AUGMENTED DOUBLY ROBUST POST-IMPUTATION INFERENCE FOR PROTEOMIC DATA

BY HAEUN MOON, JIN-HONG DU, JING LEI AND KATHRYN ROEDER

Department of Statistics and Data Science, Carnegie Mellon University

Quantitative measurements produced by mass spectrometry proteomics experiments offer a direct way to explore the role of proteins in molecular mechanisms. However, analysis of such data is challenging due to the large proportion of missing values. A common strategy to address this issue is to utilize an imputed dataset, which often introduces systematic bias into downstream analyses if the imputation errors are ignored. In this paper, we propose a statistical framework inspired by doubly robust estimators that offers valid and efficient inference for proteomic data. Our framework combines powerful machine learning tools, such as variational autoencoders, to augment the imputation quality with high-dimensional peptide data, and a parametric model to estimate the propensity score for debiasing imputed outcomes. Our estimator is compatible with the double machine learning framework and has provable properties. Simulation studies verify its empirical superiority over other existing procedures. In application to both single-cell proteomic data and bulk-cell Alzheimer's Disease data our method utilizes the imputed data to gain additional, meaningful discoveries and yet maintains good control of false positives.

1. Introduction. Recently single-cell RNA sequencing technology has fueled a revolution in our ability to study biological processes. However, mRNA transcript abundances are only a weakly correlated precursor to protein abundances (Vogel and Marcotte, 2012; Liu et al., 2016; Tasaki et al., 2022). And it is the protein that carries out the more fundamental roles of molecular mechanisms in cellular processes. Developments in mass spectrometry proteomic technology have greatly enhanced the quantitative analysis of proteins related to human health and disease. Nevertheless, such analyses often encounter challenges due to a high rate of missingness, especially for single-cell data, resulting from various technological factors (Vanderaa and Gatto, 2023). While missingness significantly impacts the validity and efficiency of downstream tasks, the optimal method for handling missing data in proteomics remains a subject of active debate, and is an area in need of novel methodological advancements (Shen et al., 2022).

Currently, a major focus of discussion in the field is on the choice of imputation method (Vanderaa and Gatto, 2023; Wei et al., 2018), which is used to infer peptide abundance, a subunit of proteins. It is common practice that imputed values are directly plugged into the original dataset, followed by downstream analyses as if the imputed values were the original observed ones ("Plugin method"). With this method, the assumption is that the imputed data accurately represents the original data. Therefore, the precision of the imputed result is crucial for a valid downstream analysis. A substantial ongoing research effort is to search and experiment with numerous imputation methods to determine the optimal one, including sample matching methods (Stuart and Satija, 2019), matrix factorization methods (Hastie et al., 2015), deep learning methods (Yoon et al., 2018; Qiu et al., 2020; Du et al., 2022) and more (Wang et al., 2016; Chen et al., 2017). See Harris et al. (2023); Välikangas et al. (2018); Liu and Dongre (2021) for a comprehensive review.

Keywords and phrases: proteomic data, post-imputation inference, double robustness, variational autoencoder.

Most of the aforementioned methods rely on a high-dimensionality and robust intercorrelation structure of the measured peptides. Such characteristics of proteomic data provide a solid foundation for various imputation algorithms; however, this approach may not be ideal when the downstream analysis plan is based on the Plugin method. There are two reasons for this. First, the aim of retrieving the original outcomes via imputation is not optimal in some downstream analyses. Consider a linear model in which we regress each peptide abundance in some low-dimensional covariates. In this context, the optimal choice for imputation is the conditional mean abundance based on these covariates. When the Plugin method is combined with high-dimensional imputation models, we are attempting to get closer to the original outcome, rather than the conditional mean, which may introduce additional variance into estimated regression coefficients. Second, when the full high-dimensional dataset is used for imputation, a systematic bias can be introduced into the imputed data, causing false discovery due to confounding. A recent paper by Agarwal et al. (2020) investigates this issue using transcriptomic datasets. They show that if the dataset contains a number of differentially expressed genes, a naive application of the Plugin method results in notably inflated False Discovery Rates (FDR). This inflation does not occur when none of the genes are differentially expressed, which indicates that the source of the FDR inflation is the cross-use of high-dimensional data for imputation. More discussions on this can be found in Andrews and Hemberg (2018); Ly and Vingron (2022). Similar post-imputation inference issues remain for proteomic studies.

One approach that can circumvent these issues is to use only complete data for analysis and simply ignore missingness ("Complete method"). This provides a simple and valid way to prevent problems from imputation under certain missingness assumptions. However, it discards any indirect information on missing outcomes and is especially vulnerable to low power with small sample sizes. Multiple imputation (Rubin, 1987) is another possible approach, which provides a general framework for obtaining valid statistical inferences while incorporating the imputation uncertainty. This technique avoids denoising and involves generating multiple complete datasets by filling in missing data with several plausible imputations. The resulting test statistic incorporates variances both within and between datasets to compute the total variance. There have been some attempts to apply this framework to proteomic data (Yin et al., 2016; Gianetto et al., 2020). Some noticeable challenges involved in using this approach include its empirical conservativeness (Chion et al., 2022), computational burden (Brini and van den Heuvel, 2023), and the lack of a straightforward expression for test statistics (Meng, 1994).

In this paper, we propose an alternative framework motivated by doubly robust estimation (Scharfstein et al., 1999), a widely used procedure to estimate mean outcomes. Our purpose is to establish a valid and efficient inference framework that is well-harmonized with high-dimensional imputation models. Estimating mean outcomes is a significant area of research, especially when certain outcomes are not observable and a propensity score (probability of observation) depends on measured covariates. Then observed outcomes do not accurately represent the entire population due to the covariate mismatch. Therefore, instead of simply averaging the observed outcomes, one first constructs an outcome model by regressing the outcomes on covariates related to the propensity score and averaging the fitted values over the entire population. A doubly robust estimator incorporates an additional term to correct for the first-order bias of the fitted outcomes. While two nuisance estimators – an outcome estimator and a propensity score estimator – are employed, this approach enjoys a "double robustness" property, which means that the statistic remains consistent as long as at least one of the nuisance estimators is consistent (Robins and Rotnitzky, 1995). Several recent papers extend this strategy to estimation problems beyond the mean outcome (Kennedy, 2023; Fisher and Fisher, 2023; Díaz et al., 2018; Qiu and Messer, 2023). In particular, Kennedy (2023) uses

each summand of the doubly robust estimator as a pseudo-outcome to measure a conditional average treatment effect in a nonparametric regression setting.

Adopting this strategy to a linear regression setting, we utilize the summands of the aforementioned doubly robust estimator as pseudo-outcomes and transfer its favorable properties to regression coefficients. Moreover, the availability of high-dimensional proteomic data offers us the opportunity to augment our estimator by using this additional information. We show that the asymptotic variance of the estimated coefficients is further reduced by augmenting the imputation model. Our strategy is to use the entire proteomic data as an auxiliary variable and use their intercorrelated structure for imputation. To illustrate the usefulness of this approach, we provide a simple experiment. Assume that there exists an auxiliary variable that is correlated with the outcome of interest. Then the outcome model with the auxiliary variable (Model UW) provides better statistical power compared to a model without it (Model W) in a downstream task, and the gap increases as the auxiliary variable becomes more informative for the outcome variable (Figure 1). Further details of implementation and interpretation are provided in Section 2.2.

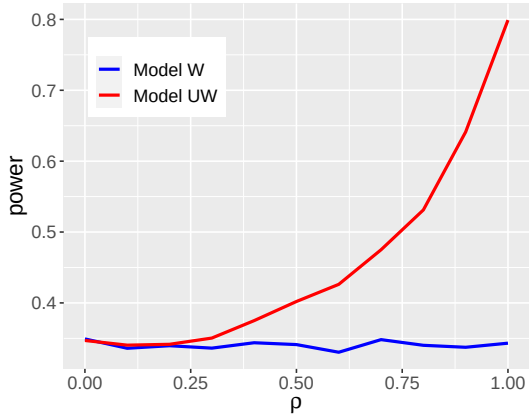

Figure 1: Statistical power of rejecting $\boldsymbol{\beta} = 0$ at different levels of correlation $\rho \in \{0.1, 0.2, \cdots, 1\}$ between an outcome Y_i and an auxilary variable U_i. Further implementation details are provided in Section 2.2.

In our framework, the propensity score is estimated through a conventional logit model to enjoy a fast rate of parametric convergence, while the outcome model is estimated through a flexible machine-learning method that can handle high-dimensional variables and their complex relationships. Our framework not only calls for, but also deliberately invites powerful modern methods because it includes a built-in mechanism to push the estimator towards achieving $\sqrt{n}$-consistency, even when the employed imputation method fails to achieve a sufficiently fast rate. In our simulations and data study, we use a variant of VAE models called VAEIT (Du et al., 2022) to fit the outcomes; see Appendix B for more details. The VAEIT model utilizes both low-dimensional covariates and high-dimensional proteomic data, and offers enough flexibility to handle missing data as well as non-linear dependency.

Other related works. In high-dimensional nuisance parameter estimation, Jiang et al. (2022) and Yadlowsky (2022) derived consistency results for estimated conditional treatment effect with sparsity or distributional assumptions. Double machine learning, proposed by Chernozhukov et al. (2018), provides a framework for building an efficient estimator of low-dimensional parameters, with nuisance functions estimated using a high-dimensional black-

box model. More papers based on semiparametric nuisance estimation are summarized in Davidian (2022). Most of the aforementioned references use the same set of high-dimensional variables for both nuisance functions. Other lines of investigation, including Berrevoets et al. (2023), Zhao and Ding (2022), and Little et al. (2012), explore an estimator for average treatment effect when some data are missing. They measure the effect size by adjusting the covariate distributions of treatment and control groups separately and computing the outcome difference.

The rest of our paper is organized as follows. In Section 2, we formally introduce the doubly robust estimator, and our procedure for estimating a regression coefficient drawn from doubly robust pseudo-outcomes. We then motivate the use of augmented imputation, define the augmented doubly robust estimator, and establish its asymptotic properties. In Section 3, we describe a multiple testing procedure as an example of downstream applications of the proposed estimators and demonstrate their favorable finite sample performance compared to benchmark methods. Next, we apply the proposed method to analyze a real proteomic dataset. In Section 4, we analyze a single-cell peptide dataset with cell-specific covariates, identifying peptides whose abundance is related to the cell size. In Section 5, we apply the proposed method to a bulk-cell dataset annotated with a range of Alzheimer's Disease symptoms. Section 6 summarizes the paper and discusses possible issues in the application of the proposed method. The results presented in Section 4 and 5 can be reproduced using the code provided at https://github.com/HaeunM/peptide-imputation-inference.

2. Method.

2.1. *Background.* Suppose n identically and independently distributed samples $(\mathbf{W}_1, Y_1)$, $\ldots, (\mathbf{W}_n, Y_n) \in \mathbb{R}^q \times \mathbb{R}$ are drawn from a linear model:

$$Y_i = \mathbf{W}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is the coefficient vector and $\epsilon_i \in \mathbb{R}$ is a zero-mean noise. We consider the missing data problem when some of the outcomes Y_i's are not observable. Specifically, we denote the observability of Y_i by a binary random variable $C_i \in \{0, 1\}$, such that one can only observe $(\mathbf{W}_i, C_i, C_i Y_i)$ for $i = 1, \ldots, n$. Under the missing data setting, we are interested in testing the hypothesis:

$$H_0 : \boldsymbol{\beta} = 0 \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq 0.$$

If every outcome is observable ($C_i = 1$ for all $i \in \{1, \cdots, n\}$), the ordinary least square regression (OLS) is arguably the most common method for estimating $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mathbf{W}_i^T \boldsymbol{\beta})^2 = \left(\sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^T \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{W}_i Y_i \right). \tag{2}$$

A test statistic can be obtained based on its asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \mathbb{E}[\mathbf{W}_i \mathbf{W}_i^T]^{-1} \mathbb{E}[\epsilon_i^2 \mathbf{W}_i \mathbf{W}_i^T] \mathbb{E}[\mathbf{W}_i \mathbf{W}_i^T]^{-1}),$$

where the asymptotic covariance can be approximated by a plugin estimator

$$\left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{W}_i \hat{\boldsymbol{\beta}}_{OLS})^2 \mathbf{W}_i \mathbf{W}_i^T \right) \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^T \right)^{-1}.$$

This is one of the most well-known inference frameworks in statistics. When some outcomes are not observed, the least squares estimate is not applicable.

If the rate of missingness is only related to a measured covariate ($C_i \perp Y_i | \mathbf{W}_i$), a simple strategy of excluding missing samples provides valid inferential results (Little, 1992); but it comes at the expense of a reduced sample size. Therefore, we consider the pseudo-outcome approach, which can offer better statistical efficiency. In an ideal scenario, when the conditional mean $\mathbb{E}[Y_i \mid \mathbf{W}_i]$ is available, replacing the outcome data with this value will provide valid and efficient inference. This approach has been explored in causal inference studies. While the typical average treatment effect estimates $\mathbb{E}[Y_i]$, the conditional average treatment effect seeks an individualized conditional outcome $\mathbb{E}[Y_i \mid \mathbf{W}_i]$, especially when Y_i is not observable for counterfactual cases. Several recent papers address this issue by utilizing pseudo-outcomes, which have the same conditional means as the original outcomes, and fitting a regression against them as if they were observed data (Kennedy, 2023; Fisher and Fisher, 2023; Semenova and Chernozhukov, 2021; Díaz et al., 2018). This approach allows for achieving desirable properties such as robustness and efficiency through a selection of appropriate pseudo-outcomes.

Inspired by these studies, we further extend the pseudo-outcome framework in linear regression. We especially focus on the doubly robust estimator suggested by Scharfstein et al. (1999), which is extensively used to estimate the mean outcome $\mathbb{E}[Y_i]$. This estimator is defined upon the construction of two nuisance functions;

(Outcome model) $$\mu(\mathbf{w}) = \mathbb{E}[Y_i \mid \mathbf{W}_i = \mathbf{w}]$$

(Propensity model) $$\delta(\mathbf{w}) = \mathbb{P}(C_i = 1 \mid \mathbf{W}_i = \mathbf{w}),$$

and is formulated as $\frac{1}{n}\sum_{i=1}^{n} g(Y_i, C_i; \hat{\mu}, \hat{\delta})$, where

$$g(Y_i, C_i; \mu, \delta) = \mu(\mathbf{W}_i) + \frac{C_i}{\delta(\mathbf{W}_i)}(Y_i - \mu(\mathbf{W}_i)),$$

with estimated outcome/propensity model $\hat{\mu}$ and $\hat{\delta}$. Appealing properties of the estimator arise from its second-order nuisance estimation error, that is,

$$\mathbb{E}[g(Y_i, C_i; \hat{\mu}, \hat{\delta}) - g(Y_i, C_i; \mu, \delta) \mid \hat{\mu}, \hat{\delta}] = (\mu(\mathbf{W}_i) - \hat{\mu}(\mathbf{W}_i)) \left(1 - \frac{\delta(\mathbf{W}_i)}{\hat{\delta}(\mathbf{W}_i)}\right). \tag{3}$$

Then, under some weak assumptions on convergence rates of $\hat{\mu}$ and $\hat{\delta}$, the bias of the estimator from the nuisance estimation error becomes negligible (Kennedy, 2023). Moreover, consistency of the estimator is achieved if either μ or δ is consistently estimated, which is referred to as the doubly robust property.

In the regression setting, pseudo-outcomes can be introduced as follows:

$$\hat{Y}_i^W = \hat{\mu}_i + \frac{C_i}{\hat{\delta}_i}(Y_i - \hat{\mu}_i).$$

Here, and in the rest of this paper, we write $\mu_i = \mu(\mathbf{W}_i)$ and $\delta_i = \delta(\mathbf{W}_i)$ and similarly for the estimated versions. Regressing $(\hat{Y}_1^W, ..., \hat{Y}_n^W)$ on $(\mathbf{W}_1, ..., \mathbf{W}_n)$ yields a least squares estimator given by:

$$\hat{\boldsymbol{\beta}}_W = \left(\sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^T\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{W}_i \hat{Y}_i^W\right).$$

As we will show in Section 2.2, the estimator $\hat{\boldsymbol{\beta}}_W$ also has the doubly robust property and can lead to more efficient inference.

2.2. *An augmented doubly robust estimator* $\hat{\boldsymbol{\beta}}_{UW}$. In peptide abundance analysis, there are often a large collection of peptides measured and analyzed together. For each peptide, one can predict its value using not only the low-dimensional covariate $\mathbf{W}$ but also the other peptides, which can be regarded as a high-dimensional covariate $\mathbf{U}$. Our strategy is to recover Y as accurately as possible through an augmented outcome model that incorporates both $\mathbf{W}$ and $\mathbf{U}$ as predictors for the response Y. If the augmented outcome model will result in a significant reduction in the variance of the regression residual $Y - \mathbb{E}[Y|\mathbf{W}, \mathbf{U}]$, then we may expect to have a smaller asymptotic variance for the estimated regression coefficient using the augmented pseudo-outcome.

Formally, our proposed estimator is defined as

$$\hat{\boldsymbol{\beta}}_{UW} = \left(\sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^T\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_i \left(\hat{\nu}_i + \frac{C_i}{\hat{\delta}_i}(Y_i - \hat{\nu}_i)\right) \tag{4}$$

for nuisance estimators $\hat{\nu}(\mathbf{w}, \mathbf{u}) = \hat{\mathbb{E}}[Y_i|\mathbf{W}_i = \mathbf{w}, \mathbf{U}_i = \mathbf{u}]$ and $\hat{\delta}(\mathbf{w}) = \hat{\mathbb{E}}[C_i = 1|\mathbf{W}_i = \mathbf{w}]$.

Before providing a rigorous analysis, we provide a simple example to illustrate the variance reduction effect of augmentation. Consider a linear regression model $Y_i = \beta W_i + \epsilon_i$ for $\beta, W_i \in \mathbb{R}$. An auxilary variable $U_i \in \mathbb{R}$ is defined as $U_i = \beta W_i + \epsilon_{u_i}$, where $\mathrm{Cor}(\epsilon_{u_i}, \epsilon_i) = \rho$. Since U_i partly explains the residual term ϵ_i, the outcome $\nu_i = \mathbb{E}[Y_i \mid W_i, U_i]$ provides a higher resolution estimate of Y_i than $\mu_i = \mathbb{E}[Y_i \mid W_i]$. We compare two pseudo-outcomes $\hat{Y}_i^W = \hat{\mu}_i + \frac{C_i}{\hat{\delta}_i}(Y_i - \hat{\mu}_i)$ (Model W) and $\hat{Y}_i^{UW} = \hat{\nu}_i + \frac{C_i}{\hat{\delta}_i}(Y_i - \hat{\nu}_i)$ (Model UW) in their downstream performance. Specifically, we perform a linear regression against each pseudo-outcome on $\mathbf{W}_i$ and their statistical powers in rejecting $\boldsymbol{\beta} = 0$ are compared. The outcome Y_i has random missingness with a known observation probability $\delta_i = 0.7$, the true coefficient is $\beta = 0.2$, and the sample size is $n = 200$. The results are averaged over 5000 repetitions. The result shows that Model UW outperforms Model W, and it provides increasing power as the auxiliary variable becomes more informative for the outcome ($\rho \to 1$, Fig. 1). In real applications, it is less probable that a single protein exhibits such a substantial correlation with an outcome. Instead, high-dimensional proteomic data may collectively contribute to recovering the outcome.

Next, we derive asymptotic properties of the proposed estimator $\hat{\boldsymbol{\beta}}_{UW}$ rigorously. Here, we prove that $\hat{\boldsymbol{\beta}}_{UW}$ possesses a doubly robust property (Theorem 2.2) and asymptotic normality (Theorem 2.3), and its asymptotic variance is smaller than that of $\hat{\boldsymbol{\beta}}_W$ (Theorem 2.4).

Notation. We denote the L_2 norm of a vector, or a random variable, or a function of a random variable as $\|\cdot\|_2$. For example, for a random vector $\mathbf{W}$ and its function $\nu = \nu(\mathbf{W})$, $\|\nu\|_2$ is defined as $(\int \|\nu(\mathbf{W})\|_2^2 \mathrm{d}P_W)^{1/2}$. L-infinity norm of a vector, or a random variable, or a function of a random variable is denoted as $\|\cdot\|_\infty$. For matrices M_A and M_B, we write as $M_A \preccurlyeq M_B$ if $(M_B - M_A)$ is positive semidefinite.

ASSUMPTION 2.1. (a) *Missing at random :* $Y_i \perp C_i \mid (\mathbf{W}_i, \mathbf{U}_i)$
(b) *The propensity score:* $\delta(\mathbf{W}_i) = \mathbb{P}(C_i = 1 \mid \mathbf{W}_i) = \mathbb{P}(C_i = 1 \mid \mathbf{W}_i, \mathbf{U}_i) \in (0, 1]$ *is bounded away from 0 by some constant with probability 1.*
(c) *Noise :* $\mathbb{E}[\epsilon_i \mid \mathbf{W}_i] = 0$, $\mathbb{E}[Y_i - \nu_i \mid \mathbf{W}_i, \mathbf{U}_i] = 0$, $\|\epsilon_i\|_2$ *and* $\|Y_i - \nu_i\|_\infty$ *are bounded.*
(d) *Covariate :* $\|\mathbf{W}_i\|_\infty$ *is bounded,* $\mathbb{E}[\mathbf{W}_i \mathbf{W}_i^T]$ *is a full-rank matrix.*

The second equality of Assumption 2.1(b) requires conditional independence between C and $\mathbf{U}$ given $\mathbf{W}$. This is the key assumption that allows us to use an augmented outcome model to improve efficiency.

Under the above assumptions and some additional mild assumptions on nuisance estimations, the doubly robust property follows, as shown in the following theorem.

THEOREM 2.2 (Double robustness). *Assume Assumption 2.1 (a)-(d). If one of the nuisance parameters is consistent, i.e.,* $\|\frac{\delta_i}{\hat{\delta}_i} - 1\|_2 = o_{\mathbb{P}}(1)$ *or* $\|\hat{\nu}_i - \nu_i\|_2 = o_{\mathbb{P}}(1)$*, then the estimator* $\hat{\boldsymbol{\beta}}_{UW}$ *defined in* (4) *is consistent, i.e.,* $\hat{\boldsymbol{\beta}}_{UW} \xrightarrow{P} \boldsymbol{\beta}$.

Theorem 2.2 guarantees the consistency of the proposed estimator. If further, the product of the nuisance estimation errors is small, we can derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

THEOREM 2.3 (Asymptotic normality). *Under the same conditions in Theorem 2.2, further assume that* $\|(1 - \delta_i/\hat{\delta}_i)(\hat{\nu}_i - \nu_i)\|_2 = o_{\mathbb{P}}(n^{-1/2})$*. Then the estimator* $\hat{\boldsymbol{\beta}}_{UW}$ *defined in* (4) *is asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{UW} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \Sigma_{UW})$$

where $\Sigma_{UW} = \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}\mathbb{E}[(\epsilon_i^2 + (\frac{1}{\delta_i} - 1)(Y_i - \nu_i)^2)\mathbf{W}_i\mathbf{W}_i^T]\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}$*. The asymptotic covariance* Σ_{UW} *can be consistently estimated by a plug-in estimator*

(5)
$$\hat{\Sigma}_{UW} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{UW} - \mathbf{W}_i\hat{\boldsymbol{\beta}}_{UW})^2\mathbf{W}_i\mathbf{W}_i^T\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^T\right)^{-1}$$

The asymptotic variance Σ_{UW} in Theorem 2.3 is identical to the variance obtained with oracle nuisance functions. That is, when both nuisance estimates are consistent and $\|\left(1 - \frac{\delta_i}{\hat{\delta}_i}\right)(\hat{\nu}_i - \nu_i)\|_2 = o_{\mathbb{P}}(n^{-1/2})$, the proposed estimator $\hat{\boldsymbol{\beta}}_{UW}$ is as efficient as the estimator derived using the true nuisance functions. In the Plugin method, the same property would require $\|\hat{\nu}_i - \nu_i\|_2 = o_{\mathbb{P}}(n^{-1/2})$, which is even not achievable by typical parametric estimators.

Theorem 2.4 asserts that the estimator $\hat{\boldsymbol{\beta}}_{UW}$ is asymptotically more efficient than $\hat{\boldsymbol{\beta}}_W$.

THEOREM 2.4. *Assume that conditions in Theorem 2.3 holds for* $\hat{\mu}$ *and* μ *in places of* $\hat{\nu}$ *and* ν*. Then,* $\sqrt{n}(\hat{\boldsymbol{\beta}}_W - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \Sigma_W)$ *and* $\Sigma_{UW} \preccurlyeq \Sigma_W$.

REMARK 2.5. *The results presented in this section assume that the nuisance functions* $\hat{\nu}$ *and* $\hat{\delta}$ *are estimated from samples independent of* $(Y_i, C_i, \mathbf{W}_i, \mathbf{U}_i)$*. This assumption is used for the brevity of the presentation. There are two standard approaches to improve the sample efficiency loss due to data splitting. The first is cross-fitting (Chernozhukov et al., 2018; Kennedy, 2023), which swaps the subsamples used for nuisance estimation and regression inference, and combines the test statistics from different folds to obtain the final inference. Alternatively, if the nuisance estimates belong to a Donsker class, then one can use empirical process theory to establish the asymptotic normality without sample splitting (see Lemma 19.24 of Van der Vaart, 2000, for example). Both approaches can be combined with the method proposed in this paper in a straightforward manner. In our numerical experiments and data analyses, we used the same data for nuisance estimation and post-imputation OLS inference. The good performance of our method suggests that the nuisance estimates in these settings are probably regular enough for the empirical process theory to work.*

3. Multiple testing procedure for peptides. The p-values derived in Section 2, combined with a multiple testing procedure, allow us to make discoveries of important peptides associated with a covariate of interest. Section 3.1 provides a detailed algorithm, and Section 3.2 investigate its performance compared to the benchmark methods. The same algorithm is applied to real data studies in Sections 4 and 5.

3.1. *The input data and the algorithm.* The observed abundance of p peptides from n samples can be written as an $n \times p$ matrix $\mathbf{C} \odot \mathbf{Y} \in \mathbb{R}^{n\times p}$, where $\mathbf{C} \in \{0,1\}^{n\times p}$ indicates the entry-wise missingness and $\mathbf{Y} \in \mathbb{R}^{n\times p}$ is the full data matrix without missing. Here "$\odot$" stands for the component-wise product. Only $\mathbf{C}$ and $\mathbf{C} \odot \mathbf{Y}$ are available. Also observed is a covariate data matrix $\mathbf{W} \in \mathbb{R}^{n\times q}$. The inference task is to test the significance of regression coefficients for the low dimensional covariates in $\mathbf{W}$ on each peptide. To this end, we will obtain individual p-values for each peptide using the asymptotic results presented in the previous section, and then apply a multiple testing framework such as the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

This procedure involves regressing each column of $\mathbf{Y}$ on both the low-dimensional covariate $\mathbf{W}$ and the other peptides as the high-dimensional auxiliary covariate $\mathbf{U}$. The fitted regression function is used as $\hat{\nu}$ in the imputation method described in Section 2. An additional challenge is that each column of $\mathbf{Y}$ has many missing entries, even when used as a covariate in the regression problem. To address this issue, we use variational autoencoder, a deep neural network tool that allows for flexible input and simultaneous estimation of the multi-response regression. The algorithm used in our simulation and data examples is VAEIT (Du et al., 2022, see Appendix B for details). The whole procedure is summarized in Algorithm 1.

Algorithm 1 Multiple testing procedure for peptides

Require: Observed outcome $\mathbf{C} \odot \mathbf{Y} \in \mathbb{R}^{n\times p}$; Observability $\mathbf{C} \in \mathbb{R}^{n\times p}$; Covariates $\mathbf{W} \in \mathbb{R}^{n\times q}$
Estimate $\hat{\boldsymbol{\nu}} \in \mathbb{R}^{n\times p}$ by running VAE on $(\mathbf{C}, \mathbf{C} \odot \mathbf{Y}, \mathbf{W})$.
for $j = 1, \cdots, p$ **do**
Rewrite $\mathbf{Y}_i = (Y_i, \mathbf{U}_i) \in \mathbb{R}^1 \times \mathbb{R}^{p-1}$ where $Y_i = \mathbf{Y}_{ij}$, $\mathbf{U}_i = \mathbf{Y}_{i(-j)}$ and $C_i = \mathbf{C}_{ij}$.
Estimate $\hat{\delta}_i$ by regressing $C_1, \cdots, C_n$ on $\mathbf{W}$ by logistic regression.
Compute pseudo-outcomes $\hat{Y}_i^{UW} = \frac{C_i}{\hat{\delta}_i} Y_i + (1 - \frac{C_i}{\hat{\delta}_i})\hat{\nu}_{ij}$.
Regress $\hat{Y}_1^{UW}, \cdots \hat{Y}_n^{UW}$ on $\mathbf{W}_1, \cdots, \mathbf{W}_n$ and compute a p-value (P_j) for the covariate of interest based on asymptotic distribution given in Theorem 2.3.
end for
return $P_1, \cdots, P_p$
Transform $P_1, \cdots, P_p$ to Benjamini-Hochberg's q-values and select indices whose q-values are less than a predefined cutoff.

3.2. *Simulation study.* We investigate the performance of our method compared to several other methods on simulated data. Six methods are compared; Full, Complete, MICE, DR_W, DR_UW (proposed), and Plugin, where they differ in the approach to obtain $P_1, \ldots, P_p$ in Algorithm 1. The Full method uses the practically unavailable data $\mathbf{Y}$ without missingness. We perform a linear regression for each column of $\mathbf{Y}$ on low-dimensional covariates $\mathbf{W} = (a, x)$, where a is the variable of interest and x represents any other covariates. We then use a linear regression t-test to decide if the coefficient of a equals zero. The Complete method works in the same way as the Full method, but uses only observed samples. The MICE method uses a multiple imputation method to impute missing values and then performs the test if the coefficient of a is equal to zero using a statistic proposed by Rubin (1987). MICE is not computationally feasible when high-dimensional auxiliary variables are used for imputation. For this reason, missing values are imputed on the basis of low-dimensional covariates only. The DR_W method is similar to Algorithm 1, but the columns of $\hat{\boldsymbol{\nu}}$ are fitted by a linear regression model only with low-dimensional covariates. The DR_UW method follows Algorithm 1. The Plugin method regresses each column of the fitted outcomes $\hat{\boldsymbol{\nu}}$ in

Algorithm 1 on the low-dimensional variables and performs a linear regression t-test for the coefficient of a. All five methods, except for the MICE method, require a choice of variance estimator to perform the linear regression t-test. For the Full, Complete, and Plugin methods, the usual OLS variance estimator is used. For the DR_W and DR_UW method, either the usual OLS variance estimator (in Models 1 and 2 below) or the heteroskedastic-consistent estimator (5) (in Models 3 and 4) is used. The six methods repeat the same procedure to obtain the p-values for each column of the outcome matrix. Then we transform the p-values into the Benjamini-Hochberg q-values and select the indices whose q-values are less than $\alpha = 0.3$ to identify the discoveries. For each of the six methods, the fraction of false discoveries over the number of total discoveries (FDR; False Discovery Rate) and the fraction of true discoveries over the number of signal peptides (TPR; True Positive Rate) are reported. An ideal method would control FDR within α, and have a TPR close to one. Two sample sizes of $n = 200, 500$ and a dimension $p = 1000$ are considered. The number of repetitions is 200.

The simulation data are generated as follows. For the jth peptide and the ith sample, the outcome y_i^j is formulated as

$$y_i^j = \beta_{x,j} x_i + \beta_{a,j} a_i + \epsilon_i^j$$

for $j \in \{1, \cdots, p\}$ and $i \in \{1, \cdots, n\}$. A case-control label a_i is generated by selecting the $0.5n$ indices from $\{1, \cdots, n\}$ and setting $a_i = 1$ for the cases. Otherwise, $a_i = 0$ for controls. To introduce differential abundance, we randomly select $0.1p$ peptides and inject positive signals into the case data. We denote $s^j = 1$ if j is selected and call it a signal peptide; otherwise, $s^j = 0$ and we call it a null peptide. A coefficient of interest $\beta_{a,j}$ is positive if $s^j = 1$, and zero otherwise.

Four scenarios are considered, including missing patterns, Gaussian and skewed distributions of abundance data, and various forms of the true regression model:

Model 1. Gaussian data without X (MCAR); $y_i^j = 0.3 s^j a_i + \epsilon_i^j$
Model 2. Gaussian data (MCAR); $y_i^j = x_i + 0.3 s^j a_i + \epsilon_i^j$
Model 3. Gaussian data (MAR); $y_i^j = x_i + 0.3 s^j a_i + \epsilon_i^j$
Model 4. Skewed data (MAR); $y_i^j = x_i + 0.08 s^j a_i + \epsilon_i^j$

Correlation between peptides is simulated using a realistic covariance structure to model the noise terms associated with each peptide; The covariance (Σ) was estimated from peptides measured in brain tissue (MacDonald et al., 2017). For Models 1, 2 and 3, we simulate n i.i.d. vectors $(\epsilon_i^1, \ldots, \epsilon_i^p)$, $i = 1, \ldots, n$, using a multivariate normal distribution with zero mean and covariance Σ. For Model 4, we generate skewed noise as follows: simulate multivariate normal variables as before, for each peptide add a constant to ensure that all entries are positive, apply a log transformation, and finally recenter each peptide at zero. Covariates $x_1, \cdots, x_n \in \mathbb{R}$ are generated independently from a uniform distribution in $(0, 1)$. After generating covariates and noise, each outcome y_i^j is randomly masked with the probability determined from the missingness model. In Models 1 and 2, each y_i^j is missing completely at random (MCAR) with equal probability: $\mathbb{P}(C_{ij} = 0) = 0.3$. In Models 3 and 4, y_i^j is missing at random (MAR): $\mathbb{P}(C_{ij} = 0) = e^{x_i}/\{2(1 + e^{x_i})\}$.

Figure 2 summarizes the FDR and TPR of six methods applied to Model 3. As expected, FDR is well controlled for the Full, Complete, MICE, DR_W, and DR_UW methods, whereas the Plugin method inflates the FDR. This occurs because the differences between the case and control data in signal peptides bleed into correlated null peptides during the imputation procedure. When no signals are injected into any of the peptides or when the signal is carried for sets of correlated peptides, the Plugin method is also well-controlled (results not shown).

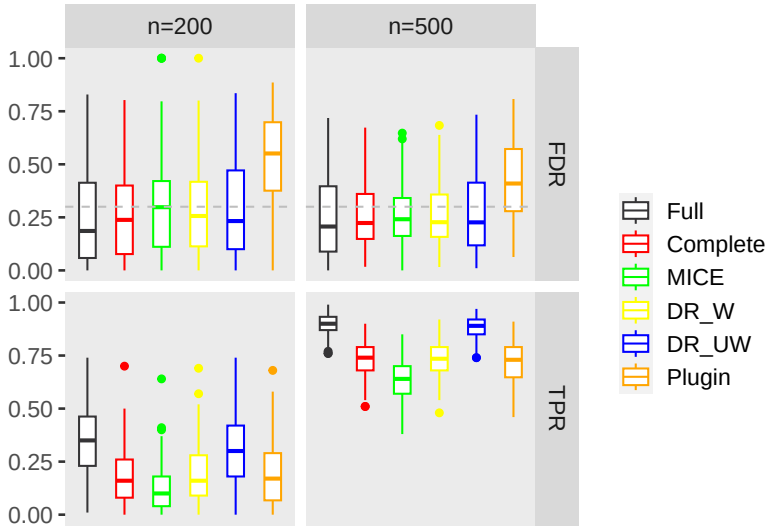

Figure 2: Performance of different methods on simulated data according to Model 3.

Naturally, the Full method demonstrates the highest TPR, representing the optimal performance achievable in this setting without missing data. DR_UW shows the second-best TPR and it becomes similar to the Full method when n=500. Complete, MICE, DR_W, and Plugin attain smaller TPR, with MICE being the most conservative. For the Complete method, this is expected because it excludes samples with missing data. MICE and DR_W perform an imputation based on low-dimensional variables, which is less accurate, leading to greater variance. Models 1,2 and 4 produce similar results (see Appendix C).

4. Single-cell protein abundance varies with cell size. We analyze a single-cell proteomic dataset based on mass spectrometry (Leduc et al., 2022) with the primary interest of detecting peptides whose abundance varies strongly with cell size. A peptide is a short chain of amino acids that constitute proteins and is a useful unit for quantitative analysis. The impact of cell size on cell physiology is of interest in two domains: large cell size may be a cause rather than a consequence of cell senescence (Lanz et al., 2022; Jones et al., 2023); and cell size is a determinant of stem cell fate (Lengefeld et al., 2021). The covariates for the analysis of these data are cell type (melanoma or monocyte) and three continuous variables (diameter, digestion, and elongation). These four variables form a low-dimensional covariate $\mathbf{W}$.

The proportion of observed cells differs significantly between the peptides, and for 85.6% of the peptides, the observation rate is smaller than 0.5. When the observation rate is too low, it is not reasonable to expect that any method will perform satisfactorily; therefore, we focus on peptides with a missingness of no more than 50%. Because the threshold for the observation rate is controversial, for the main analysis, we provide a range of results with respect to different thresholds (0.5, 0.6, 0.7, 0.9). For other parts of the analysis, including exploratory data analysis and realistic simulation, we focus on a threshold of 0.7. After removing peptides whose observation rates are less than 0.7, there are a total of 753 remaining peptides.

We first present exploratory data analysis to provide a rational basis for applying our method. The distribution of cell-wise peptide abundance data for peptides with more than 70% of observed rates reveals a Gaussian-like distribution (Fig. 3 A); thus, these data are well suited to our imputation model, which is a VAE model tailored to a Gaussian distribution; see Appendix B for more details. The distribution of pairwise distances between cells before and

after imputation shows a noticeable reduction in distances after imputation, indicating that the overall variance of abundance data has decreased (Fig. 3B). Next, we check the assumption of MAR by examining the relationship between the measured variables and the observed cell-wise rate among the peptides (propensity score). The lack of a relationship between the residuals of the estimated propensity score and the mean abundance in cells, after regressing each of them into four covariates, shows that the observed relationship between the abundance of the peptide and the propensity score is largely explained by the measured covariates (Fig. 3C). However, the joint distribution of the cell-specific propensity score and each covariate, along with its marginal distribution, illustrates that each covariate is related to the propensity score to some extent, supporting an analysis under the assumption of MAR (Fig. 3 D). Furthermore, a reasonable imputation model can be built upon the robust relationship between peptides. Examining the quantile value of 0.9 of the absolute correlation coefficient for each peptide with other peptides reveals a strong correlation pattern. These values generally fall between 0.1 and 0.5, providing a good foundation for a high-dimensional imputation model (see Figure C4 in Appendix C for details).

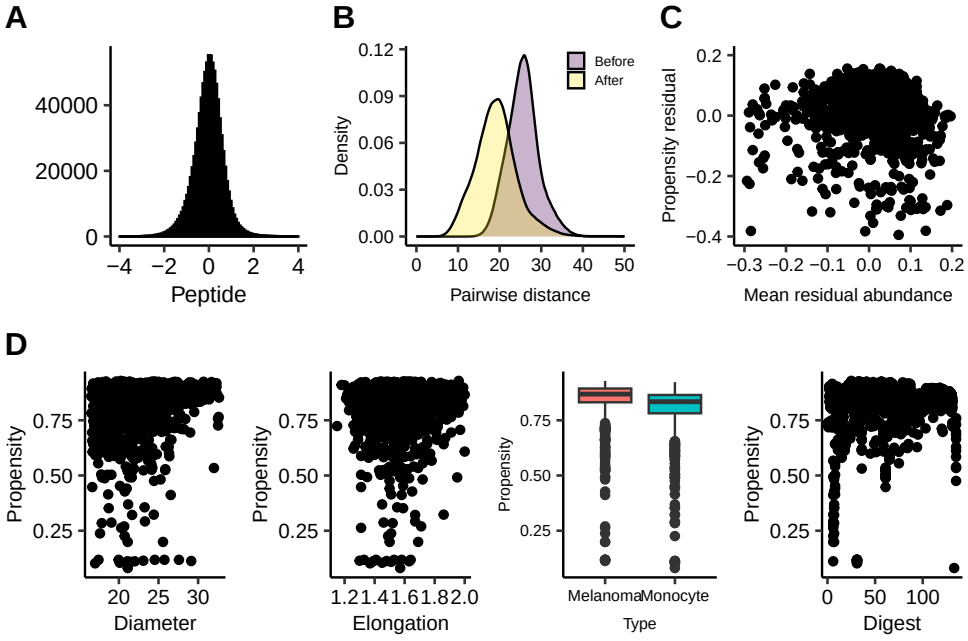

Figure 3: (A) Histogram of peptide abundance data (B) Distribution of pairwise distance between cells before and after an imputation (C) Scatter plot between propensity residual and Mean residual abundance (D) Scatterplot between Propensity score and other cell-level covariates.

This section is organized as follows. First, we conduct realistic simulations to see how the proposed method works on this dataset with some artificially generated ground truth. Next, we present the results for the main analysis, where the primary focus of our analysis is to identify peptides whose abundance varies with the diameter of the cell. For all settings, we compare the proposed DR_UW method with the Complete, DR_W, and Plugin methods. The Full method is not considered due to a lack of data.

4.1. *Realistic simulation.* Before going into the main analysis, we check the performance of the proposed method with some artificially generated ground truth incorporated into this dataset and compare different methods in terms of TPR and FDR. Specifically, we artificially generate the type variable (case control) while keeping all other variables unchanged. In Setting 1, we randomly permute the measured type variable among the cells. In Setting 2, the type variable is randomly generated from a Bernoulli distribution with a probability proportional to the propensity score. Signal peptides are randomly selected for 10% of the total considered peptides, and then a positive signal generated from a normal distribution with mean 0.2 and variance 0.05 is added to the case cells in the signal peptides. After imputation, we identify peptides with different cutoffs for the q-value, 0.01, 0.05, and 0.3.

The Complete, DR_W, and DR_UW methods detect a reasonable proportion of signal peptides while maintaining control of FDR in both settings. For FDR=0.01 and 0.05, the DR_UW method provides better TPR than the Complete and DR_W method. The Complete method is better than the DR_W method in TPR, indicating that low-dimensional imputation is too noisy for satisfactory results. For FDR=0.3, all three methods, Complete, DR_W, and DR_UW, achieve near-perfect TPR. The Plugin method severely inflates FDR for all FDR levels, and the TPR is lower than the other three methods. These results are summarized in Figure C5 and C6 in Appendix C.

4.2. *Main analysis.* One objective of this analysis is to detect peptides whose abundance varies strongly with cell size. We first filter the peptides by applying varying thresholds (0.5, 0.7, 0.9) to the observation rates of the peptides and focus on analyzing those peptides. A larger proportion of peptides, whose observed rates are greater than 0.2, is used to feed the imputation procedure. After imputation, peptides are selected based on linear regression models:

$$\text{Peptide abundance} \sim \text{Diameter+Type+Digest+Elongation}$$

where we compute the p-values associated with the diameter variable. The p-values are transformed into q-values using the BH procedure. Based on estimated coefficients and the corresponding q-values, the Complete, DR_W, and DR_UW methods exhibit roughly similar distribution patterns; however, the Plugin method presents inflation of q-values compared to the other three methods due to the signal bleeding effect (Fig. 4).

Peptide discoveries vary according to the different methods and the different threshold settings. For the purpose of comparison, we derive a robustness metric under the assumption that meaningful relationships between peptide abundance and cell diameter are strictly positive. It follows that the estimated beta coefficients tend to be positive for the signal peptides and symmetrically distributed about zero for the null peptides. This assumption appears to be valid as there is a notably distinct pattern of positive and negative signs of the beta values (Figure 4). Several recent papers (Dai et al., 2023; Guo et al., 2023; Du et al., 2023) derive an FDR metric from a statistic that possesses such an asymmetric structure of null and non-null (signal) scenarios. Motivated by these, we calculate empirical FDR as the number of significant peptides with $\hat{\boldsymbol{\beta}} < 0$ divided by the number of significant peptides with $\hat{\boldsymbol{\beta}} > 0$. We find that this metric gives a reasonable interpretation throughout the various settings. An ideal discovery result has a lower value of empirical FDR and a larger number of discoveries.

The performance of each method is evaluated in settings with different thresholds for observation rate and q-values (Table 1). We first fix the q-value cutoff at 0.05 and apply different thresholds to the observation rates of peptides. For a threshold of 0.9, empirical FDR is well-controlled for the Complete, DR_W and DR_UW methods, but the number of discoveries is relatively small because many peptides are excluded from the analysis. The Plugin method provides the largest discoveries, but its empirical FDR is inflated. When the threshold is lowered to 0.7, 0.6, and 0.5, the number of discoveries becomes larger, and empirical FDR tends

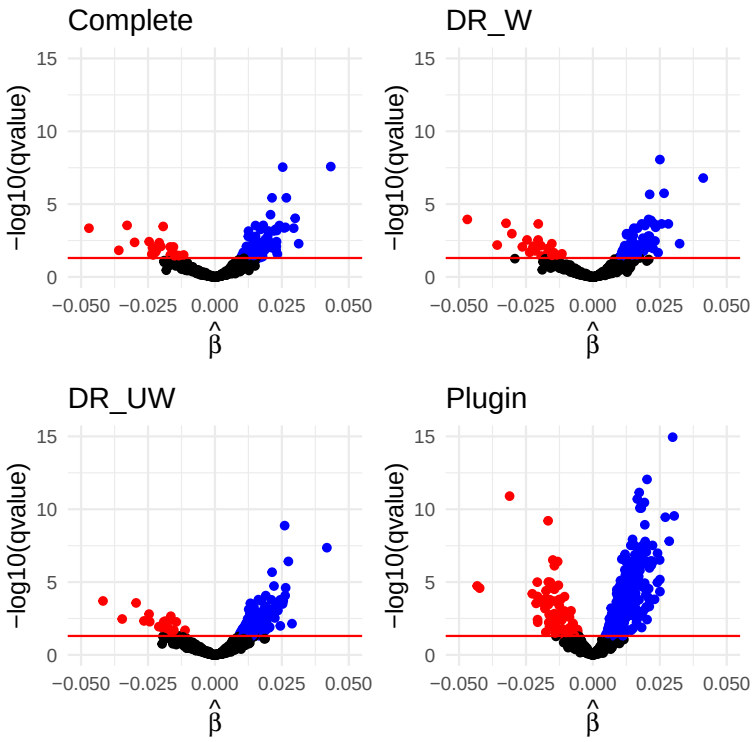

Figure 4: Volcano plot of peptide discoveries by different methods in the single-cell proteomics dataset analyzed in Section 4 when an observability threshold of 0.7 is applied. The red line indicates the q-value cutoff of 0.05.

to increase. This is natural because if the peptides with a high rate of missingness are introduced, the inference problem becomes more challenging. However, compared to the Plugin method, the other three methods consistently give better control of the empirical FDR, the proposed method DR_UW being the best. In addition, DR_UW provides a larger number of discoveries. This is consistent with what we observed from the simulations. It controls the FDR well while achieving greater TPR. The plugin method provides the largest number of discoveries, but it generally inflates the empirical FDR. Similar results hold when we fix the threshold to 0.7 and apply different q-value cutoffs.

Observability threshold	q-value cutoff	Empirical FDR				Number of selected peptides			
		Com	DR_W	DR_UW	Plugin	Com	DR_W	DR_UW	Plugin
0.9	0.05	0.05	0.05	0.04	0.10	40	44	51	86
0.7		0.20	0.22	0.13	0.26	111	106	149	303
0.6		0.31	0.33	0.25	0.40	133	128	186	419
0.5		0.41	0.45	0.35	0.55	158	152	218	535
0.7	0.01	0.20	0.25	0.15	0.20	46	44	73	225
	0.05	0.20	0.22	0.13	0.26	111	106	149	303
	0.1	0.20	0.22	0.19	0.28	148	144	189	334
	0.3	0.27	0.31	0.29	0.33	267	258	304	390

TABLE 1

Empirical FDR and number of peptides selected with each method under different combinations of thresholds applied to an observed rate of peptides and q-value cutoffs.

To further verify the robustness of the result, we map the peptides discovered by DR_W, DR_UW, and the Plugin method to the corresponding proteins, and check their overlaps with the discoveries of the Complete method (a q-value cutoff 0.05) at the protein level. We assess their contributions by only considering additional discoveries beyond those made by the Complete method with a q-value cutoff of 0.01. A threshold 0.7 is applied to the observation rate of peptides. When applying a q-value of 0.01, the additional discoveries of the DR_UW method are largely robust at the protein level; 90% of them overlap with those discovered by the Complete method. As we increase the q-value cutoff to 0.05, 0.1, and 0.3, the proportion of such overlaps tends to decrease, but the number of additional discoveries is much larger. Across all q-value cutoffs, the DR_UW method consistently provides more additional discoveries than the DR_W method, and has a slightly smaller proportion of overlaps with the Complete method. This aligns with theory and simulation results in the sense that both doubly robust methods exhibit reasonable control of the false discovery rate, while DR_UW demonstrates better efficiency. The Plugin method discovers the largest number of additional peptides, but as expected, its findings do not largely overlap with the discoveries of the Complete method at the protein level. The detailed results are summarized in Table C1 in Appendix C

5. Peptide abundance in key proteins is associated with Alzheimer's disease. Alzheimer's disease (AD) is a prominent neurodegenerative disorder among older adults. Numerous environmental and genetic factors are known to contribute to the disease, and related biological pathways have yet to be fully discovered. In this section, we apply the proposed method to identify important peptides associated with AD and related dementias. A bulk peptide-level dataset offers an opportunity to illustrate these methods in an important scientific setting (Merrihew et al., 2023). While samples in this dataset are annotated with a range of disease severity, we group them into two types; cases (samples with autosomal dominant/sporadic AD dementia) and controls (samples without dementia, with or without a high AD histopathologic burden). Two other covariates, the brain region, and PMI, are also used in our analysis.

In this analysis, we focus on peptides whose observed rates are between 0.5 and 1 in each of the four brain regions. If the observed rate is 1, the Complete method and the DR methods will provide the same selection result. The propensity scores for each sample are mostly around 0.9, and we assume the MCAR missing pattern. Some of the samples have missing covariates on PMI. After further removal of these samples, we have 488 peptides and 220 samples, including 139 cases and 81 controls. The abundance distributions of the peptides vary significantly between the brain regions. Therefore, we apply the VAE model separately to each brain region. Although VAE is usually applied to a dataset with a large sample size, when data have a Gaussian-like distribution, as in peptides, VAE gives a robust imputation outcome. The final selection of peptides is based on linear regression models:

$$\text{Peptide abundance} \sim \text{Type+Region+PMI}$$

where we compute p-values associated with the type variable. BH procedure is then applied to convert them into q-values. The final discoveries are determined by applying a cutoff of 0.05 to them. Following this procedure, the Complete method selects 55, DR_W selects 50, DR_UW selects 58, and Plugin selects 79 peptides. The Complete, DR_W, and DR_UW methods have a similar number of discoveries. Only peptides with low missingness rates were recorded for these data, so it is reasonable that the number of discoveries for each method does not vary too much. Plugin selects more peptides than the other methods.

Seven peptides are selected by the DR_UW method but not by the Complete method (Table 2). To determine whether these discoveries are meaningful, we examine the corresponding

protein and gene annotations. Six have been linked to AD and related literature (reference papers are listed in the final column of the table). Specifically, the genes PDE2A, NEUM, MX1, and CGT have revealed a direct connection with AD in the corresponding reference papers. The ANK2 gene is associated with autism and the reference paper links Drosophila ANK2 (human ANK1) to the characteristics of AD. The protein sp|Q9H305|, associated with the CDIP1 gene, plays a major role in controlling cell death, a feature of AD, and the gene is highly expressed in the brain, which implies a possible connection to AD at some level. The peptide CALD1 is only found to have an indirect connection to AD. It belongs to a group of pathway genes that change with age and are reversed by Riluzole, which is related to synaptic transmission and plasticity.

Peptide	Protein	Gene	Reference
MPLYGLHLWLPK	sp|P03905|	NU4M	Bhatia et al. (2022); Wesseling et al. (2017)
NLFTHLDDVSVLLQEIITEAR	sp|O00408|	PDE2A	Sheng et al. (2022); Shi et al. (2021); Delhaye et al. (2024)
TTHRPHPAASPSLK	sp|Q01484|	ANK2	Kumari et al. (2022); Higham et al. (2019)
MQNDTAENETTEKEEK	sp|Q05682|	CALD1	Pereira et al. (2017)
LGVSFLVLPK	sp|Q16880|	CGT	Tang et al. (2023); Moll et al. (2020); Ryckman et al. (2020)
NFEEFFNLHR	sp|P20591|	MX1	Prakash et al. (2024); Widjaya et al. (2023); Ma et al. (2012)
DVTHTC[+57]PSC[+57]K	sp|Q9H305|	CDIP1	Dileep et al. (2023); Inukai et al. (2021)

TABLE 2
Peptides discovered by DR_UW and not by the Complete method. A q-value cutoff of 0.05 is used.

6. Discussion. In this paper, we present a statistical framework for analyzing proteomic data with missing values. Our proposed estimator $\hat{\boldsymbol{\beta}}_{UW}$ is established in a doubly robust framework and achieves reduced asymptotic variance leveraging correlations between different peptides. Through simulations, we demonstrate that the proposed estimator offers highly competitive statistical decisions in discovering signal peptides.

In particular, we show that the proposed method possesses improved properties compared to other imputation-based methods, such as the Plugin method and the DR_W method. However, a final choice between imputation-based methods and the Complete method should depend on the quality of imputation achievable. Simulations and real data considered in Sections 4 and 5 provide a rich foundation for high-quality imputation with the VAE model. These data have Gaussian-like distributions, sufficient sample size, and robust correlation structure between peptides. If these conditions are not sufficiently satisfied, even the DR_UW method will not perform as well as the Complete method. For example, a simulation experiment reveals that if the outcome model in Section 4 produces a completely noisy imputation, then the doubly robust method will yield fewer discoveries compared to the complete method (Figure C7). However, even in such cases, the estimate $\hat{\boldsymbol{\beta}}_{UW}$ obtained by the DR method is similar to the estimate obtained by the Complete method (Figure C8). This follows because the consistency of $\hat{\boldsymbol{\beta}}_{UW}$ is still guaranteed by the doubly robust property, provided the estimated propensity score is consistent.

The value of the proposed method depends on the quality of the imputation procedure. Although the approach is applicable regardless of the choice of imputation algorithm, we chose a refined VAE procedure that uses masking to robustly handle missing values as an integrated part of the procedure (Du et al., 2022). In the proteomic literature, one of the most commonly

applied imputation procedures is a version of k-nearest neighbors (kNN). The standard kNN procedure in use involves imputing the missing values based on the mean of the k closest peptides (Harris et al., 2023). Close peptides are used instead of close samples because it is very difficult to define close samples when there is excessive missingness. As a consequence, this kNN approach does not utilize low-dimensional covariates $\mathbf{W}_i$ in imputation and is potentially biased with fewer discoveries. An alternative approach is to adopt a two-step method: initially, we impute missing values based on the closest peptides and then, once the missing entries are filled, we impute the entire dataset based on the closest samples. When we apply this two-step approach to the AD dataset in Section 5, DR_W selects 65, DR_UW selects 51, and the Plugin method selects 121. Although we cannot reach an exact conclusion, the result of the VAE model in Section 5 is more aligned with theoretical expectations: DR_UW method provides more discoveries than the DR_W method.

Acknowledgments. We acknowledge Bernie Devlin for his help in contextualizing the findings in Section 5 within the scientific literature. The first author thanks Chan Park for assistance in identifying important reference papers.

Funding. This project was funded by National Institute of Mental Health (NIMH) grant R01MH123184 and NSF DMS-2015492.

REFERENCES

Agarwal, D., Wang, J., and Zhang, N. R. (2020), "Data Denoising and Post-Denoising Corrections in Single Cell RNA Sequencing," *Statistical Science*, 35, 112–128.

Andrews, T. S. and Hemberg, M. (2018), "False signals induced by single-cell imputation," *F1000Research*, 7.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, 57, 289–300.

Berrevoets, J., Imrie, F., Kyono, T., Jordon, J., and van der Schaar, M. (2023), "To impute or not to impute? missing data in treatment effect estimation," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 3568–3590.

Bhatia, S., Rawal, R., Sharma, P., Singh, T., Singh, M., and Singh, V. (2022), "Mitochondrial dysfunction in Alzheimer's disease: opportunities for drug development," *Current Neuropharmacology*, 20, 675.

Brini, A. and van den Heuvel, E. R. (2023), "Missing data imputation with high-dimensional data," *The American Statistician*, 1–13.

Chen, L. S., Wang, J., Wang, X., and Wang, P. (2017), "A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments," *The annals of applied statistics*, 11, 114.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning," *The Econometrics Journal*, 21.

Chion, M., Carapito, C., and Bertrand, F. (2022), "Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics," *PLoS Comput Biol*, 18, e1010420.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023), "A scale-free approach for false discovery rate control in generalized linear models," *Journal of the American Statistical Association*, 1–15.

Davidian, M. (2022), "Methods based on semiparametric theory for analysis in the presence of missing data," *Annual Review of Statistics and Its Application*, 9, 167–196.

Delhaye, S., Jarjat, M., Boulksibat, A., Sanchez, C., Tempio, A., Turtoi, A., Giorgi, M., Lacas-Gervais, S., Baj, G., Rovere, C., Trezza, V., Pellegrini, M., Maurin, T., Lalli, E., and Bardoni, B. (2024), "Defects in AMPAR trafficking and microglia activation underlie socio-cognitive deficits associated to decreased expression of phosphodiesterase 2 a," *Neurobiol Dis*, 191, 106393.

Díaz, I., Savenkov, O., and Ballman, K. (2018), "Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes," *Biometrika*, 105, 723–738.

Dileep, V., Boix, C. A., Mathys, H., Marco, A., Welch, G. M., Meharena, H. S., Loon, A., Jeloka, R., Peng, Z., Bennett, D. A., et al. (2023), "Neuronal DNA double-strand breaks lead to genome structural variations and 3D genome disruption in neurodegeneration," *Cell*, 186, 4404–4421.

Du, J.-H., Cai, Z., and Roeder, K. (2022), "Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT," *Proc Natl Acad Sci U S A*, 119, e2214414119.

Du, L., Guo, X., Sun, W., and Zou, C. (2023), "False discovery rate control under general dependence by symmetrized data aggregation," *Journal of the American Statistical Association*, 118, 607–621.

Fisher, A. and Fisher, V. (2023), "Three-way Cross-Fitting and Pseudo-Outcome Regression for Estimation of Conditional Effects and other Linear Functionals," *arXiv preprint arXiv:2306.07230*.

Gianetto, Q. G., Wieczorek, S., Couté, Y., and Burger, T. (2020), "A peptide-level multiple imputation strategy accounting for the different natures of missing values in proteomics data," *bioRxiv*, 2020–05.

Guo, X., Ren, H., Zou, C., and Li, R. (2023), "Threshold selection in feature screening for error rate control," *Journal of the American Statistical Association*, 118, 1773–1785.

Harris, L., Fondrie, W. E., Oh, S., and Noble, W. S. (2023), "Evaluating proteomics imputation methods with improved criteria," *Journal of Proteome Research*, 22, 3427–3438.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015), "Matrix completion and low-rank SVD via fast alternating least squares," *The Journal of Machine Learning Research*, 16, 3367–3402.

Higham, J. P., Malik, B. R., Buhl, E., Dawson, J. M., Ogier, A. S., Lunnon, K., and Hodge, J. J. L. (2019), "Alzheimer's Disease Associated Genes Ankyrin and Tau Cause Shortened Lifespan and Memory Loss in Drosophila," *Front Cell Neurosci*, 13, 260.

Inukai, R., Mori, K., Kuwata, K., Suzuki, C., Maki, M., Takahara, T., and Shibata, H. (2021), "The Novel ALG-2 Target Protein CDIP1 Promotes Cell Death by Interacting with ESCRT-I and VAPA/B," *Int J Mol Sci*, 22.

Jiang, K., Mukherjee, R., Sen, S., and Sur, P. (2022), "A New Central Limit Theorem for the Augmented IPW Estimator: Variance Inflation, Cross-Fit Covariance and Beyond," *arXiv preprint arXiv:2205.10198*.

Jones, I., Dent, L., Higo, T., Roumeliotis, T., Arias Garcia, M., Shree, H., Choudhary, J., Pedersen, M., and Bakal, C. (2023), "Characterization of proteome-size scaling by integrative omics reveals mechanisms of proliferation control in cancer," *Science Advances*, 9, eadd0636.

Kennedy, E. H. (2023), "Towards optimal doubly robust estimation of heterogeneous causal effects," *Electronic Journal of Statistics*, 17, 3008–3049.

Kumari, A., Rahaman, A., Zeng, X.-A., Farooq, M. A., Huang, Y., Yao, R., Ali, M., Ishrat, R., and Ali, R. (2022), "Temporal Cortex Microarray Analysis Revealed Impaired Ribosomal Biogenesis and Hyperactivity of the Glutamatergic System: An Early Signature of Asymptomatic Alzheimer's Disease," *Frontiers in Neuroscience*, 16, 966877.

Lanz, M. C., Zatulovskiy, E., Swaffer, M. P., Zhang, L., Ilerten, I., Zhang, S., You, D. S., Marinov, G., McAlpine, P., Elias, J. E., et al. (2022), "Increasing cell size remodels the proteome and promotes senescence," *Molecular cell*, 82, 3255–3269.

Leduc, A., Huffman, R. G., Cantlon, J., Khan, S., and Slavov, N. (2022), "Exploring functional protein covariation across single cells using nPOP," *Genome Biol*, 23, 261.

Lengefeld, J., Cheng, C.-W., Maretich, P., Blair, M., Hagen, H., McReynolds, M. R., Sullivan, E., Majors, K., Roberts, C., Kang, J. H., et al. (2021), "Cell size is a determinant of stem cell potential during aging," *Science Advances*, 7, eabk0271.

Little, R. J. (1992), "Regression with missing X's: a review," *Journal of the American statistical association*, 87, 1227–1237.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012), "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, 367, 1355–1360.

Liu, M. and Dongre, A. (2021), "Proper imputation of missing values in proteomics datasets for differential expression analysis," *Briefings in Bioinformatics*, 22, bbaa112.

Liu, Y., Beyer, A., and Aebersold, R. (2016), "On the Dependency of Cellular Protein Levels on mRNA Abundance," *Cell*, 165, 535–50.

Ly, L.-H. and Vingron, M. (2022), "Effect of imputation on gene network reconstruction from single-cell RNA-seq data," *Patterns*, 3.

Ma, S. L., Huang, W., Tang, N. L., and Lam, L. C. (2012), "MxA polymorphisms are associated with risk and age-at-onset in Alzheimer disease and accelerated cognitive decline in Chinese elders," *Rejuvenation Research*, 15, 516–522.

MacDonald, M. L., Alhassan, J., Newman, J. T., Richard, M., Gu, H., Kelly, R. M., Sampson, A. R., Fish, K. N., Penzes, P., Wills, Z. P., Lewis, D. A., and Sweet, R. A. (2017), "Selective Loss of Smaller Spines in Schizophrenia," *Am J Psychiatry*, 174, 586–594.

Meng, X.-L. (1994), "Multiple-imputation inferences with uncongenial sources of input," *Statistical science*, 538–558.

Merrihew, G. E., Park, J., Plubell, D., Searle, B. C., Keene, C. D., Larson, E. B., Bateman, R., Perrin, R. J., Chhatwal, J. P., Farlow, M. R., McLean, C. A., Ghetti, B., Newell, K. L., Frosch, M. P., Montine, T. J., and

MacCoss, M. J. (2023), "A peptide-centric quantitative proteomics dataset for the phenotypic assessment of Alzheimer's disease," *Sci Data*, 10, 206.

Moll, T., Shaw, P. J., and Cooper-Knock, J. (2020), "Disrupted glycosylation of lipids and proteins is a cause of neurodegeneration," *Brain*, 143, 1332–1340.

Pereira, A. C., Gray, J. D., Kogan, J. F., Davidson, R. L., Rubin, T. G., Okamoto, M., Morrison, J. H., and McEwen, B. S. (2017), "Age and Alzheimer's disease gene expression profiles reversed by the glutamate modulator riluzole," *Molecular psychiatry*, 22, 296–305.

Prakash, P., Erdjument-Bromage, H., O'Dea, M. R., Munson, C. N., Labib, D., Fossati, V., Neubert, T. A., and Liddelow, S. A. (2024), "Proteomic profiling of interferon-responsive reactive astrocytes in rodent and human," *Glia*, 72, 625–642.

Qiu, Y. and Messer, K. (2023), "An Efficient Doubly-robust Imputation Framework for Longitudinal Dropout, with an Application to an Alzheimer's Clinical Trial," *arXiv preprint arXiv:2305.02849*.

Qiu, Y. L., Zheng, H., and Gevaert, O. (2020), "Genomic data imputation with variational auto-encoders," *GigaScience*, 9, giaa082.

Robins, J. M. and Rotnitzky, A. (1995), "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90, 122–129.

Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys*, John Wiley & Sons.

Ryckman, A. E., Brockhausen, I., and Walia, J. S. (2020), "Metabolism of glycosphingolipids and their role in the pathophysiology of lysosomal storage disorders," *International Journal of Molecular Sciences*, 21, 6881.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association*, 94, 1096–1120.

Semenova, V. and Chernozhukov, V. (2021), "Debiased machine learning of conditional average treatment effects and other causal functions," *The Econometrics Journal*, 24, 264–289.

Shen, M., Chang, Y.-T., Wu, C.-T., Parker, S. J., Saylor, G., Wang, Y., Yu, G., Van Eyk, J. E., Clarke, R., Herrington, D. M., et al. (2022), "Comparative assessment and novel strategy on methods for imputing proteomics data," *Scientific reports*, 12, 1067.

Sheng, J., Zhang, S., Wu, L., Kumar, G., Liao, Y., Gk, P., and Fan, H. (2022), "Inhibition of phosphodiesterase: A novel therapeutic target for the treatment of mild cognitive impairment and Alzheimer's disease," *Frontiers in Aging Neuroscience*, 14, 1019187.

Shi, J., Li, Y., Zhang, Y., Chen, J., Gao, J., Zhang, T., Shang, X., and Zhang, X. (2021), "Baicalein Ameliorates Aβ-Induced Memory Deficits and Neuronal Atrophy via Inhibition of PDE2 and PDE4," *Front Pharmacol*, 12, 794458.

Stuart, T. and Satija, R. (2019), "Integrative single-cell analysis," *Nature reviews genetics*, 20, 257–272.

Tang, X., Tena, J., Di Lucente, J., Maezawa, I., Harvey, D. J., Jin, L.-W., Lebrilla, C. B., and Zivkovic, A. M. (2023), "Transcriptomic and glycomic analyses highlight pathway-specific glycosylation alterations unique to Alzheimer's disease," *Sci Rep*, 13, 7816.

Tasaki, S., Xu, J., Avey, D. R., Johnson, L., Petyuk, V. A., Dawe, R. J., Bennett, D. A., Wang, Y., and Gaiteri, C. (2022), "Inferring protein expression changes from mRNA in Alzheimer's dementia using deep neural networks," *Nat Commun*, 13, 655.

Välikangas, T., Suomi, T., and Elo, L. L. (2018), "A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation," *Briefings in bioinformatics*, 19, 1344–1355.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, vol. 3, Cambridge university press.

Vanderaa, C. and Gatto, L. (2023), "Revisiting the thorny issue of missing values in single-cell proteomics," *Journal of Proteome Research*, 22, 2775–2784.

Vogel, C. and Marcotte, E. M. (2012), "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nat Rev Genet*, 13, 227–32.

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., Cox, N. J., Nicolae, D. L., and Chen, L. S. (2016), "Imputing gene expression in uncollected tissues within and beyond GTEx," *The American Journal of Human Genetics*, 98, 697–708.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., and Ni, Y. (2018), "Missing value imputation approach for mass spectrometry-based metabolomics data," *Scientific reports*, 8, 663.

Wesseling, H., Xu, B., Want, E. J., Holmes, E., Guest, P., Karayiorgou, M., Gogos, J., and Bahn, S. (2017), "System-based proteomic and metabonomic analysis of the Df (16) A+/- mouse identifies potential miR-185 targets and molecular pathway alterations," *Molecular psychiatry*, 22, 384–395.

Widjaya, M. A., Cheng, Y.-J., Kuo, Y.-M., Liu, C.-H., Cheng, W.-C., and Lee, S.-D. (2023), "Transcriptomic Analyses of Exercise Training in Alzheimer's Disease Cerebral Cortex," *J Alzheimers Dis*, 93, 349–363.

Yadlowsky, S. (2022), "Explaining Practical Differences Between Treatment Effect Estimators with High Dimensional Asymptotics," *arXiv preprint arXiv:2203.12538*.

Yin, X., Levy, D., Willinger, C., Adourian, A., and Larson, M. G. (2016), "Multiple imputation and analysis for high-dimensional incomplete proteomics data," *Statistics in medicine*, 35, 1315–1326.

Yoon, J., Jordon, J., and Schaar, M. (2018), "Gain: Missing data imputation using generative adversarial nets," in *International conference on machine learning*, PMLR, pp. 5689–5698.

Zhao, A. and Ding, P. (2022), "To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates," *Journal of the American Statistical Association*, 1–11.

APPENDIX A: PROOF

A.1. Proof of Theorem 2.2.

PROOF. By plugging in $\hat{Y}_i^{UW} = \frac{C_i}{\hat{\delta}_i} Y_i + \left(1 - \frac{C_i}{\hat{\delta}_i}\right) \hat{\nu}_i$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T\right)^{-1} (\sum_{i=1}^n \mathbf{W}_i (\hat{Y}_i^{UW} - \mathbf{W}_i^T \boldsymbol{\beta})) \\
&= (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T)^{-1} (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i (\epsilon_i + (1 - \frac{C_i}{\hat{\delta}_i})(\hat{\nu}_i - \nu_i + Y_i - \nu_i)) \\
&= (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T)^{-1} (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i (\epsilon_i + (1 - \frac{C_i}{\hat{\delta}_i})(Y_i - \nu_i) \\
&\qquad + (1 - \frac{C_i}{\delta_i})(\hat{\nu}_i - \nu_i) + C_i(\frac{1}{\delta_i} - \frac{1}{\hat{\delta}_i})(\hat{\nu}_i - \nu_i)) \qquad (6)
\end{aligned}$$

By assumption, $\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T$ is nonsingular with probability approaching one, so its inverse exists. Let $D = \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T - \mathbb{E}[\mathbf{W}\mathbf{W}^T]$ which is $o_{\mathbb{P}}(1)$ by the law of large number. Then,

$$\begin{aligned}
(\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T)^{-1} &= (\mathbb{E}[\mathbf{W}\mathbf{W}^T] + D)^{-1} \\
&= \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1}(I + \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1} D)^{-1} \\
&= \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1}(I - \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1} D(I + \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1} D)^{-1}) \\
&= \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1} + o_{\mathbb{P}}(1) \qquad (7)
\end{aligned}$$

For the third step of Equation 7, we used $(M_1 + M_2)^{-1} = M_1^{-1} - M_1^{-1} M_2 (M_1 + M_2)^{-1}$ proved in Henderson and Searle (1981). Therefore

$$(\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T)^{-1} \xrightarrow{p} \mathbb{E}[\mathbf{W}\mathbf{W}^T]^{-1} \qquad (8)$$

Also, we have

$$\begin{aligned}
&\mathbb{E}(\mathbf{W}_i(\epsilon_i + (1 - \frac{C_i}{\hat{\delta}_i})(Y_i - \nu_i) + (1 - \frac{C_i}{\delta_i})(\hat{\nu}_i - \nu_i))) \\
&= \mathbb{E}(\mathbf{W}_i \mathbb{E}(\epsilon_i + (1 - \frac{C_i}{\hat{\delta}_i})(Y_i - \nu_i) + (1 - \frac{C_i}{\delta_i})(\hat{\nu}_i - \nu_i) | \mathbf{W}_i, \mathbf{U}_i)) \\
&= \mathbb{E}(\mathbf{W}_i(\mathbb{E}(1 - \frac{C_i}{\hat{\delta}_i} | \mathbf{W}_i, \mathbf{U}_i)\mathbb{E}(Y_i - \nu_i | \mathbf{W}_i, \mathbf{U}_i) + \mathbb{E}(1 - \frac{C_i}{\delta_i} | \mathbf{W}_i, \mathbf{U}_i)\mathbb{E}(\hat{v}_i - \nu_i | \mathbf{W}_i, \mathbf{U}_i))) \\
&= 0,
\end{aligned}$$

where we used $\mathbb{E}(\epsilon_i | \mathbf{W}_i, \mathbf{U}_i)$, $\mathbb{E}(Y_i - \nu_i | \mathbf{W}_i, \mathbf{U}_i) = 0$ and $\mathbb{E}(1 - \frac{C_i}{\delta_i} | \mathbf{W}_i, \mathbf{U}_i)$ in a last step.

Then, by a weak law of large number, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i(\epsilon_i + (1 - \frac{C_i}{\hat{\delta}_i})(Y_i - \nu_i) + (1 - \frac{C_i}{\delta_i})(\hat{\nu}_i - \nu_i))$$

$$\xrightarrow{P} \mathbb{E}(\mathbf{W}_i(\epsilon_i + (1 - \frac{C_i}{\hat{\delta}_i})(Y_i - \nu_i) + (1 - \frac{C_i}{\delta_i})(\hat{\nu}_i - \nu_i)))$$

$$= 0 \tag{9}$$

Also, for each W_{ij} which denotes jth entry of a vector $\mathbf{W}_i$, below inequality is obtained by using a Cauchy-Schwarz inequality

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_{ij}C_i(\frac{1}{\delta_i} - \frac{1}{\hat{\delta}_i})(\hat{\nu}_i - \nu_i) &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} W_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{\delta} - \frac{1}{\hat{\delta}_i})^2(\hat{\nu}_i - \nu_i)^2} \\ &\leq \sqrt{\|W_{ij}\|_\infty^2}\sqrt{\|(\frac{1}{\delta} - \frac{1}{\hat{\delta}_i})(\hat{\nu}_i - \nu_i)\|_2^2} \\ &= o_{\mathbb{P}}(1) \end{aligned} \tag{10}$$

Plugging in Equation 8, 9 and 10 to Equation 6, we have a desired result. □

A.2. Proof of Theorem 2.3.

PROOF. **Part 1: limiting distribution of $\boldsymbol{\beta}$**
By plugging in the formula for $\hat{\boldsymbol{\beta}}$, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^T\right)^{-1}\sum_{i=1}^{n}\mathbf{W}_i(Y_i + \left(\frac{C_i}{\hat{\delta}_i} - 1\right)(Y_i - \hat{\nu}_i)) - \boldsymbol{\beta} \\ &= \left(\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^T\right)^{-1}\sum_{i=1}^{n}\mathbf{W}_i(\epsilon_i + (\frac{C_i}{\hat{\delta}_i} - 1)(Y_i - \hat{\nu}_i))\end{aligned}$$

Then $\sqrt{n}$-scaled difference is expressed as

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^T)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i(\epsilon_i + (\frac{C_i}{\hat{\delta}_i} - 1)(y_i - \hat{\nu}_i)) \tag{11}$$

As proved in Equation 8, we have

$$(\frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^T)^{-1} \xrightarrow{P} \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} \tag{12}$$

On the other hand, we can decompose

$$\begin{aligned}&(\frac{C_i}{\hat{\delta}_i} - 1)(Y_i - \hat{\nu}_i) - (\frac{C_i}{\delta_i} - 1)(Y_i - \nu_i) \\ &\quad= (\frac{C_i}{\hat{\delta}_i} - 1)(Y_i - \hat{\nu}_i) - (\frac{C_i}{\hat{\delta}_i} - 1)(Y_i - \nu_i) + (\frac{C_i}{\hat{\delta}_i} - 1)(Y_i - \nu_i) - (\frac{C_i}{\delta_i} - 1)(Y_i - \nu_i) \\ &\quad= (\frac{C_i}{\hat{\delta}_i} - 1)(\nu_i - \hat{\nu}_i) + C_i(\frac{1}{\hat{\delta}_i} - \frac{1}{\delta_i})(Y_i - \nu_i) \\ &\quad= (\frac{C_i}{\hat{\delta}_i} - \frac{C_i}{\delta_i})(\nu_i - \hat{\nu}_i) + (\frac{C_i}{\delta_i} - 1)(\nu_i - \hat{\nu}_i) + C_i(\frac{1}{\hat{\delta}_i} - \frac{1}{\delta_i})(Y_i - \nu_i)\end{aligned}$$

which leads to the equation below

(13)
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i(\frac{C_i}{\hat{\delta}_i}-1)(Y_i-\hat{\nu}_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i(\frac{C_i}{\hat{\delta}_i}-\frac{C_i}{\delta_i})(\nu_i-\hat{\nu}_i)$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i(\frac{C_i}{\delta_i}-1)(\nu_i-\hat{\nu}_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_iC_i(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta_i})(Y_i-\nu_i)$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i)$$

Let W_{ij} denote a jth component of a vector $\mathbf{W}_i$. Then,

(14)
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}(\frac{C_i}{\hat{\delta}_i}-1)(Y_i-\hat{\nu}_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}(\frac{C_i}{\hat{\delta}_i}-\frac{C_i}{\delta_i})(\nu_i-\hat{\nu}_i)$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}(\frac{C_i}{\delta_i}-1)(\nu_i-\hat{\nu}_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}C_i(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta_i})(Y_i-\nu_i)$$
$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i)$$

The limit properties of each component in equation (14) are as follows.

For a first term, by Assumption 2.1 and a Cauchy-Schwarz inequality, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}(\frac{C_i}{\hat{\delta}_i}-\frac{C_i}{\delta_i})(\nu_i-\hat{\nu}_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{ij}\frac{C_i}{\delta_i}(\frac{\delta_i}{\hat{\delta}_i}-1)(\nu_i-\hat{\nu}_i)$$
$$\leq\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}W_{ij}^2}\sqrt{\sum_{i=1}^{n}\frac{C_i^2}{\delta_i^2}(\frac{\delta_i}{\hat{\delta}_i}-1)^2(\nu_i-\hat{\nu}_i)^2}$$
$$=\sqrt{\frac{1}{n}\sum_{i=1}^{n}W_{ij}^2}\sqrt{n\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{C_i^2}{\delta_i^2}(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta})^2(\nu_i-\hat{\nu}_i)^2}$$
$$\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}W_{ij}^2}\sqrt{n\cdot\|\frac{C_i}{\delta_i}\|_\infty^2\cdot\|(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta})(\nu_i-\hat{\nu}_i)\|_2^2}$$
$$=o_{\mathbb{P}}(1) \tag{15}$$

for $j=1,\ldots,q$. We used $\mathbb{E}[\|\mathbf{W}_i\|_2^2]=o_{\mathbb{P}}(1)$, $\|\frac{C_i}{\delta_i}\|_\infty=o_{\mathbb{P}}(1)$, and $\|(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta})(\nu_i-\hat{\nu}_i)\|_2=o_{\mathbb{P}}(n^{-1/2})$

For a second term of Equation 14, since $W_{ij}(\frac{C_i}{\delta_i}-1)(\nu_i-\hat{\nu}_i)$ are i.i.d, we use CLT.

$$\mathbb{E}[W_{ij}(\frac{C_i}{\delta_i}-1)(\nu_i-\hat{\nu}_i)]=\mathbb{E}[W_{ij}(\nu_i-\hat{\nu}_i)\mathbb{E}[\frac{C_i}{\delta_i}-1\mid\mathbf{W}_i,\mathbf{U}_i]]$$
$$=0$$

Therefore, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_{ij}(\frac{C_i}{\delta_i}-1)(\nu_i-\hat{\nu}_i) &= o_{\mathbb{P}}(\|W_{ij}(\frac{C_i}{\delta_i}-1)(\nu_i-\hat{\nu}_i)\|_2)\\ &\leq o_{\mathbb{P}}(\|W_{ij}\|_\infty\|\frac{C_i}{\delta_i}-1\|_\infty\|\nu_i-\hat{\nu}_i\|_2)\\ &= o_{\mathbb{P}}(1) \qquad (16)\end{aligned}$$

Lastly, for the third term of the Equation 14,

$$\begin{aligned}\mathbb{E}[W_{ij}C_i(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta_i})(Y_i-\nu_i)] &= \mathbb{E}[W_{ij}(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta_i})\mathbb{E}[Y_i-\nu_i \mid \mathbf{W}_i,\mathbf{U}_i]\mathbb{E}[C_i \mid \mathbf{W}_i,\mathbf{U}_i]]\\ &= 0\end{aligned}$$

by Assumption 2.1 (a). Therefore,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_{ij}C_i(\frac{1}{\hat{\delta}_i}-\frac{1}{\delta_i})(Y_i-\nu_i) &= o_{\mathbb{P}}(\|W_{ij}\frac{C_i}{\delta_i}(\frac{\delta_i}{\hat{\delta}_i}-1)(Y_i-\nu_i)\|_2)\\ &\leq o_{\mathbb{P}}(\|W_{ij}\|_\infty\|\frac{\delta_i}{\hat{\delta}_i}-1\|_2\|Y_i-\nu_i\|_\infty) = o_{\mathbb{P}}(1). \qquad (17)\end{aligned}$$

Putting Eqations 15, 16, 17 together, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_{ij}(\epsilon_i+(\frac{C_i}{\hat{\delta}_i}-1)(Y_i-\hat{\nu}_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_{ij}(\epsilon_i+(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i))+o_{\mathbb{P}}(1).$$

which naturally leads to

(18)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{W}_i(\epsilon_i+(\frac{C_i}{\hat{\delta}_i}-1)(Y_i-\hat{\nu}_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{W}_i(\epsilon_i+(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i))+o_{\mathbb{P}}(1).$$

The remaining task is to get a limiting distribution of Equation 18. Since $\mathbf{W}_i(\epsilon_i+(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i))$ are i.i.d variable with mean zero, its variance is

$$\begin{aligned}&\mathbb{E}[(\epsilon_i^2+(\frac{C_i}{\delta_i}-1)^2(Y_i-\nu_i)^2+2\epsilon_i(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i))\mathbf{W}_i^T\mathbf{W}_i]\\ =&\mathbb{E}[\epsilon_i^2\mathbf{W}_i^T\mathbf{W}_i]+\mathbb{E}[\mathbb{E}[(\frac{C_i}{\delta_i}-1)^2(Y_i-\nu_i)^2+2\epsilon_i(\frac{C_i}{\delta_i}-1)(Y_i-\nu_i)) \mid \mathbf{W}_i,\mathbf{U}_i]\mathbf{W}_i^T\mathbf{W}_i]\\ =&\mathbb{E}[\epsilon_i^2\mathbf{W}_i^T\mathbf{W}_i]+\mathbb{E}[\mathbb{E}[(\frac{C_i}{\delta_i}-1)^2 \mid \mathbf{W}_i,\mathbf{U}_i]\mathbb{E}[(Y_i-\nu_i)^2 \mid \mathbf{W}_i,\mathbf{U}_i]\mathbf{W}_i^T\mathbf{W}_i]\\ &+2\mathbb{E}[\mathbb{E}[\epsilon_i(Y_i-\nu_i) \mid \mathbf{W}_i,\mathbf{U}_i]\mathbb{E}[\frac{C_i}{\delta_i}-1 \mid \mathbf{W}_i,\mathbf{U}_i]\mathbf{W}_i^T\mathbf{W}_i]\end{aligned}$$

(19)

$$=\mathbb{E}[\epsilon_i^2\mathbf{W}_i^T\mathbf{W}_i]+\mathbb{E}[(\frac{1}{\delta_i}-1)(Y_i-\nu_i)^2\mathbf{W}_i^T\mathbf{W}_i]$$

Therefore, by CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i\left(\epsilon_i+(\frac{C_i}{\delta_i}-1)(Y_i-\hat{\nu}_i)\right) \xrightarrow{D} \mathcal{N}(0,\mathbb{E}[(\epsilon_i^2+(\frac{1}{\delta}-1)(Y_i-\nu_i)^2)\mathbf{W}_i^T\mathbf{W}_i]) \qquad (20)$$

holds. Then we plug in the result of Equation 12, 20 to Equation 11 and apply Slutsky Theorem to get the desired result.

Part 2: Variance consistency

Let us denote $\epsilon_i^* = \epsilon_i + (\frac{C_i}{\delta_i} - 1)(Y_i - v_i)$ and $\hat{\epsilon}_i^* = \hat{Y}_i^{UW} - \mathbf{W}_i^T \hat{\boldsymbol{\beta}}$. We claim that $\|\Sigma - \hat{\Sigma}\|_2 = o_{\mathbb{P}}(1)$, where

$$\hat{\Sigma} = (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}\frac{1}{n}\sum_{i=1}^n(\hat{\epsilon}_i^{*2}\mathbf{W}_i\mathbf{W}_i^T)(\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}$$
$$\Sigma = \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}\mathbb{E}[\epsilon_i^{*2}\mathbf{W}_i\mathbf{W}_i^T]\mathbb{E}[\mathbf{W}_i\mathbf{W}_i]^{-1}$$

Note that the above expression for Σ use the fact that $\mathbb{E}[\epsilon_i^{*2}\mathbf{W}_i\mathbf{W}_i^T] = \mathbb{E}[(\epsilon_i^2 + (1 - \frac{1}{\delta_i})(Y_i - v_i)^2)\mathbf{W}_i\mathbf{W}_i^T]$ whose derivation is in Equation 19.

Then,

$$\hat{\Sigma}_{UW} - \Sigma_{UW} = (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}\{\frac{1}{n}\sum_{i=1}^n(\hat{\epsilon}_i^{*2}\mathbf{W}_i\mathbf{W}_i^T) - \mathbb{E}[\epsilon_i^{*2}\mathbf{W}_i\mathbf{W}_i^T]\}(\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}$$
$$- \{\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} - (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}\}\mathbb{E}[\epsilon_i^{*2}\mathbf{W}_i\mathbf{W}_i^T]\{\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} - (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}\} \tag{21}$$

First we show that $\mathbb{E}[\epsilon_i^{*2}W_{ij}W_{ik}]$ is bounded for $j, k \in \{1, ..., q\}$;

$$\begin{aligned}\mathbb{E}[\epsilon_i^{*2}W_{ij}W_{ik}] &= \mathbb{E}[(\epsilon_i + (\frac{C_i}{\delta_i} - 1)(Y_i - v_i))^2 W_{ij}W_{ik}] \\ &= \mathbb{E}\left[(\epsilon_i^2 + (\frac{C_i}{\delta_i} - 1)^2(Y_i - v_i)^2 + 2\epsilon_i(\frac{C_i}{\delta_i} - 1)(Y_i - v_i)W_{ij}W_{ik}\right] \\ &= \mathbb{E}[\epsilon_i^2 W_{ij}W_{ik}] + \mathbb{E}[(\frac{1}{\delta_i} - 1)\mathbb{E}[(Y_i - v_i)^2 \mid \mathbf{W}_i, \mathbf{U}_i]W_{ij}W_{ik}]\end{aligned} \tag{22}$$

Each term in expectation is bounded by Assumption 2.1. Combining this to the fact that $\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} - (\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1} = o_{\mathbb{P}}(1)$, second term of the Equation 21 is $o_{\mathbb{P}}(1)$. Also, $(\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\mathbf{W}_i^T)^{-1}$ is bounded by a full-rank assumption of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]$. Therefore, it is suffice to show that $\frac{1}{n}\sum_{i=1}^n((\hat{\epsilon}_i^{*2}\mathbf{W}_i\mathbf{W}_i^T) - \mathbb{E}[\epsilon_i^{*2}\mathbf{W}_i\mathbf{W}_i^T]) = o_{\mathbb{P}}(1)$.

To this end, we show that every component of them are $o_{\mathbb{P}}(1)$. That is, for $j, k = 1, .., q$, we claim that

$$\left|\frac{1}{n}\sum_{i=1}^n(\hat{\epsilon}_i^{*2}W_{ij}W_{ik}) - \mathbb{E}[\epsilon_i^{*2}W_{ij}W_{ik}]\right| = o_{\mathbb{P}}(1). \tag{23}$$

A left-hand side of Equation 23 can be bounded by

$$\begin{aligned}&|\frac{1}{n}\sum_{i=1}^n(\hat{\epsilon}_i^{*2}W_{ij}W_{ik}) - \mathbb{E}[\epsilon_i^{*2}W_{ij}W_{ik}]| \\ &\leq |\frac{1}{n}\sum_{i=1}^n(\hat{\epsilon}_i^{*2}W_{ij}W_{ik} - \epsilon_i^{*2}W_{ij}W_{ik})| + |\frac{1}{n}\sum_{i=1}^n(\epsilon_i^{*2}W_{ij}W_{ik}) - \mathbb{E}[\epsilon_i^{*2}W_{ij}W_{ik}]|\end{aligned} \tag{24}$$

Second term of Equation 24 is $o_{\mathbb{P}}(1)$ by a law of large number. Therefore, it is suffice to show that the first term is $o_{\mathbb{P}}(1)$.

The first term can be bounded by

$$
\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^{*2}W_{ij}W_{ik}-\epsilon_i^{*2}W_{ij}W_{ik})| \\
&\qquad = |\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^{*}-\epsilon_i^{*})(\hat{\epsilon}_i^{*}+\epsilon_i^{*})W_{ij}W_{ik})| \\
&\qquad = |\frac{1}{n}\sum_{i=1}^{n}\{(\hat{\epsilon}_i^{*}-\epsilon_i^{*})W_{ij}\}\{(\hat{\epsilon}_i^{*}-\epsilon_i^{*}+2\epsilon_i^{*})W_{ik}\}| \\
&\qquad \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{(\hat{\epsilon}_i^{*}-\epsilon_i^{*})W_{ij}\}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{(\hat{\epsilon}_i^{*}-\epsilon_i^{*}+2\epsilon_i^{*})W_{ik}\}^2} \qquad (25)
\end{aligned}
$$

where the last step uses a Cauchy-Shwarz inequality.

Using a decomposition

$$
\begin{aligned}
\hat{\epsilon}_i^{*}-\epsilon_i^{*} &= \hat{Y}_i^{UW}-\mathbf{W}_i^T\hat{\boldsymbol{\beta}}-\epsilon_i^{*} \\
&= Y_i+(\frac{C_i}{\hat{\delta}_i}-1)(Y_i-\hat{\nu}_i)-\mathbf{W}_i^T\hat{\boldsymbol{\beta}}-\epsilon_i-(\frac{C_i}{\delta_i}-1)(Y_i-v_i) \\
&= \mathbf{W}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+(\frac{C_i}{\hat{\delta}_i}-1)(Y_i-\hat{\nu}_i)-(\frac{C_i}{\delta_i}-1)(Y_i-v_i) \\
&= \mathbf{W}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+\left(1-\frac{\delta_i}{\hat{\delta}_i}\right)(\hat{\nu}_i-\nu_i),
\end{aligned}
$$

and a Cauchy-Shwarz inequality, we have

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^{*}-\epsilon_i^{*})^2 &\leq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{W}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+\left(1-\frac{\delta_i}{\hat{\delta}_i}\right)(\hat{\nu}_i-\nu_i))^2 \\
&\leq \|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{W}_i^T\|_2^2+\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{\hat{\delta}_i}\right)^2(\hat{\nu}_i-\nu_i)^2 \\
&\quad +2\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{W}_i^T\|_2\left(1-\frac{\delta_i}{\hat{\delta}_i}\right)(\hat{\nu}_i-\nu_i) \\
&= o_{\mathbb{P}}(1)\,, \qquad (26)
\end{aligned}
$$

where the last step uses the fact that $\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2=o_{\mathbb{P}}(1)$, $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{W}_i\|_2^2\leq p\|\mathbf{W}_i\|_\infty^2$, which is bounded by a constant, and that $(1-\delta_i/\hat{\delta}_i)^2(\hat{\nu}_i-\nu_i)^2$ has vanishing L_1 norm for each i according to Assumption 2.1(e).

Then, since $\|\mathbf{W}_i\|_\infty$ bounded,

$$
\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{(\hat{\epsilon}_i^{*}-\epsilon_i^{*})W_{ij}\}^2}=o_{\mathbb{P}}(1). \qquad (27)
$$

Also, applying $(\hat{\epsilon}_i^{*}-\epsilon_i^{*})^2=o_{\mathbb{P}}(1)$, we have

$$
\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{(\hat{\epsilon}_i^{*}+\epsilon_i^{*})W_{ik}\}^2}\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{2(\hat{\epsilon}_i^{*}-\epsilon_i^{*})^2W_{ik}^2+2(\epsilon_i^{*})^2W_{ik}^2\}}
$$

$$= \mathcal{O}_{\mathbb{P}}(1) \tag{28}$$

because $\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon_i^*} - \epsilon_i^*)^2 W_{ik}^2 = o_{\mathbb{P}}(1)$ as Equation 27 and $\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^*)^2 W_{ik}^2 = \mathcal{O}_{\mathbb{P}}(1)$ by Assumption 2.1(d). Plugging in the Equation 27, 28 to Equation 25 completes the proof.

□

A.3. Proof of Theorem 2.4 .

PROOF.

$$\begin{aligned}
\Sigma_U - \Sigma_{UW} &= \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}\mathbb{E}[(\epsilon_i^2 + (\frac{1}{\delta_i} - 1)(Y_i - \mu_i)^2)\mathbf{W}_i\mathbf{W}_i^T]\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} \\
&\quad - \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}\mathbb{E}[(\epsilon_i^2 + (\frac{1}{\delta_i} - 1)(Y_i - \nu_i)^2)\mathbf{W}_i\mathbf{W}_i^T]\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} \\
&= \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}\mathbb{E}[(\frac{1}{\delta_i} - 1)((Y_i - \mu_i)^2 - (Y_i - \nu_i)^2)\mathbf{W}_i\mathbf{W}_i^T]\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}[(Y_i - \nu_i)(\nu_i - \mu_i)|\mathbf{W}_i] &= \mathbb{E}_{\mathbf{U}_i}[\mathbb{E}[(Y_i - \nu_i)(\nu_i - \mu_i)|\mathbf{W}_i, \mathbf{U}_i]] \\
&= \mathbb{E}_{\mathbf{U}_i}[(\nu_i - \mu_i)\mathbb{E}[(Y_i - \nu_i)|\mathbf{W}_i, \mathbf{U}_i]] \\
&= 0,
\end{aligned}$$

by Assumption 2.1(c), we have

$$\begin{aligned}
\mathbb{E}[(Y_i - \mu_i)^2 - (Y_i - \nu_i)^2|\mathbf{W}_i] &= \mathbb{E}[\{(Y_i - \nu_i) + (\nu_i - \mu_i)\}^2 - (Y_i - \nu_i)^2|\mathbf{W}_i] \\
&= \mathbb{E}[(\nu_i - \mu_i)^2|\mathbf{W}_i]
\end{aligned}$$

Combining with $\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T] \succcurlyeq 0$ and $(\frac{1}{\delta_i} - 1)(\nu_i - \mu_i)^2 \geq 0$, we have

$$\begin{aligned}
\Sigma_U - \Sigma_{UW} &= \mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1}\mathbb{E}[(\frac{1}{\delta_i} - 1)(\nu_i - \mu_i)^2\mathbf{W}_i\mathbf{W}_i^T]\mathbb{E}[\mathbf{W}_i\mathbf{W}_i^T]^{-1} \\
&\succcurlyeq 0
\end{aligned}$$

□

APPENDIX B: IMPUTATION METHODS

B.1. Probabilistic modeling of peptide datasets. The semiparametric inference results established in the main paper allow us to use more flexible non-parametric machine learning and deep learning models to estimate the mean regression nuisance function and improve the imputation quality. Inspired by recent advancements in conditional variational inference (Kingma and Welling, 2014; Sohn et al., 2015; Ivanov et al., 2018) in the machine learning community, Du et al. (2022) propose a variational autoencoder (VAE) model for imputation of single-cell multi-omics data by utilizing a masking procedure to inform the missing patterns and help the model to learn conditional distributions among features, which we refer to VAEIT in the current section.

Specifically, VAEIT models the missing features as a conditional probability estimation problem. For each individual, we denote its measurements of p peptides by a random vector $\mathbf{Y} = (Y_1, \ldots, Y_p) \in \mathbb{R}^p$. We introduce a binary mask $\mathbf{M} \in \{0,1\}^p$ for $\mathbf{Y}$ and its bitwise complement $\mathbf{M}^c$, such that the jth entry of the observed sample $\mathbf{Y}_{M^c}$ is Y_j if $M_j = 1$ and 0 otherwise. We use an authentic missing pattern $\mathbf{M}_a = \mathbf{1}_p - \mathbf{C}$ to represent which components of $\mathbf{Y}$ are actually missing, while the distribution of $\mathbf{M}$ can be arbitrary during training. For example, if we want to model missing completely at random, the entries of $\mathbf{M}$ could be independent Bernoulli random variables. Furthermore, we can incorporate extra structural information to model the situation of missing modality. To model the conditional distribution of the missing peptides given the observed values, we consider the following maximum likelihood problem:

$$\max_{\theta} \mathbb{E}_{\mathbf{Y},\mathbf{M}} \log p_\theta(\mathbf{Y}_\mathbf{M} \mid \mathbf{Y}_{\mathbf{M}^c}, \mathbf{M}, \mathbf{W}).$$

In other words, we aim to determine the conditional distribution of $\mathbf{Y}_\mathbf{M}$ given $\mathbf{Y}_{\mathbf{M}^c}$, $\mathbf{M}$ and the low-dimensional covariate $\mathbf{W} \in \mathbb{R}^q$. We utilize the flexibility of neural networks to jointly model all conditional distributions at once.

Because the above condition density itself is hard to formulate and optimize, we follow the variational Bayesian approach (Blei et al., 2017) to maximize the negative evidence lower bound (ELBO):

$$\log p_\theta(\mathbf{Y}_\mathbf{M} \mid \mathbf{Y}_{\mathbf{M}^c}, \mathbf{M}, \mathbf{W}) \geq \underbrace{\mathbb{E}_{q_\psi(\mathbf{Z}|\mathbf{Y},\mathbf{M},\mathbf{W})} \log p_{\theta_2}(\mathbf{Y}_\mathbf{M} \mid \mathbf{Z}, \mathbf{Y}_{\mathbf{M}^c}, M, \mathbf{W})}_{\mathcal{L}_{\text{impute}}}$$

$$- \beta_{\text{kl}} \cdot KL(q_\psi(\mathbf{Z} \mid \mathbf{Y}, \mathbf{M}, \mathbf{W}) \| p_{\theta_1}(\mathbf{Z} \mid \mathbf{Y}_{\mathbf{M}^c}, \mathbf{M}, \mathbf{W})) =: \mathcal{L}_M, \tag{29}$$

where $\mathbf{Z} \in \mathbb{R}^m$ is a latent variable with approximate posterior distribution q_ψ, $\beta_{\text{kl}} = 1$ is the regularization strength, KL denotes the Kullback–Leibler divergence, and $\theta = (\theta_1, \theta_2)$. Increasing the regularization strength β_{kl} usually improves the representation learning, which gives rise to the so-called β-VAE. We specify the distributions for data as follows.

Under the target distribution p_{θ_1}, we assume that the latent variables are normally distributed:

$$\mathbf{Z} \mid \mathbf{Y}_{M^c}, \mathbf{M}, \mathbf{W} \sim \mathcal{N}(\mu_{\theta_1}, \text{diag}(\sigma^2_{\theta_1,1}, \ldots, \sigma^2_{\theta_1,m})). \tag{30}$$

Ideally, we want $\mathbf{Z}$ generated from p_{θ_1} to be as close as possible to the one generated from the proposal distribution q_ψ when $\mathbf{Y}$ is fully observed except for its authentic missing entries $\mathbf{M}_a = \mathbf{1}_p - \mathbf{C}$:

$$\mathbf{Z} \mid \mathbf{Y}_{\mathbf{M}_a^c}, \mathbf{M}_a, \mathbf{W} \sim \mathcal{N}(\mu_\psi, \text{diag}(\sigma^2_{\psi,1}, \ldots, \sigma^2_{\psi,m})). \tag{31}$$

This formulation also allows us to compute the KL divergence analytically in the ELBO (29), while it is possible to extend to normal mixtures to model more complex latent structures (Du et al., 2020). In our implementation, we simply set $q_\psi(\mathbf{Z} \mid \mathbf{Y}_{\mathbf{M}_a^c}, \mathbf{M}_a, \mathbf{W}) = p_{\theta_1}(\mathbf{Z} \mid$

$\mathbf{Y}_{\mathbf{M}^c}, \mathbf{M}, \mathbf{W})$ to reduce computational complexity. Finally, q_ψ and p_{θ_2} are modeled as two fully-factorized Gaussian distributions, whose mean and variance are estimated by two neural networks, respectively. The generative distribution p_{θ_1} are also assumed to be fully-factorized for $\mathbf{Y}_{\mathbf{M}}$ given $\mathbf{Z}, \mathbf{Y}_{\mathbf{M}^c}$, $\mathbf{M}$ and $\mathbf{W}$. We use normal distributions to model the peptide abundances. We assume that the intensities are generated based on $\mathbf{Z}$ as follows

$$Y_j \mid \mathbf{Z}, \mathbf{M}, \mathbf{W} \sim \mathcal{N}(\lambda_j, \theta_j), \tag{32}$$

which are independent of $\mathbf{M}$ given $\mathbf{Z}$. Here the parameters λ_j and θ_j are the expected intensity and the variance of the normal distribution. The posterior expectations λ_j's are outputted by the decoder and sample-specific, while the dispersion parameters θ_j's are treated as trainable variables. These parameters are learned from the data.

The aforementioned probabilistic modeling (29) emphasizes missing features imputation. On the other hand, we not only want to impute the unobserved quantities but also denoise the observed quantities. Therefore, we also attempt to maximize the reconstruction likelihood

$$\mathcal{L}_{rec} := \mathbb{E}_{p_{\theta_2}(\mathbf{Z}|\mathbf{Y}_{\mathbf{M}^c},\mathbf{M})} \log p_{\theta_1}(\mathbf{Y}_{\mathbf{M}^c} \mid \mathbf{Z}, \mathbf{M}, \mathbf{W}). \tag{33}$$

B.2. Network architecture. VAEIT is implemented using the Tensorflow (version 2.4.1) Python library (Abadi et al., 2015). VAEIT consists of three main branches, the mask encoder, the main encoder, and the main decoder. For each sample, a missing mask $\mathbf{M}$ is embedded as $\mathbf{E}$ to a dense vector of dimension 128 through the mask encoder, which greatly reduces the input dimension to the main encoder and decoder. Then, the encoder takes data $\mathbf{Y}$ (log-normalized peptide abundance), a mask embedding vector $\mathbf{E}$, and (optional) covariates $\mathbf{W}$ as input, and outputs the estimated posterior mean and variance of the distribution of the latent variable $\mathbf{Z}$. Next, a realization is drawn from this posterior distribution and fed to the decoder along with the mask embedding vector $\mathbf{E}$ and the low-dimensional covariates $\mathbf{W}$. The decoder finally outputs the posterior mean of $\mathbf{Y}$.

The encoder has two hidden layers of 64 and 16 units, and the decoder has two hidden layers of 16 and 64 units. The activation functions are set to LeakyReLU with parameter 0.2. The latent dimension is set to be 4.

B.3. Model training. VAEIT is trained in an end-to-end manner. The objective function is a convex combination of the ELBO (29) and the reconstruction likelihood (33):

$$\mathcal{L} := \beta_{\text{unobs}} \mathcal{L}_M + (1 - \beta_{\text{unobs}}) \mathcal{L}_{recon},$$

where $\beta_{\text{unobs}} \in [0, 1]$ is a hyperparameter set to be 0.9 for all experiments. We set the KL regularization parameter as $\beta_{\text{kl}} = 10$. The parameters are optimized by Monte Carlo sampling to maximize the weighted average of the reconstruction likelihood and the imputation likelihood while minimizing the KL divergence between masked posterior latent variable $\mathbf{Z} \mid \mathbf{Y}_{\mathbf{M}^c}, \mathbf{M}, \mathbf{W}$ and the authentic posterior latent variable $\mathbf{Z} \mid \mathbf{Y}_{\mathbf{M}_a^c}, \mathbf{M}_a, \mathbf{W}$. During training, with equal probability, we observe the original data and the masked data. The mask is repeatedly randomly generated for each sample at the beginning of every gradient update step in each epoch during the optimization process, such that each modality is observed with equal probability, and each entry is further randomly masked out with probability 0.5. The default variable initializer in Tensorflow is used, sampling the weight matrix from a uniform distribution and setting bias vectors to be zero. We trained our model for 300 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) with full batches and a learning rate of 1e-3 and a weight decay of 1e-4. We also use batch normalization to aid in training stability.

APPENDIX C: SUPPLEMENTARY FIGURES/TABLES

Supplementary figures for Section 3.2

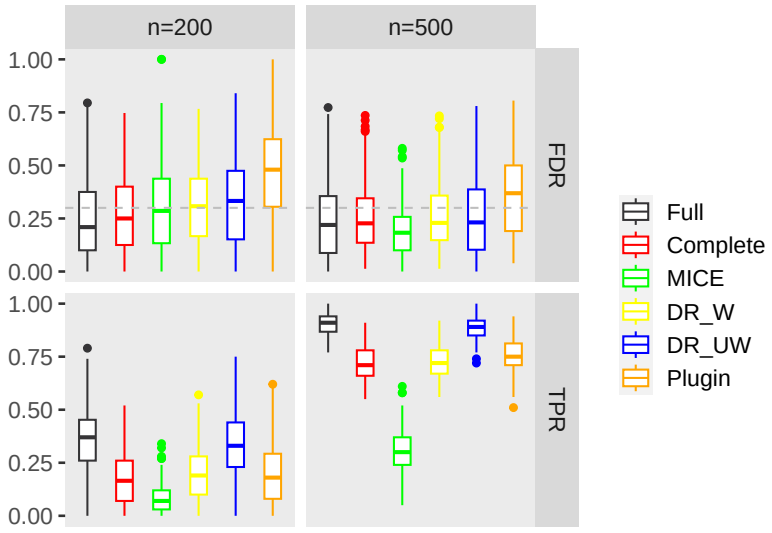

Figure C1: Simulation result of Model 1.

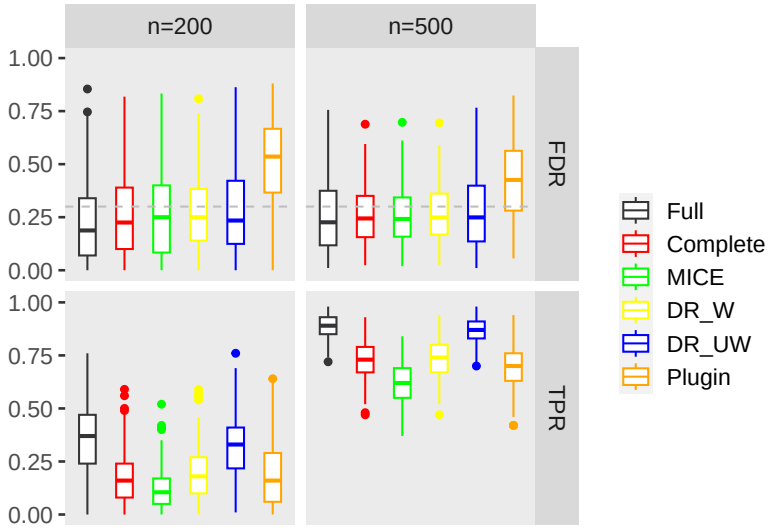

Figure C2: Simulation result of Model 2.

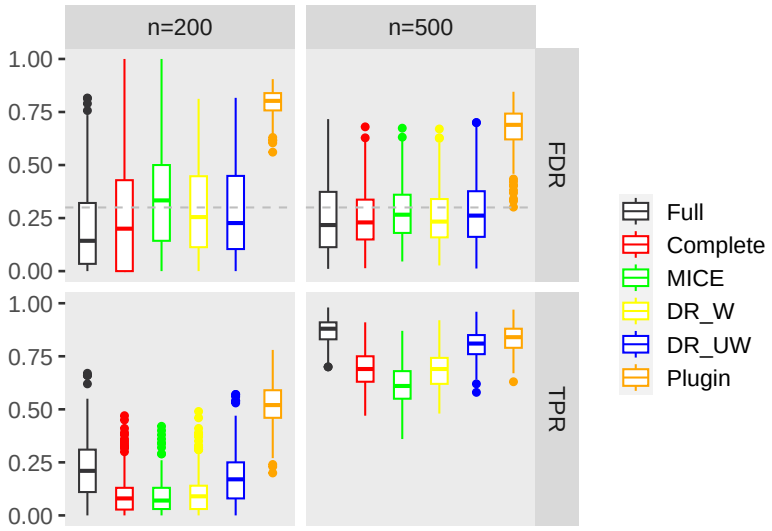

Figure C3: Simulation result of Model 4.

Supplementary figures and tables for Section 4

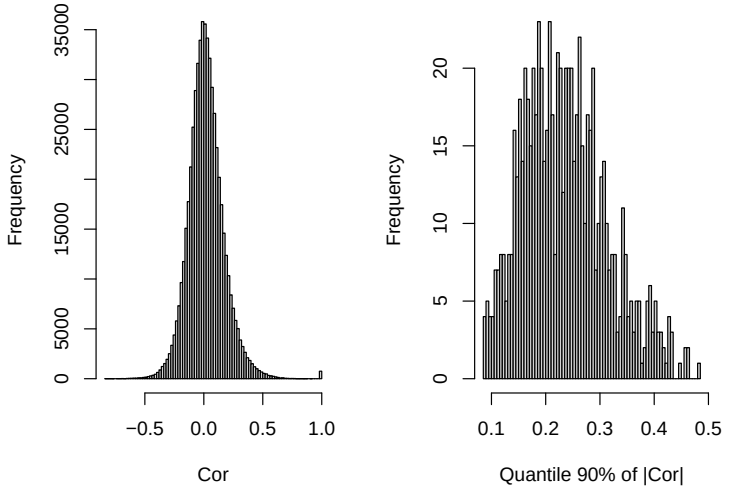

Figure C4: Histogram of correlation coefficient between peptides (left) and a 90% quantile absolute value of correlation coefficients computed for each peptide (right)

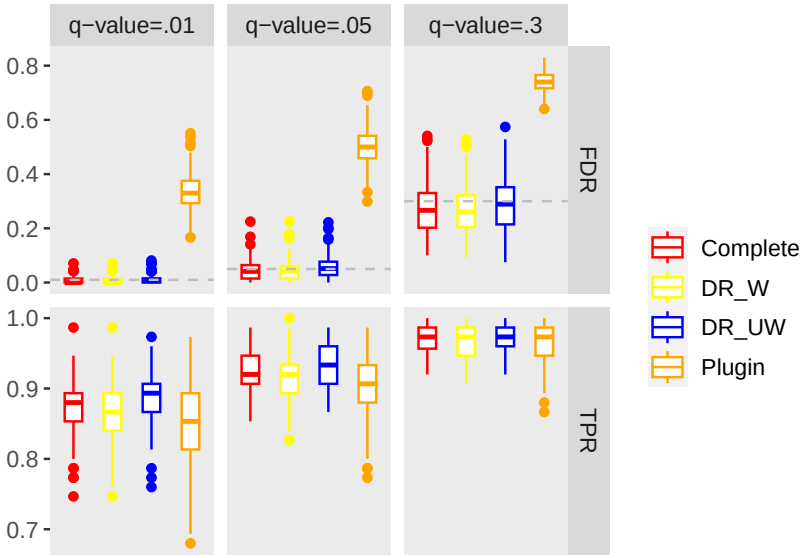

Figure C5: A result of realistic simulation with a single-cell dataset (Setting 1); FDR and TPR are summarized under different q-value cutoffs (0.01, 0.05 and 0.3)

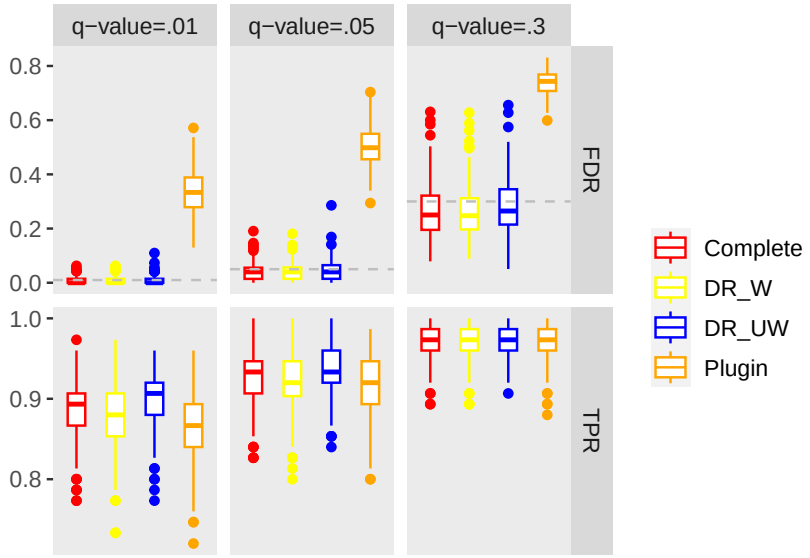

Figure C6: A result of realistic simulation with a single-cell dataset (Setting 2); FDR and TPR are summarized under different q-value cutoffs (0.01, 0.05 and 0.3)

q-value cutoff	% Protein overlaps			Number of additional peptides		
	DR_W	DR_UW	Plugin	DR_W	DR_UW	Plugin
0.01	1	0.90	0.51	6	31	179
0.05	0.90	0.71	0.44	62	104	257
0.1	0.74	0.62	0.41	100	143	288
0.3	0.51	0.47	0.38	214	258	344

TABLE C1

The proportions of peptides, whose corresponding proteins are included in protein lists corresponding to the peptides selected by the Complete method with a q-value cutoff of 0.05. Only the peptides additionally selected by each method compared to the Complete method (with a q-value cutoff of 0.01) are considered. A threshold 0.7 is applied to the observation rate of peptides.

Supplementary figures for Section 6

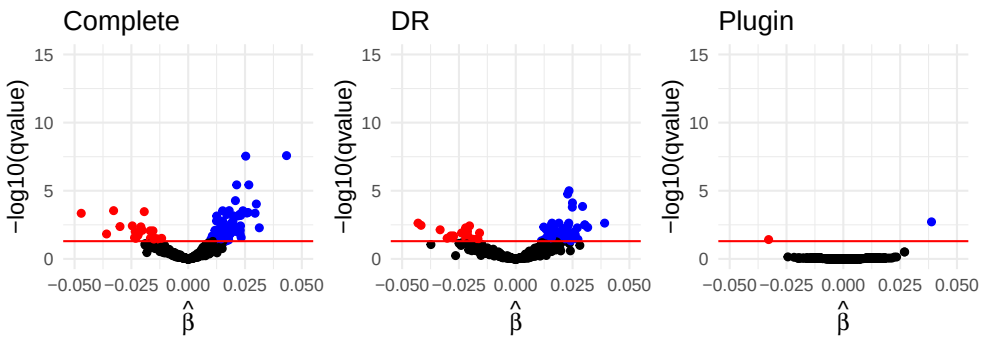

Figure C7: Volcano plot with a completely noisy imputation (a q-value cutoff=0.05)

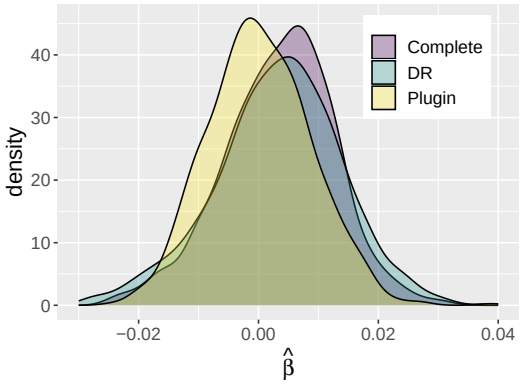

Figure C8: Distribution of estimated coefficient $\hat{\beta}$ for the diameter variable under a completely noisy imputation

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015), "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," Software available from tensorflow.org.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational inference: A review for statisticians," *Journal of the American statistical Association*, 112, 859–877.

Du, J.-H., Cai, Z., and Roeder, K. (2022), "Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT," *Proc Natl Acad Sci U S A*, 119, e2214414119.

Du, J.-H., Chen, T., Gao, M., and Wang, J. (2020), "Model-based trajectory inference for single-cell rna sequencing using deep learning with a mixture prior," *bioRxiv*, 2020–12.

Henderson, H. V. and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *Siam Review*, 23, 53–60.

Ivanov, O., Figurnov, M., and Vetrov, D. (2018), "Variational Autoencoder with Arbitrary Conditioning," in *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014), "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, eds. Bengio, Y. and LeCun, Y.

Loshchilov, I. and Hutter, F. (2017), "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*.

Sohn, K., Lee, H., and Yan, X. (2015), "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, 28.