-condition bulk and single-condition scRNA-Seq data to estimate cell-type-specific expression profiles across the conditions observed in the bulk data. To accomplish this goal, we build on ideas from three methodological approaches: bulk deconvolution [4–14], variational autoencoder (VAE)[15] models for perturbation prediction[16–22], and disentanglement methods[18,23–27].

Bulk deconvolution methods unify single-cell and bulk data types by attempting to deconvolve an observed bulk expression profile as a sum of cell-type-specific expression profiles[4–14,28]. One key limitation of this deconvolution approach is that most methods assume the bulk expression profile is similar to the reference single-cell profiles. BayesPrism[13] addresses this problem using a Bayesian framework to directly account for differences between the observed bulk and single-cell data for one cell type among those with fixed profiles. We account for not only the differences between the bulk and single-cell data but additionally other sources of variation, such as sample variability and perturbation response. Furthermore, we seek to independently perturb each source of variation to simulate cell-type-, sample-, and perturbation-specific differences. We would also like our deconvolution method to be flexible and easily integrated into a larger generative model, similar in structure to Scaden, a VAE-based bulk deconvolution method[7].

There exist several generative methods to learn interpretable latent spaces that decompose the input single-cell expression profiles into relevant sources of variation. These methods can be directly trained to capture a specific source of variation[29–35] or post-hoc-interpreted after training[36–40]. Furthermore, there exist several methods to learn a latent space such that shifts within the latent space represent specific perturbation effects on an unobserved cell or cell type[4–14,28]. Instead of leveraging perturbation responses in other cells or cell types, we would like to leverage complex bulk expression profiles, not only cell lines or single-cell profiles, to infer the cell-type-specific perturbation response.

However, to simulate accurate perturbation responses, it is key that perturbing one latent space does not affect another latent space, i.e., changing the latent space that represents cell type proportion should only affect the variability related to cell type proportions, and not other sources of variability related to the sample identity or sequencing technology. This concept is related to domain invariance, where latent representations are invariant to changes in a domain. One difference between our proposal and typical domain invariance approaches is that our main goal is not for our method to be invariant of unseen domains, but invariant to observed domains within our dataset of interest. In our case, we would like to model each latent representation to be independent of one another, which could also be phrased as having latent representations that are disentangled. This framework can be used to learn classifiers invariant to a specific confounding factor[24,27] or to analyze the latent spaces to interrogate the sources of variability within the data[23,25,26]. Our use case requires

the generative aspect of the model to predict cell-type-specific perturbation effects similar to MichiGAN[18], except we will infer the perturbation response from bulk data, not single-cell.

BuDDI combines strategies to learn domain-invariant representations that capture cell type proportions, perturbation effects, and experimental variability. BuDDI not only learns interpretable latent representations to understand the data better but can also compose changes in each latent space to predict cell-type-specific perturbation responses.

Results

The model structure of BuDDI

BuDDI's VAE structure (Fig 1) reflects the belief that our observed gene expression data is generated from at least four sources of variability: sample or technical variability (z_e), condition-specific variability (z_p), differences in cell-type proportion (z_y), and other sources of noise (z_x). To ensure each latent space is specific to its source of variability, an auxiliary loss is added to BuDDI to predict the labels related to the sample, technology, condition, and cell-type proportion. Since BuDDI learns from bulk and single-cell RNA-Seq data, the cell-type proportions are not always known; therefore, z_y is trained semi-supervised, and z_e and z_p are trained fully supervised. z_x is unrestricted but is the same dimensionality as z_e and z_p. A more detailed description of the training procedure and model is given in **Methods**.

BuDDI utilizes the generative model structure introduced in DIVA[27], a method to identify disentangled latent representations in cellular images. Similarly, BuDDI treats each of these sources of variability as specific and invariant domains. Domain invariance is key to BuDDI learning cell-type-specific perturbation effects since we can independently learn representations for the perturbation and cell type and compose them together to learn a cell-type-specific effect.

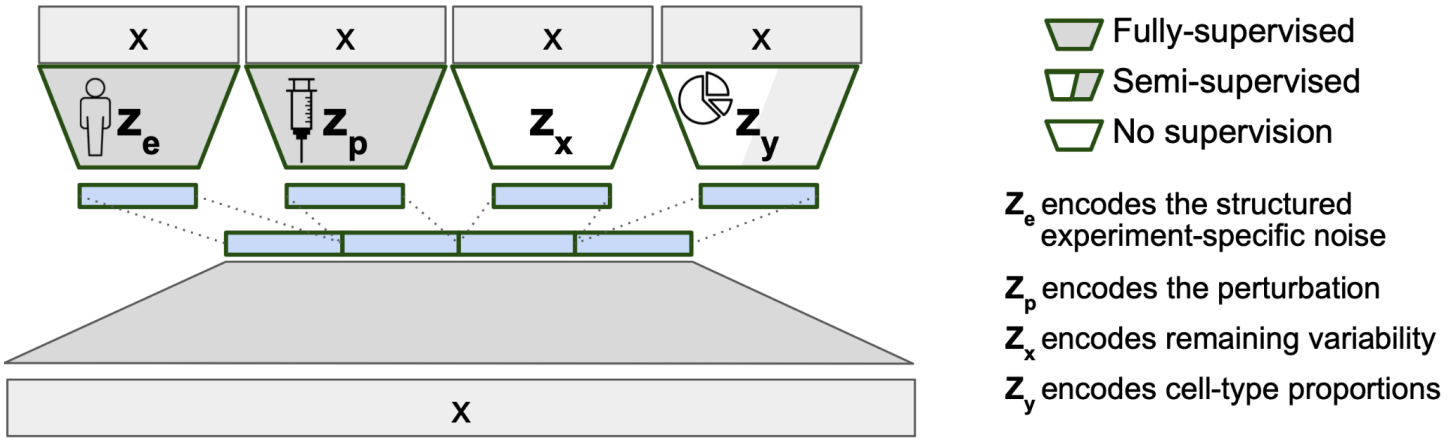


Figure 1. VAE structure of BuDDI. X is our bulk or pseudobulk. We apply an auxiliary loss on each latent code for them to encapsulate a specific source of variability. Since our model is generative, we can later sample from each latent space to simulate experimental changes to our input expression profile. To simulate cell-type-specific effects, we can sample a cell-type proportion where the cell type of interest is the predominant cell type.

While the generative structure of BuDDI encourages each latent space to be invariant, real biological data is unlikely to have training data with independent sources of variability. Specifically, cell-type proportions are likely dependent on the sample or perturbation status. To break this dependence, we simulate pseudobulk data used in training to have random cell-type proportions. This allows us to break the dependence between cell-type proportions and the other sources of variation. The approach assumes the observed expression data is sufficiently independent for the remaining latent spaces to learn descriptive and domain-invariant

representations. In the following sections, we evaluate this assumption, finding that BuDDI works on data with increasing levels of interdependence across the latent representation. Firstly, we validate BuDDI on the simplest experimental design using only pseudobulks, where we have matched samples across each source of variability. Next, in a more realistic setting, we still use pseudobulks but now have no matched samples between bulk and single-cell. Finally, we test BuDDI on real single-cell and bulk data from Tabula Muris Senis[41,42], where there are no matched samples across any source of variation.

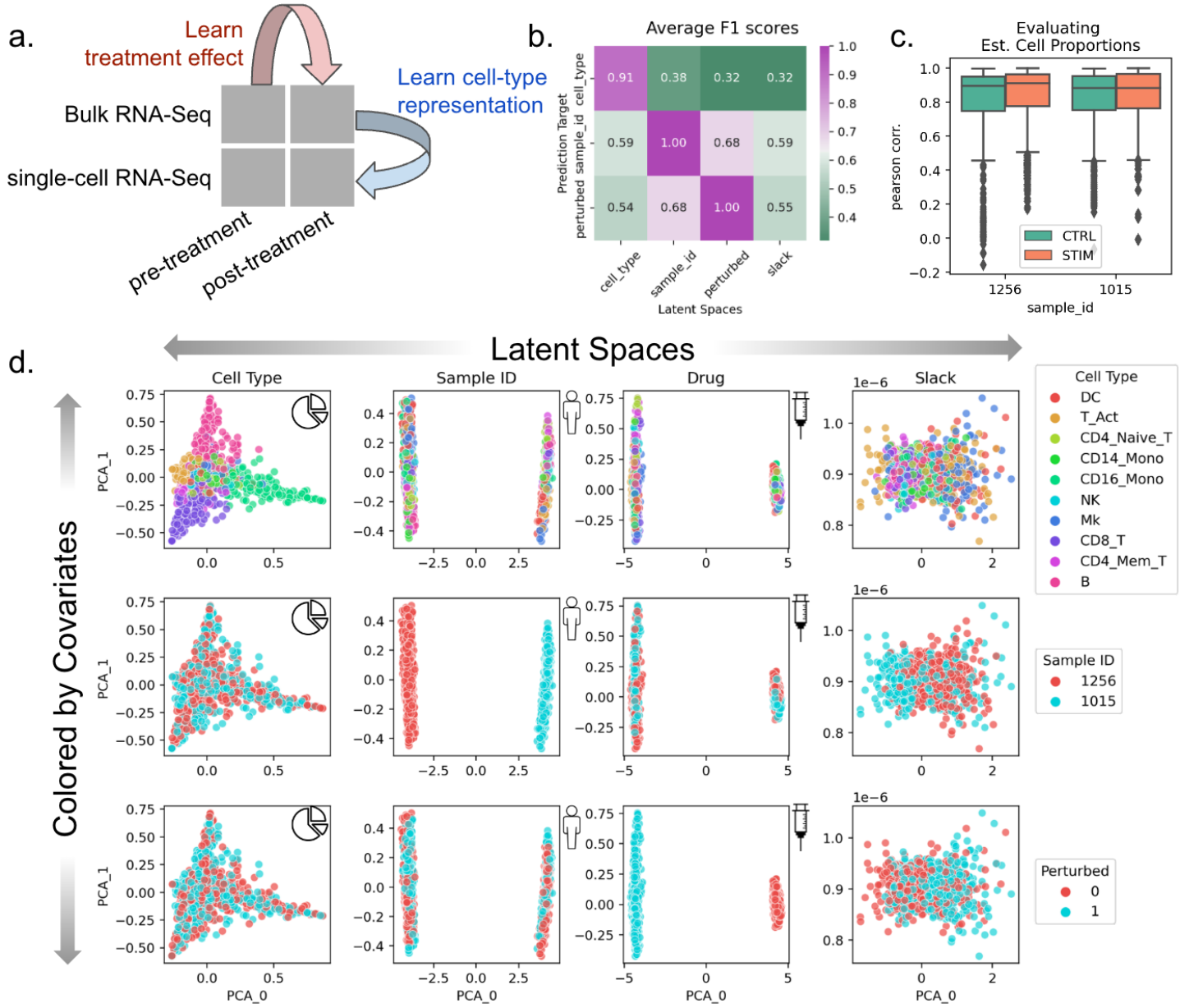


Figure 2. Evaluation of BuDDI on pseudobulk data with matched samples across each source of variability. **Panel a** depicts a schematic of the experimental design. **Panel b** depicts a heatmap of the average F1 score using each latent space to predict each source of variability. A high F1 score along the matched latent space and source of variability, and a low F1 score where the latent space does not match the source of variability is a measure of disentanglement across the latent spaces. **Panel c** shows the performance of BuDDI at predicting the cell-type proportions. **Panel d** visualizes the first two principal components (PCs) of each latent space (columns) and colors them by different sources of variation (rows).

BuDDI learns descriptive and domain-invariant latent representations

To validate that BuDDI works as expected, we first tested the simplest experimental design, where we have matched observations across each source of variability. We used a dataset created by Kang et al.[43] of peripheral blood mononuclear cells from two of the eight lupus patients with matched samples that either had

interferon-Beta stimulation or no stimulation. To simulate bulk samples, we omitted cell-type proportions from half of the pseudobulks during training. An overview of the data included in our experimental design is shown in **Figure 2a**.

After training BuDDI, we measured the extent of domain invariance across latent spaces. We compared the predictive accuracy of each latent space in predicting its intended and unintended targets on a held-out test set. This is similar to the Separated Attribute Predictability (SAP) score[44], except we compare distinct latent spaces to one another instead of an individual latent dimension. Each latent space approximated domain invariance: the accuracy of each latent space to predict its intended source of variability was significantly higher than a mismatched source of variability (**Figure 2b**). This indicated that each latent space was specific to only its intended target, not targets described by another latent space. Furthermore, we observed that each latent space was not only relatively accurate in predicting its intended target but generally accurate; each latent space was predictive of its intended source of variability with a very high F1 score (>0.9). We also observed that BuDDI can learn the cell-type proportions of the pseudobulk data accurately, as shown by the strong correspondence between ground truth and predicted cell-type proportions (**Figure 2c**).

After quantitative evaluation, we also qualitatively evaluated the specificity of each latent space. We observed that the first two principal components (PCs) divide each latent space by its target value, demonstrated in the plots along the diagonal of **Figure 2d**. Furthermore, along the off-diagonal, the non-target sources of variability are well mixed. This indicated that most of the variance in the latent spaces specifically captures the target source of variability. In the slack latent space, each target is well mixed, indicating that it is not capturing variability from explicitly modeled sources. We also observe a lack of clear structure in the slack latent space, indicating that there is little remaining structured variability to be explained by the slack.

BuDDI accurately predicts cell-type-specific perturbation response

After validating that BuDDI learns specific latent space representations, we examined the extent to which BuDDI predicts cell-type-specific perturbation responses when perturbation measurements are only available in bulk data. Again, we used the data from Kang et al.[43] to generate our simulated data, except used all eight available samples. To make the bulk data more comparable with actual data, we simulated realistic cell-type proportions that were again omitted during training. Furthermore, to examine the method's ability to identify a cell-type specific effect and not simply a global shift, we only use stimulated CD14 monocytes for simulation (**Figure 3a**).

First, we determined whether or not BuDDI could capture the perturbation response in our dataset when not explicitly modeled. We trained an augmented version of BuDDI (BuDDI-noPert), where we removed the perturbation latent space. The BuDDI-noPert slack latent space captured the perturbation response (**Figure 3b**). Once the perturbation space was reintroduced, the slack space no longer separated the samples by perturbation status (**Supp. Figure 1a**; the slack space was not strongly predictive of the perturbation status; mean F1 score: 0.52). Additionally, the latent spaces were still generally predictive of and specific to their specific source of variation, although as expected performance was degraded in comparison with the experiment where paired samples were supplied across each source of variability (**Supp. Figure 1a-c**).

Next, we identified if BuDDI could predict the expression and effect size of the perturbation for each cell type. We compared BuDDI against PCA with latent space projections and a conditional VAE (CVAE)[45]. To get cell-type specific expressions for PCA and CVAE, we used the pseudobulks generated primarily from one cell type, then applied the perturbation. For PCA, we learned a sample-specific linear translation to simulate the perturbation. For CVAE, the perturbation and sample IDs were included in the conditions, so we only had to change the condition status in the CVAE on the pseudobulks with primarily one cell-type to simulate a cell-type

specific perturbation effect. We evaluated each method on pseudobulks generated from held-out single-cell RNA-Seq profiles. Full details of the experimental design are given in **Methods**. Across all metrics and cell types, BuDDI outperformed all other methods (**Figure 3c**). Since our experimental design only perturbs CD14 monocytes, it is unsurprising that we see performance degradation in that cell type; however, BuDDI still outperforms all other methods and maintains a relatively high Pearson correlation for the predicted stimulated expression (mean > 0.8) and log2 fold change (mean > 0.65). We then examined if performance was degraded in more lowly expressed genes. We observed that CVAE performance increases for more highly expressed genes (**Supp. Figure 1d**). BuDDI also performs better with higher levels of expression, but the performance increase if not as drastic. BuDDI's performance was comparable to PCA for lowly expressed genes and comparable to CVAE on highly expressed genes, with BuDDI outperforming all models when considering all levels of expression (**Supp. Figure 1d**).

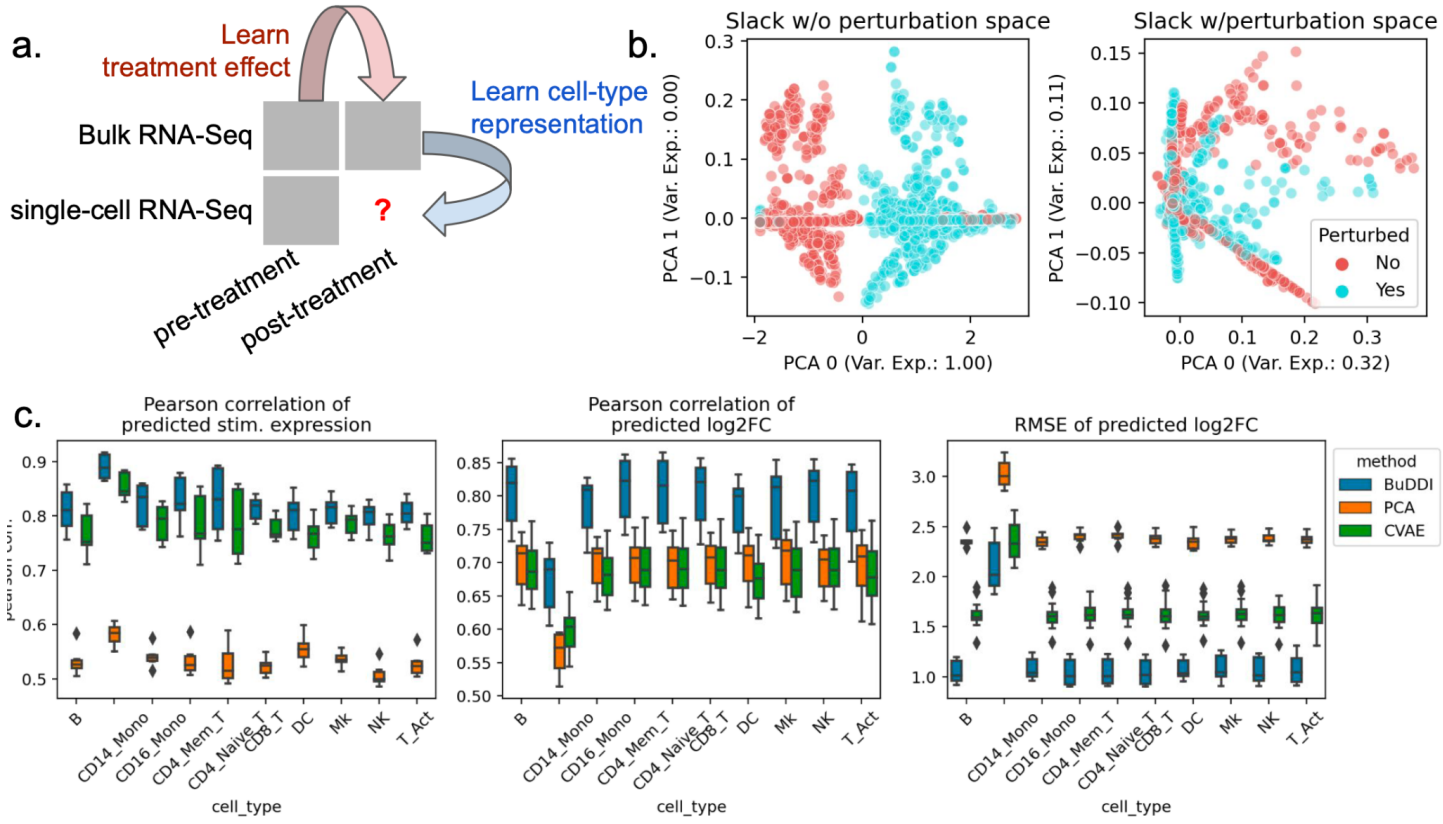


Figure 3. Evaluation of BuDDI on cell-type-specific perturbation simulation. BuDDI on pseudobulk data with matched samples across each source of variability. **Panel a** depicts a schematic of the experimental design; we no longer include the single-cell perturbation response during training. **Panel b** depicts the slack space when training BuDDI without (left) and with the perturbation latent space (right). Here we observe that when we train BuDDI without the perturbation space, the slack space picks up the perturbation response. This effect is greatly diminished once we include the perturbation latent space. **Panel c** depicts the performance of BuDDI, PCA, and CVAE in predicting the cell-type-specific expression and log2 fold change. In this experiment, only CD14 monocytes are stimulated. To evaluate the model variability of BuDDI and CVAE, each model was trained and evaluated three independent times and is included in **Panel c**.

BuDDI accurately identifies cell-type-specific sex differences

Finally, we examined the extent that BuDDI predicted cell-type-specific sex differences in the Tabula Muris Senis dataset[41,42]. Tabula Muris Senis consists of male and female mice's bulk and single-cell expression data in several organs. We restricted our analysis to the liver, a sexually dimorphic organ. The challenge of this dataset is that there are no matched samples across any source of variability. There were no technical replicates for any samples nor matched bulk and single-cell samples. Furthermore, we do not have matched perturbation effects to examine sex differences because each mouse was either male or female. This

experimental design implies that each source of variability is highly entangled with each other. We evaluated predictions using a held-out single-cell female mouse sample (**Figure 4a**).

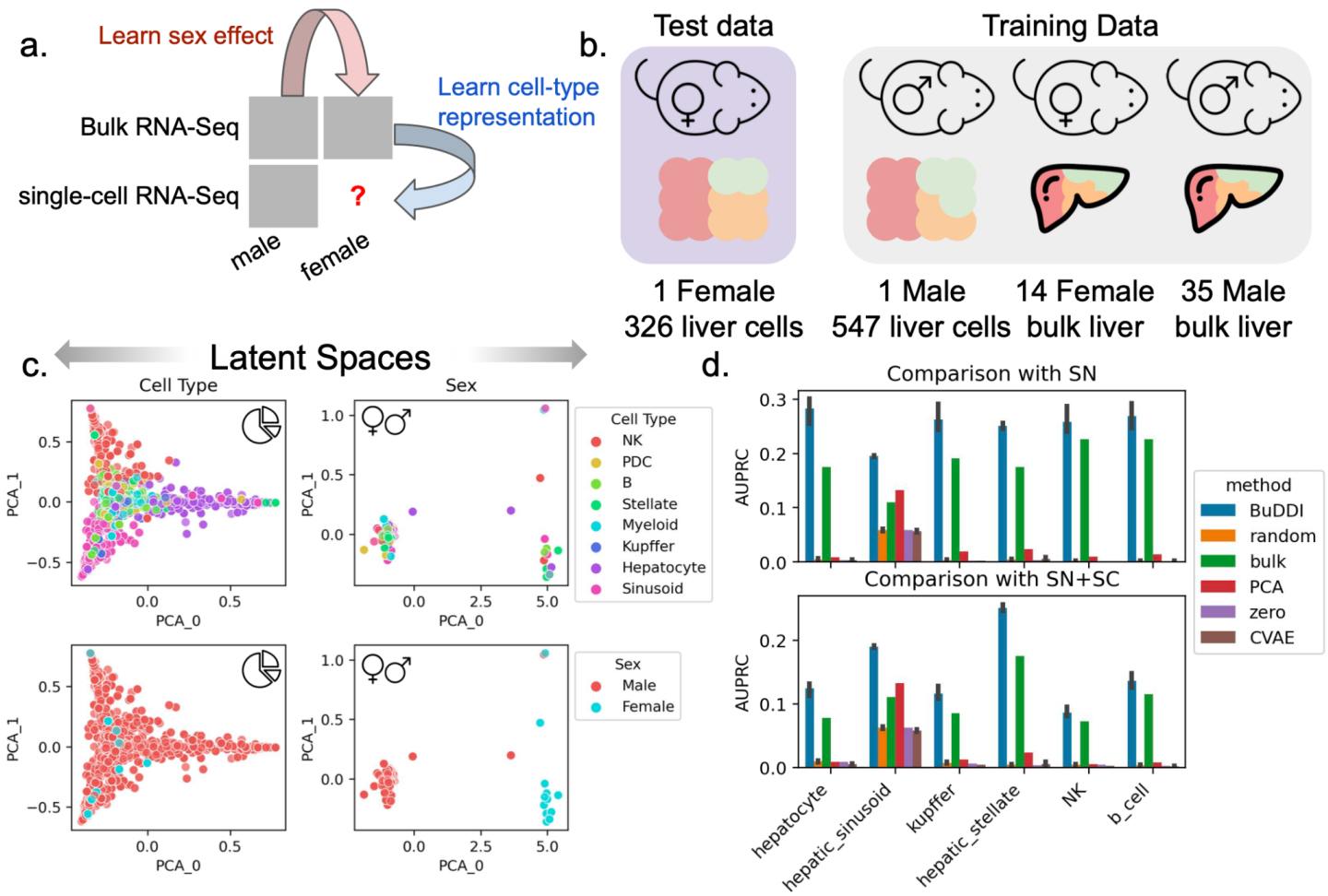


Figure 4. Evaluation of BuDDI to predict cell-type-specific differences in the mouse liver. **Panel a,b** depict a schematic of the experimental design and data used for training and evaluation. **Panel c** depicts the cell type and sex latent spaces colored by either the most abundant cell type or sex. **Panel d** depicts the area under the Precision-Recall curve in predicting the differential gene between the sexes for each cell type. **Panel d**, top, uses differentially expressed genes identified by an independent single-nucleus experiment analyzing sex-specific differences in the liver. **Panel d**, bottom, uses the union of differentially expressed genes from the aforementioned single-nucleus experiment and the Tabula Muris Senis[41,42] single-cell experiment. Bar height represents the mean area under the precision-recall curve (AUPRC) and the black lines indicate the 95% confidence interval. To consider the model variability of BuDDI and CVAE, each model was trained and evaluated three independent times.

First we examined whether or not BuDDI separated the sources of variability in this highly correlated dataset. We visually found that each latent space was specific to its target source of variability (**Figure 4c** and **Supp. Figure 2**). Importantly, we observed a clear separation between the cell type and the sex, the two latent factors required predict cell-type-specific sex differences (**Figure 4c**). However, some entanglement remained between the slack and cell-type latent spaces (**Supp. Figure 2**).

Next, we aimed to predict genes with the largest sex differences in each cell type. In contrast to experiments using perturbation data, obtaining matching expression data across sexes is impossible. Because it is not possible to validate predictions by predicting each sample's exact gene expression value for each sample since we have no ground truth, we identified the top genes predicted to have the largest difference in expression between the sexes. In addition to CVAE and PCA, we also compare against: random, a baseline of the shuffled predicted values; zero, a baseline of the majority label (0); and bulk, a baseline of the differentially expressed genes between the bulk samples. The bulk baseline represents the global shift in expression;

therefore, outperforming the bulk baseline indicates that the model identifies cell-type-specific differences. We compared our results against two validation sets. The first set is the differentially expressed genes between the single female and male mice provided by Tabula Muris Senis[41,42]. We provide full details of the data processing and differential expression pipeline in **Methods**. The second validation set is from an independent study of sex differences using single-nucleus RNA-Seq data[46]. We included this secondary study since it has more biological replicates and is from a complementary sequencing platform.

BuDDI outperforms all other methods and baselines in each cell type, including the bulk baseline, indicating that BuDDI can identify cell-type-specific sex differences beyond a global shift in expression (**Figure 4d** and **Supp. Figure 3)**. PCA with a latent transformation is the only method to outperform the bulk expression in only one cell type, hepatic stellate cells. In all other cell types, PCA and CVAE perform similarly and are better than random, but are significantly outperformed by BuDDI.

Discussion

We introduce BuDDI, a method to learn cell-type-specific perturbation responses using reference single-cell and multi-condition bulk data. BuDDI learns latent representations specific to a single source of variation and independent of all other sources of variation. This model design enables BuDDI to individually perturb one or more latent spaces and compose them to simulate cell-type specific perturbations. In most experimental designs, it is impossible to have data that has matched samples across all sources of variability. We successively evaluated BuDDI on increasingly entangled data, moving from data that had all, some, and then no matched samples across the sources of variability. We found that BuDDI outperforms all competitor models and baselines in each instance. BuDDI can help researchers interrogate the sources of variability within their data. The model's slack space, z_x, captures remaining variability that was not directly modeled, allowing researchers to identify unaccounted confounders.

BuDDI can be tuned in different ways. There is an inherent tradeoff between the accuracy of latent representation and the reconstruction, which leads to significant degradation of the cell-type proportion estimator when the experimental design has more entangled sources of variability (**Supp Figure 2b)**. In our evaluations, we optimized the reconstruction accuracy of BuDDI to predict cell-type-specific perturbation response. Depending upon the use case, the end-user can specifically train BuDDI to have a better cell type proportion estimator, but at the cost of reconstruction accuracy.

While we evaluated BuDDI on expression data, this implementation is conceptually extendable to other data types. The approach can be applied to other data modalities as long as it is possible to generate augmented training data that separates the cell-type specific signal from the other sources of variation. Furthermore, other than cell-type proportion, we have currently implemented BuDDI to represent sources of variability only as discrete values. Conceptually, BuDDI could model continuous sources of variability, such as age, perturbation time, or drug concentration.

BuDDI provides a methodological solution to a missing data pattern that is common in genomic analyses of publicly available data. Without needing to sequence more, BuDDI can leverage one technologies' depth in its cellular profiles with another's breadth in the heterogeneity of profiles. BuDDI has several potential use cases, such as providing a way to analyze tissues whose cells are difficult to dissociate at a single-cell resolution, to leverage difficult-to-obtain data from patients with rare diseases, or to re-analyze the tens of thousands of heterogeneous existing bulk samples. BuDDI strives to make the most out of existing bulk datasets in the era of large-scale single-cell reference atlases.

Methods

BuDDI model description

BuDDI extends the VAE framework[15] and uses a similar conceptual structure as DIVA[27]. The entire VAE structure attempts to find a latent representation (z) that is likely to reconstruct the original data (x). The goal is to maximize the marginal likelihood [15,47]

$$p_\theta(x) \;=\; \int p_\theta(x|z)\, p_\theta(z)\, dz$$

$p_\theta(x|z)$ is the decoder and uses a neural network to learn the parameters θ, where given z we reconstruct x. Unfortunately, learning $p_\theta(x)$ is intractable, since it requires integrating over all possible latent representations z. Instead, we estimate it by learning a lower bound to $p_\theta(x)$, by learning an approximate posterior $q_\phi(z|x)$. $q_\phi(z|x)$ is our encoder, where ϕ are learned parameters of the encoder neural network. We can rewrite $p_\theta(x)$ as

$$log\, p_\theta(x) \;=\; \mathbb{E}_{q_\phi(z|x)}[log(\frac{p_\theta(x,z)}{q_\phi(z|x)})] \;+ \mathbb{E}_{q_\phi(z|x)}[log(\frac{q_\phi(z|x)}{p_\theta(z|x)})]$$

$$= L_{\theta,\phi}(x) \;+\; D_{\mathbb{KL}}(q_\phi(z|x)\,||\,p_\theta(z|x))$$

Since $D_{\mathbb{KL}}(q_\phi(z|x)\,||\,p_\theta(z|x))$ is non-negative, $L_{\theta,\phi}(x)$ is a lower bound on $log\, p_\theta(x)$. Now we learn parameters to maximize $L_{\theta,\phi}(x)$, which can be rewritten as

$$L_{\theta,\phi}(x) \;= \mathbb{E}_{q_\phi(z|x)}[log(p_\theta(x|z)] \;-\; \beta D_{\mathbb{KL}}(q_\phi(z|x)\,||\,p_\theta(z))$$

where β is a weighting term to constrain the amount of variability that can be explained by the latent space[48]. Unlike a VAE with a single latent space (z), DIVA and BuDDI learn independent latent spaces to capture different sources of variability (experimental z_e, perturbation z_p, and remaining variability z_x). This is done through learning separate encoders, $q_{\phi_e}(z_e|x)$, $q_{\phi_p}(z_p|x)$, and $q_{\phi_x}(z_x|x)$, and a single decoder. To capture variability due to cell type proportions, we directly append the observed cell type proportion to the latent space when it is available or use a predicted cell type proportion from an auxiliary predictor when the cell type proportion is not available. This implies that $z_y \approx y$, instead of being predictive of y as done in the other latent spaces. The auxiliary predictor takes the gene expression x as input and predicts the cell type proportion, y, and it's weights are only updated when the cell type proportions are known. This is how BuDDI is able to predict the cell type proportions in a semi-supervised fashion. The loss without the auxiliary proportion loss, but including the additional latent spaces is the following:

$$L_{\theta,\phi}(x) \;= \mathbb{E}_{q_{\phi_e}(z_e|x)q_{\phi_p}(z_p|x)q_{\phi_x}(z_x|x)q_{\phi_y}(z_y|x)}[log(p_\theta(x|z_e,\, z_p,\, z_x,\, y))] \;-\; \beta_e D_{\mathbb{KL}}(q_{\phi_e}(z_e|x)\,||\,p_\theta(z_e))$$
$$-\; \beta_p D_{\mathbb{KL}}(q_p(z_p|x)\,||\,p_\theta(z_p)) \;-\; \beta_x D_{\mathbb{KL}}(q_{\phi_x}(z_x|x)\,||\,p_\theta(z_x))$$

Unlike DIVA, we do not use conditional priors to separate the latent spaces from one another and instead only use auxiliary classifiers on the experiment and perturbation specific latent spaces, $q_{\omega_e}(e|z_e)$ and $q_{\omega_p}(p|z_p)$, to constrain the latent spaces to their intended source of variability. The full loss is

$$L_{BuDDI}(x) = L_{\theta,\phi}(x) + \alpha_e \mathbb{E}_{q_{\phi_e}(z_e|x)} [log(q_{\omega_e}(e|z_e))] + \alpha_p \mathbb{E}_{q_{\phi_p}(z_p|x)} [log(q_{\omega_p}(p|z_p))] + \alpha_y \mathbb{E} [log(p_{\theta_y}(y|x))]$$

A detailed diagram of the BuDDI implementation is provided in **Supp. Fig 5**.

BuDDI training and implementation details

In generating the pseudobulks used for testing and training, cells were divided into two even sets stratified by each source of variation: perturbation status, cell type, and sample ID. Therefore, pseudobulks used in training will not have any cells seen in testing. BuDDI was implemented in Keras version 2.12.0, and was trained using the Adam optimizer[49], with a learning rate of 0.005. The non-slack β terms are always set to 100 and β_x is set to 0.1. This parameter choice encourages the non-slack latent representations to be biased towards fully capturing the source of variability, since a larger β term creates a stronger bottleneck on the latent representation and encourages stronger disentanglement within the latent space[48]. The number of epochs [50, 100] and the classifier weights [100, 1000, 10000] were identified using grid search. We wanted to minimize reconstruction loss and the Spearman correlation of the true and estimated cell-type proportions on a training validation set, which is 20% of the training set held out during training. After the initial set of classifier weights was identified, they were further adjusted using the training set to encourage further disentanglement of the latent spaces. For all models, we used 64 dimensions for each latent representation and a batch size of 500. We used internal dimensions of 512 and 256 for the cell type proportion predictor. We used a single 512-dimensional dense layer for the perturbation and experimental predictors.

To train BuDDI cell-type proportions in a semi-supervised manner, we created two separate encoder models with shared weights. When the cell-type proportions are not known, the cell-type proportion predictor weights are not updated, and its predictions are used in the latent space during training. When the cell-type proportions are known, the cell-type proportion predictor weights are updated, but the predictions are not used in the latent space. Instead, the true value is directly input into the latent space during training. This is depicted as two separate model diagrams in **Supp. Fig 5.** During training, BuDDI switches between the supervised and unsupervised models within each epoch. In both cases, the auxiliary classifiers for predicting the sources of variation, excluding the cell-type proportions, are always supervised, and their weights are updated throughout the entire epoch.

The structure of each latent space is identical to one another, with two hidden layers of dimensions 512 and 256. In all experiments, we have two latent spaces representing experiment-specific variability, z_e, one that is predictive of the sample ID and the other that predicts whether the data comes from a pseudobulk sample or a real bulk sample. For the BuDDI-noPert experiment, the perturbation latent space z_x is excluded from the entire model.

BuDDI simulation of perturbation response

BuDDI learns a separate latent space for each source of variability, allowing us to modify a specific latent space to simulate a change related to that latent space. To do this, we use our training data to sample latent codes that predict a specific source of variability. We can perturb a single latent space or several latent spaces and combine them to produce the desired latent representation. To generate a cell-type-specific perturbation effect, we use a y with the highest cell-type proportion for the cell type of interest. We will combine this with latent codes related to unperturbed and perturbed samples. Combining these two latent codes with the remaining latent codes relevant to the experiment, we compared the gene expression differences between the

perturbed and unperturbed samples for a specific cell type. Depending on the desired analysis, the additional latent spaces could be sampled randomly or specific to a sample of interest. For the Kang et al.[43] data with matched samples, we sampled latent codes specific to each sample. For the sex-dependent liver analysis, we jointly sampled the latent slack, sample, perturbation, and bulk codes. When the latent spaces were observed to have high amounts of independence between them, each latent space could be sampled more independently. Conversely, if high dependence between latent spaces is observed, it is recommended to jointly sample the latent spaces that are not directly relevant to the perturbation of interest.

CVAE model description

The CVAE[45] learns a latent representation conditioned on specific variables; in our case, we implemented a CVAE conditioned on the sample ID, perturbation status, and whether the input data is pseudobulk or a real bulk. The CVAE differs from a VAE in its implementation by appending a 1-hot-encoded vector representing the sources of variation to the input to both the encoder and the decoder. After training, new data is generated by changing the appended vector to represent the perturbation of interest. However, unlike BuDDI the vector representing the source of variation cannot be trained in a semi-supervised manner. Therefore, it is impossible to learn a model that is conditional on the cell-type proportions and the perturbation status since we only have perturbed observations from the bulk data, which has no cell-type proportion estimate. To get around this limitation, we instead learn a latent space that captures the cell-type proportions and is independent of all other sources of variation. This enables us to calculate cell-type-specific perturbation changes by sampling from regions in the latent space specific to a cell type, then appending our latent code that represents our perturbation of interest.

The CVAE was implemented in Keras. For consistency, we maintained the same latent code dimension as BuDDI and the same dimension of encoder and decoder layers. We also used the same optimizer, ADAM, with a learning rate of 0.005. The β term was set to 1 in all experiments. β values were grid searched [0.1, 1, 10] to minimize the reconstruction error and identify a latent space that was predictive of the cell-type proportions.

PCA model description

PCA was used to learn a low-dimensional data representation. We then learned a linear transformation between the perturbed and non-perturbed samples in the low-dimensional representation. To learn a cell-type-specific perturbation response, we used pseudobulks with a cell-type proportion where the cell type of interest was the majority cell type. Next, we summed its low-dimensional representation with the perturbation vector and projected the sample back into the original dimensionality of the data. Since we had matched samples for the Kang et al.[43] data, we also learned a sample translation vector and the perturbation vector to simulate a sample-, cell-type-, and perturbation-specific effect. The number of latent dimensions used for PCA was 20, which explained >90% of the variability in both datasets.

Data processing

The single-cell data used in each experiment was processed using scanpy[50]. For all experiments, the cell-type labels were taken from the original manuscript. The Kang et al.[43] analysis data was downloaded from SeuratData[51] and converted to h5ad format for downstream processing in scanpy. In the Kang et al.[43] analysis, we removed outlier cells with less than 500 or more than 2500 genes expressed. We removed genes expressed in less than five cells. The total number of cells used by cell type and sample are shown in **Supp. Table 1**.

The data for the sex-specific liver differences were downloaded from the Tabula Muris Senis[41,42] project (https://figshare.com/articles/dataset/Processed_files_to_use_with_scanpy_/8273102/2), hosted by FigShare [https://doi.org/10.6084/m9.figshare.8273102.v2]. Due to a low number of cells and expressed genes in the liver dataset, we could only analyze one male and one female mouse sample. Two male mice samples had a sufficient number of cells for each cell type, but we restricted our analysis to post-pubescent mice (3 months or older), which resulted in the filtering of one of the male mice. Furthermore, hepatic stellate cells were very rarely observed (<27 cells per sample, 3.25 on average) and therefore combined with endothelial cells of the hepatic sinusoid, a more abundant cell type with a similar expression profile. We did not filter cells, but we removed genes expressed in less than three cells. **Supp. Table 2** provides the counts of cells by sample.

The bulk liver data was downloaded from Gene Expression Omnibus under accession ID GSE132040. We filtered samples that were less than three months old. The total number of samples by age and sex are provided in **Supp. Table 3**. We did not perform any additional count processing on the single-cell data before pseudobulk generation for each dataset. Additional processing was only done for identifying differentially expressed genes in the single-cell data. Raw counts were used for differential expression analysis of the bulk data, as needed for pyDESeq2[52].

Pseudobulk generation

After processing the data, as described in the **Data processing** section, we performed a 50/50 split of the cells, stratified by sample and cell type. This ensured we did not observe any pseudobulks with shared cells between the training and testing sets. To create the pseudobulks, we summed over sampled cells from each individual dependent upon a specific cell-type proportion. We generated three types of cell-type proportions: random, cell-type specific, and realistic. Random proportions were sampled from a lognormal distribution, with a mean of 5 and a variance uniformly sampled between 1 and 3. All proportions were scaled to sum to 1. The cell-type specific proportions were generated by first creating a vector of the length of cell types where the cell-type of interest had a proportion of $1 - ((\#\, cell\, types) * 0.01)$, and the remaining cell types had a proportion of 0.01. Lognormal noise with mean 0 and variance 1 was added to the cell-type proportions and then rescaled such that they sum to 1. Suppose the new cell-type proportion did not have a Pearson correlation coefficient > 0.95 with the original cell-type proportion vector before the noise was added. In that case, noise vector was discarded, and a new one was sampled. The realistic cell-type proportion estimator calculated the sample-specific cell-type proportion observed from the single-cell data. Noise was added in the same way as for the random cell-type proportions. After the cell-type proportions were sampled, we sampled a total of 5000 cells dependent upon the cell-type proportion and sum over the counts to generate the pseudobulk values. **Supp. Figure 5** depicts the generated pseudobulks with each type of sampled proportion.

Differential expression of single-cell and bulk liver data

Differential single-cell expression was done using scanpy[50] and pyDESeq2[52]. We first generated cell-type-specific pseudobulks, generating ten samples and 30 cells sampled per cell type. Using these pseudobulks, we used pyDESeq2 to identify the genes that were differentially expressed between the sexes for each cell-type. For the bulk and pseudobulk pyDESeq2 analyses, genes with a mean expression across all samples < 1 were removed from the analysis. We considered genes with adjusted p-value < 0.01 as differentially expressed for all downstream analyses. The single-nucleus differentially expressed genes were taken from[46].

Pseudobulk normalization

After the pseudobulk data was generated, it was uniformly processed for each experiment and model. First, we identified 7000 genes that form the union between CIBERSORTx-identified signature genes[4] and the genes we calculated to have the highest coefficient of variance. These genes were highly overlapping (**Supp. Figure 6**). Next, we MinMax scaled the gene expression. Since gene counts typically have long-tailed expression profiles, we clipped the expression at the 90th quantile before scaling.

Predicting source of variability using each latent space

To predict each source of variability, we used a Naive Bayes classifier. We reported the average F1 score on a held-out test set of 10% of the data. We performed this classification task 30 times for each model. To take into account the variability of BuDDI, we independently trained three separate BuDDI models and averaged their performance.

Evaluation of genes predicted to be sex-dependent

Since we could not have matched samples from different sexes, we could not directly compare sample- and cell-type-specific changes in gene expression due to sex. Instead, we predicted the genes most affected by sex differences for each cell type. We compared the simulated male and female gene expression for each model for each cell type. We then reported the median rank difference between male and female simulated data. To calculate the area under the precision-recall curve (AUPRC), we used the absolute value of the median rank difference. Our true values were either from an independent single-nucleus experiment[46] that identified sex-dependent genes, or from the genes identified as sex-dependent from the Tabula Muris Senis data[41,42] used to generate the pseudobulks. The comparative baselines were 1) random: shuffled ranks; 2) zero: a predictor that only reported zero, the majority label; and 3) bulk: the sex-dependent genes identified by analyzing the bulk Tabula Muris Senis data.

Data and code availability

All code is available on GitHub. The BuDDI model code is available at https://github.com/greenelab/buddi, and the code to recreate all analyses is available at https://github.com/greenelab/buddi_analysis. The trained models and processed data needed to recreate the analyses is available on figshare under the DOI: 10.6084/m9.figshare.23721336

Works Cited

1. Walsh, A. M. *et al.* Triple DMARD treatment in early rheumatoid arthritis modulates synovial T cell activation and plasmablast/plasma cell differentiation pathways. *PLoS One* **12**, e0183928 (2017).
2. Zhang, F. *et al.* Cellular deconstruction of inflamed synovium defines diverse inflammatory phenotypes in rheumatoid arthritis. *bioRxiv* 2022.02.25.481990 (2022) doi:10.1101/2022.02.25.481990.
3. Zhang, F. *et al.* Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat. Immunol.* **20**, 928–942 (2019).
4. Steen, C. B., Liu, C. L., Alizadeh, A. A. & Newman, A. M. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods Mol. Biol.* **2117**, 135–157 (2020).
5. Frishberg, A. *et al.* Cell composition analysis of bulk genomics using single-cell data. *Nat. Methods* **16**, 327–332 (2019).
6. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
7. Menden, K. *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv* **6**, eaba2619 (2020).
8. Wang, Z. *et al.* Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, 451–460 (2018).
9. Dong, M. *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2021).
10. Lin, Y. *et al.* DAISM-DNNXMBD: Highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks. *Patterns (N Y)* **3**, 100440 (2022).
11. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).
12. Torroja, C. & Sanchez-Cabo, F. DigitaldIsorter: Deep-Learning on scRNA-Seq to Deconvolute Gene Expression Data. *Front. Genet.* **10**, 978 (2019).
13. Chu, T., Wang, Z., Pe’er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism

enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* **3**, 505–517 (2022).

14. Song, L., Sun, X., Qi, T. & Yang, J. Mixed model-based deconvolution of cell-state abundances (MeDuSA) along a one-dimensional trajectory. *Nature Computational Science* 1–14 (2023).
15. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).
16. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
17. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
18. Yu, H. & Welch, J. D. MichiGAN: sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome Biol.* **22**, 158 (2021).
19. Ghahramani, A., Watt, F. M. & Luscombe, N. M. Generative adversarial networks simulate gene expression and predict perturbations in single cells. *bioRxiv* 262501 (2018) doi:10.1101/262501.
20. Bunne, C. *et al.* Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *bioRxiv* 2021.12.15.472775 (2021) doi:10.1101/2021.12.15.472775.
21. Stark, S. G. *et al.* SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**, i919–i927 (2020).
22. Marouf, M. *et al.* Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **11**, 166 (2020).
23. Weinberger, E., Lin, C. & Lee, S.-I. Isolating salient variations of interest in single-cell data with contrastiveVI. *bioRxiv* (2021) doi:10.1101/2021.12.21.473757.
24. Aliee, H., Kapl, F., Hediyeh-Zadeh, S. & Theis, F. J. Conditionally Invariant Representation Learning for Disentangling Cellular Heterogeneity. *arXiv [cs.LG]* (2023) doi:10.48550/ARXIV.2307.00558.
25. Jones, A., William Townes, F., Li, D. & Engelhardt, B. E. Contrastive latent variable modeling with application to case-control sequencing experiments. *arXiv [stat.ME]* (2021) doi:10.48550/ARXIV.2102.06731.
26. Weinberger, E., Lopez, R., Hütter, J.-C. & Regev, A. Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI. *bioRxiv* 2022.12.13.520349 (2022)

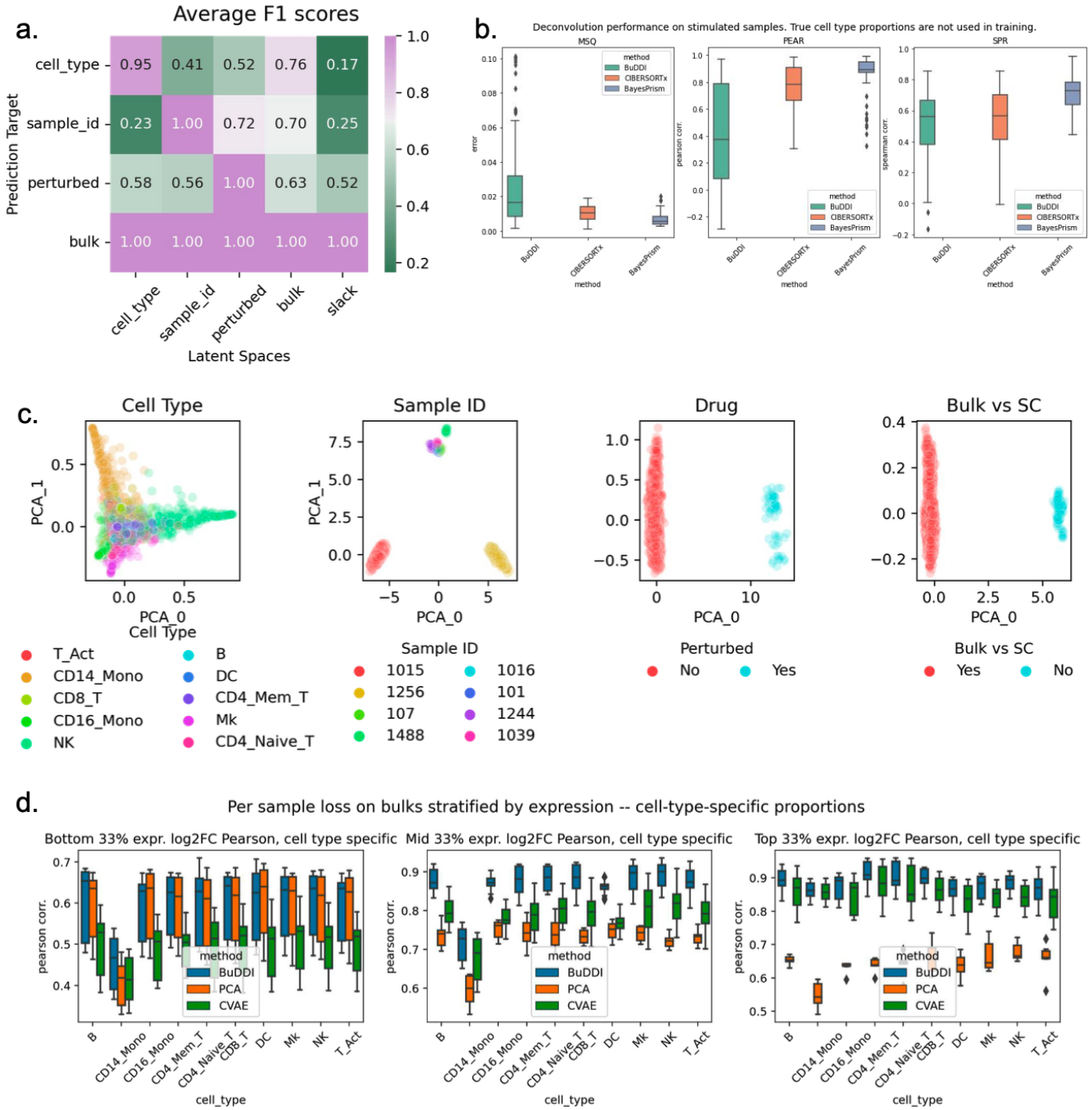
doi:10.1101/2022.12.13.520349.

27. Ilse, M., Tomczak, J. M., Louizos, C. & Welling, M. DIVA: Domain Invariant Variational Autoencoders. in *Proceedings of the Third Conference on Medical Imaging with Deep Learning* (eds. Arbel, T. et al.) vol. 121 322–348 (PMLR, 06--08 Jul 2020).
28. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).
29. Lotfollahi, M. *et al.* Biologically informed deep learning to query gene programs in single-cell atlases. *Nat. Cell Biol.* **25**, 337–350 (2023).
30. Gut, G., Stark, S. G., Rätsch, G. & Davidson, N. R. PmVAE: Learning interpretable single-cell representations with pathway modules. *bioRxiv* (2021) doi:10.1101/2021.01.28.428664.
31. Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **16**, 607–610 (2019).
32. Pividori, M. *et al.* Projecting genetic associations through gene expression patterns highlights disease etiology and drug mechanisms. *bioRxiv* (2021) doi:10.1101/2021.07.05.450786.
33. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18**, 212 (2017).
34. Rybakov, S., Lotfollahi, M., Theis, F. J. & Alexander Wolf, F. Learning interpretable latent autoencoder representations with annotations of feature sets. *bioRxiv* 2020.12.02.401182 (2020) doi:10.1101/2020.12.02.401182.
35. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat. Commun.* **12**, 5684 (2021).
36. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. in *Biocomputing 2018* (WORLD SCIENTIFIC, 2018). doi:10.1142/9789813235533_0008.
37. Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S. & Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol.* **21**, 109 (2020).
38. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via

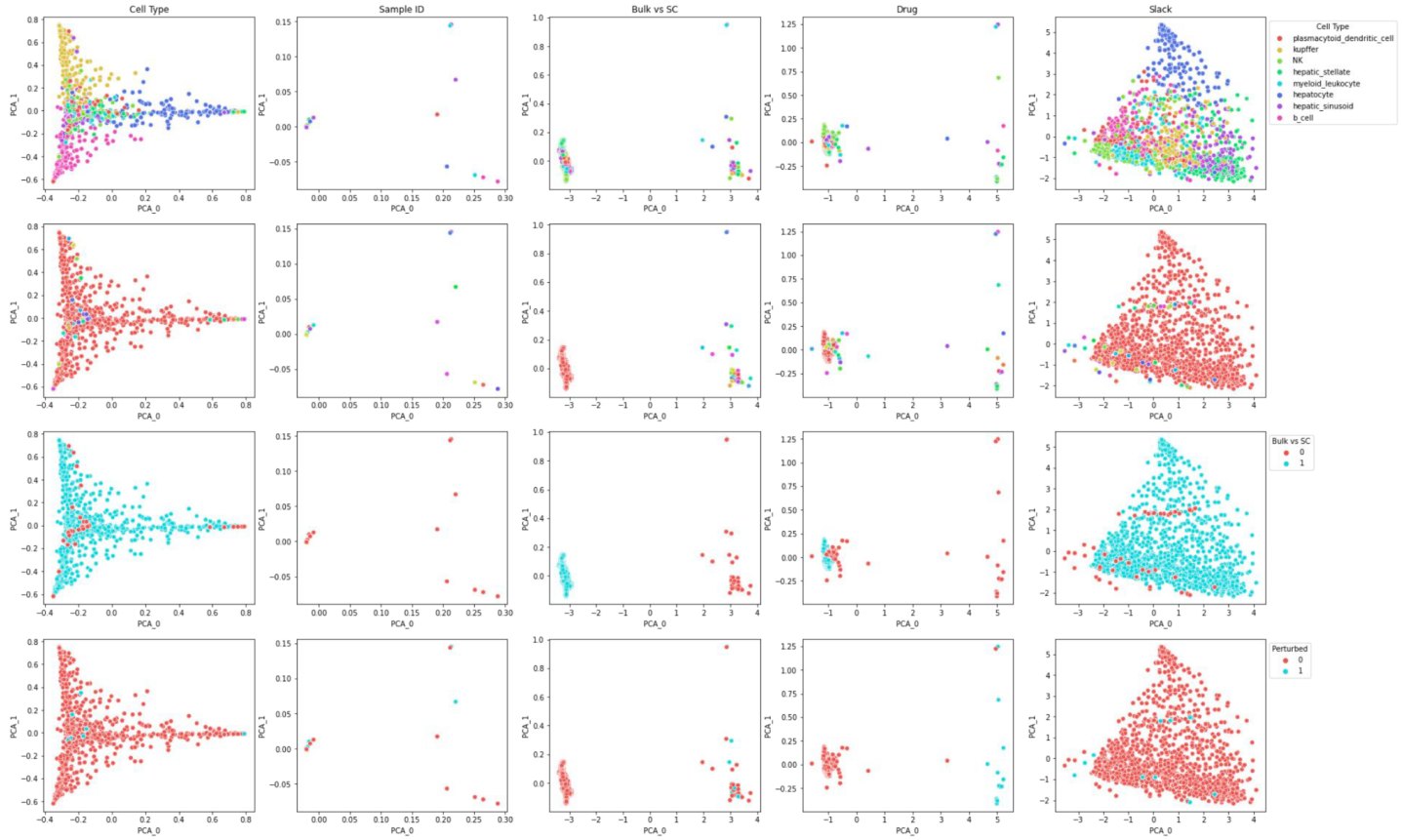
variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).

39. Choi, Y., Li, R. & Quon, G. siVAE: interpretable deep generative models for single-cell transcriptomes. *Genome Biol.* **24**, 29 (2023).
40. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. Commun.* **12**, 5261 (2021).
41. Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
42. Schaum, N. *et al.* Ageing hallmarks exhibit organ-specific temporal signatures. *Nature* **583**, 596–602 (2020).
43. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
44. Kumar, A., Sattigeri, P. & Balakrishnan, A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. *arXiv [cs.LG]* (2017) doi:10.48550/ARXIV.1711.00848.
45. Learning structured output representation using deep conditional generative models. https://proceedings.neurips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.
46. Goldfarb, C. N., Karri, K., Pyatkov, M. & Waxman, D. J. Interplay Between GH-regulated, Sex-biased Liver Transcriptome and Hepatic Zonation Revealed by Single-Nucleus RNA Sequencing. *Endocrinology* **163**, (2022).
47. Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics*. (MIT Press, 2023).
48. Higgins, I. *et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. (2016).
49. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
50. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
51. Satija, R., Hoffman, P. & Butler, A. SeuratData: Install and manage seurat datasets. *R package*.
52. Muzellec, B., Teleńczuk, M., Cabeli, V. & Andreux, M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *bioRxiv* 2022.12.14.520412 (2022) doi:10.1101/2022.12.14.520412.

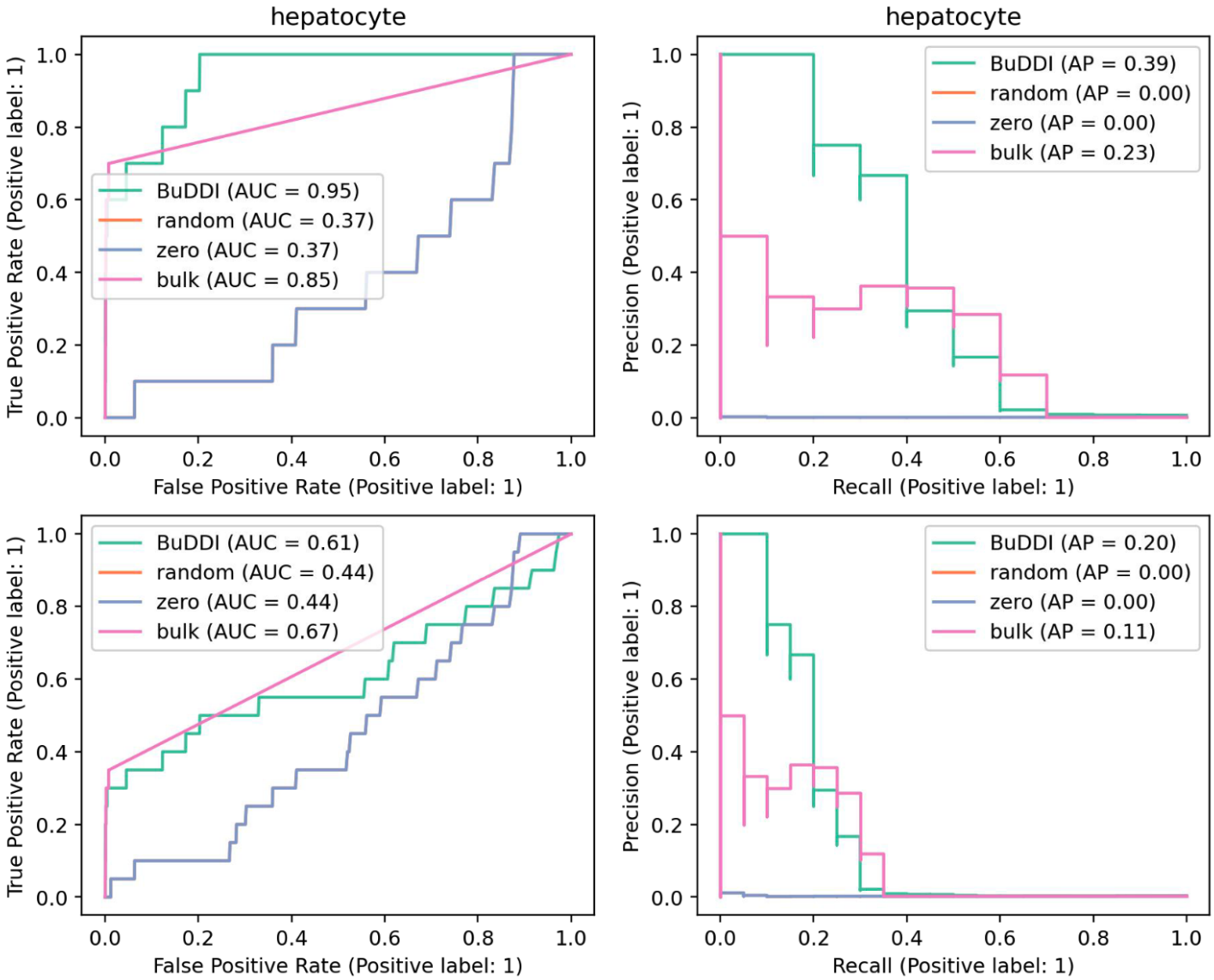
Supplementary Figures and Tables



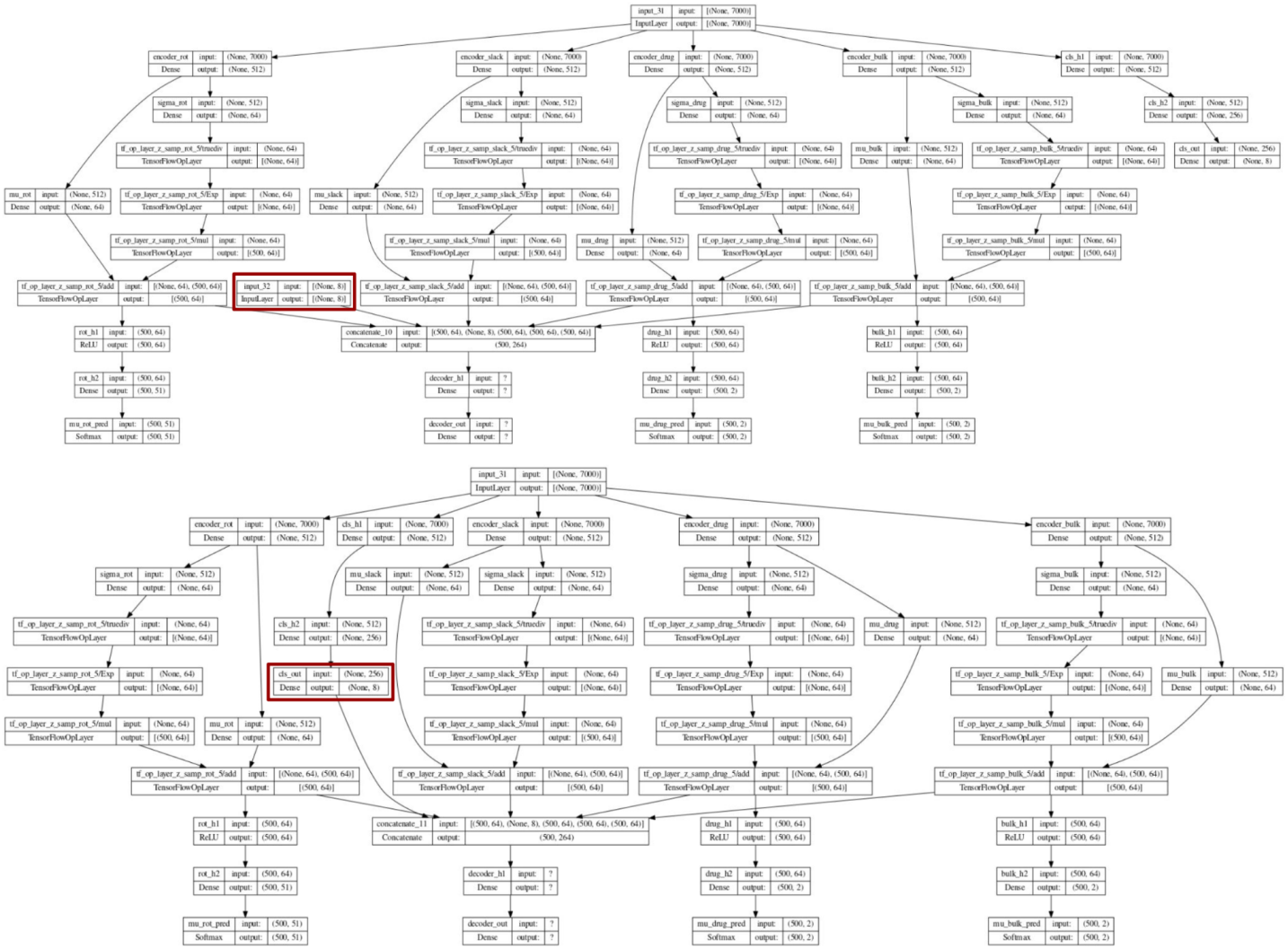
Supp Figure 1. Latent space analysis of BuDDI on Kang et al. data set with an experimental design where bulk samples are correlated with the sample IDs and perturbation status. **Panel a** depicts that average F1 score of each latent space to predict each source of variation. Midpoint coloration is the average across all observed F1 scores. Panel b compares the performance of BuDDI, CIBERSORTx, and BayesPrism, in estimating the cell type proportions. Panel c depicts each of BuDDI's latent spaces, colored by source of variation. Panel d depicts the Pearson correlation of the simulated perturbation expression, stratified by expression level.



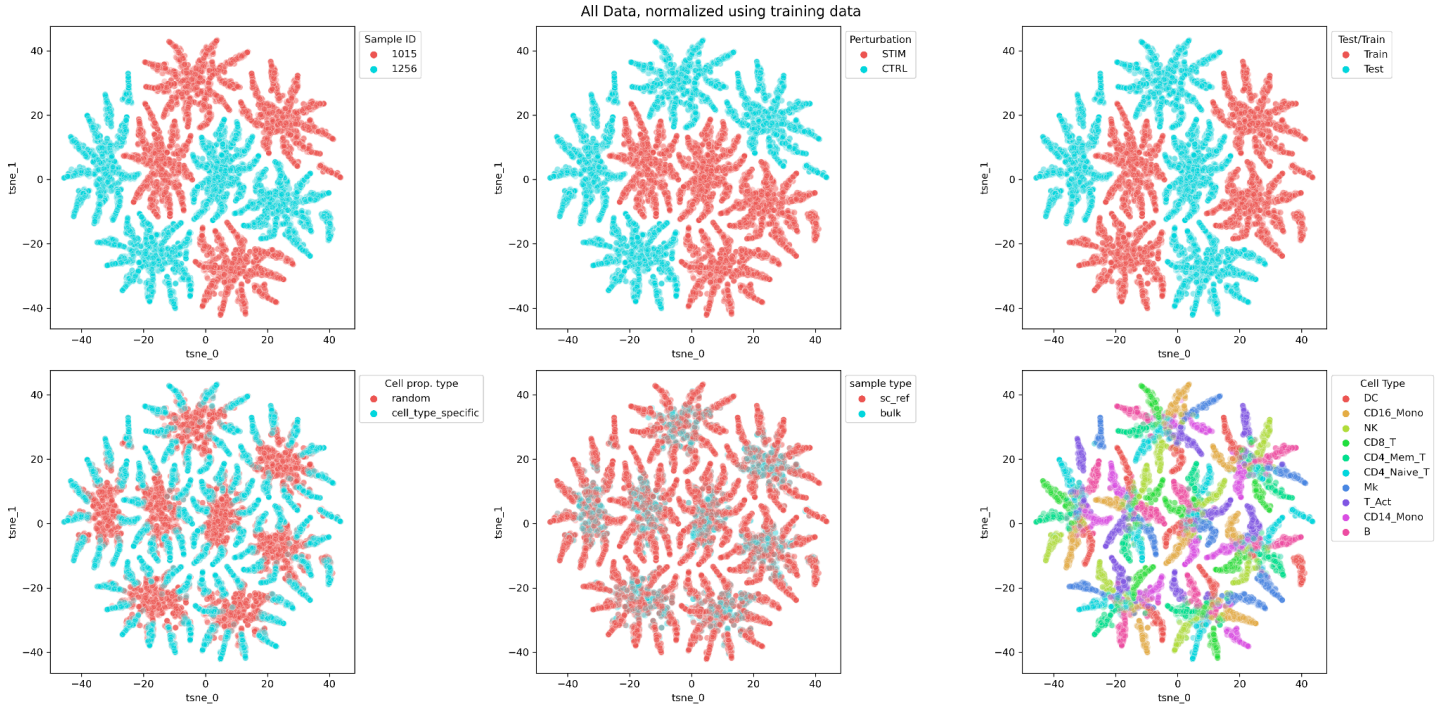
Supp Figure 2. Latent space analysis of BuDDI on Tabula Muris Senis dataset. Each column is a latent space and each row is colored by a source of variation. The second row is colored by sample ID, but due to the number of bulk samples, we omit the sample ID legend.



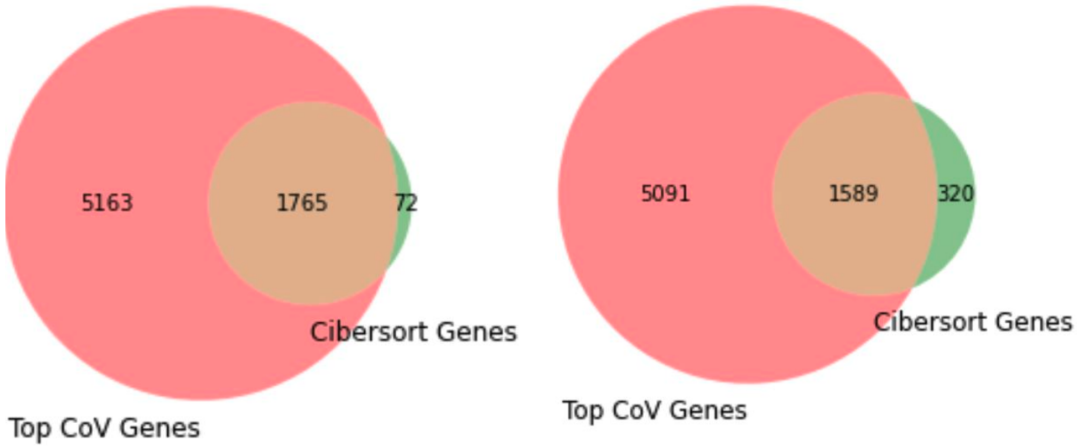
Supp Figure 3. ROC and PR curves for predicting differentially expressed genes between sexes in hepatocytes using BuDDI. Top row uses the differential expressed genes form an independent single-nucleus experiment[46] as the ground truth, bottom row uses the union of the single-nucleus and our calculated single-cell results from Tabula Muris Senis[41,42] as the ground truth.



Supp Figure 4. BuDDI model overview for the supervised (top) and unsupervised (bottom) models. The red box highlights the true or estimated cell-type proportions used in BuDDI.



Supp Figure 5. Pseudobulk data generated and colored by source of variation. Our generated data shows independence between, each source of variation, including cell-type proportion.



Supp Figure 6. Overlap of top coefficient of variation genes and CIBERSORTx signature genes used in the Kang et al. (left) and sex-dependent liver (right) analyses.

Cell Type in Kang et al.	Number of cells after filtering
CD14 Monocyte	4361
CD4 Naive T cells	2504
CD4 Memory T cells	1762
B cells	1363
CD16 Monocytes	1044
CD8 T cells	813
T Activated	633
NK	619
CD	604
Mk	236

Sample ID in Kang et al.	Number of cells after filtering
1015	3177
1256	2396
1499	2280
1244	2031
1016	1484
101	1224
1039	679
107	668

Supp Table 1. Number of cells by cell type and by sample ID in the Kang et al. dataset after filtering.

cell_ontology_class mouse.id	B cell	Kupffer cell	NK cell	endothelial cell of hepatic sinusoid	hepatic stellate cell	hepatocyte	myeloid leukocyte	plasmacytoid dendritic cell
1-M-62	1	5	2	23	2	412	1	1
1-M-63	11	489	206	320	26	827	40	14
3-F-56	0	8	2	6	1	305	0	0
3-F-57	0	1	0	4	0	152	0	0
3-M-8/9	0	0	0	4	0	371	0	0
18-F-51	22	195	59	25	2	194	42	8
21-F-54	39	21	31	12	2	229	4	1
24-M-58	0	0	0	1	0	62	0	0
24-M-59	3	18	7	5	1	2	5	0
30-M-3	33	1608	55	3	0	1	178	7
30-M-4	14	2	10	5	0	0	13	1
30-M-5	51	40	88	74	5	22	37	9

Supp Table 2. Number of cells by sample ID and cell type after filtering and before combining the two cell types “endothelial cell of hepatic sinusoid” and “hepatic stellate cell”

characteristics: sex characteristics: age	f	m
12	2.0	4.0
15	2.0	4.0
18	2.0	4.0
21	2.0	4.0
24	NaN	3.0
27	NaN	4.0
3	2.0	4.0
6	2.0	4.0
9	2.0	4.0

Supp Table 3. Number of bulk liver samples used in analysis by sample ID and age.