Deconvolving Bulk Proteomics Datasets via Deep Adaptation Network

Anonymous Author(s)
Submission Id: 57

Abstract

Accurate estimation of cellular composition in heterogeneous tissues remains a critical yet challenging task in biomedical research. We present Deconv-DAN, a novel deep–learning deconvolution framework that utilize single-cell omics data as references to infer cell-type proportions of tissues with high accuracy. Our approach begins by applying an unbiased mix-up strategy to synthesize a richly diverse set of pseudo-bulk training samples from single-cell profiles. We then employ a deep adaptation network to bridge the distribution gap between synthetic and target measurements, ensuring the accuracy on target domain. Through extensive benchmarks, we demonstrate Deconv-DAN consistently outperforms state-of-the-art methods in accuracy and stability. Our results underscore its broad applicability for deconvolution of bulk proteomics and DNA methylation data, paving the way for more precise characterization of tissue ecosystems.

CCS Concepts

• **Applied computing → Computational proteomics**.

Keywords

Deconvolution, Proteomics, Heterogeneity

ACM Reference Format:
Anonymous Author(s). 2025. Deconvolving Bulk Proteomics Datasets via Deep Adaptation Network. In *Proceedings of 16th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB' 25).* ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Proteins constitute a fundamental class of macromolecules that execute a wide array of biological functions within cells. In contrast to mRNA measurements or genomic analyses, direct quantification of protein abundance offers a more faithful reflection of cellular state. Over the past two decades, mass-spectrometry-based proteomics has seen tremendous advances in accuracy, robustness, and throughput [9]. As a result, large-scale proteomic datasets now play a pivotal role in elucidating disease mechanisms. For instance, proteomic profiling of cancer tissues enables direct assessment of the downstream effects of somatic mutations and DNA-methylation alterations, thereby informing the development of targeted immunotherapies and refining prognostic models [17].

Bulk proteomic assays measure the average protein abundance across a heterogeneous tissue, which can be viewed as the weighted sum of cell type specific profiles, with weights corresponding to each cell type's proportion. Accurately estimating these proportions is critical, since distinct cell populations contribute unique functional roles to tissue biology. The advent of single-cell proteomic technologies provides an invaluable reference for inferring the cellular composition of bulk samples.

Recently, several proteomics-specific deconvolution tools have been introduced. ProteoMixture [19] is a protein-signature–based method; however, its signature set is restricted to high-grade serous ovarian cancer (HGSOC) tumors, limiting its broader applicability. scpDeconv [22] leverages single-cell proteomics data as a reference for deconvolving bulk proteomics data, but its performance varies considerably across cell types, which constrains its practical utility. GraphDEC[3], a graph neural network–based approach, was more recently proposed; yet, as we will later demonstrate, it performs poorly on certain tasks, especially when cell type proportion distribution of target data does not match that of training data.

To overcome these limitations, we present Deconv-DAN, a novel machine learning framework that integrates single-cell proteomic references with a deep adaptation network to robustly infer cell-type proportions for bulk proteomics data. By generating diverse pseudo-bulk training samples through an unbiased mix-up strategy and aligning feature distributions between training and target data, Deconv-DAN achieves superior accuracy and stability compared to existing methods.

2 Method

Proteomics deconvolution can be considered as a matrix decomposition problem:

$$X = S\,P + E, \tag{1}$$

where X is the measured proteomics abundance matrix (proteins × samples), S is the reference signature matrix (proteins × cell-types), P is the unknown proportion matrix (cell-types × samples), and E captures noise. Instead of using statistical methods to explicitly model signature matrix and noise, we utilize a machine learning model to solve this matrix decomposition problem.

Deconv-DAN is a two-stage, supervised deep–learning pipeline for estimating cell–type composition from bulk proteomic profiles (Fig. 1). The model takes both pseudo–bulk training samples (with known cell fractions) and unlabeled target bulk samples as inputs, and outputs predicted cell–type proportions for each target.

BCB' 25, Philadelphia, PA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

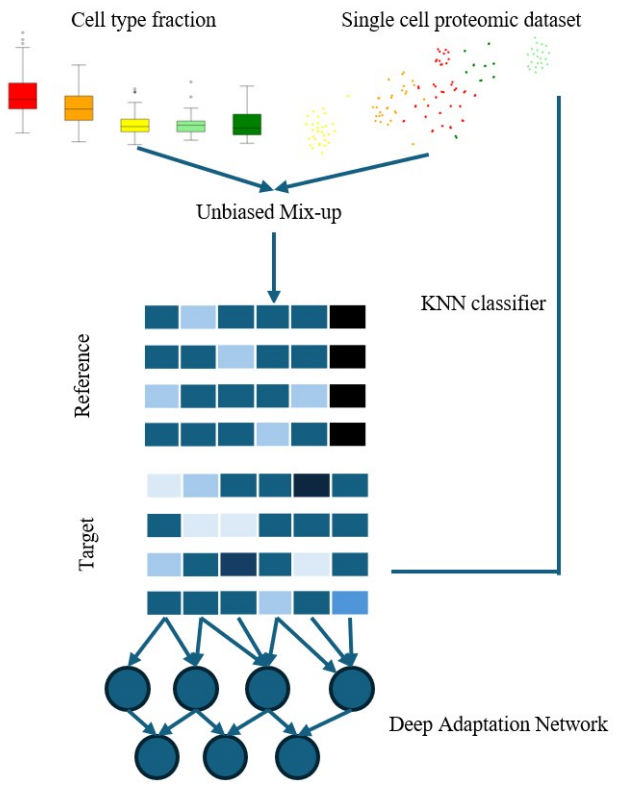

Figure 1: Overall architecture of Deconv-DAN. Pseudo bulk training samples are first generated and a deep adaptation network is then trained on such samples.

Algorithm 1: Mixup-fraction sampling for training

Input: N ; // number of cell types
$\mathbf{p} = (p_1, \ldots, p_N)$; // probability of each cell type to be the most abundant in test data
Output: $\mathbf{f} = (f_1, \ldots, f_N)$; // mix-up fractions, $\sum_i f_i = 1$
$m \leftarrow \text{CategoricalSample}(\mathbf{p})$;
// pick the most abundant cell type
$\alpha \leftarrow \text{Uniform}(0, 1)$;
// fraction for the most abundant type
for $i \leftarrow 1$ **to** $N - 1$ **do**
  $r_i \leftarrow \text{Uniform}(0, 1)$;
  // raw weight for non-most-abundants
$S \leftarrow \sum_{i=1}^{N-1} r_i$;
// normalizer
for $i \leftarrow 1$ **to** $N - 1$ **do**
  $f_i \leftarrow (r_i/S)\,(1 - \alpha)$;
  // normalize to leftover mass
$f_m \leftarrow \alpha$;
// assign most abundant mass
shuffle$(f_1, \ldots, f_N)$;
// random permutation of fractions
$j \leftarrow \arg\max_{1 \le i \le N} f_i$;
// find current max
if $j \neq m$ **then**
  swap(f_j, f_m);
  // ensure the most abundant one has the max fraction
return $\mathbf{f}$;

2.1 Stage 1: Pseudo–Bulk Generation via Unbiased Mix-Up

Starting from an $m \times n$ matrix of single-cell abundances (with m denoting the number of proteins and n the number of cells), we generate a diverse set of pseudo-bulk samples. To construct these, mixing weights are drawn from a heavy-tailed distribution, which increases the likelihood that mixtures are dominated by one or two cell types and thus avoids the overly uniform proportions typically produced by standard mix-up strategies. Our approach (Algorithm 1) yields a greater fraction of high-purity samples while preventing the generation of excessively sparse ones, enabling strong performance across a broad spectrum of target distribution patterns, including both high- and low-purity cases.

When there are in total N different cell types and assume the expert proportion of every cell type equal to 1/N, the distribution that mix-up proportions of one certain cell type would follow:

$$f_{f_k}(x) = \frac{1}{N} + \frac{(N-1)(N-2)}{N} \int_0^{1-x} \frac{u^{N-3}}{1-u}\, du, \quad 0 < x < 1. \tag{2}$$

Prior to sampling, we optionally apply a k-nearest neighbor (KNN) classifier to each target bulk sample to estimate its most abundant cell type, which can then guide the selection of dominant components when simulating training data. Alternatively, users may provide expert-derived proportion priors when available. Specifically, we train a KNN classifier on reference single-cell protein profiles annotated with cell-type labels. For each target bulk sample x, the classifier predicts a cell type $\hat{c}(x)$, which we treat as the sample's dominant cell type. Aggregating predictions across all target samples yields an empirical categorical prior p over dominant cell types. In practice, we applied this strategy only when deconvolving real bulk data and non-uniform distribution task (Section 4). When dealing with simulated data (Section 3), the KNN-based prior was replaced with an expert-defined prior that assumes each cell type is equally likely to dominate the target data.

2.2 Stage 2: Deep Domain Adaptation

Proteomic data from different sources often differ due to protocol– or instrument-specific biases. To bridge this domain gap, we adopt a Deep Adaptation Network (DAN) [15].The advantage of applying a DAN is that it explicitly minimizes the Maximum Mean Discrepancy

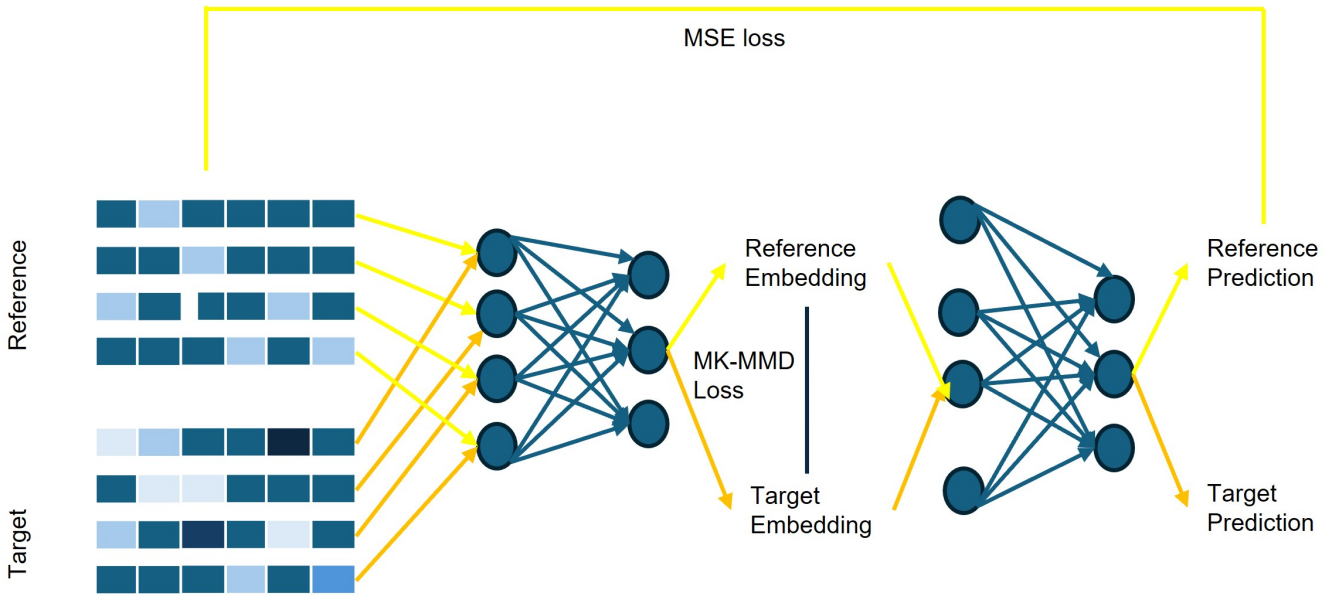

Figure 2: Deep adaptation network architecture in Deconv-DAN. The neural network is trained on the sum of MK-MMD loss and L2 prediction loss.

(MMD) between source and target embeddings and thus is generally stable and effective. The network is trained simultaneously on:

(1) **Prediction Loss** ($\mathcal{L}_{\text{pred}}$): mean squared error between predicted and true cell–type fractions on pseudo–bulk samples.
(2) **Adaptation Loss** ($\mathcal{L}_{\text{adapt}}$): multiple–kernel maximum mean discrepancy (MK–MMD) between the feature distributions of pseudo–bulk (source) and real bulk (target) samples.

We define the loss functions as follows:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(\hat{f}_{i,c} - f_{i,c}\right)^2. \tag{3}$$

where $\hat{f}_{i,c}$ denotes the model's predicted proportion of the c-th cell type for the i-th target sample, and $f_{i,c}$ denotes the ground–truth proportion of that same cell type and sample. By averaging the squared differences across all samples, $\mathcal{L}_{\text{pred}}$ computes the L_2 prediction loss (mean squared loss).

$$\mathcal{L}_{\text{adapt}} = \sum_{u=1}^{U}\beta_u \,\text{MMD}^2_{k_u}(\mathcal{D}^s, \mathcal{D}^t), \tag{4}$$

where $\mathcal{D}^s$ and $\mathcal{D}^t$ are the encoder embeddings of the training and test protein samples, respectively; U is the number of Gaussian kernels; and β_u is the weight of the u-th kernel. By summing the Maximum Mean Discrepancy (MMD) loss across several kernels, $\mathcal{L}_{\text{adapt}}$ computes the multiple–kernel variant of MMD (MK–MMD) between the training and target embeddings.

$$\text{MMD}^2_{k_u}(\mathcal{D}^s, \mathcal{D}^t) = \frac{1}{n_s^2}\sum_{i,i'=1}^{n_s} k_u(x_i^s, x_{i'}^s) + \frac{1}{n_t^2}\sum_{j,j'=1}^{n_t} k_u(x_j^t, x_{j'}^t) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k_u(x_i^s, x_j^t). \tag{5}$$

where k_u is one Gaussian kernel, n_s and n_t are the numbers of training and target samples, respectively, x_i^s is the embedding of the i-th training sample, and x_j^t is the embedding of the j-th target sample. One kernel MMD loss between training and target embeddings can be calculated from this equation.

The overall objective is:

$$\min_{\theta}\ \mathcal{L}_{\text{pred}}(\theta) + \lambda\,\mathcal{L}_{\text{adapt}}(\theta), \tag{6}$$

where λ balances prediction accuracy against domain alignment. By jointly minimizing these losses, Deconv-DAN learns feature representations that generalize effectively from training mixtures to target bulk assays, yielding robust and accurate cell–type proportion estimates.

However, it is important to notice that DAN aligns feature distributions under covariate shift, where

$$p_s(y \mid x) = p_t(y \mid x), \quad p_s(x) \neq p_t(x), \tag{7}$$

but it cannot correct for prior shift where

$$p_s(x \mid y) = p_t(x \mid y), \quad p_s(y) \neq p_t(y). \tag{8}$$

In the above formula, $p_s(x)$ and $p_t(x)$ denote the marginal feature distributions in the source and target domains, while $p_s(y)$ and $p_t(y)$ denote the corresponding marginal label distributions. $p_s(y \mid x)$ and $p_t(y \mid x)$ represent the conditional label distributions given x, and $p_s(x \mid y)$ and $p_t(x \mid y)$ represent the conditional feature

distributions given y. As DAN cannot correct for prior shift, a KNN classifier is essential for guiding the mix-up process during training. Without such guidance, feature alignment would instead bias target predictions toward the source label distribution. The KNN-based dominant component estimate provides a lightweight form of prior correction, mitigating this issue.

3 Benchmarking of Deconv-DAN's Performance on Simulated Pseudo-Bulk Data

To quantify Deconv-DAN's deconvolution accuracy, we require bulk proteomic profiles with known cell–type proportions. In the absence of such annotated bulk datasets, we generate both training and target "bulk" samples by mixing single–cell proteomics measurements according to predefined proportions. For the target set, we adopt the mix-up procedure from scpDeconv [22] to minimize bias and directly compare with other methods (Algorithm 2).

Algorithm 2: Mixup-Fraction Sampling (scpDeconv)

Input: N ; `// number of cell types`
Output: $\mathbf{f} = (f_1, \ldots, f_N)$
for $i \leftarrow 1$ **to** N **do**
  $r_i \leftarrow \text{Uniform}(0, 1)$;
$S \leftarrow \sum_{i=1}^{N} r_i$;
for $i \leftarrow 1$ **to** N **do**
  $f_i \leftarrow r_i / S$;
return $\mathbf{f}$;

Once the target pseudo-bulk samples are generated, we apply Deconv-DAN alongside scpDeconv [22], Scaden [18], GraphDEC [3], and a baseline non-negative least squares (NNLS) method to estimate cell-type fractions. For NNLS, we first compute the average abundance profile for each cell type, then estimate proportions by minimizing the squared error between the reconstructed target sample (obtained from the predicted cell-type proportions and the average profiles) and the observed target sample. It is important to note that, in simulated tasks, each cell type is typically represented in equal proportions on average. Consequently, we exclude the KNN component of Deconv-DAN in this setting, focusing instead on evaluating the contributions of the novel mix-up strategy and the deep adaptation network. We provide Deconv-DAN with an expert-defined prior, assuming that each cell type dominates in $1/N$ of the target samples.

We benchmark each method against the ground truth using Lin's concordance correlation coefficient (CCC) and root mean squared error (RMSE). These metrics can be expressed as follows:

$$\rho_c = \frac{2\, s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \tag{10}$$

where x_i is the predicted cell type proportion for sample i and y_i is the corresponding ground-truth value. The variable n is the number of target samples; s_{xy} is the sample covariance between x and y; s_x and s_y are the standard deviations of the prediction and ground truth matrices, respectively; and $\bar{x}$ and $\bar{y}$ are their corresponding averages.

3.1 Evaluation on Simple *In Silico* Tasks

We first benchmark different algorithms on simple *in silico* tasks. We design three progressively challenging tasks:

(a) **Matched protocol, cell-type resolution.** Reference and target mixtures are derived from the same experimental protocol and are annotated at the major cell–type level.
(b) **Matched protocol, subtype resolution.** As in (a), but annotations distinguish finer cell subclusters, making deconvolution more difficult due to high similarity between subtypes.
(c) **Cross-protocol.** Reference and target mixtures come from different proteomic platforms, introducing substantial domain shift.

For cases (a) and (b), we use the human breast cancer atlas from Gray *et al.* [5], which quantifies 39 proteins across nearly one million epithelial and stromal single cells from 38 patients. We conduct train-test split by patient and select 22 for training (carriers of germline mutations in BRCA1, BRCA2, or RAD51C), 16 for target (non-carriers), ensuring no data leakage. As training and testing data are obtained from the same dataset and the same protocol, there would be minimal domain adaptation issue. Major cell types comprise three epithelial subgroups (alveolar, hormone-sensing, basal/myoepithelial) and three stromal subgroups (fibroblast, vascular/lymphatic, immune), with finer Leiden-cluster subtypes also provided by Gray *et al.*[5] for Task (b).

For Task (c), we assemble two murine cell line datasets (C10, RAW, SVEC cell lines) profiled on the N2 [6] and nanoPOTS [8] platforms, each measuring over 1,000 proteins per single cell. The cross-platform design challenges methods to overcome protocol-specific biases. This is considered much more difficult than previous two cases but is also current reality of proteomics data as the accuracy of mass-spectrometry-based proteomics still have room for improvement.

As shown in Fig. S1, Deconv-DAN achieves a high overall concordance correlation coefficient (CCC > 0.9) on the breast cancer dataset at the cell-type resolution. Although it is slightly outperformed by Scaden and scpDeconv on this task, all three methods reach CCC values above 0.9, indicating strong performance. Deconv-DAN surpasses most competing methods on the breast cancer dataset at the subtype resolution in terms of RMSE (Fig. 3), demonstrating its ability to capture finer-grained features and perform higher-resolution deconvolution. On the murine cell line dataset, Deconv-DAN again achieves very high accuracy across all cell types (CCC =0.89, Fig. S1), though it is outperformed by NNLS and GraphDEC, which exceed the 0.9 threshold. While Deconv-DAN is not the top-performing method in any of these three benchmark tasks, its predictions remain consistently reliable across different tasks, underscoring its broad applicability.

When considering deconvolution at the cell-type resolution, Deconv-DAN exhibits stable performance, as reflected by the low variance in RMSE across cell types within the same task. By contrast, on the murine cell line dataset, methods such as scpDeconv and NNLS perform better on the RAW cell line but substantially

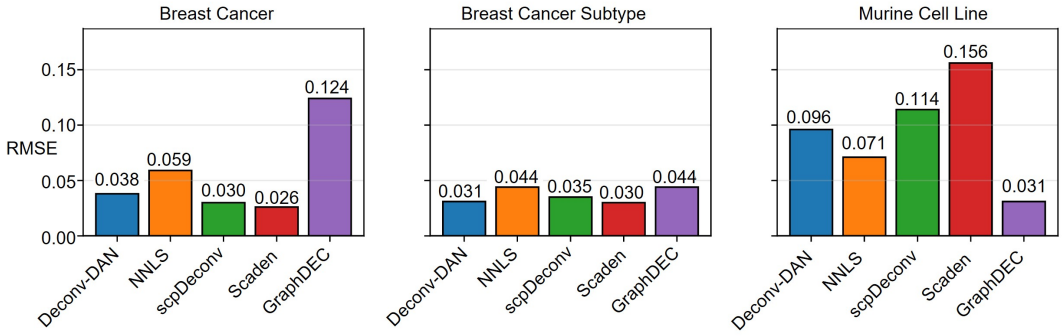

Figure 3: Comparison of Deconv-DAN and other methods on simple *in silico* tasks. RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

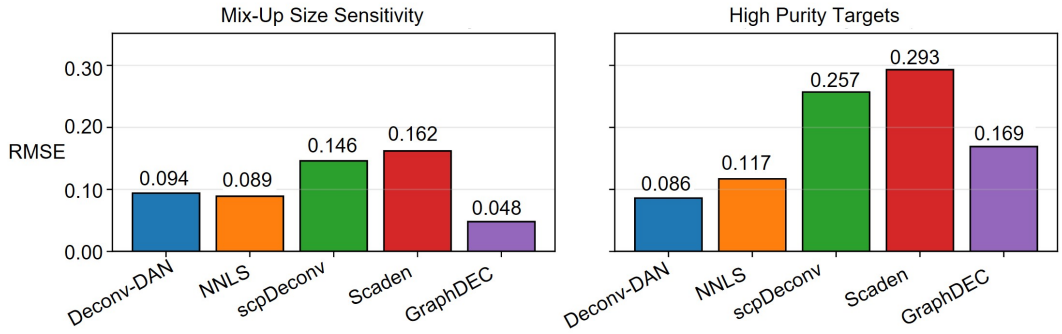

Figure 4: RMSE of Deconv-DAN, scpDeconv, and Scaden on two complex *in silico* tasks: (a) sensitivity to number of cells per target mixture, (b) high-purity samples. RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

worse on C10 and SVEC cell lines, limiting their overall applicability (Table 1).

Table 1: Benchmark of deconvolution methods on murine cell line. RMSE values of predictions on different cell types, calculated by averaging RMSE of 20 independent target data, are reported.

Method	C10	SVEC	RAW
Deconv-DAN	0.096	0.096	0.096
NNLS	0.077	0.087	0.038
scpDeconv	0.133	0.115	0.090
Scaden	0.161	0.156	0.151
GraphDEC	0.032	0.033	0.028

3.2 Evaluation of Inherent Model Bias through More Complicated *In Silico* Tasks

Models developed for proteomics deconvolution may inherently carry certain biases. In particular, hidden assumptions, if left unexamined, can lead to substantial performance degradation when models are applied across different settings. Here, we investigate two potential sources of bias in proteomics deconvolution models:

1. **Mix-Up Size.** We vary the number of single cells used to form each pseudo–bulk sample in target dataset to see whether models trained on mixtures of large number of cells degrade when applied to targets with fewer cells.

2. **Sample Purity.** We assess performance when targets have very high purity (one cell type>70 %) as models may perform better on target samples with lower purity, resulting from the fact that in some mix-up methods many training samples mimic samples with lower purity. However, in real bulk datasets, it is possible that target samples have high purity (e.g. granulocytes may comprise 55-70% of leukocytes).

We use the murine cell-line datasets (C10, RAW, SVEC) as described above for the above two tasks.

Deconv-DAN exhibits minimal variation across different mix-up sizes, highlighting its unbiased and reliable behavior. In contrast, we observed that the performance of scpDeconv declines markedly when target samples are composed of 20 single cells while reference samples are generated from 100 single cells. This sensitivity to cell count limits its applicability in real-world settings, where the number of cells contributing to bulk samples is often unknown. Although the performance of NNLS and GraphDEC also shows

slight degradation under this setting, both methods continue to produce reliable predictions, as reflected by their low RMSE values.

Deconv-DAN further maintains high accuracy on high-purity samples, underscoring its robustness in realistic deconvolution scenarios. By comparison, the performance of all other methods drops sharply in this case (Fig. 4). This decline can likely be attributed to the overly uniform mix-up strategies employed by methods such as scpDeconv and GraphDEC, which fail to generate sufficient high-purity training data. Without such data, these models struggle to accurately predict proportions when presented with high-purity target samples.

3.3 Ablation Studies

Deconv-DAN introduces a different mix-up strategy as well as applies a novel deep-learning-based regression model for the purpose of deconvolution. To disentangle the contributions of our novel mix-up strategy and the deep adaptation network, we performed two sets of ablation experiments on the murine cell-line dataset:

3.3.1 Mix-Up Strategy Ablation. Our mix-up emphasizes mixtures with a clearly dominant cell type while retaining a broad spectrum of non-sparse compositions (Fig. 5). However, scpDeconv's [22] mix-up tends to yield proportions clustered around the mean, reducing diversity. For such mix-up scheme, issues may arise when target data is of high purity, as in the case of Fig. 4. Although Scaden's [18] mix-up can generate a wide range of proportions, it often produces extreme proportions of 0 or 1 (Fig. S3) when there are 3 different cell types, which can potentially degrade performance. To evaluate the influence of mix-up method on overall performance, we replaced our novel mix-up method in turn with:

- The mix-up from scpDeconv [22],
- Scaden's mix-up [18]

For each method, we generated 1 000 target pseudo-bulk samples with the method adopted by scpDeconv[22] under two target scenarios and evaluated RMSE: (a) uniform proportions across all cell types, and (b) high-purity mixtures ($\geq$ 70% of one type).

As shown in Fig. S3, when predicting samples with high purity, mix-up strategy adopted by Deconv-DAN and Scaden [18] significantly outperformed scpDeconv's approach [22], which over-represents more balanced mixtures. Although the performance of Scaden's mix-up method is slightly better than that of Deconv-DAN in high purity samples, its performance on samples without clearly dominant cell type is much worse than Deconv-DAN (Fig. 6), likely due to its over-representation of 0 and 1 in training sample proportions (Fig. S3). In conclusion, the mix-up method proposed in Deconv-DAN achieves good performance consistently.

3.3.2 Network Architecture Ablation. Next, we replaced Deconv-DAN's deep adaptation network with two alternatives:

- A simple multilayer perceptron (as in Scaden [18]), and
- A domain-adversarial neural network (as in scpDeconv [22]).

We have found when coupled with the novel mix-up method we proposed, the neural network Scaden and scpDeconv both performed poorly when dealing with samples of high purity. The performance of three different networks does not vary by much when predicting normal samples. Thus, we conclude that Deconv-DAN's network would be superior as it performed much better in samples with high purity and comparably on normal samples.

By evaluating mix-up strategy and network architecture independently, we can conclude both the mix-up method and the application of deep adaptation network into proteomics deconvolutio each confer improvements over existing deconvolution approaches.

3.4 Application of Deconv-DAN to Cell State Deconvolution

Beyond cell-type deconvolution, estimating the proportions of cells in different cell-cycle states (G_1, S, G_2/M) is also of great importance. Unlike cell-type labels, cell-state annotated single cells are often unavailable for many cell types of interest, and reference profiles may come from distinct cell types. Since many proteins vary more strongly between cell types than between cell cycle states, this prediction task is more complicated and selecting features that correlate specifically with cell-cycle phase is critical.

Here, we adopt the dataset of Leduc *et al.* [14], which measures protein abundances in melanoma and monocyte cell lines across the three cell-cycle stages. As Leduc *et al.* suggested, we first identify proteins with significant changes in abundance across the CDC phases. Next, we construct a CDC marker vector that represents the average abundance of several proteins. The markers for a specific cell state include only those proteins whose abundances peak in the same CDC phase. Finally, the CDC markers derived from the monocyte dataset are evaluated on the melanoma dataset. Those markers that exhibit the expected correlation pattern (i.e. showing high correlation for cells in the same state and low correlation for cells in different states) are selected as the final markers.

Fig. 7 compares the performance of Deconv-DAN, NNLS, scpDeconv, Scaden and GraphDEC on two cross-cell-type tasks:

(a) **Melanoma→Monocyte.** Train on melanoma references, deconvolve monocyte mixtures.
(b) **Monocyte→Melanoma.** Train on monocyte references, deconvolve melanoma mixtures.

Deconv-DAN achieves a CCC of 0.90 (Fig. S4) and RMSE of 0.09 in both scenarios, outperforming scpDeconv, Scaden and NNLS. RMSE of GraphDEC is lower than Deconv-DAN, but both methods achieve higher than 0.9 CCC, indicating satisfactory performance. These results demonstrate that Deconv-DAN generalizes beyond cell-type deconvolution to accurately resolve cell-state proportions across different lineages.

3.5 Application of Deconv-DAN to Other Omics Data

Beyond proteomics, DNA methylation profiling is a cornerstone of epigenomic research. DNA methylation regulates transcription process where genome areas highly methylated are usually silent and not transcribed into mRNA. Deficiencies of DNA replication machinery may unexpectedly activate areas of the genome, generating mRNA not present in normal cells and leading to cancer. Bulk DNA methylation sequencing datasets can be deconvoluted with corresponding experimentally purified bulk cell population datasets

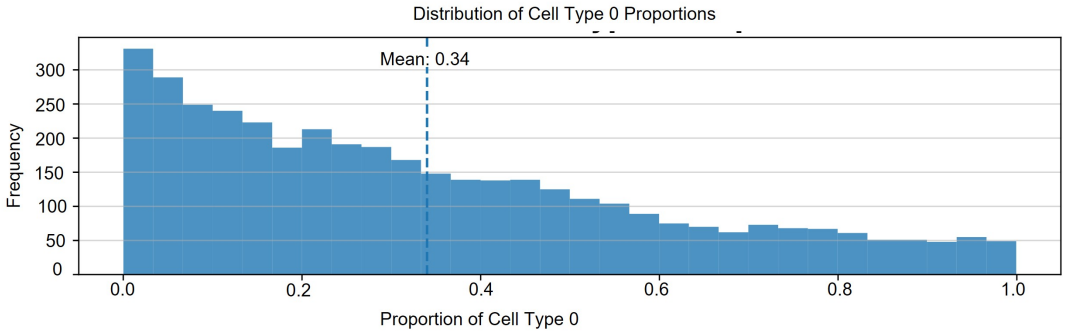

Figure 5: Histograms illustrate the distribution of one selected cell type's proportions in simulated data. Compared to scpDeconv and Scaden, Deconv-DAN's mix-up produces more high-purity samples (> 0.7 proportion) and fewer extreme-sparse samples (≈ 0 proportion).

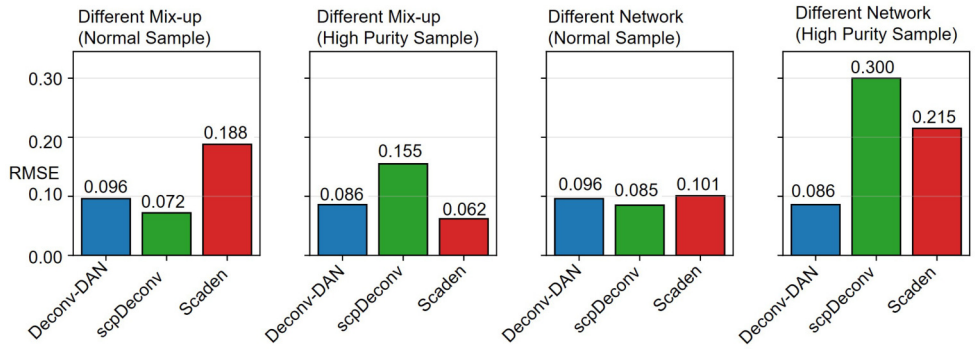

Figure 6: Ablation results: (left) RMSE for different mix-up methods; (right) RMSE for different network backbones.RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

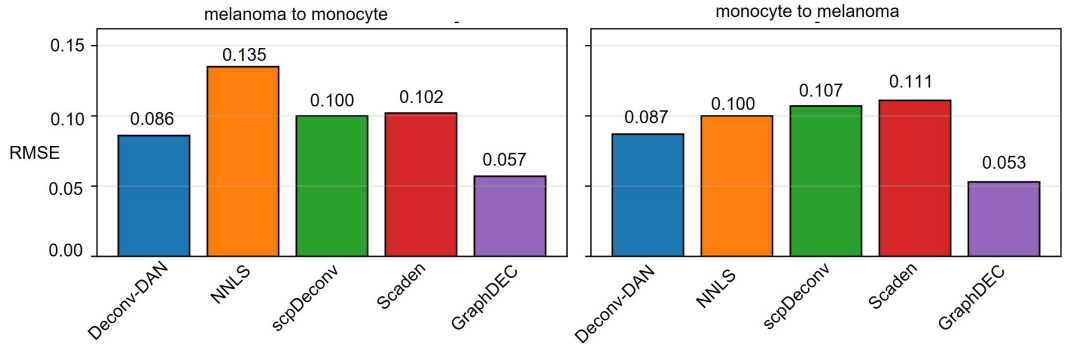

Figure 7: Performance of Deconv-DAN and other methods on cell-state deconvolution. Left: Melanoma→Monocyte. Right: Monocyte→Melanoma.RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

as reference to reveal patterns in different tissues. 450K/850K microarrays and whole-genome bisulfite sequencing (WGBS) are three methods quantifying DNA methylation.

We assembled six independent datasets across six immune cell types (B cell, $CD4^+$ T cell, $CD8^+$ T cell, neutrophil, natural killer cell, monocyte):

- Two datasets with experimentally purified bulk cell population assayed on Illumina 450K platforms.
- Two datasets with experimentally purified bulk cell population assayed on Illumina EPIC platforms.
- Two datasets with experimentally purified bulk cell population sequenced via whole genome bisulfide sequencing.

From each reference, we selected the top 100 CpG sites (450K/850K microarray) or 100 non-overlapping genomic regions flanking 100bp (WGBS) most predictive of cell identityas markers (600 total markers), following the instruction provided by DeRidder *et al.* [20]. We initially filter CpGs for each of the different cell types, selecting only those that display Benjamini–Hochberg adjusted significant Welch two-sample t-test p-values (with a significance level of 0.05) computed between the methylation values of the target group and all other groups. From this set of CpGs, we select the 100 CpGs with the highest mean methylation differences between the target group and the other groups to compile the complete marker list.

We benchmarked Deconv-DAN against five recent methylation-deconvolution methods: EMeth (normal variant) [24], MethylResolver [1], Iced-t [23], FARDEEP [10], and DCQ [4] and a baseline NNLS. As shown in Fig. 8, Deconv-DAN achieves acceptably low RMSE when deconvolving 450K microarray data, 850K microarray data, and WGBS data. Although Deconv-DAN is developed for the main purpose of proteomics deconvolution and is not specifically adapted for the use of methylation settings, its performance is better than DCQ and NNLS on all three datasets, and better than FARDEEP, emeth_normal, methylresolver, and Iced-t on at least one benchmarking dataset.

It is noteworthy that when predicting proportion of B cells in terms of WGBS dataset, the RMSE of Deconv-DAN is much lower than other methods (Fig. 9), except NNLS. This might be explained by the fact that the correlation between reference and target B cell methylation beta value is lower, compared with other cell types (Fig. S6). Therefore, methods which do not take such domain shift into consideration would have lower accuracy while Deconv-DAN achieves better results as it can explicitly handle the domain adaptation problem. Although NNLS performs quite well on B cells in WGBS data, its overall performance is not as well as Deconv-DAN, as shown in Fig. 8.

Also, in terms of runtime, Deconv-DAN ranks among the fastest deconvolution tools. Processing 1,000 bulk methylation samples requires only around 25 s, substantially faster than emeth [24], methylresolver [1], Iced-t [23], and FARDEEP [10], thereby enabling rapid, large-scale methylation deconvolution.

4 Benchmarking of Deconv-DAN's Performance on Real Bulk Data

4.1 Proteomics Deconvolution

In silico benchmarks often assume that each cell type contributes equally on average (i.e. each has mean proportion $1/C$ for C types) and models may exploit this uniform-mixup "shortcut" instead of learning true deconvolution signals. In real bulk samples, however, cell-type frequencies vary widely. To reveal such biases, we generate a new set of pseudo–bulk targets in which each sample contains at least 50 % of one designated cell type (e.g. C10). This enforces a nonuniform distribution of cell–type proportions across the target dataset.

As shown in Fig.10, the performance of scpDeconv, Scaden and GraphDEC drops significantly when the target mixtures contain non-uniform cell-type proportions, whereas Deconv-DAN and NNLS remain largely unaffected. This robustness of Deconv-DAN stems from its KNN-guided mix-up. By estimating the dominant cell type in each target sample, it tailors the training pseudo–bulk generation to better match the true target distribution. When this KNN step is omitted, Deconv-DAN's predictions on the murine cell–line data become highly inaccurate (Fig. S5), highlighting the necessity of aligning the training distribution with the target.

We also tested scpDeconv and Scaden with KNN priors provided to guide the mix-up step. RMSE of scpDeconv improved from 0.217 to 0.127 and that of Scaden improved from 0.207 to 0.175, but both remained worse than Deconv-DAN.

We further evaluated Deconv-DAN on a glioma proteomics dataset which requires prediction of proportions of five spatial niches (cellular tumor (CT), infiltrating region (IT), microvascular proliferation (MVP), tumor cells around necrosis (PAN), and adjacent leading-edge brain tissue (LE)) in real bulk tumor samples. The target dataset comprises 100 samples from patients and an additional 10 samples from healthy individuals[7]. While no ground-truth proportions exist, biological expectations are clear: patient samples should have high CT, controls should have high LE.[12][13]

Table 2: Predicted mean LE (controls) and CT (patients)

Method	LE (controls)	CT (patients)
Deconv-DAN	0.54	0.45
Scaden	0.65	0.24
scpDeconv	0.37	0.21
GraphDEC	0.50	0.20
NNLS	0.46	0.27

As shown in Table 2, Deconv-DAN provided the best overall agreement with biological prior knowledge. Although Scaden's and GraphDEC's prediction for the average LE proportion in control samples is more accurate, it fails to estimate the CT proportion among patients effectively. scpDeconv and NNLS both fail to accrately predict average CT proportion among patients.

4.2 Benchmarking on real DNA Methylation Data

We also benchmarked Deconv-DAN against established methylation deconvolution methods on a dataset which profiled DNA methylation status of whole blood and reconstructed mixtures of purified leukocytes isolated from human adult blood (GSE77797[11]). The reference dataset integrated purified bulk profiles three datasets (GSE65097[2], GSE71244[16], and GSE110554[21]. Six leukocyte types (CD4+ T cells, CD8+ T cells, B cells, natural killer cells, monocytes, and neutrophils) were quantified via flow cytometry.

As shown in Fig. 11, Deconv-DAN achieves competitive accuracy, outperforming IceDT and emeth_normal. Deconv-DAN is slightly

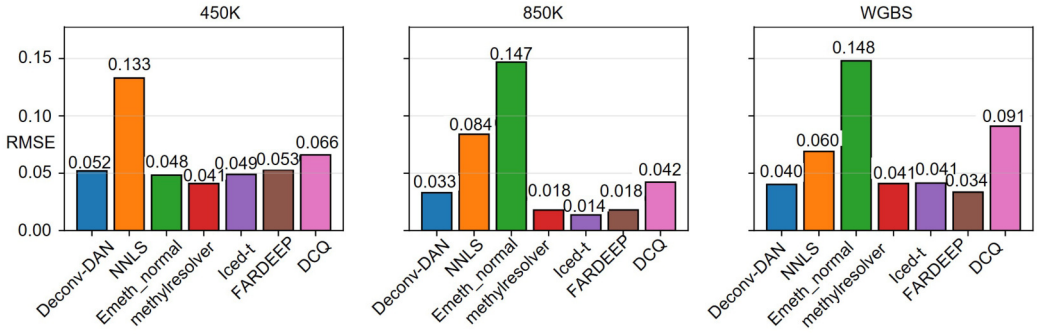

Figure 8: Benchmarking of Different Models on DNA methylation deconvolution. Left: perfomance on 450K microarray deconvolution. Middle: performance on 850K microarray deconvolution. Right: performance on WGBS deconvolution.RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

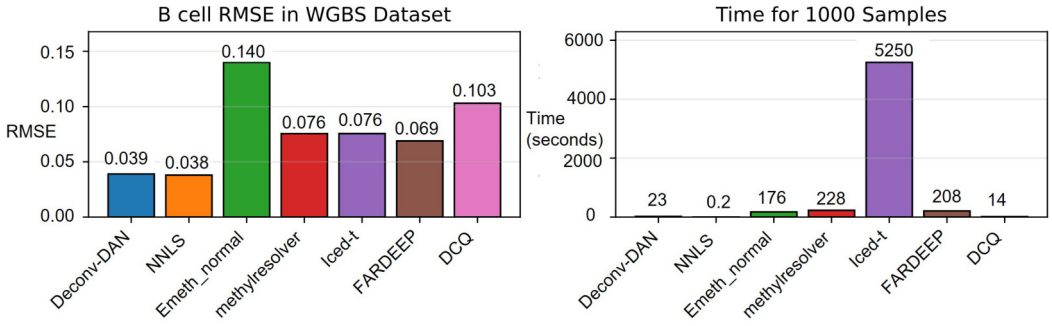

Figure 9: Benchmarking of Different Models on DNA methylation deconvolution. Left: RMSE when predicting B cell proportions in terms of WGBS data (RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.). Right: time taken to deconvolute 1000 target WGBS pseudo bulk DNA methylation samples.

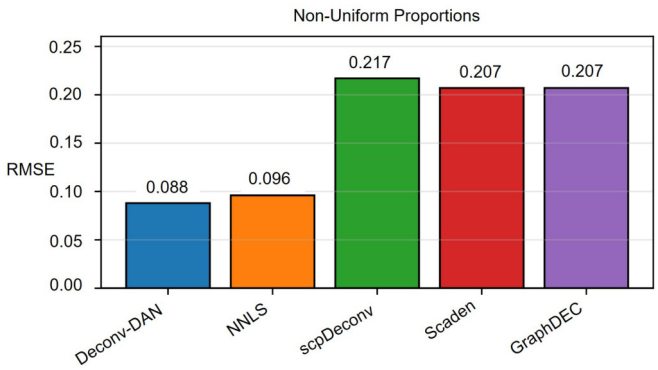

Figure 10: Performance of Deconv-DAN and other methods on non-uniform proportion task. RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

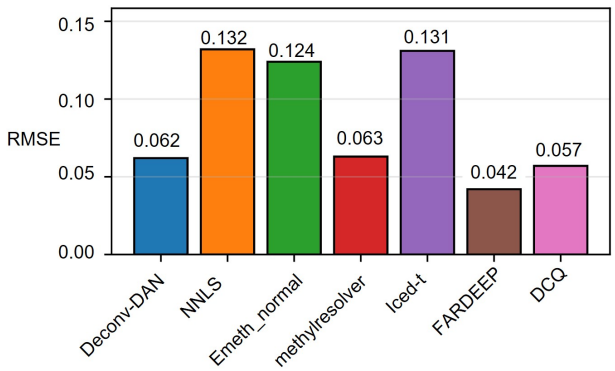

Figure 11: Performance of Deconv-DAN and other methods on real 450K bulk methylation profile deconvolution. RMSE is computed per target dataset and reported RMSE is the average value on 20 simulated target datasets.

better compared with methylresolver, but it performed worse compared with DCQ and FARDEEP. We also explored whether suboptimal classification made by KNN would impact deconvolution results. The true dominant cell type was neutrophil for 14 samples, CD8+ T cell for 2, NK cell and B cell each for 1. KNN predicted 15

neutrophil and 3 NK. The resulting RMSE (0.0619) was similar to the perfect-classification case (0.0629) but far better than a deliberately poor KNN in which each cell type is predicted to be dominant in 3 target samples (0.15). Thus, small KNN errors have negligible effect, but large errors degrade performance significantly.

5 Discussion

We have presented Deconv-DAN, a novel deep–learning framework for deconvolving bulk proteomic profiles using single-cell references. Through comprehensive benchmarks on proteomics mixtures, Deconv-DAN outperforms existing methods in accuracy in many cases, and at the same time addressed several potential biases in proteomics deconvolution models. Although originally designed for proteomic data, our experiments demonstrate its applicability to DNA-methylation deconvolution with high fidelity.

Looking forward, we plan to explore the possibility to adapt Deconv-DAN to bulk RNA deconvolution. We are also exploring potential strategies to prevent the KNN step from producing highly inaccurate predictions. However, at present, our only mitigation is to monitor KNN outputs and, when necessary, replace them with an estimated most abundant cell type based on biological insight.

6 Data and Code Availability

All datasets and code supporting this study are publicly accessible:

- **Human breast cancer atlas.** Processed data are available at https://data.mendeley.com/datasets/vs8m5gkyfn/1.
- **Murine cell-line proteomics.** Processed data are obtained from Wang *et al.*[22] and can be downloaded from https://github.com/TencentAILabHealthcare/scpDeconv.
- **Cell-cycle proteomics.** Data from Leduc *et al.*[14], and description of suggestted processing methods are available at https://scp.slavovlab.net/Leduc_et_al_2022.
- **Single-cell DNA methylation.** Processed 450K, 850K and WGBS datasets from DeRidder *et al.* [20] can be obtained from source data of their article (see Source Data Sheets 3A, 4A, 7A for references; S3A, S5A, S10A for targets).
- **Real glioma proteomics** Proteomics data can be obtained from the supplementary material section for both reference [12] and target [7] datasets.
- **Real DNA methylation data** Four DNA methylation datasets used can be obtained via GSE access code GSE65097, GSE71244, GSE77797, and GSE110554.

All code relevant to this manuscript can be accessed at: https://anonymous.4open.science/r/deconv-dan-2D8F/.

References

[1] Douglas Arneson, Xia Yang, and Kai Wang. 2020. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Communications Biology* 3 (12 2020). Issue 1. doi:10.1038/s42003-020-01146-2

[2] Patrick Coit, Srilakshmi Yalavarthi, Mikhail Ognenovski, Wenpu Zhao, Sarfaraz Hasni, Jonathan D. Wren, Mariana J. Kaplan, and Amr H. Sawalha. 2015. Epigenome profiling reveals significant DNA demethylation of interferon signature genes in lupus neutrophils. *Journal of Autoimmunity* 58 (4 2015), 59–66. doi:10.1016/j.jaut.2015.01.004

[3] Zhiming Dai, Yujie Song, Tuoshi Qi, Hongyu Zhang, Huiying Zhao, Zheng Wang, Yuedong Yang, and Yuansong Zeng. 2025. Deciphering Cell Type Abundance in Proteomics Data Through Graph Neural Networks. *Advanced Science* (2025). doi:10.1002/advs.202502987

[4] Samuel A. Danziger, David L. Gibbs, Ilya Shmulevich, Mark McConnell, Matthew W.B. Trotter, Frank Schmitz, David J. Reiss, and Alexander V. Ratushny. 2019. AdApTS: Automated deconvolution augmentation of profiles for tissue specific cells. *PLoS ONE* 14 (11 2019). Issue 11. doi:10.1371/journal.pone.0224693

[5] G. Kenneth Gray et al. 2022. A human breast atlas integrating single-cell proteomics and transcriptomics. *Developmental Cell* 57 (6 2022), 1400–1420.e7. Issue 11. doi:10.1016/j.devcel.2022.05.003

[6] Jongmin Woo et al. 2021. High-throughput and high-efficiency sample preparation for single-cell proteomics using a nested nanowell chip. *Nature Communications* 12 (12 2021). Issue 1. doi:10.1038/s41467-021-26514-2

[7] Liang Bo Wang et al. 2021. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* 39 (4 2021), 509–528.e20. Issue 4. doi:10.1016/j.ccell.2021.01.006

[8] Maowei Dou et al. 2019. High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform. *Analytical Chemistry* 91 (10 2019), 13119–13127. Issue 20. doi:10.1021/acs.analchem.9b03349

[9] Tiannan Guo, Judith A Steen, and Matthias Mann. 2025. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature* 638 (2 2025), 901–911. Issue 8052. doi:10.1038/s41586-025-08584-0

[10] Yuning Hao, Ming Yan, Blake R. Heath, Yu L. Lei, and Yuying Xie. 2019. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS Computational Biology* 15 (5 2019). Issue 5. doi:10.1371/journal.pcbi.1006976

[11] Devin C. Koestler, Meaghan J. Jones, Joseph Usset, Brock C. Christensen, Rondi A. Butler, Michael S. Kobor, John K. Wiencke, and Karl T. Kelsey. 2016. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics* 17 (3 2016). Issue 1. doi:10.1186/s12859-016-0943-7

[12] K. H.Brian Lam and Phedias Diamandis. 2022. Niche deconvolution of the glioblastoma proteome reveals a distinct infiltrative phenotype within the proneural transcriptomic subgroup. *Scientific Data* 9 (12 2022). Issue 1. doi:10.1038/s41597-022-01716-5

[13] K. H.Brian Lam, Alberto J. Leon, Weili Hui, Sandy Che Eun Lee, Ihor Batruch, Kevin Faust, Almos Klekner, Gábor Hutóczki, Marianne Koritzinsky, Maxime Richer, Ugljesa Djuric, and Phedias Diamandis. 2022. Topographic mapping of the glioblastoma proteome reveals a triple-axis model of intra-tumoral heterogeneity. *Nature Communications* 13 (12 2022). Issue 1. doi:10.1038/s41467-021-27667-w

[14] Andrew Leduc, R. Gray Huffman, Joshua Cantlon, Saad Khan, and Nikolai Slavov. 2022. Exploring functional protein covariation across single cells using nPOP. *Genome Biology* 23 (12 2022). Issue 1. doi:10.1186/s13059-022-02817-5

[15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. (2 2015). http://arxiv.org/abs/1502.02791

[16] Shimrat Mamrut, Nili Avidan, Elsebeth Staun-Ram, Elizabeta Ginzburg, Frederique Truffault, Sonia Berrih-Aknin, and Ariel Miller. 2015. Integrative analysis of methylome and transcriptome in human blood identifies extensive sex- and immune cell-specific differentially methylated regions. *Epigenetics* 10 (2015), 943–957. Issue 10. doi:10.1080/15592294.2015.1084462

[17] D R Mani, Karsten Krug, Bing Zhang, Shankha Satpathy, Karl R Clauser, Li Ding, Matthew Ellis, Michael A Gillette, and Steven A Carr. 2022. Cancer proteogenomics: current impact and future prospects. *Nature Reviews Cancer* 22 (2022), 298–313. Issue 5. doi:10.1038/s41568-022-00446-5

[18] Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. 2020. Deep learning-based cell composition analysis from tissue expression profiles. *Science Advances* 6 (7 2020), eaba2619. https://scaden.ims.bio

[19] Pang ning Teng et al. 2024. ProteoMixture: A cell type deconvolution tool for bulk tissue proteomic data. *iScience* 27 (3 2024). Issue 3. doi:10.1016/j.isci.2024.109198

[20] Kobe De Ridder, Huiwen Che, Kaat Leroy, and Bernard Thienpont. 2024. Benchmarking of methods for DNA methylome deconvolution. *Nature Communications* 15 (12 2024). Issue 1. doi:10.1038/s41467-024-48466-z

[21] Lucas A. Salas, Devin C. Koestler, Rondi A. Butler, Helen M. Hansen, John K. Wiencke, Karl T. Kelsey, and Brock C. Christensen. 2018. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biology* 19 (5 2018). Issue 1. doi:10.1186/s13059-018-1448-7

[22] Fang Wang, Fan Yang, Longkai Huang, Wei Li, Jiangning Song, Robin B. Gasser, Ruedi Aebersold, Guohua Wang, and Jianhua Yao. 2023. Deep domain adversarial neural network for the deconvolution of cell type mixtures in tissue proteome profiling. *Nature Machine Intelligence* 5 (11 2023), 1236–1249. Issue 11. doi:10.1038/s42256-023-00737-y

[23] Douglas R. Wilson, Chong Jin, Joseph G. Ibrahim, and Wei Sun. 2020. ICeD-T Provides Accurate Estimates of Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene Expression Patterns. *J. Amer. Statist. Assoc.* 115 (7 2020), 1055–1065. Issue 531. doi:10.1080/01621459.2019.1654874

[24] Hanyu Zhang, Ruoyi Cai, James Dai, and Wei Sun. 2021. EMeth: An EM algorithm for cell type decomposition based on DNA methylation data. *Scientific Reports* 11 (12 2021). Issue 1. doi:10.1038/s41598-021-84864-9

A Supplementary Figures

Below we would like to provide supplementary figures to further illustrate the results. Please check them in the next pages.

Received ; revised ; accepted

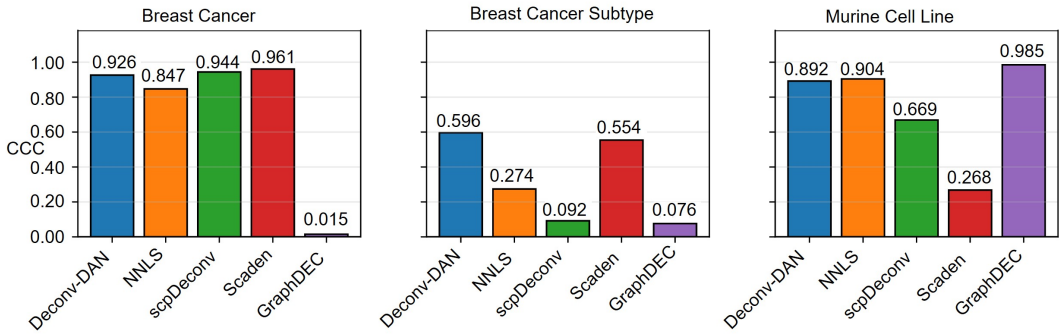

Figure S1: Comparison of Deconv-DAN and other methods on simple *in silico* tasks. CCC is computed per target dataset and reported CCC is the average value on 20 simulated target datasets.

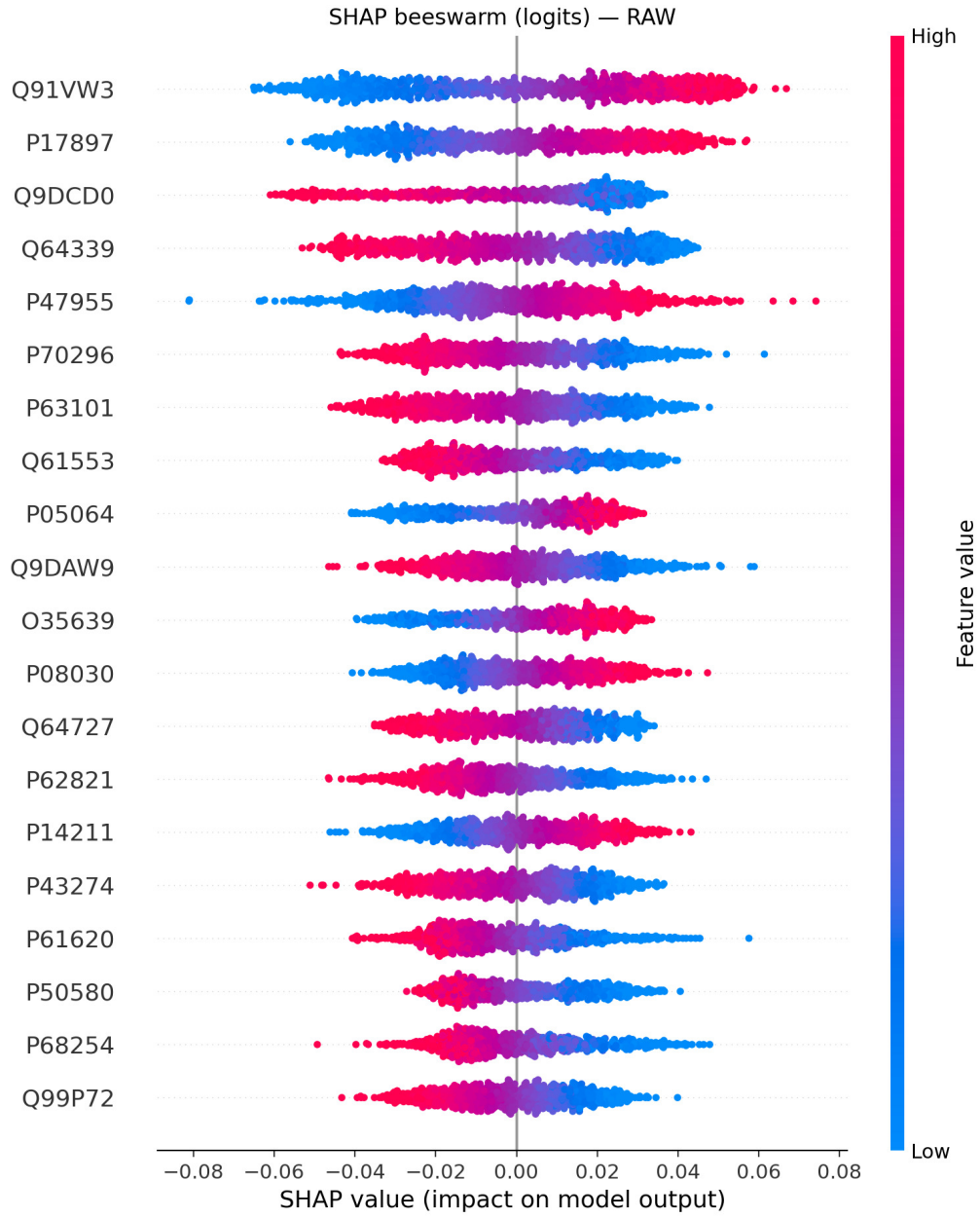

Figure S2: Top 20 proteins most strongly influencing Deconv-DAN's prediction of RAW proportion. Positive values indicate association with higher predicted RAW proportion. Seven proteins are positively associated with higher predicted RAW proportions. Q91VW3 (SH3 domain-binding glutamic acid-rich-like protein 3) exerts the strongest positive influence, consistent with the two datasets as in both datasets it is upregulated in RAW cells. P17897 (Lysozyme C-1), also positively associated with RAW proportion, matches the expected biology of macrophages. Other positively associated proteins include P47955, P05064, O35639, P08030, and P14211, covering biological functions such as translation, the pentose phosphate pathway, AMP formation, and post-translational modification.

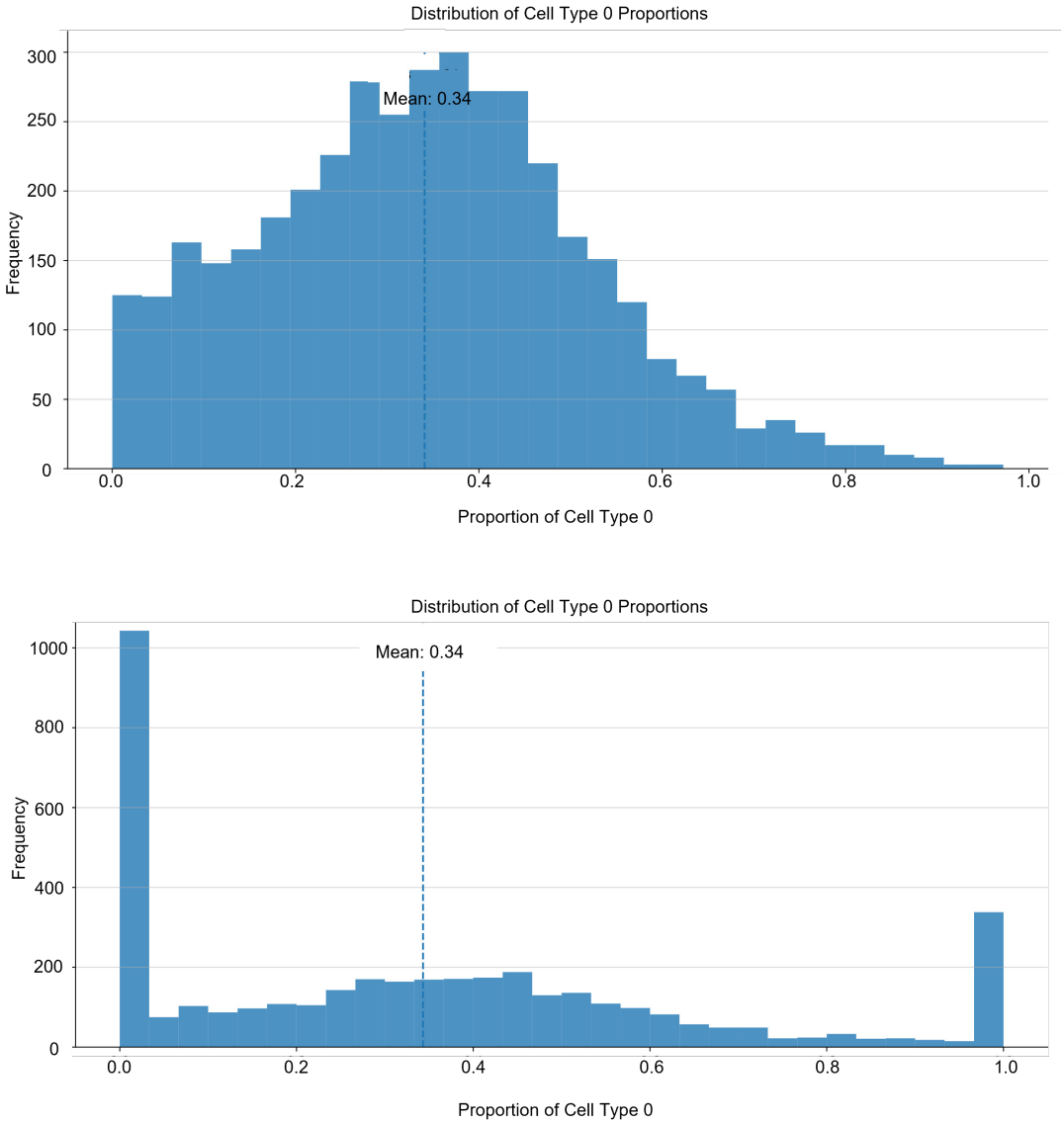

Figure S3: Comparison of mix-up strategies. Histograms show the distribution of the mix-up proportions of one cell type over 4 000 samples and 3 cell types. Top: scpDeconv[22] Bottom: Scaden[18]

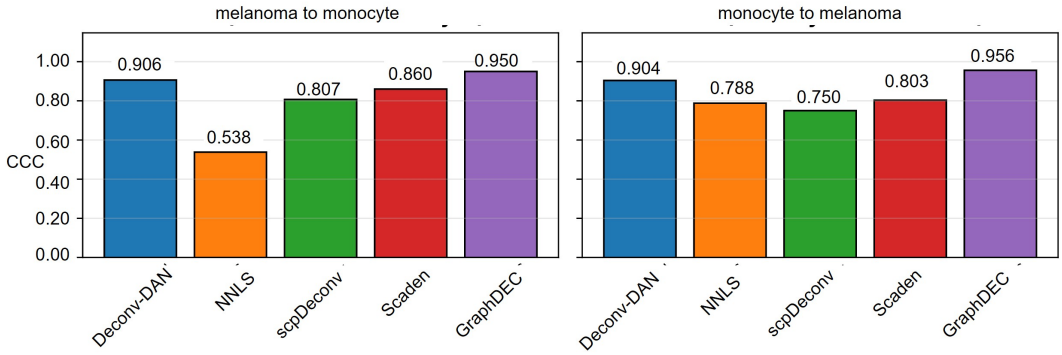

Figure S4: Performance of Deconv-DAN and other methods on cell-state deconvolution. Left: Melanoma→Monocyte. Right: Monocyte→Melanoma.CCC is computed per target dataset and reported CCC is the average value on 20 simulated target datasets.

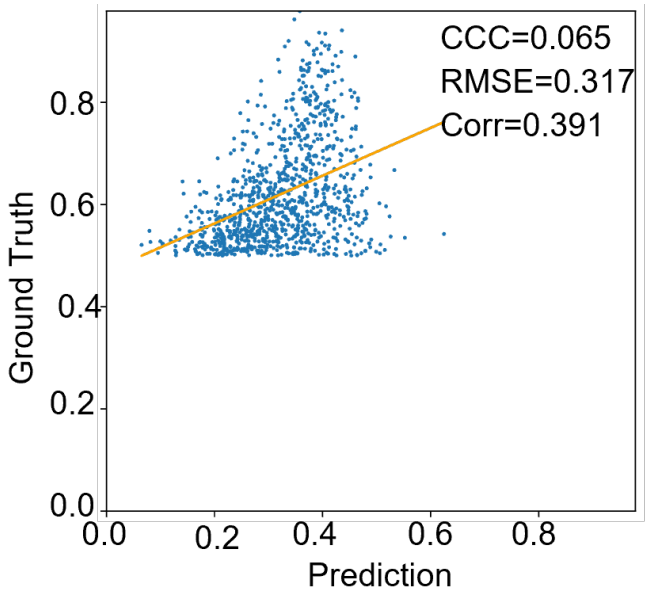

Figure S5: Deconv-DAN performance without the KNN-guided mix-up step, illustrating the need to match training and target distributions.

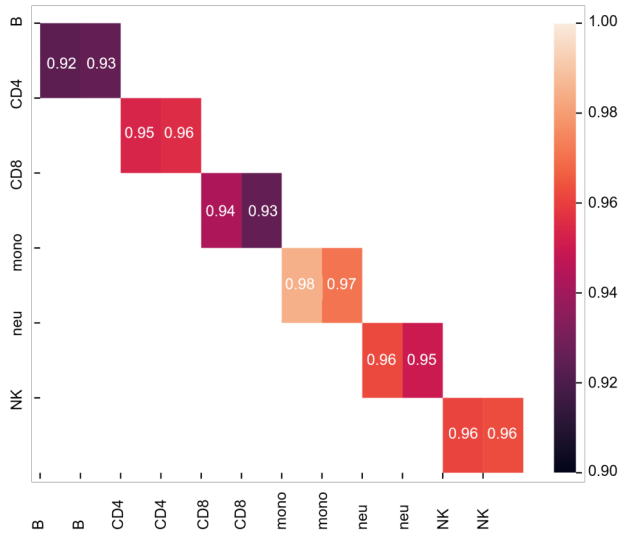

Figure S6: Pearson correlation between reference and target single-cell methylation profiles, illustrating domain shifts.The reported correlations as shown in the heatmap are computed using the 600 selected markers.