Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Bulk Proteome Deconvolution

Bachelor Thesis

Samuel Francis Gair

August 1, 2025

Advisors: Prof. V. Boeva, T. Scheithauer, A. Kraft

Department of Computer Science, Institute for Machine Learning, Computational Cancer Genomics Lab, ETH Zürich

Abstract

Lymphoma is a type of blood cancer affecting the white blood cells, called lymphocytes, which are an important part of the human immune system. To analyse this disease, many RNA-based algorithms have been proposed to estimate the cell composition of tumor environments. However, approximately one third of patients who develop a drug resistance to a subtype of lymphoma exhibit a short expected overall survival of approximately a year and a half. There remains a gap in accurate diagnosis and choosing the right strategy for an individual patient. We demonstrate the efficacy and limitations of various proteomic deconvolution methods with respective normalization strategies and provide a testing framework that generates realistic data and evaluates proteomic profile estimations to extend our current understanding of this disease. The framework serves as a foundation for developing and benchmarking future tools that potentially generate more accurate estimations for proteomic profiles, which in turn contributes to improved fraction estimation, ultimately advancing personalized treatment strategies for lymphoma patients.

Contents

Chapter 1

Introduction

Deconvolution is the computational process that aims to disentangle mixed signals into their constituent components. This approach is valuable for a variety of complex tasks where direct, high-resolution measurement of individual elements within a heterogeneous mixture is challenging or costly. The applications range from wastewater-based epidemiology for the assessment of illicit drug consumption [10] to disentangling overlapping light profiles from spatially coincident sources for measuring the brightness or size of galaxies in a cluster, or binary stars [2]. In fact, our brain effortlessly performs a form of deconvolution when separating and following individual conversations despite numerous people speaking simultaneously in a noisy room [8].

In biomedical research, cellular deconvolution addresses a fundamental challenge: bulk tissue samples contain mixtures of diverse cell types, each with distinct molecular signatures, but standard analytical techniques measure the aggregate signal from all cells simultaneously. This limitation is particularly relevant in the study of B-cell lymphoid malignancies, including chronic lymphocytic leukemia, Waldenström macroglobulinemia, mantle cell lymphoma, and marginal zone lymphoma, complex and heterogeneous diseases where understanding cellular composition is crucial for clinical decision-making [13]. When analyzing a patient's blood sample using mass-spectrometry-based proteomics, we do not measure the protein content of each cell type individually. Instead, we capture the combined signals of a diverse array of cell types that possess distinct protein expression patterns and specialized functions. It is difficult to determine which cell types contributed what proportions to the bulk sample data. This deconvolution is critical for gaining cell-type-resolved insights into the tumor microenvironment of B-cell lymphomas, which is essential for understanding disease progression and the mechanisms of drug resistance, particularly to Bruton tyrosine kinase inhibitors (BTKi) [9].

While numerous computational methods have been developed for RNA-based cellular deconvolution, relatively few approaches have been specifically designed or thoroughly validated for proteomic data. The distinct properties of protein expression, including different intensity distributions, and measurement technologies may require different analytical considerations than those developed for transcriptomic applications. To address this gap, we implement a comprehensive framework for generating biologically inspired synthetic bulk samples and optimized single-cell reference profiles, enabling systematic evaluation of various deconvolution methods across different normalization approaches. We also employ a Bayesian inference model that estimates the reference profiles of cell types and provide a methodology to benchmark and visualize those signature matrix estimations. Through this systematic approach, we aim to establish best practices for proteomic deconvolution and identify the most promising computational strategies. A full implementation of the framework, including all code and evaluation tools, is available as open-source; detailed usage and repository information are provided in the Methods section.

Chapter 2

Background

Lymphomas represent a significant health burden as approximately two percent of men and women will be diagnosed with with non-Hodgkin lymphoma at some point during their lifetime [6], where B-cell lymphomas make up a large majority of NHL [1]. In the USA, UK, Japan, France, Germany, Italy and Spain, diagnosed cases are projected to increase by 15% in the next decade [5]. Over the past years, a novel treatment option for B cell malignancies has been developed. Ibrutinib, a Bruton tyrosine kinase inhibitor, prolongs progression-free and overall survival when compared with many traditional treatment options for patients with B-cell malignancies. Despite the substantial efficacy of BTKi's in multiple B cell malignancies, the Achilles heel of this drug class is either primary or acquired resistance. Acquired ibrutinib resistance has been reported in 11% to 38% of patients with a subtype of B cell lymphoma. Initial subsets of patients who developed resistance exhibited a short expected overall survival of 4 to 18 months. Because of the poor clinical outcomes of this population, investigation into the mechanism of resistance and alternative treatment strategies to circumvent resistance is crucial to prolong the survival of these patients [13]. This resistance often stems from specific cellular and molecular changes, including mutations in BTK, but also from paracrine mechanisms. Therefore, accurate deconvolution of bulk data to charactersize the cellular composition, particularly the proportions of B cell, NK cell and T cell subsets within these complex blood cancer samples is essential.

Using patients' peripheral blood samples, it is possible to physically separate the different cell types. However, this process is laborious and subjects significant stress on the sample, which is at high risk of altering measurements [14].

A frequently used approach to better understand cell distributions is to perform bulk RNA deconvolution. However, this approach may miss crucial information for the study of cell composition. Proteins are the functional

molecules that directly influence essential cellular behaviours that play an important role in cancer, such as growth, division, and death. As a result of the imperfect correlation between the proteome and transcriptome [12], cell proportions derived from RNA data may insufficiently capture cellular cancer processes on the level of proteins. First, unlike RNA, proteins undergo various post-translational modifications (PTMs) that regulate their structure and function, which are not necessarily reflected in transcriptomic data. Second, proteins are selectively degraded, thereby regulating cell functionality. This degradation is captured by proteome data but not necessarily by the transcriptome. Finally, not every RNA molecule is translated into a protein, due to post transcriptional modifications. This phenomenon might lead to misleading signals in the transcriptome.

As the amount of proteomics data in public repositories increases at an unprecedented rate [4], there is a rising need for digital tools for deconvolving bulk proteomic samples.

Chapter 3

Methods

The goal of this thesis is to explore the possibility of utilizing available proteomic data for a more accessible and accurate diagnosis of immune environment health, by benchmarking the performance of deconvolution methods adapted to proteomic data.

To tackle the problem of proteomic deconvolution, we propose to adapt methods inspired by reference-based RNA-seq deconvolution. Some of the more prominent methods are CIBERSORT [7] and BayesPrism [3], which are optimized for data following the distribution of RNA-seq data. Our main contribution is adapting and testing variants of such methods on biologically inspired, generated proteomic bulk samples with optimized reference signature matrices. Subsequently, we benchmark said methods against classical baselines such as Non Negative Least Squares, Mean and methods not using randomized signature matrices.

Mathematically the problem can be described as follows: Given a bulk data sample b defined as

$$S_{real} * p_{real} = b, \tag{3.1}$$

where S_{real} is the signature matrix containing the exact proteomic profile of each cell type of the patient producing the sample and p_{real} is the composition vector denoting the proportion of each cell type in the bulk sample. The entries of p_{real} are non-negative and sum to one. We seek to infer p_{real} in order to identify and characterize the patient's cancer status.

In a real-world setting we do not know what the cellular proteomic profile S_{real} is for individual patients, due to the technical complexities and costs associated with high-resolution quantitative proteomics of single-cell protein profiling and general challenges with single-cell sample preparation from tissues, which make it impractical for generating individual signature matrices. The technical complexities involve the large dynamic range of the cellular

proteome ($>$ 6 orders of magnitude) and the scarcity and limited amounts of certain subtypes of immune cells [12].

However, the proposed reference-based methods require a signature matrix $S_{estimated}$, and a bulk sample b, both of which we generate to simulate realistic data. In the this chapter we discuss strategies for generating bulk samples, explain how signature matrices can be produced, then propose methods for tackling the proteomic deconvolution problem, and lastly show the evaluation metrics used for benchmarking.

3.1 Synthetic Mixture Generation

The work of *Rieckmann et al.* provides us with an extensive and high-resolution proteomic dataset [12], which maps the social architecture of the human immune system. Specifically, we use Label-Free Quantification (LFQ) values from four individuals, covering approximately 10,000 proteins across 28 immune cell types. We focus on predicting the abundance of 26 relevant lymphocytes out of 28 individual cell types in blood samples. We then further aggregate these predictions based on their classification to the 4 main cell types of interest, B cells, T cells, Natural Killer cells, and Myeloid cells.

Figure 3.1: Heatmap of Missing Values in Proteomic Profiles of Cell Samples Provided by the MaxQuant Algorithm.

Since most methods, such as BayesPrism and CIBERSORT, assume strictly positive input values, we are only concerned with imputed data, provided by *Rieckmann et al.* This ensures that low-abundance proteins are still captured, even if the MaxQuant algorithm, used for identifying LFQ intensities, falsely gives them a zero value.

3.1.1 Bulk Samples

To accurately model the performance of different methodologies under realistic scenarios, we construct bulk LFQ intensity mixtures that closely reflect true biological conditions. We incorporate insights from real cancerous and healthy cell distributions [11] with healthy proteomic profiles to combine the the LFQ intensity vectors of the cell types into a bulk sample.

Cell Type-Specific Proteomic Profile Generation

Every individual has their unique protein distributions in their cells. For example, B cells derived from two different individuals might vary significantly in their composition, as discussed in the Results Chapter 4.1a.

There are several approaches to approximate this distribution. The simplest method is to assume a static profile for each cell type across samples. In this case, a perfect reconstruction is always possible, assuming we use the same static profile for the signature matrix. However, each individual having the same proteomic profile is far from realistic, as mentioned bevore. We suspect a signature matrix, that better estimates the average proteomic profile of individuals, to be constructed from convex combinations of the four reference profiles of each cell type provided by *Rieckmann et al.* [12]. To construct such a matrix, we need to be able to generate biologically inspired synthetic proteomic profiles:
Let $prot_x$ denote a generated proteomic profile of cell type x, where x is an element of the list $\mathcal{C}$ composed of 26 cell subtypes, in the order of the list of subtypes in Appendix A.1 . Let $prot_{y,j}$ denote the j'th reference proteomic profile from the publication of *Rieckmann et al.* with $y \in \mathcal{C}$. Note that we have four reference profiles for each cell type, meaning $j \in \{1,2,3,4\}$.

We model the random proteomic profile of sample as:

$$prot_x = \sum_{i=1}^{4} \delta_i \cdot prot_{x,i} \text{ , where } \delta \sim \text{Dirichlet}(1,1,1,1) \tag{3.2}$$

Depending on the type of normalization chosen, as discussed in the section on normalization, the bulk sample could also be computed according to the preprocessing step defined in Section 3.1.3.

$$prot_x = \sum_{i=1}^{4} \delta_i \cdot \log(prot_{x,i}), \text{ where } \delta \sim \text{Dirichlet}(1,1,1,1) \tag{3.3}$$

We define this probability distribution from which $prot_x$ is drawn as $\mathcal{P}rot_x$.

Cell Distributions

The distributions of cell types are the core of what our methods predict. To test their ability to accurately estimate healthy and cancerous distributions, we test our models with exactly these types of compositions. Similar to the generation of cell-type-specific profiles, we compute the random Dirichlet combination of 10 healthy distributions of lymphocytes, which consist of normalized averages of healthy blood sample compositions from the work of *Piatosa et al.* [11]. To obtain a realistic cancerous cell type distribution we use the same technique, except that we draw from a Dirichlet combination of

52 real cancerous distributions [1]. We model a healthy cell type proportion vector $p_{healthy}$ as follows:

$$p_{healthy} = \sum_{i=1}^{10} \delta_i \cdot p_{i,healthy} \text{ , where } \delta \sim \text{Dirichlet}(1_{10}) \tag{3.4}$$

Respectively, for cancerous cell distributions:

$$p_{cancerous} = \sum_{i=1}^{52} \delta_i \cdot p_{i,cancerous} \text{ , where } \delta \sim \text{Dirichlet}(1_{52}) \tag{3.5}$$

where $p_{i,cancerous}$ denotes the cell type distribution vector of patient i from the list of cancerous references and equivalently $p_{i,healthy}$ denotes the cell type distribution vector of patient i from the list of healthy references.

Bulk Sample Generation

Protein expression data of bulk tissue can be viewed as the weighted average of the protein expression profiles of different cell types in the tissue [9]. Specifically, we model the LFQ intensity values of the bulk sample as a weighted sum according to formula:

$$b_{healthy} = 100 * \sum_{i=1}^{26} p_{healthy}[i] * prot_{C[i]} \tag{3.6}$$

and respectively:

$$b_{cancerous} = 100 * \sum_{i=1}^{26} p_{cancerous}[i] * prot_{C[i]} \tag{3.7}$$

where $p_{healthy}[i]$ and $p_{cancerous}[i]$ are the ith entries of the proportions vector, meaning the fraction that the i'th cell type $C[i]$ takes up in the bulk sample.

We multiply the sample values by a factor of 100 to improve numerical stability during deconvolution. Without this scaling, several methods either fail to converge or produce poor results. This practical adjustment was determined empirically through trial and error.

3.1.2 Signature Matrices

All reference-based methods require a signature matrix $S_{estimated}$. This Matrix serves as a prior, in the case of BayesPrism, or as the assumed ground truth of the protein profile of all involved cell types. The simplest approach to

[1] Unpublished data, spectral flow cytometry-derived

modeling such a matrix is to assume every individual has the same cell type specific proteomic composition across individuals. In that case, we would simply use the reference profile from individual i from the work of *Rieckmann et al.* as our signature matrix $S_{estimated} = S_{ref_i}, \forall i \in \{1,2,3,4\}$, as under this assumption the reference profiles are assumed to be equal by the deconvolution method. We show in the Results section that the biological proteomic composition of cell types varies across patients, and this variance is reflected in the generated bulk sample, which is why we propose a randomized approach for generating signature matrices. To recreate the probability distribution from which proteomic data is found in real individuals, we draw the cell type profiles for the signature matrix form the same distribution as in the formula 3.2 or 3.3 for generating random samples. With this approach the profiles in the signature matrix and the ones used for the sample generation are randomly drawn from the same probability distribution, yielding a fully randomized approach.

$$S_{\text{fixed}} = \begin{bmatrix} | & | & & | \\ prot_{x_1}^{\text{sig}} & prot_{x_2}^{\text{sig}} & \cdots & prot_{x_n}^{\text{sig}} \\ | & | & & | \end{bmatrix}, \quad \forall x_i = \mathcal{C}[i], \quad prot_{x_i}^{\text{sig}}, prot_{x_i}^{\text{sample}} \sim \mathcal{P}rot_x \tag{3.8}$$

3.1.3 Preprocessing and Normalization

Usually, normalization in genomics and proteomics refers to methods aiming to reduce variance across technical replicates and to make measurements from different samples comparable. A good normalization technique should ensure that any variation is due to biological differences rather than technical artifacts from the measurement process or sample handling. This is especially important to note when deconvolving real samples. Since our bulk samples are based on the same references as the signature matrix, the technical artifacts are not our focus; nevertheless, normalization can stabilize variance, and adapt the distribution to better fit the methods which were originally optimized for RNA data [8]. To that end we tested our methods by using various transformations on the input data.

Logarithmic Transformation

Protein intensities measured by MS are evaluated by the number of peptides measured, as the peptide count varies across proteins and the probability of peptides being captured differs across peptide types, preprocessing to capture the real measure of proteins is essential. One approach to preprocessing the data is to log-transform it.

$$\text{LFQ values} \rightarrow \log(\text{LFQ values} + 1)$$

This would reduce the high signals produced by B cells. In addition, log transformation is a common preprocessing step in transcriptomics

An important open question is whether the normalization step should be applied before or after computing the weighted average, as discussed by *Petralia et al.* [9]. This distinction is crucial because the operations are not mathematically interchangeable: $\alpha \log(\text{prot}_i) + \beta \log(\text{prot}_j) \neq \log(\alpha \cdot \text{prot}_i + \beta \cdot \text{prot}_j)$ To investigate this, we evaluated both approaches.

Inlogged Before any proteomic profiles are combined into a bulk sample using weighted averages, they undergo log transformation. The signature matrix gets the same transformation. The objective is to mitigate the disproportionate influence of high-abundance proteins on the weighted average. This is equivalent to computing the log of a weighted geometric mean.

$$w_1 \cdot \log(prot_1) + w_2 \cdot \log(prot_2) = \log(prot_1^{w_1} \cdot prot_2^{w_2}) \tag{3.9}$$

While it's unlikely that bulk expression profiles result from a well-defined nonlinear transformation of individual cell-type signals, the possibility cannot be entirely ruled out. In practice, this assumption is consistent with standard preprocessing workflows and may still yield useful insights for method development and exploratory analysis.

Outlogged After all proteomic profiles are combined into a bulk sample using weighted averages they undergo log transformation. This may be useful if preserving absolute differences is important.

Quantile Normalization

CIBERSORT has its own recommended preprocessing step, recommended by its authors, by making it a built-in option: quantile normalization.

3.2 Deconvolution Methods

3.2.1 CIBERSORT

CIBERSORT is a method based on ν-support vector regression for estimating relative subsets of RNA transcripts [7]. It takes a signature matrix and number of bulk samples. The objective function includes an L2 penalty function, and the data used is expected to be quantile normalized RNA data. Since we are working with proteomic inputs other normalization methods could potentially be better suited. Our main contribution in adapting the methodology lies in optimizing normalization or transformation of the data.

The primary objective of SVR is to minimize both a loss function and penalty function given a defined set of constraints. The former measures the error

associated with fitting the data, whereas the latter determines model complexity. More specifically, SVR solves an optimization problem that minimizes the following two quantities

i a linear ε-insensitive loss function, which outperforms other common loss functions in noisy samples

ii an L2-norm penalty function (the same as that used in ridge regression), which penalizes model complexity while minimizing the variance in the weights assigned to highly correlated predictors (for example, closely related cell types), thereby combating multicollinearity [7].

This gives CIBERSORT the ability to perform well despite noise, which is crucial in our proteomic deconvolution model, since noise is inherently built in by randomizing the proteomic profiles in bulk samples.

3.2.2 BayesPrism

BayesPrism is a Bayesian cell proportion reconstruction method that uses statistical marginalization to predict cellular composition and gene expression from bulk RNA Seq using patient-derived, scRNA-seq as prior information. While it is optimized for raw count data, it provides flexibility for other input formats, specifically gene expression profiles (GEP). BayesPrism introduces an additional layer of complexity compared to CIBERSORT by attempting to infer a sample-specific signature matrix for each bulk RNA-seq sample. In our work, we initially employed a fixed (static) signature matrix across all samples. We then extended this approach by introducing variability into the signature matrix, sampling it from the same distribution as the bulk expression data. BayesPrism represents a further advancement in this regard, aiming to infer an improved, individualized signature matrix by combining the bulk data with single-cell reference profiles used as a prior. Based on this adapted signature matrix BayesPrism then estimates the composition of each bulk sample.

Technically, the deconvolution is achieved by firstly estimating the proportion of reads derived from each cell type, assumed to be proportional to the cell type fraction, and secondly, the expression level of genes within each cell type. The first step is to find the joint probability distribution over the gene expression per cell state matrix (U_n) and the cell state proportions vector μ_n for the n'th sample X_n conditioned on the observed single-cell reference φ and the bulk sample.

$$p(U_n, \mu_n | \varphi, X_n) \tag{3.10}$$

Where φ is an average over the reference profiles, which is why providing more reference profiles as the paper suggests does not hold for our case since it would merely create more values from the Dirichlet distribution of the four original references we use from *Rieckmann et al.* [12] and might skew the

average estimation. Since directly calculating the exact probability of every possible combination of cell proportions and gene expression values (the full joint posterior distribution p_{full}) is computationally infeasible, BayesPrism uses Gibbs sampling, to infer p_{full} based on the marginal distributions:

$$p(U_n|\mu_n, \varphi, X_n) \text{ and } p(\mu_n|U_n, \varphi, X_n) \tag{3.11}$$

The first is based on the assumption that the proportion of each cell state, follows a multinomial distribution. For a given cell state s, the probability of observing g reads for a specific gene is determined by $\varphi_{s,g}$ which encodes the event probability of that gene in that cell state s. Specifically U_{ns} is modeled as a multinomial distribution $U_{ns} \sim$ Multinomial(φ_s, R_{ns})

where R_{ns} is the proportion assigned to that cell state s. In other words, the read count of a specific gene for a specific cell state in the updated posterior is equal to the estimated proportion $\mu_s \cdot \varphi_{s,g}$. The event probability $\varphi_{s,g}$ of that specific gene for the specific cell state times the cell proportion is the expected contribution of this cell state for that gene. If we normalize to the total contribution we get the estimated percentage of contribution of that gene from the cell state. Based on these probabilities, a sampling step occurs. This is done independently for every gene. Now we have a new expression per cell state matrix. The initial choice of the parameters for μ is drawn from a Dirichlet distribution with a small alpha of 10^{-8}.

Filtering

For an effective analysis of the data, we computed the cell type specificity score for every protein and removed those of low specificity and high magnitude before the deconvolution. The specificity score was computed by counting how frequently a protein was significantly (FDR 5%) more abundant in one cell type compared to all other cell types, and then normalizing this count by the maximum count. We translated every protein to a Ensembl Gene ID using the UniProt library (https://www.uniprot.org) to identify the mitochondrial and ribosomal proteins/genes for visualization purposes. Proteins expressed at high magnitude, such as ribosomal proteins and mitochondrial proteins, may dominate the distribution and bias the inference. These proteins are often not informative in distinguishing cell types and can be a source of large spurious variance. As a result, they can be detrimental to deconvolution [3]. We filter out the proteins that are expressed from genes with low specificity-expression ratios, by using the built-in filtering function of the BayesPrism algorithm.

Deconvolution

Within the primary BayesPrism deconvolution module, the user has the option to specify the format of the reference. The recommended input type

for the single-cell RNA-seq reference consists of raw count data (UMIs). However, the LFQ intensities from *Rieckmann et al.* are protein expression data obtained from sorted cell populations, and do not consist of raw count data. The second option, which we found to be more well-suited for this type of data, uses a gene expression profile (GEP) as an input and assumes the data to be reference derived from other assays, such as sorted bulk data [3], similar to the data we use for the deconvolution framework. The paper recommends adding at least 20 or 50 reference profiles per cell type. Even though we only use 4, namely the 4 reference values provided by *Rieckmann et al.* the results are comparable[2] as when we generate the missing 16-46 profiles from a Dirichlet distribution as described in Section 3.1.1.

3.2.3 Baseline Models

Non Negative Least Squares

Non Negative Least Squares (NNLS) is a classical way to approach deconvolution problems. We applied this model to every sample using the LinearRegression(positive = True) implementation from scikit-learn.

$p_{nnls} = \arg\min_{\mathbf{p} \geq 0} \|\mathbf{S}\mathbf{p} - \mathbf{b}\|_2^2$

Mean

The mean method statically predicts the mean of the healthy fractions and the mean of the lymphoma patient's cell fractions. It uses prior knowledge about whether the cell distribution, with which the sample constructed with, was based on healthy references or cancerous ones and it utilizes those reference distributions to estimate the cell composition of the sample. The cell type fractions of healthy samples are predicted as the mean cell type fraction of all reference healthy fractions. To compute the mean fraction for a healthy sample we compute the average of all the reference healthy cell type compositions and then L1 normalize it, such that the components add up to 1.

$$\tilde{p}_{\text{mean},c} = \frac{1}{m} \sum_{j=1}^{m} p_{c,j} \qquad p_{\text{mean},c} = \frac{\tilde{p}_{\text{mean},c}}{\sum_{i=1}^{m} \tilde{p}_{\text{mean},c}(i)} = \begin{pmatrix} \text{B cells:} & 0.8252 \\ \text{T cells:} & 0.1231 \\ \text{Myeloid cells:} & 0.0361 \\ \text{NK cells:} & 0.0156 \end{pmatrix}$$

$$\tilde{p}_{\text{mean},h} = \frac{1}{n} \sum_{i=1}^{n} p_{h,i} \qquad p_{\text{mean},h} = \frac{\tilde{p}_{\text{mean},h}}{\sum_{i=1}^{n} \tilde{p}_{\text{mean},h}(i)} = \begin{pmatrix} \text{B cells:} & 0.0913 \\ \text{T cells:} & 0.6410 \\ \text{Myeloid cells:} & 0.2255 \\ \text{NK cells:} & 0.0423 \end{pmatrix}$$

[2]Unpublished results

3.3 Evaluation Methods

3.3.1 Mean Absolute Error (MAE) over Cell Types and Samples

For every sample and accumulated cell type we analyzed the difference between the predicted fraction and the actual fraction. With this evaluation method, it is easy to see what the bias and variance of the prediction error are for different cell types.

3.3.2 Pearson Correlation

We show the Pearson correlation of the predicted and actual cell type fractions.

$$r = \frac{\sum_{i=1}^{4}(p_{predicted,i} - \bar{p}_{predicted})(p_{real,i} - \bar{p}_{real})}{\sqrt{\sum_{i=1}^{n}(p_{predicted,i} - \bar{p}_{predicted})^2}\sqrt{\sum_{i=1}^{n}(p_{real,i} - \bar{p}_{real})^2}} \tag{3.12}$$

This serves as a mathematical tool to analyze the efficacy of the measured methods.

3.3.3 Spearman Correlation

Additionally, we evaluate the methods according to the Spearman correlation to understand how well they predict the order of the sizes of cell fractions. Let $R_{\text{predicted},i}$ be the rank (not the value) of the predicted fraction of cell type i, and $R_{\text{real},i}$ be the rank of the true (ground truth) fraction of cell type i. Let $\bar{R}_{\text{predicted}}$ and $\bar{R}_{\text{real}}$ be the average ranks across all cell types. The Spearman correlation ρ is given by:

$$\rho = \frac{\sum_{i=1}^{n}(R_{predicted,i} - \bar{R}_{predicted})(R_{real,i} - \bar{R}_{real})}{\sqrt{\sum_{i=1}^{n}(R_{predicted,i} - \bar{R}_{predicted})^2}\sqrt{\sum_{i=1}^{n}(R_{real,i} - \bar{R}_{real})^2}}$$

3.3.4 RMSE of Signature Matrix Estimates

BayesPrism includes functionality to iteratively update the reference signature matrix based on observed bulk data and prior reference proteomic profiles. To evaluate this process in a proteomic context, we compared three versions of the cell-type-specific protein intensity profiles:

- The original reference signature matrix
- The real per-cell protein intensities that generated the bulk data
- The BayesPrism updated signature matrix

We calculated the root mean squared error (RMSE) between each profile and the real data to quantify how much the updated signatures diverged from the ground truth.

$$\text{RMSE} = \sqrt{\frac{1}{n \cdot d} \sum_{i=1}^{n} \sum_{j=1}^{d} \left(\text{prot}_{i,j}^{\text{real}} - \text{prot}_{i,j}^{\text{pred}} \right)^2}$$

where:

- n: Number of proteins
- d: Number of cell types
- $\text{prot}_{i,j}^{\text{real}}$: The real expression value of protein i in cell type j
- $\text{prot}_{i,j}^{\text{pred}}$: The predicted expression value of protein i in cell type j

It is noteworthy that the RMSE may be dominated by high expression proteins. To further investigate the distance, we also visualize the distance between the first two principal components of the proteomic profiles captured by the signature matrices.

3.4 Implementation and Availability

All scripts and data used for simulation, deconvolution, and benchmarking are publicly available at:

https://github.com/Samuelisusername/Proteomic-Deconvolution-Framework

The repository includes installation instructions and reproducible workflows for testing various deconvolution methods on proteomic mixtures.

Chapter 4

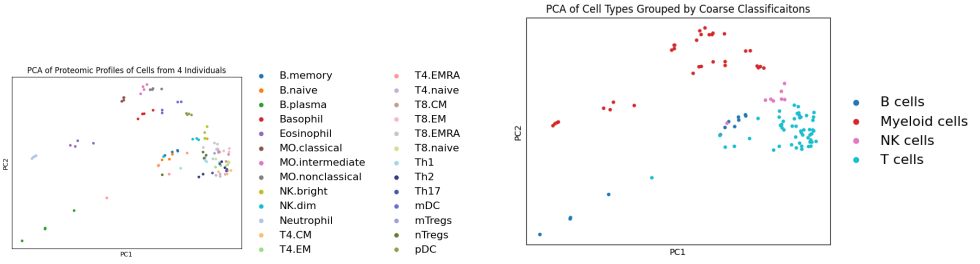
Results

In this Chapter, we begin by analyzing the proteomic profiles of cells and then healthy and cancerous cell type distributions used to generate the bulk samples and the signature matrix, with a focus on the variability of cell-type-specific profiles. We then evaluate the predictive performance of different methods under the various conditions introduced in Chapter 3 and assess the results using the evaluation metrics defined in Section 3.3.

4.1 Initial Data Exploration

4.1.1 Proteomic Profiling of Immune Cell Subsets

In order to realistically simulate patient cell type specific protein signals, we must understand the distribution of those proteomes across patients. Since the proteomic profiles of cell types lie in about 10,000 dimensional space we use a PCA plot to only view the two largest principal components, allowing us to observe dominant patterns and variance structures. A notable



(a) 2 Largest PC's of Proteome of all 26 Lymphocytes

(b) Proteomes Sorted and Grouped into 4 Main Cell Types

Figure 4.1: Proteomic PCA of Immune Cell Subsets. (a) PCA of Individual Cell Types. (b) Coarsely Grouped Cell Types.

observation in Figure 4.1a, is the general pattern of proteome signals of cells clustering by cell types rather than by individual patient, which indicates that the intrinsic variability within cell types might be approximately captured by generating random signals as convex combinations from the four reference profiles.

Similarly, we can also observe in Figure 4.1b that the proteomes cluster according to broader cell types. Specifically, B cells (including memory, naive, and plasma) are proximally located in the PCA plot, indicating their shared proteomic identity. This pattern suggests that proteomic data successfully captures fundamental biological distinctions between these major immune cell lineages.

Cell State

Determining the appropriate cell types for deconvolving real samples is an important consideration. As researchers tend to subdivide broader cell types into different levels of granularity and states, one must be careful not to overlook cells or cell states that can have an impact on the proteomic mixture of the bulk sample.

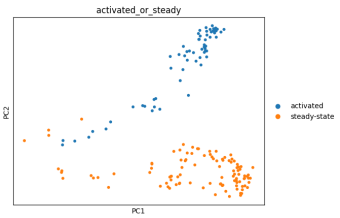


Figure 4.2: Two Largest Principal Components of Cells: Activated (Blue) and Steady State (Orange)

For example, the following plot shows that lymphocytes' protein composition drastically changes when in an activated state, meaning when exposed to an immune threat. To simplify benchmarking and data generation, we only consider cells that are in steady state, but for clinical use we recommend to be mindful of these shifts in cell states, as they could significantly impact model performance.

4.1.2 Cell Type Composition of Healthy and Cancerous Samples

Using the distribution described in a reference Figure adapted from *Piatosa et al.*'s work [11], we determine the typical composition of lymphocytes in healthy blood samples. A comparison with distributions observed in cancerous samples reveals a marked shift in cellular makeup.

The cancerous distributions were derived from unpublished data. These Figures illustrate a substantial difference in lymphocyte composition between healthy and cancerous samples. In healthy blood, T cells constitute the majority of the lymphocyte population, whereas in cancerous samples, B cells dominate, comprising approximately 80% of the lymphocytes. Despite this shift, intra-group variance remains relatively low for both conditions. While

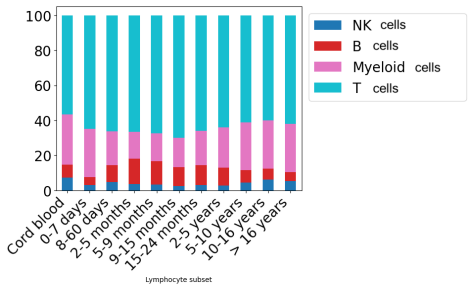


Figure 4.3: Healthy Lymphocyte Composition

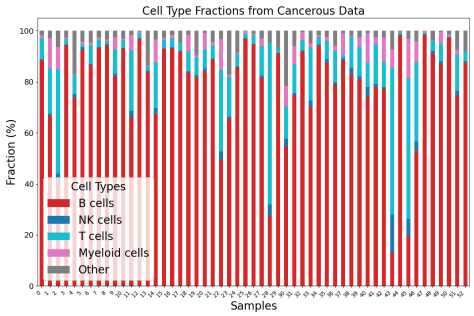


Figure 4.4: Cancerous Lymphocyte Composition

these compositional differences alone may not be sufficient for diagnostic classification, they offer valuable insight and may serve as a contributing factor in broader diagnostic frameworks.

4.2 Benchmarking Results

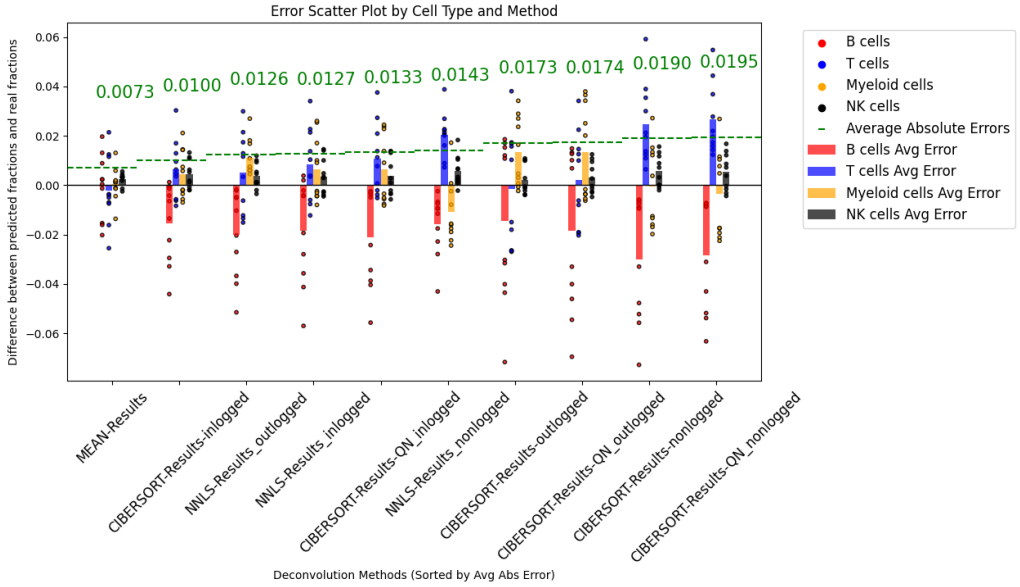


Figure 4.5: MAE of Proteomic Deconvolution Methods and Their Respective Normalizations

Figure 4.5 presents the different methods evaluated on the same randomized signature matrix and samples with various types of normalization. The y-axis

represents the difference between the predicted and real cell type fraction of the generated sample per coarse cell type. The mean method performs the best, as expected, since the variance in cell proportions in healthy samples and cancerous ones is low. Among the methods that do not assume prior knowledge of which distributions (healthy or cancerous) the generated cell distribution stems from, CIBERSORT and NNLS perform best. All methods exhibit reduced accuracy when the cell type distribution is highly skewed towards B cells. Notably, we observe consistent biases of certain cell types to be under- or overestimated across all methods, which are further examined in Section 4.3. This finding aligns with results from bulk RNA literature, such as the study by *Nguyen et al.* [8] which reports that all methods they tested were biased toward certain cell types, consistently underestimated or overestimated the proportion of specific cell types. The authors also report that at least 10 cell types were significantly under or overestimated with an absolute mean difference of more than 10 percent. While the magnitude of the bias is smaller in our case, we observe similar consistent prediction biases in our results 4.3. Due to the relatively poor performance of BayesPrism in our setup, its results will be discussed separately and excluded from direct comparisons with other methods.

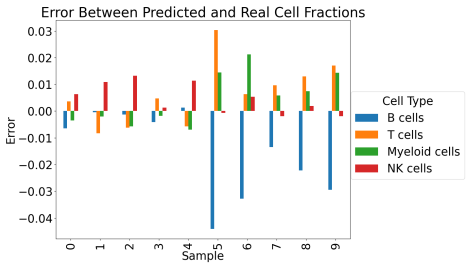
4.2.1 CIBERSORT

To assess the applicability of CIBERSORT to proteomics based deconvolution, the algorithm was applied using a signature matrix created as explained in Section 3.1.2. The optimal data processing approach used inlogged transformation as described in Section 3.1.3 and yielded the best average error of 1.00%. Outlogged results will also be presented as it is likely that bulk sample addition represents an arithmetic rather that geometric mean, in which case the inlogged results would be unrealistic.

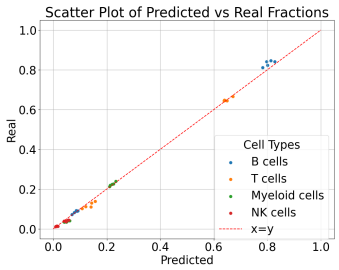
Outlogged transformations yielded intermediate results (1.73% and 1.74% average error) and for both inlogged and nonlogged data, using the QN algorithm consistently increased the average prediction error to 1.33% for QN inlogged, 1.74 for QN outlogged, and 1.95% for QN nonlogged.

The prediction error ranges from 1.00% to 1.95%, indicating that while CIBERSORT can achieve relatively low error rates under optimal preprocessing, certain standard preprocessing steps recommended for transcriptomic data are suboptimal for proteomics. This suggests that normalization procedures should be carefully tailored to the characteristics of proteomic data, as assumptions valid for transcriptomic measurements may not hold in this context.

Analysis reveals that the error is larger for the cancerous samples (samples 5-9), where the distribution is highly skewed towards B cells. Comparison of inlogged and outlogged prediction errors, demonstrates that outlogged

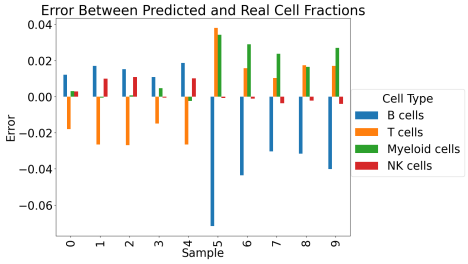


(a) Prediction Error of CIBERSORT With Inlogged Normalization

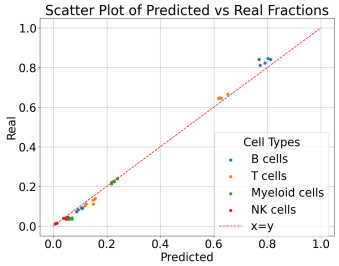


(b) Predicted and Actual Fractions of Bulk Samples Against Real Fractions

Figure 4.6: Comparison of CIBERSORT Inlogged Predictions and Their Errors for Randomized Samples.



(a) Prediction Error of CIBERSORT with Nonlogged Normalization



(b) Predicted and Actual Fractions of Bulk samples

Figure 4.7: Comparison of CIBERSORT outlogged Predictions and Their Errors for Randomized Samples.

errors are not only generally larger but also there is a smaller difference in errors between healthy and cancerous samples. The results demonstrate close alignment with actual fractions. Furthermore, B cells for the skewed distributions are consistently underestimated, while NK and T cells are typically overestimated, with this bias being more pronounced in nonlogged samples.

As shown in Figure 4.6b the predicted fractions exhibit a strong alignment with the actual values, supporting the high Spearman and Pearson correlation coefficients. For the outlogged approach the Spearman coefficient is approximately 0.913 and the Pearson correlation is 0.994. For inlogged they are 0.932 and 0.933, respectively. For more information, please refer to A.3 and A.2 in the appendix.

These high correlation values are significant because they indicate that the model is not only capturing the overall trends in the data, as shown by Spearman's rank-based correlation, but also preserving the actual magnitude relationships between predicted and true cell type fractions, as measured by

Pearson correlation.

4.2.2 BayesPrism

BayesPrism is a probabilistic deconvolution framework that jointly estimates cell-type fractions and the underlying signature matrix. A key feature of BayesPrism involves filtering out genes that are highly expressed but low in cell-type specificity, as these can bias the model and obscure true cell-type differences. To utilize this recommended feature for proteomic deconvolution, we mapped protein identifiers to their corresponding genes and annotated the proteomic LFQ intensities with Ensembl gene annotations from the UniProt database. The resulting gene-level intensities were then passed to BayesPrism's cleanup.genes function, which automatically selects genes to filter.

Filtering of Proteins

The model identified 471 out of 10'320 genes to exclude, most of which were ribosomal proteins, proteins essential to cell survival, and as such highly expressed and ubiquitously present. A visual reference to the expression specificity and log mean expression is provided in Appendix A.2.

Cell Fraction Estimation

BayesPrism shows mixed performance across all cell types. While NK cells are consistently predicted with low error, the algorithm struggles with other cell types, particularly in cases where cell type fractions are large. Particularly noteworthy are cancerous samples 5-9 in Figure 4.8, where the fraction of B cells is substantially larger. In these cases, predicted B cell fractions do increase noticeably, indicating that BayesPrism captures the overall trend, even if the magnitude is often misestimated.

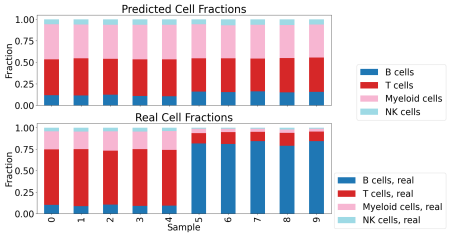


Figure 4.8: Predicted and Real Fractions

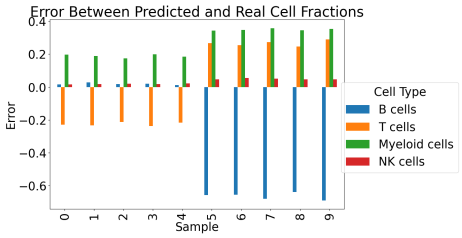


Figure 4.9: Errors Produced by BayesPrism Estimations

Correlation analyses further support this observation.

- B cells — Pearson: 0.975, Spearman: 0.915
- T cells — Pearson: 0.944, Spearman: 0.842
- Myeloid cells — Pearson: 0.793, Spearman: 0.697
- NK cells — Pearson: –0.132, Spearman: 0.176

For the majority of cell types, especially B and T cells, there is a strong positive correlation between predicted and actual values. This suggests that the model is particularly reliable at ranking and proportionally estimating those two cell types, even if some predictions overshoot or undershoot in absolute terms. For myeloid cells, the correlation is moderate, suggesting weaker but still detectable trend alignment. Notably, for NK cells, the near zero and negative correlation indicates that the model fails to capture the correct directional patterns despite low absolute error. It should be noted that without the highly skewed cancerous proportions, the model would achieve similar prediction errors to CIBERSORT or NNLS.

Signature Matrix Estimation

Quantitative comparisons revealed the following RMSE values:

- Reference vs. Real: 1.65×10^{-6}
- Updated vs. Real: 8.77×10^{-6}

Contrary to expectation, the updated signature matrix showed greater deviation from the real profiles than the original reference. This suggests that the BayesPrism adjustment algorithm, while directional, systematically overcorrects in the presence of proteomic data. This tendency is illustrated in the PCA plot (Figure 4.10), where most updated profiles shift towards the ground truth but consistently overshoot.

Analysis across several runs shows that in at least 63% of cases, the direction of the update, with an appropriate scaling factor, would lead to a signature matrix with lower RMSE. As visualized in Figure 4.10 while the red line overshoots, it tends in the direction of the blue dot more often than not.

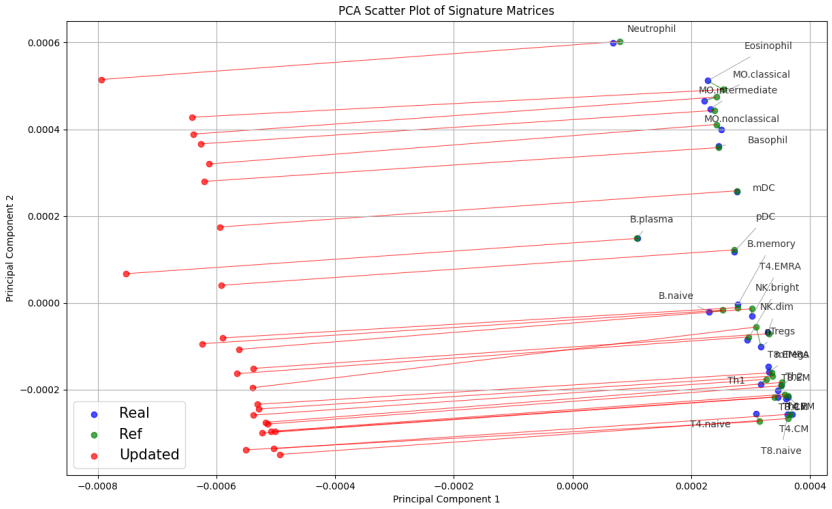


Figure 4.10: 2 PC's of Estimated (Red), Real (Blue) and Reference (Green) Proteomic Profiles

4.2.3 Non-Negative Least Squares (NNLS)

Assuming non-linear dependency among all proteomic reference profiles, and therefore full rank of the signature matrix, perfect reconstruction of the cellular composition would theoretically be possible, provided that the exact proteomic profiles of the constituent cells are known. However, these assumptions are not realistic when applied to patient data, nor are they fully met in our synthetic data generation process.

Nevertheless, the stability and simplicity of NNLS, which requires no hyper-parameter tuning, make it a strong baseline method. NNLS with inlogged normalization achieves competitive performance, ranking as the third best method overall. While it does not surpass the Mean baseline (0.73% error) or CIBERSORT inlogged (1.00% error), it demonstrates robust performance across different normalization approaches, with outlogged (1.27% error) and nonlogged (1.43% error) variants showing only modest increases in error. Like BayesPrism and CIBERSORT, NNLS performs particularly well for samples with healthy, less skewed cell distributions.

To evaluate prediction accuracy, we calculated both Pearson and Spearman correlations between the real and NNLS-predicted cell-type proportions. As shown in Table 4.1, NNLS achieves consistently high correlations across most cell types. B cells and T cells show near-perfect agreement, with Pearson correlations above 0.999 and strong rank-based correlations as well. Myeloid cells also perform robustly, while NK cells, which are typically harder to predict due to their lower abundance and higher variance, still achieve a

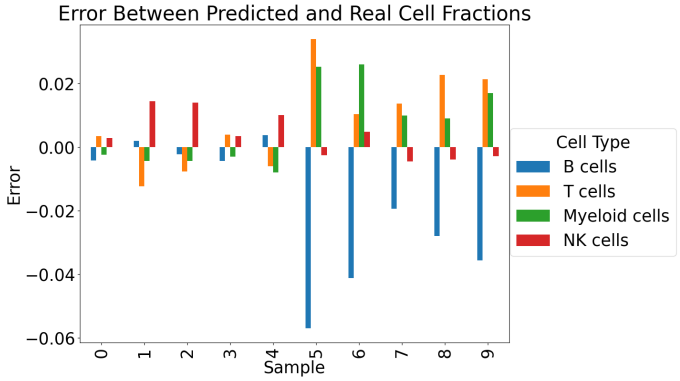


Figure 4.11: Prediction Error of NNLS with in-logged normalization across all samples and cell types.

Pearson correlation of 0.978. These results demonstrate that despite its simplicity, NNLS is a competitive and reliable approach, especially when paired with appropriate normalization.

Table 4.1: Pearson and Spearman correlations between real and NNLS-predicted proportions for each cell type.

Cell Type	Pearson Correlation	Spearman Correlation
B cells	0.999	0.939
T cells	0.999	0.903
Myeloid cells	0.997	0.987
NK cells	0.978	0.866

4.2.4 Mean Estimator

The mean estimator achieved strong performance, primarily because the variance between healthy and unhealthy cell type compositions in the bulk samples was low. However, this performance relies on an unrealistic assumption: it requires prior knowledge of the health status and corresponding reference distributions of the samples. In practical applications, these are precisely the unknowns we aim to predict.

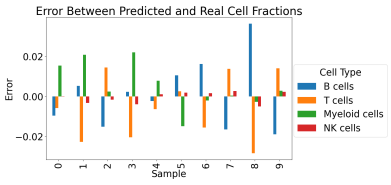


Figure 4.12: Prediction Errors of The Mean Estimator across Different Samples and Cell Types.

In contrast to other methods, the mean estimator does not exhibit systematic biases toward over- or underestimating specific cell

types, regardless of health condition.

4.3 Bias Exploration

Our analysis revealed a consistent underestimation of B cell proportions, particularly in cancerous samples, where B cells often constitute a large fraction of the cellular composition. This systematic error has led us to hypothesize that the observed bias may not stem solely from RNA-seq-specific assumptions, but rather from intrinsic properties of the proteomic profiles. For instance, if the expression signatures of certain cell types significantly overlap, especially in high-dimensional space, this could lead to systematic misattribution, where the shared signal is assigned to the wrong population, particularly when one cell type dominates the mixture.

Since the discrepancy in performance between healthy and cancerous samples appeared to be driven largely by differences in cell composition, we hypothesized that estimation error might correlate with cell type abundance. Specifically, we examined whether large cell fractions tend to be systematically underestimated and smaller cell fractions overestimated. Additionally, we explored whether certain proteomic profiles are inherently harder to estimate, independent of their abundance. To explore these hypotheses we generated synthetic mixtures with controlled compositions. In the first 10 samples, all four cell types were assigned equal proportions (25% each). In samples 11 to 20, one cell type was assigned a dominant proportion (97%) while the others were each set to 1%. Since CIBERSORT and NNLS produced comparable results, we used NNLS with the non-logged method for the following experiments.

Large NK Cell Fraction

Figure 4.13 shows that in the balanced case, B cells, Myeloid cells and NK cells tend to be similarly underpredicted, while T cells are consistently overestimated.

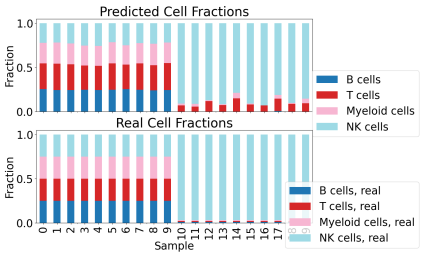
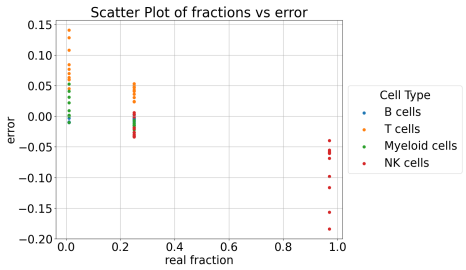


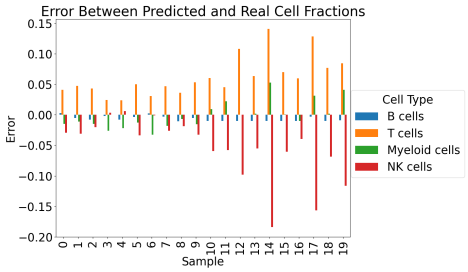
Figure 4.13: Real (Bottom) and Predicted (Top) Fractions per Sample

In the unbalanced case (Figure 4.14a) NK cells, which make up the largest fraction, are systematically underestimated, while T cells are more strongly overestimated. B cells remain underpredicted, and Myeloid cells show only mild overprediction. There is a strong trend where smaller cell fractions are overpredicted and larger underpredicted. Interestingly, despite all cell types having the same true fractions in

the balanced case, some consistently receive higher predicted values than others. Note that even though NNLS is a deterministic algorithm, predictions vary across samples due to the randomly generated proteomic profiles.



(a) Errors of Samples by Fraction (Balanced at 0.25% Each; Skewed at 97 % for NK Cells and 1% for the Others)

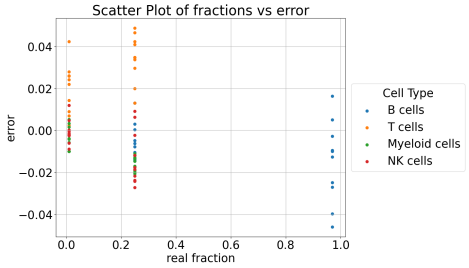


(b) Fraction Error per Cell Type per Sample where Sample (0-9 Balanced, 10-19 Skewed)

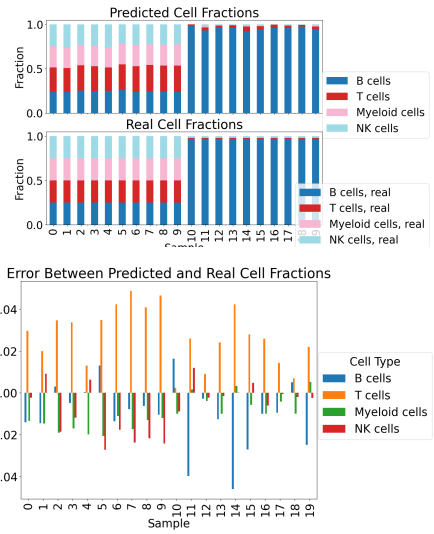
Figure 4.14

Large B Cell Fraction

The B cell-dominated case resembles the cancerous scenario where B cells are overrepresented. Unsurprisingly, we observe similar prediction errors as in CIBERSORT and NNLS, only amplified. T cells are strongly overestimated across all samples, while NK and Myeloid cells are generally underestimated. B cells, as the most dominant cell type is, underestimated in the skewed case and in all but two balanced cases.



(a) Errors Per Sample and Cell Type By Real Fraction



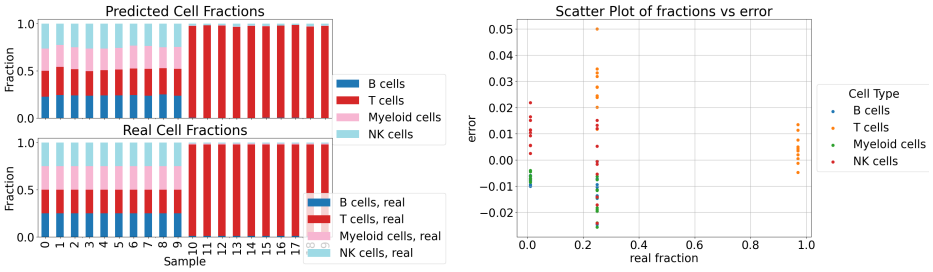
(b) Errors When B Cells Dominate Distribution

Figure 4.16: Metrics Showing The Performance When B Cells Dominate Distribution

Figure 4.19: Errors When Myeloid Cells Dominate Distribution by Real Fractions

Large T Cell Fraction

Next, we examine T cells, which are usually overpredicted, under the highly skewed distribution. Surprisingly while under-prediction of the dominant T cell type occurs in some samples, it is less pronounced than in NK or B cell skewed cases.



(a) Estimated and Real Fractions When T Cells Dominate Distribution

(b) Errors by Real Fractions When T Cells Dominate Distributions

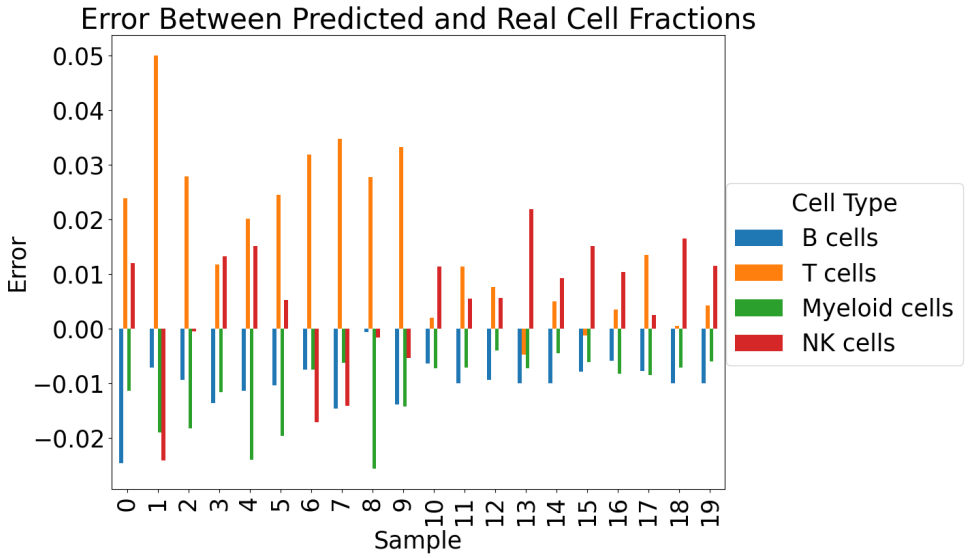
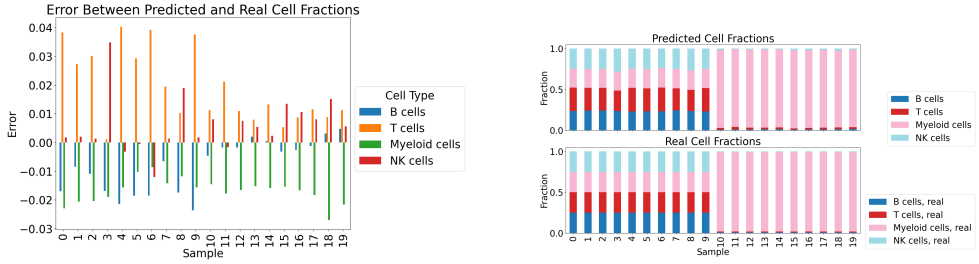


Figure 4.18: Errors When T Cells Dominate Distribution by Sample

In the balanced case, B and Myeloid cells are again underestimated, NK cells slightly less so, and T cells are consistently overestimated. Interestingly, the prediction error in the T-skewed case appears smaller than in the balanced case.

4.3.1 Large Myeloid Cell Fraction

Lastly, we examine mixtures dominated by Myeloid cells. These are slightly underestimated in the skewed scenario, while small fractions of B cells are less severely underestimated compared to balanced settings. T cells remain strongly overestimated, with NK cells following a similar under-prediction pattern.



(a) Errors for Myeloid Cells Dominating Distribution

(b) Predicted Fractions and Real Fractions

Figure 4.20

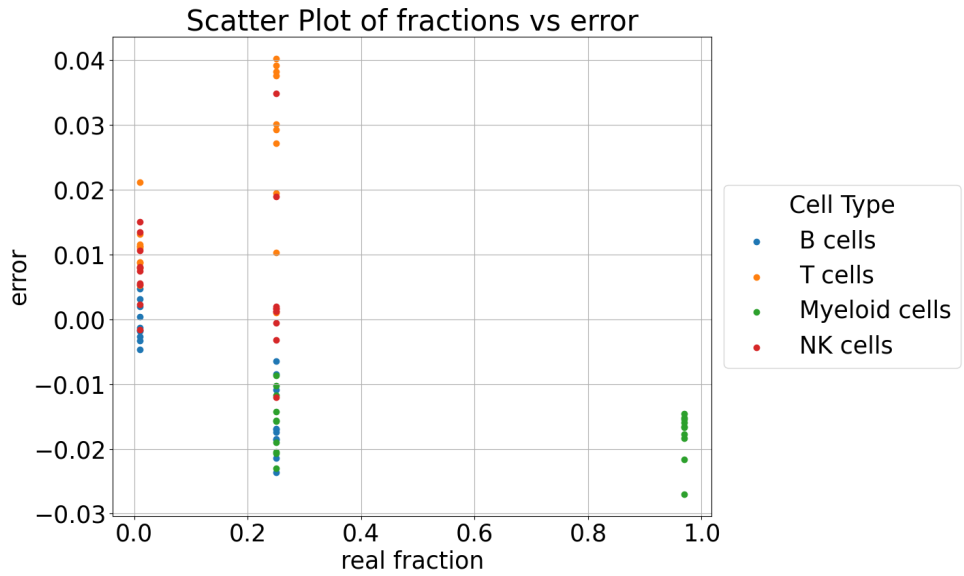


Figure 4.21: Errors of Cell Fraction Predictions of Samples by Real Fraction Size

Cell Type and Fraction Size Bias

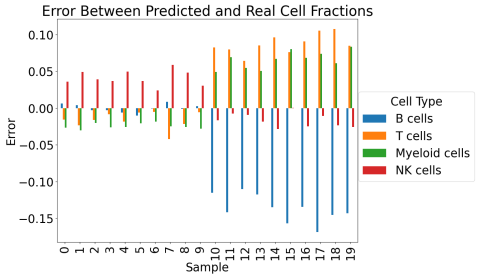
Across all cell types and configurations, we observe that larger fractions tend to be underestimated, while smaller fractions are overestimated. In addition to this fraction-size bias, there is a clear cell-type-specific bias:

- B cells and Myeloid cells are generally underestimated.
- T cells are frequently overestimated.

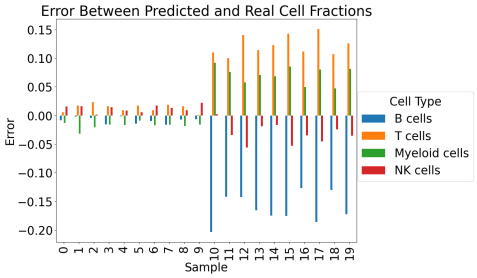
4.3.2 The Importance of Randomizing Signature Matrices

Our initial approach relied on a fixed signature derived from the proteomic profile of an individual, rather than generating a randomized signature matrix from the same distribution as the bulk samples. This approach performed very well under the strong assumption that all individuals share identical cell-type-specific proteomic profiles. However, when this assumption is relaxed, as is necessary for realistic use, which we approximate with the bulk sample being produced from the random distribution 3.1.1, the error increases significantly. Specifically, regardless of which individual's profile we choose as the reference for the deconvolution, the average prediction error for NNLS nonlogged and similarly for other methods is more than double that obtained using randomized signature matrices.

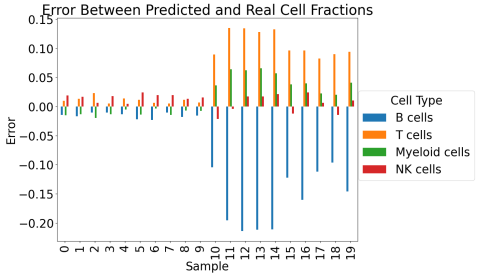
Across all individuals, we observe a consistent tendency to overestimate T cells except for the method using signature matrix derived from individual 4 when deconvolving healthy samples. Myeloid cells are overestimated



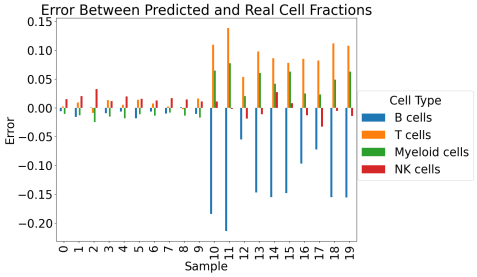
(a) fixed reference matrix based on proteomic profile of Individual 4 (avg error 4.91%)



(b) fixed reference matrix based on proteomic profile of Individual 3 (avg error 5.49%)



(c) fixed reference matrix based on proteomic profile of Individual 2 (avg error 4.72%)



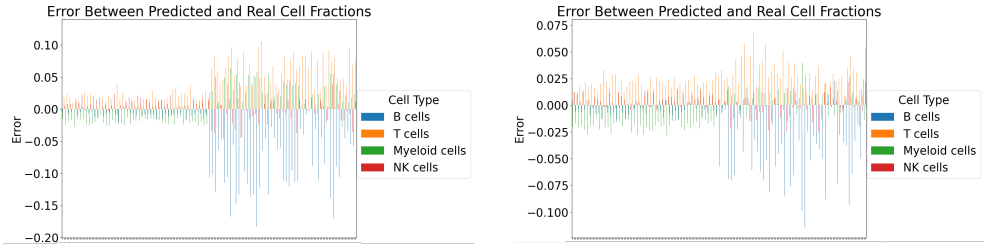
(d) fixed reference matrix based on the proteomic profile of Individual 1 (avg error 4.30%)

Figure 4.22: Comparison of signature matrices based on proteomic profiles of different individuals.

in all cancerous samples and underestimated in all healthy samples, and the inverse relationship tends to hold for NK cells. Interestingly, these biases align with those observed when using randomized signature matrices. This is somewhat counterintuitive, given that the randomized matrix is a Dirichlet-distributed weighted average of the individual profiles. One might expect errors to remain similar, but they decrease by about half. Given the significant improvements in accuracy, we adopt the exclusive use of randomized signature matrices in all subsequent analyses. While one could argue that improved performance might stem from information leakage, since the randomized matrix is derived from the same distribution as the constituent individual profiles, this explanation is insufficient to account for the observed effect. To further investigate this, we compare two settings: The first, a randomized signature matrix is constructed from two individual reference profiles and used to deconvolve bulk samples generated from the remaining wo individuals. In the second, we use a single reference profile to deconvolve bulk samples composed from the other two individuals. To showcase this we use nonlogged NNLS, but similar results can be found for the other presented methods. Despite both approaches maintaining a strict separation between reference and bulk data, the two randomized matrix

derived from two reference profiles achieves a markedly lower mean absolute error of 1.76%, compared to 3.11% for the single-reference approach (Figures 4.23b and 4.23a).

These results suggest that the superior performance of randomized matrices cannot be attributed solely to overlap in data distribution. Rather, the combination of multiple profiles in the randomized matrix likely captures a more robust and representative signal of cell-type-specific variation thereby reducing estimation bias and variance.



(a) Errors of NNLS Nonlogged with Signature Matrix based on Individual 1 Deconvolving Bulk Samples based on Individuals 3 and 4

(b) Errors of NNLS Nonlogged with Signature Matrix based on Individual 1 and 2 Deconvolving Bulk Samples based on Individuals 3 and 4

Figure 4.23: Comparison between different reference signatures

In general the performance of the NNLS nonlogged model using two proteomic reference profiles for the signature matrix (in blue in Figure 4.24) to deconvolve 200 bulk samples generated from two other proteomic profiles significantly improved compared to the case where we only use one reference profile for the signature matrix (in gray in Figure 4.24).

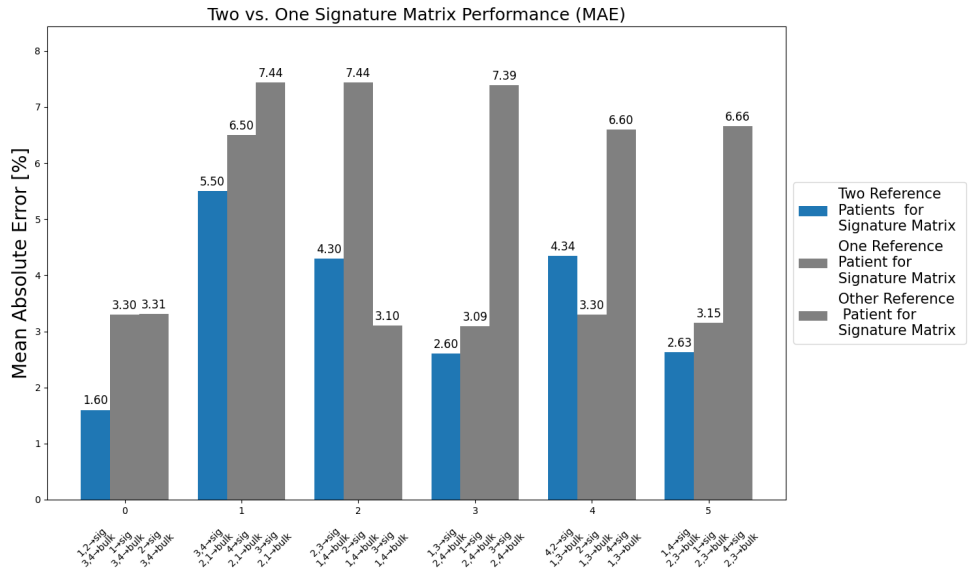


Figure 4.24: Comparing Randomized Signature Matrices to Fixed Ones

The first methodology results in an average error of approximately 3.5%, while the latter yields an error of 5.1%.

4.3.3 Randomized Signature Matrices From Three Individuals Deconvolving The Remaining One

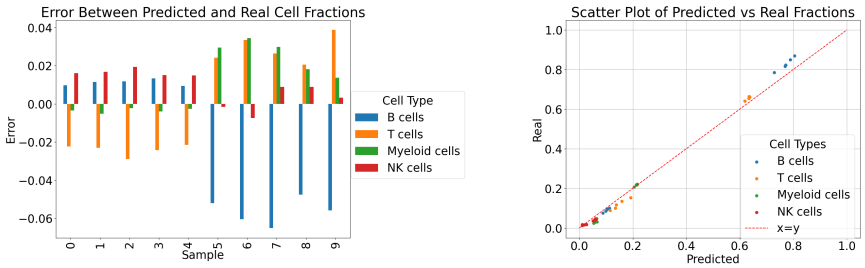
To assess the generality and robustness of randomized signature matrices, we further explored configurations where the signature matrix was constructed from three individual proteomic profiles and used to deconvolve bulk samples generated from the remaining individual. This approach eliminates any overlap between reference and target samples, ensuring a clean separation of data.

Interestingly, this three-versus-one (3v1) setup yielded even better performance than the two-versus-two (2v2) scenario. Across multiple permutations, the average absolute error consistently decreased, suggesting that incorporating additional individuals into the reference matrix helps better capture the underlying cell-type-specific variability. This likely results in a more representative and less biased basis for deconvolution.

Some representative results are as follows:

- Signature from individuals 2, 3, and 4 vs. samples from individual 1: 2.1% error
- Signature from individuals 3, 4, and 1 vs. samples from individual 2: 1.3% error

- Signature from individuals 4, 1, and 2 vs. samples from individual 3: 2.4% error
- Signature from individuals 1, 2, and 3 vs. samples from individual 4: 2.5% error



(a) Mean absolute errors using 3-individual signature matrices vs. a held-out individual.

(b) Predicted vs. true cell fractions in the 3v1 setting showing reduced bias.

Figure 4.25: Performance of 3-individual signature matrices on held-out individuals.

In Figures 4.25a and 4.25b, we present results for the example configuration where the signature matrix is constructed using data from individuals 2, 3, and 4, and used to deconvolve bulk samples generated from individual 1. The bar plot shows a notably low mean absolute error of 2.1%, while the scatter plot demonstrates strong alignment between predicted and true cell fractions, with minimal bias. These results highlight the effectiveness of using aggregated reference profiles from multiple individuals to generalize to previously unseen data—underscoring the robustness and practical applicability of this approach for realistic deconvolution tasks.

4.4 Sample Health Classifier

As the prediction accuracy is relatively high and the variance of samples in healthy and cancerous distributions is low, respectivey, it is clear that we can predict with high accuracy which samples were produced by cancerous fractions, and which ones by healthy ones. Even with the least accurate method, BayesPrism predictions can be used to infer the origin (healthy vs cancerous) of the cell type composition. As B cells undergo the greatest shift in average fraction between healthy and cancerous conditions, we have decided to use the B cell proportion as a reference for predicting the health status. If the predicted fraction of those cells is above the threshold of a certain value, it is classified as cancerous, otherwise it is classified as healthy.

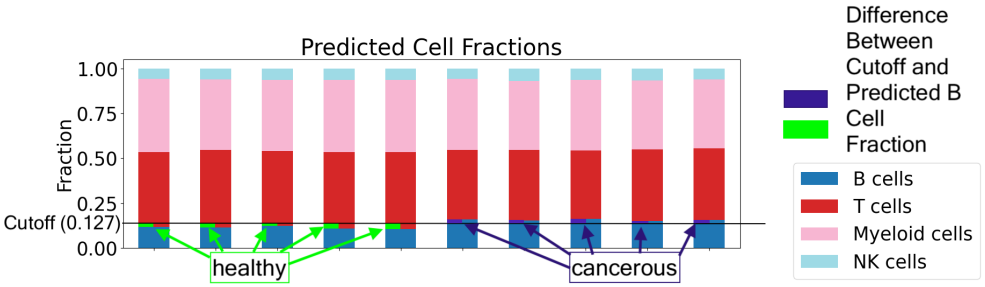


Figure 4.26: Example of How the Health of the Cell Distribution could be Measured with a B Cell Cutoff Value

Similarly a static cutoff value for the B-cell fraction can be used for every other method other than random guessing, and we will achieve the same result.

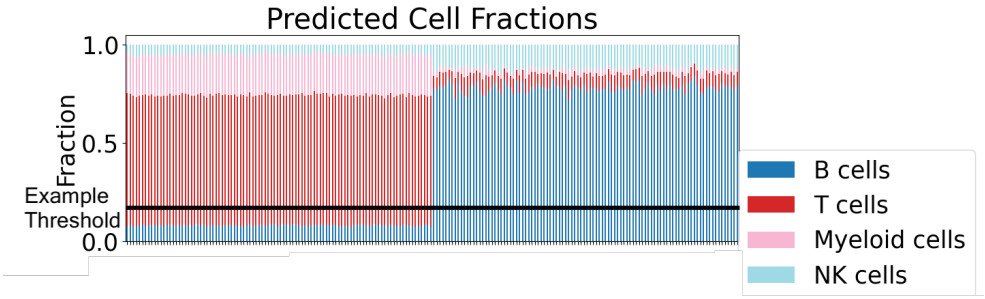


Figure 4.27: Testing the Classifier with NNLS Nonlogged Predictions

Figure 4.27 is the visualization of an example threshold with 100% prediction accuracy over 200 samples when being used on predicted cell type fractions from the NNLS nonlogged algorithm. Similar results can be found for all introduced normalization used on NNLS and CIBERSORT. Standard proteomic marker proteins such as CD19, CD20, and CD79 are well-established for identifying and quantifying B cells within complex mixtures (e.g., via cytometry or immunohistochemistry), so in that regard classification of cancerous and noncancerous fractions is not an unsolved problem and could be solved

by other means, it merely showcases the strengths of our deconvolution algorithms.

Chapter 5

Discussion

5.1 Conclusion

This study represents one of the first comprehensive benchmarking efforts for proteomic-based cellular deconvolution methods. Our analysis began with extensive data exploration, including PCA visualization of immune cell subsets across 10,000-dimensional proteomic space, characterization of cell state dependencies (activated vs. steady state), and detailed analysis of healthy versus pathological cell type distributions. This biological foundation informed our simulation framework and revealed critical insights about proteomic signature clustering patterns. Then our further analysis on deconvolution models demonstrates that while existing computational methodologies originally developed for transcriptomic data can be adapted for proteomic applications, they require careful consideration of data preprocessing and normalization strategies specific to protein expression data. In addition, we show how the reference signature matrix is constructed is vital to prediction accuracy of any of the reference based models discussed. The results reveal that NNLS and CIBERSORT achieve the highest prediction accuracy when paired with appropriate procedures. Notably, both performed sufficiently well to distinguish between healthy and cancerous samples based on predicted cell-type composition achieving 100% classification accuracy using simple B-cell fraction thresholds. However, it is important to note that the deconvolution and classification results were achieved using the protemoic profiles of exclusively healthy patients as references. This limitation highlights the need for need for future validation with authentic pathological samples.

To facilitate the reproducibility and continued research in this emerging field, we have developed and openly released comprehensive computational framework that enables researches to generate synthetic bulk samples with user-defined or predefined healthy and cancerous cell type distributions,

visualize signature matrix relationships, and systematically evaluate deconvolution methods using standardized metrics. The framework supports flexible signature matrix construction- from Dirichlet distributed combinations of multiple individual reference, of users choice, to fixed single patient matrices, and enables direct comparison of CIBERSORT, BayesPrism and NNLS with any of the mentioned normalization strategies.

Our investigation also uncovered fundamental challenges in proteomic deconvolution. All methods exhibited systematic biases toward specific cell types, with consistent overestimation of T cells and a strong tendency to underestimation of B and Myeloid cells. More critically, performance degraded substantially when cell type distributions were highly skewed, as typically observed in pathological conditions.

5.2 Model Performance

5.2.1 Methodological Contributions and Open Science Impact

Beyond the algorithmic findings, this work makes significant methodological contributions to the field though its comprehensive benchmarking framework and commitment to reproducible research. Our data exploration revealed fundamental biological insights, including the clustering of proteomic signatures of healthy cells by cell lineage rather than individual patient, the dramatic impact of cell activation states on proteomic profiles, and the lymphocyte distribution of healthy individuals compared to the ones of lymphoma patients. These biological observations directly informed our simulation strategies and highlighted considerations often overlooked in purely computational studies. The open-source computational pipeline we developed represents a significant resource for the research community. The framework enables users to generate realistic bulk proteomic samples with customizable cell type distributions, systematically explore signature matrix construction approaches and evaluate deconvolution performance using standardized metrics. Researchers can directly compare CIBERSORT, BayesPrism, and NNLS implementations with any of the presented normalization strategies, facilitating rapid method development and validation. The pipeline's flexibility in signature matrix construction, supporting both Dirichlet distributed, multi-individual references and fixed single patient matrices, allows systematic investigation of how reference diversity impacts deconvolution accuracy.

5.2.2 Methodological Strengths and Limitations

CIBERSORT with inlogged normalization emerged as the most robust method among those not requiring prior knowledge of the generating reference distributions, achieving the lowest average error (1.00%) while demonstrating

strong correlations across all cell types. However, this method revealed significant sensitivity to preprocessing choices; the dramatic difference between inlogged and nonlogged normalizations highlights the critical importance of data preprocessing in proteomic applications.

For the methods that deconvolved bulk samples, that didn't undergo log transformation before being added up, as is suspected to be biological reality, NNLS with outlogged normalization performed the best, combining simplicity with competitive performance across all evaluation metrics. Its deterministic nature and lack of hyperparameter requirements make it particularly attractive as a baselines for further benchmarking. This method achieved near-perfect correlations with true cell fractions (Pearson 0.99 for B and T cells) while maintaining computational efficiency, with by far the shortest runtime compared to BayesPrism and CIBERSORT. Based on the observed performance differences of methods with different choices of normalizations, it is evident that assumptions underlying transcriptomic normalization procedures may not translate directly to proteomic data, necessitating protocol-specific optimization.

BayesPrism, while conceptually appealing for its joint estimation of cell fractions and signature matrices, consistently underperformed relative to simpler alternatives. The method's signature matrix updates systematically overshot true profiles, suggesting that the Bayesian framework may require substantial modification for proteomic applications. Despite filtering 471 of 10'320 genes based on expression specificity, the method failed to achieve performance gains observed in transcriptomic studies. However, as visualized in the PCA plot of the signature matrices, the update direction tends to be correct, even if the magnitude is scaled to overshoot massively. This suggests that with an optimized scaling, BayesPrism could potentially be used to better estimate signature matrices and therefore produce improved results by itself or by running NNLS or CIBERSORT on the updated signature matrices.

5.2.3 Impact of Signature Matrix Design

Perhaps the most significant methodological finding was the critical importance of signature matrix construction. The use of randomized signature matrices, generated as Dirichlet-distributed combinations of individual proteomic profiles, reduced prediction errors by approximately 50% compared to fixed individual-based references. This improvement cannot be attributed solely to information leakage, as demonstrated by our controlled experiments using separate reference and test populations. The superior performance of multi-individual randomized matrices likely reflects their ability to capture population-level variation in cell type specific proteomic signatures, thereby providing more robust references for deconvolution. This finding has impor-

tant implications for future studies, suggesting that larger reference datasets incorporating biological diversity will be essential for clinical translation.

5.2.4 Bias Characterization and Clinical Implications

Our systematic bias analysis revealed concerning patterns across methodologies, also observed in transcriptomic literature. A consistent underestimation of dominant cell types and overestimation of rare populations creates fundamental tension in clinical contexts. While slight overestimation of rare cell types may be preferable for detecting clinically significant minority populations, such as cancer stem cells or immune infiltrates, the magnitude of bias observed for highly skewed distributions could lead to misinterpretation of pathological states. The cell type specific biases we observed, particularly the systematic underestimation of B cells in cancerous distributions, may reflect fundamental limitations of these algorithms. These biases seem to influence results independent of abundance, suggesting that certain cell types may have inherently overlapping proteomic profiles that confound deconvolution algorithms.

5.3 Future Research Directions

The most pressing need for advancing proteomic reference based deconvolution is the generation of larger, more diverse reference datasets. Current limitations in single-cell proteomic profiling restrict the number of reference profiles available for signature matrix construction. Future studies should prioritize: 1. Expanded single-cell proteomic atlases: Comprehensive profiling of immune cells subsets across diverse populations, age groups, and health conditions 2. Disease specific references: integration of proteomic profiles from B-cell lymphoma patients. 3. Cross platform validation: Assessments of method performance across different proteomic technologies and sample preparation protocols.

5.3.1 Methodological Development

Several methodological improvements could benefit performance:

1. Bias corrections strategies: development of post-processing algorithms to correct for systematic cell type type specific biases
2. Skewed distribution handling: novel algorithms specifically designed to handle highly imbalanced cell-type compositions.
3. Uncertainty quantification: Integration of confidence intervals and prediction of uncertainty measures.

4. Signature matrix update correction: Algorithms that determine the scaling factor of the update tendency and possibly improve the updates to more accurately estimate the real signature matrix used for sample construction.

The field of proteomic deconvolution represents a promising but nascent area requiring substantial methodological development and biological validation. While current results demonstrate feasibility for certain applications, achieving the accuracy and reliability required for clinical decision-making will require more data and research. The systematic biases and limitations identified in this study provide a roadmap for future research efforts in reference based proteomic deconvolution.

Appendix A

Appendix

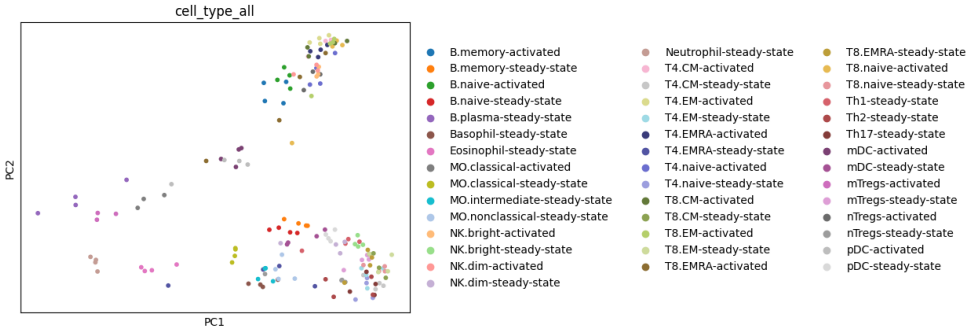


Figure A.1: Activated vs. steady state fine

Table A.1: Subtype composition percentages within major immune cell types

Main Cell Type	Subtype	Percentage of subtype in main type (%)
B cells	B.memory	45.15
	B.naive	54.18
	B.plasma	0.67
T cells	T4.CM	14.77
	T4.EM	7.17
	T4.EMRA	0.84
	T4.naive	12.66
	T8.CM	8.86
	T8.EM	9.70
	T8.EMRA	6.75
	T8.naive	10.55
	Th1	10.13
	Th17	5.91
	Th2	6.33
	mTregs	4.64
	nTregs	1.69
Myeloid cells	Basophil	8.55
	Eosinophil	6.84
	MO.classical	28.21
	MO.intermediate	4.27
	MO.nonclassical	9.40
	Neutrophil	34.19
	mDC	2.56
	pDC	5.98
NK cells	NK.bright	41.03
	NK.dim	58.97

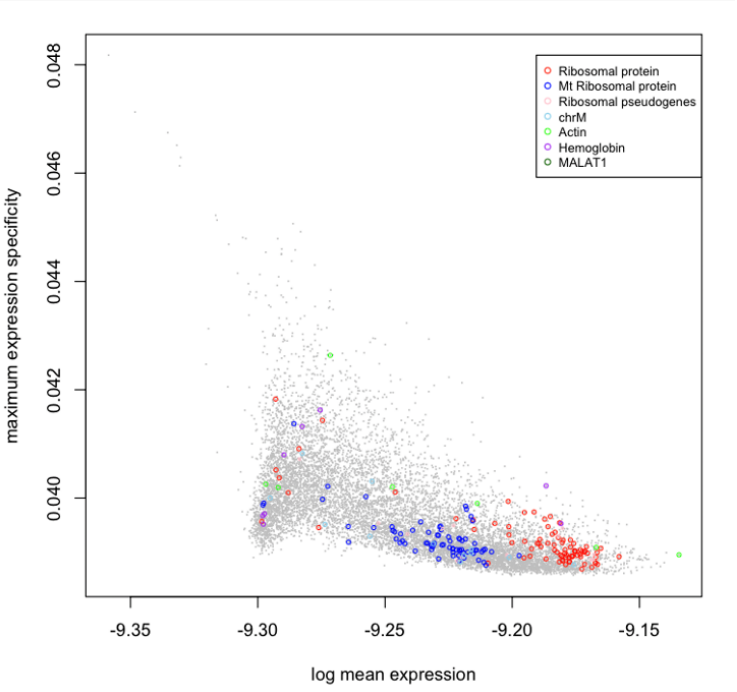


Figure A.2: Enter Caption

Table A.2: Pearson and Spearman correlations for each cell type of the inlogged cibersort

Cell Type	Pearson Correlation	Spearman Correlation
B cells	0.999	0.939
T cells	0.999	0.939
Myeloid cells	0.998	0.987
NK cells	0.983	0.866

Table A.3: Pearson and Spearman correlations for each cell type outlogged cibersort.

Cell Type	Pearson Correlation	Spearman Correlation
B cells	0.999	0.915
T cells	0.999	0.927
Myeloid cells	0.998	0.975
NK cells	0.986	0.915

Bibliography

[1] American Cancer Society. Types of b-cell lymphoma. https://www.cancer.org/cancer/non-hodgkin-lymphoma/about/types.html, 2019. Accessed: 2025-07-10.

[2] Virginie Chantry, Dominique Sluse, and Pierre Magain. COSMOGRAIL: the COSmological MOnitoring of GRAvItational lenses VIII. deconvolution of high resolution near-IR images and simple mass models for 7 gravitationally lensed quasars. *arXiv [astro-ph.CO]*, 2010.

[3] Tinyi Chu, Zhong Wang, Dana Pe'er, and Charles G Danko. Cell type and gene expression deconvolution with BayesPrism enables bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer*, 3(4):505–517, April 2022.

[4] Eric W. Deutsch, Nuno Bandeira, Vishal Sharma, Yasset Perez-Riverol, Jonah J. Carver, Dario J. Kundu, Daniel Garcia-Seisdedos, Andrzej F. Jarnuczak, Sanduni Hewapathirana, Benjamin Pullman, et al. The proteomexchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Research*, 48(D1):D1145–D1152, 2020.

[5] GlobalData. B-cell non-hodgkin's lymphoma: Epidemiology analysis and forecast to 2032. https://www.globaldata.com/store/report/b-cell-non-hodgkins-lymphoma-epidemiology-analysis/, 2024. Accessed: 2025-07-10.

[6] National Cancer Institute. Seer cancer stat facts: Non-hodgkin lymphoma. https://seer.cancer.gov/statfacts/html/nhl.html, 2024. Accessed: 2025-07-10.

[7] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A

Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, May 2015.

[8] Hung Nguyen, Ha Nguyen, Duc Tran, Sorin Draghici, and Tin Nguyen. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research*, 52(9):4761–4783, 04 2024.

[9] Francesca Petralia, Azra Krek, Anna P. Calinawan, Daniel Charytonowicz, Robert Sebra, Song Feng, Sara Gosline, Pietro Pugliese, Amanda G. Paulovich, Jacob J. Kennedy, Michele Ceccarelli, and Pei Wang. Bayesdebulk: A flexible bayesian algorithm for the deconvolution of bulk tumor data. June 2021.

[10] Jorge Pitarch-Motellón, Lubertus Bijlsma, Juan Vicente Sancho Llopis, and Antoni F Roig-Navarro. Isotope pattern deconvolution as a successful alternative to calibration curve for application in wastewater-based epidemiology. *Anal. Bioanal. Chem.*, 413(13):3433–3442, May 2021.

[11] Barbara Piatosa, Beata Wolska-Kuśnierz, Katarzyna Siewiera, Hanna Grzduk, Ewa Gałkowska, and Ewa Bernatowska. Clinical immunology-distribution of leukocyte and lymphocyte subsets in peripheral blood. age related normal values for preliminary evaluation of the immune status in polish children. *Central European Journal of Immunology*, 35(3), 2010.

[12] Jan C Rieckmann, Roger Geiger, Daniel Hornburg, Tobias Wolf, Ksenya Kveler, David Jarrossay, Federica Sallusto, Shai S Shen-Orr, Antonio Lanzavecchia, Matthias Mann, and Felix Meissner. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nature Immunology*, 18(5):583–593, May 2017.

[13] Deborah M Stephens and John C Byrd. Resistance to bruton tyrosine kinase inhibitors: the achilles heel of their success story in lymphoid malignancies. *Blood*, 138(13):1099–1109, September 2021.

[14] Susanne C van den Brink, Fanny Sage, Ábel Vértesy, Bastiaan Spanjaard, Josi Peterson-Maduro, Chloé S Baron, Catherine Robin, and Alexander van Oudenaarden. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10):935–936, September 2017.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten schriftlichen Arbeit. Eine der folgenden zwei Optionen ist **in Absprache mit der verantwortlichen Betreuungsperson** verbindlich auszuwählen:

☐ Ich erkläre hiermit, dass ich die vorliegende Arbeit eigenverantwortlich verfasst habe, namentlich, dass mir niemand beim Verfassen der Arbeit geholfen hat. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge der Betreuungsperson. Es wurden keine Technologien der generativen künstlichen Intelligenz[1] verwendet.

☐ Ich erkläre hiermit, dass ich die vorliegende Arbeit eigenverantwortlich verfasst habe. Dabei habe ich nur die erlaubten Hilfsmittel verwendet, darunter sprachliche und inhaltliche Korrekturvorschläge der Betreuungsperson sowie Technologien der generativen künstlichen Intelligenz. Deren Einsatz und Kennzeichnung ist mit der Betreuungsperson abgesprochen.

Titel der Arbeit:

Verfasst von:
Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.

Name(n):	**Vorname(n):**

Ich bestätige mit meiner Unterschrift:
- Ich habe mich an die Regeln des «Zitierleitfadens» gehalten.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu und vollständig dokumentiert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Eigenständigkeit überprüft werden kann.

Ort, Datum	**Unterschrift(en)**

Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie grundsätzlich gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.

[1] Für weitere Informationen konsultieren Sie bitte die Webseiten der ETH Zürich, bspw. https://ethz.ch/de/die-eth-zuerich/lehre/ai-in-education.html und https://library.ethz.ch/forschen-und-publizieren/Wissenschaftliches-Schreiben-an-der-ETH-Zuerich.html (Änderungen vorbehalten).