Proteomics

WILEY-VCH

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE OPEN ACCESS

Practical Impact of Imputation and Batch-Effect Correction for Proteomics/Peptidomics Differential-Abundance Analysis

Charis Gonidaki[1,2] | Agnieszka Latosinska[3] | Antonia Vlahou[1] | Rafael Stroggilos[1] | Harald Mischak[3]

[1]Center of Systems Biology, Biomedical Research Foundation of the Academy of Athens (BRFAA), Athens, Greece | [2]Department of Applied Mathematics, University of Twente, Enschede, The Netherlands | [3]Mosaiques Diagnostics GmbH, Hannover, Germany

Correspondence: Charis Gonidaki (cgonidaki@bioacademy.gr) | Harald Mischak (mischak@mosaiques-diagnostics.com)

Received: 14 July 2025 | **Revised:** 4 February 2026 | **Accepted:** 6 February 2026

Keywords: batch effects | imputation | large-scale proteomics | missing values | urine peptidomics

ABSTRACT

MS-based proteomics offers powerful opportunities for biomarker discovery; nevertheless, it is associated with technical challenges, including missing values and batch effects. Although imputation and batch-correction methods are well established in proteomics, their impact remains incompletely characterized in large-scale clinical proteomics datasets. Here, we examine the practical impact and interaction of three popular imputation methods (Gaussian, ½ LOD, KNN) in combination with three batch-effect correction approaches (ComBat, ComBat with disease covariate, MNN) on differential abundance analysis in a CE-MS urine peptidomics dataset of 1,050 samples across 13 batches from chronic kidney disease (CKD) patients and controls. Downstream effects were assessed based on peptide validation between discovery and validation sets. Imputation method choice had minimal impact on the final list of disease-associated peptides (DAPs), given the missingness structure and normalization strategy. In contrast, batch-effect correction largely affected the results: MNN and especially unadjusted ComBat removed a large proportion of DAPs (~50% and >90%, respectively), whereas inclusion of disease status in the ComBat model largely preserved biological signal. This study highlights how popular preprocessing choices can affect biological signal, showing that imputation and batch-effect correction interact and jointly influence downstream results, underscoring the need for caution when applying batch-effect correction.

1 | Introduction

One of the most effective methods in biomedical research for identifying and measuring proteins and peptides in biological samples is mass spectrometry (MS). The performance of MS has significantly improved over the past decade thanks to significant developments in instrumentation, such as enhanced sensitivity, faster acquisition rates, and higher resolution [1]. These improvements have allowed researchers to detect low-abundance proteins and peptides, resolve closely related molecular species, and analyze thousands of features across large sample cohorts. As a result, MS-based platforms now play a central role in biomarker discovery and personalized medicine [2].

Despite these advances, proteomics, in fact omics approaches in general, face several technical challenges. Among the most prominent are missing values and batch effects. Missing values leave gaps in the data that can skew statistical analyses. Their causes fall into three main categories [3, 4]. If a feature's intensity is lost purely by chance (e.g., a transient instrument

Abbreviations: CKD, chronic kidney disease; DBP, diastolic blood pressure; DAPs, disease-associated peptides; GFR, glomerular filtration rate; J, Jaccard similarity index; KNN, K-nearest neighbors; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; MNN, mutual nearest neighbors; SBP, systolic blood pressure.

Statement of Significance of the Study

Finding reliable biomarkers in clinical proteomics requires addressing the technical noise that can hide true biological signals. In this work, we examine the practical impact and interaction of commonly used imputation and batch correction methods on the list of peptides that emerge as differentially abundant. Instead of relying on simulations or small datasets, we examine a large, real-world urine-peptidomics cohort of more than 1,000 samples screened for chronic kidney disease. The results demonstrate that, in datasets such as the one used here, different preprocessing strategies can lead to substantially different outcomes. Imputation and batch-effect correction were found to be interdependent, and batch effect removal can lead to loss of meaningful biological differences, highlighting the importance of applying such corrections with caution.

glitch), it is considered missing completely at random (MCAR). When detection depends on other measured factors (such as generally low signal across runs) but not on the feature's own true abundance, data are missing at random (MAR). Finally, features with true concentrations below the detection limit produce missing not at random (MNAR) values. Although these mechanisms, in practice, often overlap, distinguishing among them guides the choice of imputation strategy [5]. In parallel, batch effects introduce non-biological variation when samples are collected at different centers (introducing center bias) or processed on different days, with different instruments, or in different laboratories. These effects can similarly obscure genuine biological signals [6].

Over the years, many imputation and batch-correction methods have been developed in order to address these distinct sources of variation. Of the most widely used imputation techniques as reflected in the current literature (see Section 2.3) are Gaussian- and LOD-based approaches, which are well suited for MNAR scenarios, and K-nearest neighbors (KNN) that can handle MAR patterns by borrowing information from similar features [7–9]. Imputation is also commonly applied as a preparatory step since many downstream methods, such as batch correction, assume complete data and cannot directly handle missing values. Likewise, batch-correction approaches like the empirical-Bayes ComBat method and the mutual nearest neighbors (MNN) algorithm make different assumptions about how non-biological variation should be modeled and removed [10, 11].

Numerous benchmarking studies independently evaluated imputation tools [8, 9] or batch-correction algorithms [12, 13], yet only a handful have attempted to consider how the two steps interact, and/or how those data manipulations affect downstream statistical analyses [14]. Most studies apply a single preprocessing pipeline, leaving it uncertain how different combinations remove technical noise while preserving biological signals [15].

To address this gap, we utilized peptidomics data from samples collected from 13 different research hubs (batches) to examine the practical impact of the three most popular imputation approaches (Gaussian, ½ LOD replacement and KNN) and three batch-effect correction methods (ComBat, ComBat with the disease status as a covariate and MNN). These summed to a total of 12 different processing pipelines which were assessed alongside a simple zero-fill baseline, in which missing values were set to zero and no batch-effect correction was applied. The focus was not to identify an optimal preprocessing approach, but to investigate how choices made at each preprocessing step and their interaction impact on downstream statistical results in this real-world dataset.

As a test case, we used urine peptide data collected via capillary electrophoresis mass spectrometry (CE-MS), in the context of chronic kidney disease (CKD). The readout chosen was significantly disease associated peptides (DAPs), comparing cases and controls. In order to illustrate the impact of the different preprocessing decisions in this dataset, we assessed how different imputation and batch correction choices affected the consistency of DAPs between discovery and validation sets, as well as their reproducibility across preprocessing scenarios. The conclusions regarding the imputation are interpreted in the context of this large CE-MS urine peptidomics dataset, which is characterized by predominantly MNAR missingness and a collagen-based normalization strategy routinely applied in CE-MS workflows.

2 | Methods

2.1 | Dataset Description

An anonymized urine-peptidomics dataset was obtained from the Mosaiques Diagnostics database. It contained data from 1,224 samples (612 healthy controls, 612 chronic kidney disease (CKD) patients) drawn from 36 independent studies (batches). Each sample included measurements for 24,346 unique Peptide IDs, acquired using capillary electrophoresis mass spectrometry (CE–MS; see Section 2.2 for details on data acquisition and normalization). For each of the 1,224 samples, clinical data were available. In particular, the disease status (healthy control vs CKD), sex (male/female), age (years), estimated glomerular filtration rate (GFR, mL/min/1.73 m^2), diastolic blood pressure (DBP, mmHg) and systolic blood pressure (SBP, mmHg) were recorded.

As reliable imputation and batch-correction require sufficient sample sizes, any study (batch) with fewer than 20 samples was excluded. This reduced the dataset to 13 batches, corresponding to 565 controls and 500 CKD cases. The retained batches ranged in size from 21 to 332 samples. After matching individuals based on age, sex, and systolic/diastolic blood pressure, the final number of samples included in the analysis was 565 for the control group and 485 for the CKD patient group (1,050 samples in total). An overview of the sample composition and data quality across the 13 batches is provided in Table S1. For downstream analysis, the samples were randomly split into discovery and validation subsets of equal size (n = 525 each), ensuring a balanced distribution of clinical and pathological variables (Table S2).

Peptides were further filtered based on the availability of sequence information, (reducing the total to 18,035 Peptide IDs)

TABLE 1 | Clinical, demographic and data-quality characteristics of the urine-peptidomics cohort. The 1,050 samples were randomly split 1:1 into a discovery (training) set ($n = 525$) and a validation (test) set ($n = 525$), preserving the distribution of key clinical variables.

	Overall Set	Discovery Subset	Validation Subset
Number of samples	1,050	525	525
Group = CKD (%)	485 (46.2%)	244 (46.5%)	241 (45.9%)
Sex = Female (%)	397 (37.8%)	193 (36.8%)	204 (38.9%)
Age (mean ± SD)	60.76 ± 10.18	60.78 ± 10.14	60.74 ± 10.24
GFR (mean ± SD)	75.12 ± 28.63	74.90 ± 28.75	75.35 ± 28.53
SBP (mean ± SD)	135.83 ±15.56	135.25 ± 15.67	136.41 ± 15.44
DBP (mean ± SD)	77.99 ± 9.40	77.71 ± 9.44	78.27 ± 9.37
Missing Percentage per sample (%) (mean ± SD)	27.92 ± 12.92	27.75 ± 12.77	28.10 ± 13.07

Continuous values are shown as mean ± standard deviation; categorical values are counts with percentages in parentheses.
Abbreviations: CKD, chronic kidney disease; GFR, estimated glomerular filtration rate (mL min^{-1} 1.73 m^{-2}); SBP, systolic blood pressure (mm Hg); DBP, diastolic blood pressure (mm Hg).

and a frequency threshold (peptides present in at least 50% of samples) which resulted in a final number of 3,262 Peptide IDs included in the analysis (Table S3). A concise overview of the final, filtered cohort is given in Table 1.

2.2 | CE-MS Data Acquisition and Normalization

CE–MS analyses were carried out on a P/ACE MDQ instrument (Beckman Coulter, Fullerton, CA, USA) coupled to a micro-TOF mass spectrometer (Bruker Daltonics, Bremen, Germany) as described in detail in prior work [16]. The resulting CE–MS data were processed with the MosaFinder software, which aligns detected signals in silico to a list of 24,346 urinary peptides, of which 18,035 have an amino acid sequence assigned [17]. Peptide intensities were normalized using 29 collagen fragments, the intensity of which, when combined into a panel, is largely unaffected by disease thus serving as internal standard [18]. The resulting normalized intensity matrix was used as input for all subsequent analyses in this study.

2.3 | Literature Survey and Selection of Imputation Methods

To identify which imputation strategies are most used in current proteomics practice, we conducted a systematic literature review. First, we screened all articles published in the journal *PROTEOMICS* from 2020 onward. Second, we searched PubMed for open-access experimental proteomics studies on biological samples published in the last five years. Third, we inspected the studies cited in the survey by Harris et al. [19] and retained only those published in 2020 or later. Benchmarking papers and studies focused on developing new imputation algorithms were excluded, as we aimed to capture imputation choices in routine experimental workflows rather than to compare specialized methods. For each eligible article, we recorded the DOI, imputation category, year of publication, and source (Proteomics, PubMed or Harris et al.).

This process yielded a final set of 131 studies. A full list of the included articles, along with DOI, imputation category and the year they were published, is provided in Table S4. The distribution of imputation categories is summarized in Figure 1. Gaussian-based imputation was the most frequently reported approach, followed by LOD-based methods, KNN and zero-fill/pseudocount schemes. Based on these findings, we selected Gaussian, ½ LOD replacement and KNN as representative imputation strategies for detailed evaluation in this study and included zero-fill as a simple baseline.

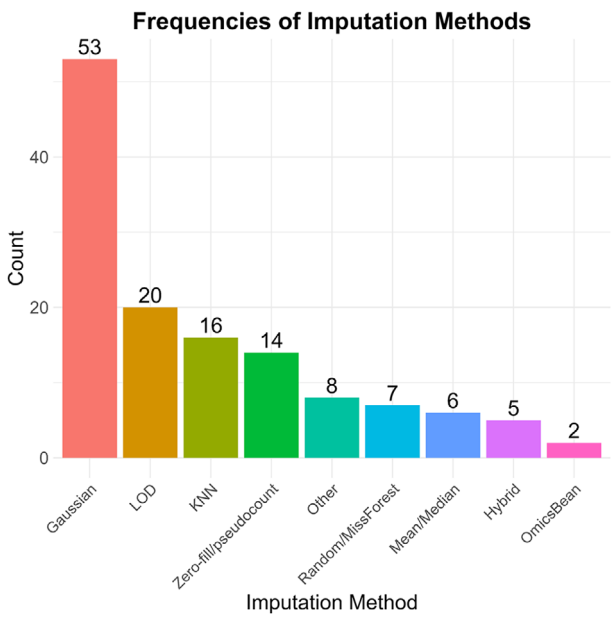

FIGURE 1 | Frequency of application of imputation methods in recent experimental proteomics studies (131 studies published from 2020 onward). The number of studies per method is provided. "Other" refers to methods used only once in the analyzed studies. "Hybrid" denotes studies that applied two imputation methods to the same dataset, typically combining MNAR- and MAR-oriented strategies according to the presumed missingness mechanism. "Random/MissForest" denotes random forest–based imputation approaches (e.g. MissForest), while "OmicsBean" refers to a software platform that applies predefined preprocessing pipelines including imputation using single value replacement or distribution-based approaches. The full results of this literature search, including the DOIs of the identified studies, are given in Table S4.

2.4 | Combinations of Applied Methods

Three imputation methods were tested individually and in combination with three batch-effect correction methods, resulting in 12 distinct preprocessing pipelines. All analyses were performed in R programming language.

2.4.1 | Imputation Methods

Gaussian imputation replaces missing values by random numbers drawn from a normal distribution that is shifted toward lower intensities and is narrower than the empirical distribution. Following the widely accepted Perseus-style characteristics, we sampled imputed values from a distribution with mean $\mu - 1.8\sigma$ and standard deviation 0.3σ, where μ and σ are the mean and standard deviation of the observed log-intensities within each sample [20]. It is a method which assumes that missing peptide intensities correspond to low-abundance signals just below the detection limit.

Half limit of detection (1/2 LOD) replacement sets all missing values to half the minimum observed intensity in the dataset. This approach assumes that missingness arises because true peptide abundances fall below the instrument's detection threshold [21].

K-Nearest Neighbors (KNN) replaces a missing value by averaging the values of the K most similar features, where "similarity" is defined via a chosen distance metric [22]. In this study, we applied kNN() function from VIM package [23] to our peptide-intensity matrix, using $K = 5$ and package's default Gower-type distance (which scales numeric variables to [0,1] before computing Euclidean distance).

2.4.2 | Batch Correction Methods

Three batch-correction variations were applied to the $\log_2$-transformed imputed datasets:

ComBat applies either parametric or non-parametric empirical Bayes frameworks to adjust data for batch effects [24]. For this analysis, we applied the parametric empirical Bayes framework (package default) using the ComBat() function from sva package [25] on our $\log_2$-transformed imputed peptide matrix. No biological covariates were included, so all between-batch variations were subject to removal.

ComBat using the CKD covariate extends the above method by supplying a design matrix with the disease-status covariate [11]. In this way, ComBat preserves differences associated with CKD status while still removing unwanted batch effects, ensuring that true case–control signal is not inadvertently washed out.

Mutual Nearest Neighbors (MNN) identifies mutual nearest neighbors between datasets to align their shared structure [26]. These neighbors are pairs of data points (samples) from different batches that are closest to each other in the high-dimensional space, representing similar biological states. For this, we applied the function batchCorrect() from batchelor package [27] to our $\log_2$-transformed imputed peptide matrix with PARAM = ClassicMnnParam($k = 15$), and cos.norm.out = FALSE.

2.5 | Assessment of Batch Effects

Principal Component Analysis (PCA) was performed on the $\log_2$-transformed peptide-intensity matrices before and after batch correction. PCA plots with the first two principal components (PC1 and PC2) were generated to visually inspect the presence and strength of batch effects and to evaluate the extent to which these correction methods reduced the unwanted variation.

Furthermore, we generated boxplots of $\log_2$ peptide intensities for every study (batch) across all pipelines. In these boxplots, each box represents the distribution of peptide intensities within one study, allowing direct comparison of median, spread and outliers between batches. In order to highlight within- vs between-batch similarities, we additionally calculated sample–sample Pearson correlation matrices and displayed them as heatmaps with samples arranged and labeled by study/batch.

2.6 | Statistical Analysis

Differential abundance analysis was performed as follows: To assess the stability of results, we applied the analysis separately to the discovery and validation sets ($n = 525$ each), which had balanced clinical and pathological characteristics. Within each subset, we performed non-parametric Mann–Whitney U-tests to compare CKD and control samples, followed by Benjamini–Hochberg adjustment to control the false discovery rate. Only peptides that were significant (adjusted p-value < 0.05) in both sets and showed the same direction of change were retained.

To further investigate the overlaps of the differentially abundant peptides and obtain an overview of the results, we calculated the Jaccard similarity index between all the different pairs of pipelines. The Jaccard similarity index measures the proportion of shared peptides between two sets relative to their total peptide number. In mathematical terms the Jaccard index between two sets A and B is thereby defined as [28],

$$J = \frac{|\, A \cap B \,|}{|\, A \cup B \,|} \tag{1}$$

In our case, $|\, A \cap B \,|$ was the number of peptides that were found significantly different by both pipelines and $|\, A \cup B \,|$ was the union of the significant peptides of these two pipelines. A higher J indicates greater overlap in significantly different peptides between the two methods.

All exploratory analyses and statistical tests were performed using standard procedures implemented in R.

3 | Results

The goal of the study was to assess how different imputation methods and their pairwise combinations with batch-effect cor-

rection methods affect differential-abundance analyses in urine peptidomics. The urine peptide dataset after filtering consisted of 3,262 Peptide IDs and 1,050 samples across 13 batches-studies.

3.1 | Sample Distribution Across Pipelines

A zero-fill baseline dataset, in which all missing values were set to zero and no batch correction was applied, served as the comparator. We initially examined using PCA plots, boxplots and heatmaps the following preprocessing pipelines: Gaussian-only, ½ LOD-only, KNN-only, and each of these imputation methods paired with ComBat or MNN.

PCA was performed on the $\log_2$-transformed peptide-intensity matrices before and after batch correction. The PCA results for the zero-fill and ½ LOD–imputed datasets (Figure S1) were visually identical. To quantify this similarity, we correlated the sample PC scores between these two analyses using Pearson's correlation (cor.test in R). PC1 and PC2 scores were almost perfectly correlated (both r ≈ 1.00, $p < 2.2 \times 10^{-16}$), which quantitatively supports that the PCA patterns for these two pipelines are essentially identical. Figure 2 displays PC1 versus PC2 for each pipeline. Across the imputation-only panels: Gaussian-imputed (Figure 2A), ½ LOD–imputed (Figure 2B) and KNN-imputed data (Figure 2C) the overall PCA structure indicates that simple imputation alone appears to have almost no effect on the global variance structure of this dataset. Notably, even though samples from the same batch tend to cluster closer to each other, the degree of batch clustering in these imputation-only panels appears relatively mild, suggesting that without explicit correction the inter-study effect is not extensive. In the batch-corrected panels (Figure 2D—I), samples appear more evenly mixed, with no single study dominating any region, especially for the ComBat corrected pipelines. Yet, the shapes denoting clinical status also appear diffusely scattered, indicating that true disease-related variance may have been reduced as well.

Boxplots of $\log_2$ peptide intensities per study showed varying between-batch differences in median and spread for the imputation-only datasets (Figures S2–S5). The zero-fill and ½ LOD baselines showed high batch-dependent median shifts, with batch-specific medians reaching over 80% of the overall median and interquartile-range (IQR) differences approximately 40%. In contrast, Gaussian imputation showed moderate between-batch median differences (up to ∼55%) and smaller differences in spread (∼15%), while KNN imputation showed minimal between-batch variation, with median and IQR differences being below ∼5%. After batch correction, between-batch median differences were reduced across all pipelines (Figures S6–S11). For ComBat-corrected data, median ranges dropped to below ∼30%, while Gaussian- and KNN-based pipelines corrected using MNN showed median differences below ∼20% and ∼10%, respectively. Similarly, differences in spread were reduced for most imputation–correction combinations, with IQR ranges generally below ∼25%, although the extent of this reduction depended on the specific imputation–correction pairing.

In the sample–sample Pearson correlation heatmaps (Figures S2–S5), samples from the same study in the imputation-only datasets consistently showed higher correlations than samples from different studies, producing some clear study-specific blocks (median within-study $r \approx 0.42$–0.63 vs between-study $r \approx 0.33$–0.55). After batch correction (Figures S6–S11), these study-specific correlation patterns were reduced. In ComBat-corrected datasets, within- and between-study correlations became nearly indistinguishable ($\Delta r \approx 0.003$), indicating strong removal of batch-associated structure, with overall median correlations ranging from approximately 0.36 to 0.60 depending on the imputation method. In contrast, MNN correction preserved uniformly high correlations across samples (median $r \approx 0.54$–0.66).

3.2 | Impact of Imputation and Batch Effect Removal on Significant Peptides

Differential abundance between CKD and healthy samples was assessed using Mann–Whitney *U*-tests with Benjamini–Hochberg correction. Only peptides that were significant in both the discovery and validation sets with the same direction of change were retained (see Table S5 for the detailed results of the statistical analysis). The table in Figure 3A summarizes the results of this analysis specifically depicting:

a. The total number of peptides that reached significance in the discovery set.
b. The number of significant peptides replicated in the validation set with the same direction of regulation.
c. The percentage of discovery-set hits that were successfully validated.

As expected for a rank-based test, replacing missing values with a constant value, either 0 or ½ LOD does not alter the relative ordering of peptide intensities. Consequently, the zero-fill baseline and the ½ LOD-only imputed pipeline produced identical lists of significant peptides (1,640 validated peptides with validation rate 79.3%). KNN imputation gave almost the same number (1,614) and proportion (79.5%) of validated peptides as zero-fill ($\chi^2 = 0.04$, $p = 0.85$). Gaussian imputation led to a slightly higher validation rate (82.4%), but the overall number of validated peptides remained similar across all three pipelines (1,614–1,661). Adding batch correction uniformly reduced the number of significant peptides (from a mean of 1,640 before to fewer than 880 after batch-effect removal). The pipelines combining Gaussian or ½ LOD imputation followed by MNN batch correction (hereafter Gaussian + MNN and ½ LOD + MNN, respectively) retained a moderate set of validated peptides ($n \approx 855$) while maintaining a high validation percentage (∼76%); these numbers declined further in the combination of KNN + MNN ($n = 330$, < 60%). The drop was even more pronounced with ComBat, regardless of the imputation method applied ($n \leq 50$, ≤ 52%).

Overlaps in findings of the different pipelines were further investigated using the pairwise Jaccard similarity index. The resulting values of the validated DAPs are displayed in Figure 3B, where each cell's color intensity reflects the degree of overlap between two pipelines (darker = greater overlap). As evident from the heatmap, the zero-fill baseline, Gaussian, ½ LOD and KNN pipelines share a high proportion of differentially abundant peptides ($J \geq 0.69$), reflecting similar results when no batch

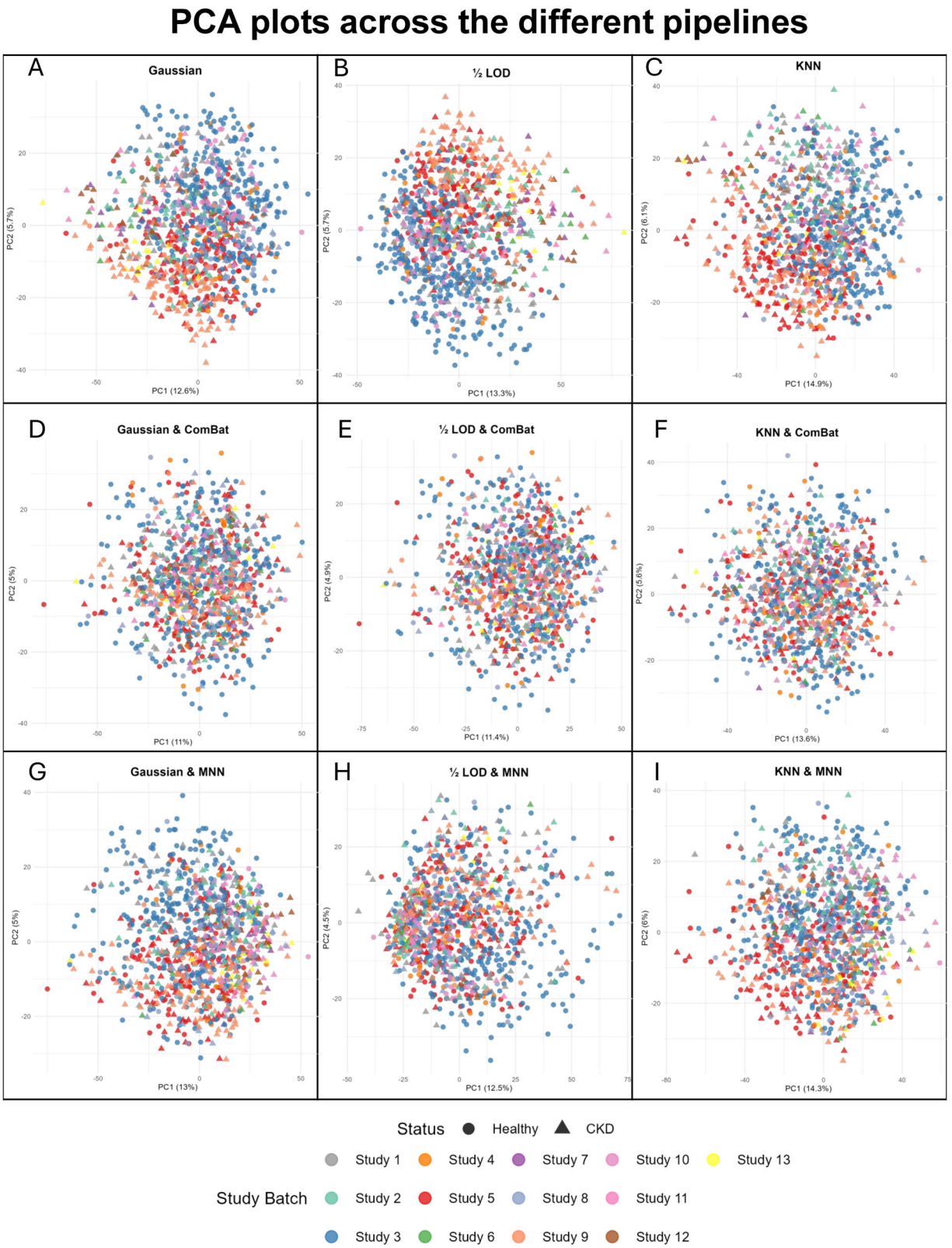

FIGURE 2 | Principal Components Analysis (PCA) plots (PC1 vs PC2) for the nine core pipelines: A) Gaussian, B) ½ LOD, C) KNN, D) Gaussian & ComBat, E) ½ LOD & ComBat, F) KNN & ComBat, G) Gaussian & MNN, H) ½ LOD & MNN and I) KNN & MNN. Each point represents a single sample, colored by its study batch and shaped by clinical status (● healthy, ▲ CKD).

A

Method	Significant peptides identified from discovery set	Significant peptides validated	Validation Rate
Zero-Fill	2,069	1,640	**79.3%**
Gaussian	2,017	1,661	**82.4%**
½ LOD	2,069	1,640	**79.3%**
KNN	2,030	1,614	**79.5%**
Gaussian + MNN	1,160	876	**75.5%**
½ LOD + MNN	1,094	834	**76.2%**
KNN + MNN	576	330	**57.3%**
Gaussian + ComBat	66	29	**43.9%**
½ LOD + ComBat	112	49	**43.8%**
KNN + ComBat	25	13	**52%**

B

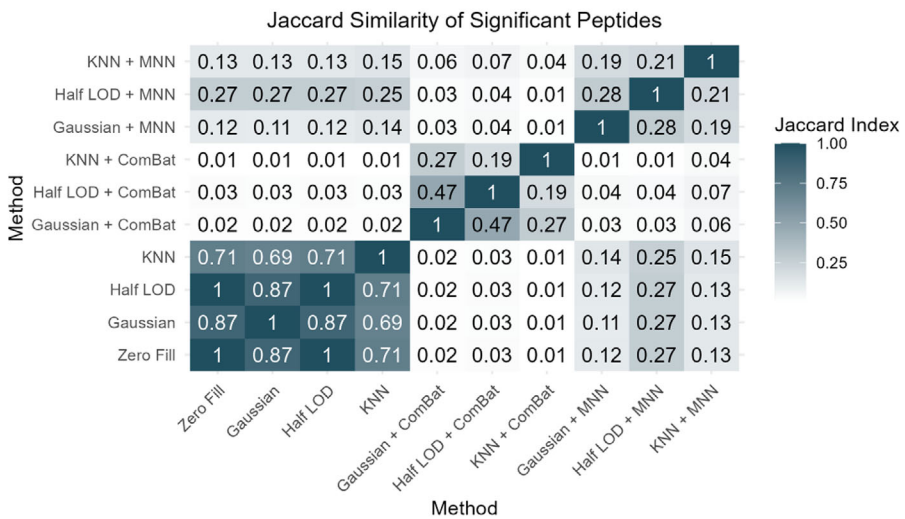

FIGURE 3 | A) Table reporting the number of peptides found significant in the discovery set, the number of them that were validated in the validation set (with the same regulation trend) and the respective percentage of discovery-set hits validated per pipeline. B) Heatmap showing the Jaccard index for the validated significant peptides, computed for every pair of the 10 core pipelines (Zero-Fill, Gaussian, ½ LOD, KNN, and their combinations with ComBat, and MNN). Each cell's value and color intensity represent the fraction of shared significant peptides (darker = greater overlap).

correction is applied. The three MNN pipelines (Gaussian + MNN, ½ LOD + MNN and KNN + MNN) show low concordance with the other pipelines ($J \approx 0.15$–0.3). This low index value is to a good extent expected, since the total number of significant peptides highlighted by pipelines using MNN is also lower. Even more pronounced, ComBat shows no considerable overlap with any pipeline ($J < 0.1$), highlighting its large impact on the dataspace, leading to the hypothesis that this approach may be removing not only batch-related variation but also biological signal.

To further investigate the hypothesis that ComBat removes biological variability, the algorithm was reapplied, with the disease status (healthy control or CKD) included as a covariate (see Figures S12–S14 for the PCA plots, boxplots and heatmaps). As shown in the table in Figure 4A, in all: Gaussian + ComBat(+CKD), ½ LOD + ComBat(+CKD) and KNN + ComBat(+CKD) pipelines, a substantial number of DAPs was identified and successfully validated. Specifically, over 750 DAPs were detected in each pipeline, with validation percentages being more than 70%. Along the same lines, all three ComBat(+CKD) pipelines exhibit moderate overlap ($J \approx 0.2$–0.8) with the rest methods (Figure 4B), however, exceeding the almost-zero concordance seen using standard (unadjusted) ComBat. This heatmap is consistent with the hypothesis that ComBat without covariate adjustment removes biological variability.

4 | Discussion

In this study, we examined how different combinations of widely applied imputation and batch-effect correction strategies influence the statistical results in a large CE-MS urine peptidomics dataset. The aim was to illustrate their practical impact and

A

Method	Significant peptides identified from discovery set	Significant peptides validated	Validation Rate
Zero-Fill	2,069	1,640	**79.3%**
Gaussian	2,017	1,661	**82.4%**
½ LOD	2,069	1,640	**79.3%**
KNN	2,030	1,614	**79.5%**
Gaussian + MNN	1,160	876	**75.5%**
½ LOD + MNN	1,094	834	**76.2%**
KNN + MNN	576	330	**57.3%**
Gaussian + ComBat	66	29	**43.9%**
½ LOD + ComBat	112	49	**43.8%**
KNN + ComBat	25	13	**52%**
Gaussian + ComBat (+CKD)	1,069	761	**71.2%**
½ LOD + ComBat (+CKD)	1,089	809	**74.3%**
KNN + ComBat (+CKD)	1,122	858	**76.5%**

B

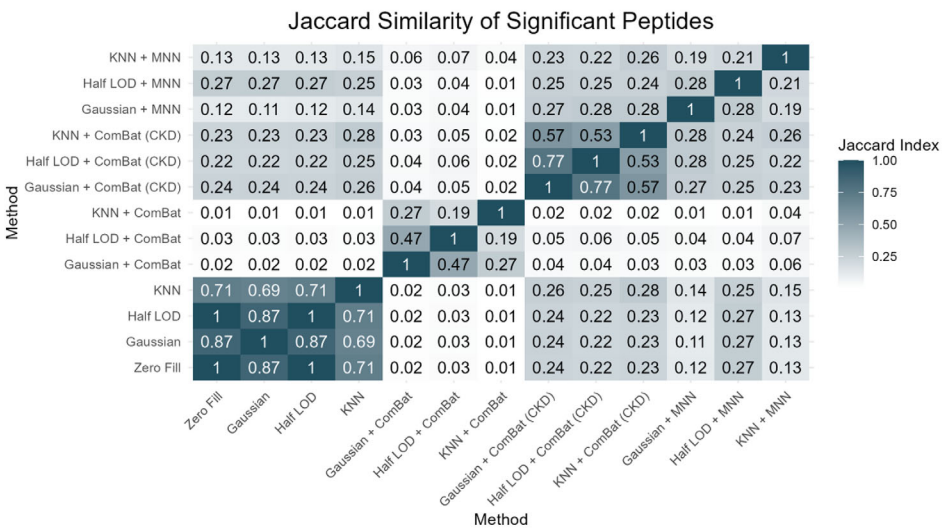

FIGURE 4 | A) Table reporting the number of peptides found significant in the discovery set, the number of them that were validated in the validation set and the percentage of validated discovery-set hits per pipeline. B) Heatmap that displays the Jaccard index for the validated DAPs calculated for every pair of the 13 pipelines (Zero-Fill, Gaussian, ½ LOD, KNN, and their combinations with ComBat, ComBat(+CKD), and MNN). Each cell's value and color intensity indicate the proportion of shared significant peptides (darker = greater overlap).

interaction in a real-world clinical setting. Even though Gaussian-based imputation and KNN led to different intensity distributions compared with ½ LOD and zero-fill (Figures S2–S5), minimal differences in the list of validated DAPs could be observed between these four imputation methods. In contrast, batch-effect correction had a remarkable influence: MNN and, in an even more pronounced manner, ComBat without covariates was associated with a substantial reduction of the number of validated DAPs and showed minimal overlap with other pipelines. Including disease status (CKD) in the model restored to some extent the signal, suggesting that it corresponded mainly to biological variability.

Importantly, although imputation-only pipelines yielded similar numbers of validated DAPs in our dataset, applying the same batch-effect correction resulted in markedly different outcomes depending on the imputation method used. Gaussian and KNN imputation applied without batch correction returned comparable numbers of validated DAPs (1,661 and 1,614, respectively), whereas after MNN correction they yielded different numbers of validated biomarkers (876 vs 330 respectively, Figure 3). This indicates that, in this dataset, the choice of imputation directly influences the downstream effect of subsequent batch correction. Therefore, a core finding of this study is that the effects of imputation and batch correction are not entirely independent and may interact in ways that affect downstream analyses.

In addition, in our dataset, where batch-related variation was relatively mild, strong correction methods like unguided ComBat removed meaningful biological signals, which was avoided by including biological covariates in the model or applying milder corrections like MNN. The results suggest that the structure of the dataset can be an important factor when making preprocessing decisions, especially in relation to the extent of the batch effects

and the availability of relevant biological information. In this well-balanced, multi-batch CE-MS urine peptidomics dataset, minimal preprocessing was sufficient to retain biological signal while limiting technical variation. This limited batch-related variation and similar performance of the tested imputation methods may be explained by the normalization routinely applied to CE-MS datasets from urine samples, which uses a panel of 29 collagen calibrant peptides as internal standards to harmonize peptide intensities across runs [29, 30].

Along these lines, when comparing the mean intensity of the 29 calibrants between CKD and control samples for each pipeline, the percent difference was below 7%, with $\log_2$(CKD/Control) ratios close to zero, indicating that the overall abundance of the calibrant peptides remained highly similar between groups, as expected. As a complementary "positive-control" assessment, when investigating the distribution of previously described CKD biomarkers (specifically, the 10 most significant peptides from CKD273 classifier [31] detected in our dataset) (Table S7), most consistent results were obtained when applying the imputation-only pipelines and batch-correction approaches with either MNN or ComBat with CKD as a covariate. In contrast, ComBat without CKD as a covariate missed most of these peptide biomarkers (9/10), supporting the interpretation that unguided batch correction can reduce the ability to detect established relevant biological signals.

The preprocessing methods examined in this study represent some of the most commonly applied approaches for addressing missing values and batch effects in omics research. This is also reflected in the literature survey we conducted of recent experimental proteomics studies (see Section 2.3). Although the performance of these approaches has been studied, their impact on large-scale proteomics and peptidomics datasets has not been thoroughly examined.

Gaussian-based imputation is widely used in proteomics when missing values are assumed to be MNAR, typically because low-abundance signals fall below the detection limit. It replaces missing values with draws from a left-shifted, low-intensity Gaussian distribution, which preserves some variability, while still encoding the assumption of low abundance; however, it can still shrink variance and influence estimated fold-changes [5, 19]. Half- LOD (½ LOD) imputation follows a similar MNAR rationale but uses a single constant replacement (half the minimum observed density). It is simple, fast, and preserves the overall data structure, but it tends to reduce variance more strongly and can bias fold-changes [5, 8]. Both Gaussian and ½ LOD therefore belong to the family of "single-value" imputation approaches that do not use information from similar peptides. KNN imputation, by contrast, assumes MAR values and estimates them based on the similarity of peptides across samples. KNN is more flexible but it can over-smooth biologically relevant variation and it is sensitive to parameter settings [8, 9]. In our study, all three methods produced similar results, indicating that the missing values were mostly MNAR, consistent with earlier findings in label-free proteomics [32].

ComBat, a popular batch-effect correction method originally developed for microarray data, has also been widely adopted in proteomics due to its apparent robustness in small-sample settings. It applies empirical Bayes shrinkage to remove non-biological variation, but when the biological groups overlap with the batch structure, it risks removing true biological signal [11, 25]. MNN corrects batch effects by aligning shared structure across datasets using local neighborhoods, preserving more complex biological variation [10]. However, if batches are very small or uneven in size, MNN may compress the dataset's dimensionality, which can weaken its statistical power [10]. This limitation may also be relevant in our dataset, given the variation in batch sizes.

Although the here applied methods are widely used and supported by most analysis platforms, we acknowledge that additional more advanced imputation and batch-correction approaches exist, particularly those based on deep learning and more complex statistical modeling. For example, several deep learning methods have shown promising results in both transcriptomics and proteomics [33, 34]. In the context of imputation, other recent developments include statistical frameworks that either avoid imputation altogether or account for the variability introduced into downstream analyses [5, 35]. In this study, we chose to focus on well-known and user-friendly methods that are commonly used in practice and easy to implement.

Under the missingness structure and normalization strategy of this CE-MS dataset, commonly used imputation choices had limited influence on downstream statistical analysis results, whereas their interaction with batch-effect correction played a more marked role. Overall, our findings reinforce the importance of carefully evaluating any preprocessing step against the raw data, as such manipulations carry the risk of introducing artifacts and potentially distorting biological signals, ultimately preventing, rather than supporting the identification of significant features. Therefore, batch effect removal deserves particular consideration, following an investigation of the dataset's specific characteristics and the severity of batch effects in comparison to the expected biological variability.

5 | Concluding Remarks

In this urine-peptidomics dataset, the imputation of missing values (whether with Gaussian, ½ LOD or KNN) had little influence on the differential abundance analysis results under the missingness structure and normalization strategy applied. Batch-effect correction, on the other hand, reduced the number of peptides identified as significantly associated with the disease (CKD) without clear evidence of added benefit in this setting. Unguided ComBat was associated with a substantial reduction of the disease signal, whereas milder approaches like MNN or CKD-guided ComBat resulted in a moderate loss of DAPs which were otherwise detected and verified. Notably, similar imputation-only results could lead to markedly different outcomes once combined with batch-effect correction.

Our findings also indicate that imputation and batch correction do not act independently, and can interact in ways that influence downstream analysis and therefore they should be considered jointly rather than in isolation. In this dataset, the impact of batch correction depended on the strength of batch effect and the inclusion of relevant biological covariates in the model. In well-balanced datasets with many batches like ours,

a minimal preprocessing approach that focuses on addressing missing values was sufficient in order to avoid unnecessary loss of meaningful biological information. All things considered, our findings highlight the importance of adapting preprocessing strategies according to the features of each dataset, rather than depending on a single pipeline for every dataset. Additionally, since preprocessing choices can influence the results in ways that it are not always easy to predict, providing a minimally processed version of the data as supplementary material may help readers better assess the impact of these choices.

Associated Data

All data needed to reproduce the results of this study are provided in the accompanying Supporting Information files.

Acknowledgments

This article/publication is based upon work conducted within a Short-Term Scientific Action from COST Action CA21165 (PerMedik) supported by COST (European Cooperation in Science and Technology).

The publication of this article in OA mode was financially supported by HEAL-Link.

Conflicts of Interest

H.M. is the co-founder and co-owner of Mosaiques Diagnostics. A.L. is employed by Mosaiques Diagnostics. All other authors declare no conflict of interest.

References

1. A. Son, W. Kim, J. Park, et al., "Mass Spectrometry Advancements and Applications for Biomarker Discovery, Diagnostic Innovations, and Personalized Medicine," *International Journal of Molecular Sciences* 25, no. 18 (2024): 9880, https://doi.org/10.3390/ijms25189880.

2. A. M. Hawkridge and D. C. Muddiman, "Mass Spectrometry–Based Biomarker Discovery: Toward a Global Proteome Index of Individuality," *Annual Review of Analytical Chemistry* 2 (2009): 265–277, https://doi.org/10.1146/annurev.anchem.1.031207.112942.

3. D. Maillardet, "Missing Data," *Statistical Consulting Centre* (2024), https://scc.ms.unimelb.edu.au/resources/preparing-your-data/missing-data.

4. W. Kong, H. W. H. Hui, H. Peng, and W. W. B. Goh, "Dealing With Missing Values in Proteomics Data," *Proteomics* 22 (2022): 2200092, https://doi.org/10.1002/pmic.202200092.

5. C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies," *Journal of Proteome Research* 15, no. 4 (2016): 1116–1125, https://doi.org/10.1021/acs.jproteome.5b00981.

6. Y. Yu, Y. Mai, Y. Zheng, and L. Shi, "Assessing and Mitigating Batch Effects in Large-scale Omics Studies," *Genome Biology* 25, no. 1 (2024): 254, https://doi.org/10.1186/s13059-024-03401-9.

7. M. L. Gardner and M. A. Freitas, "Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics," *International Journal of Molecular Sciences* 22, no. 17 (2021): 9650, https://doi.org/10.3390/ijms22179650.

8. R. Wei, J. Wang, M. Su, et al., "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data," *Scientific Reports* 8, no. 1 (2018): 663, https://doi.org/10.1038/s41598-017-19120-0.

9. O. Troyanskaya, M. Cantor, G. Sherlock, et al., "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics* 17, no. 6 (2001): 520–525, https://doi.org/10.1093/bioinformatics/17.6.520.

10. L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch Effects in Single-cell RNA-sequencing Data Are Corrected by Matching Mutual Nearest Neighbors," *Nature Biotechnology* 36, no. 5 (2018): 421–427, https://doi.org/10.1038/nbt.4091.

11. W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods," *Biostatistics (Oxford, England)* 8, no. 1 (2007): 118–127, https://doi.org/10.1093/biostatistics/kxj037.

12. X. Hu, H. Li, M. Chen, J. Qian, and H. Jiang, "Reference-Informed Evaluation of Batch Correction for Single-Cell Omics Data With Overcorrection Awareness," *Communications Biology* 8, no. 1 (2025): 521, https://doi.org/10.1038/s42003-025-07947-7.

13. Y. Yu, N. Zhang, Y. Mai, et al., "Correcting Batch Effects in Large-Scale Multiomics Studies Using a Reference-Material-Based Ratio Method," *Genome Biology* 24, no. 1 (2023): 201, https://doi.org/10.1186/s13059-023-03047-z.

14. H. W. H. Hui, W. X. Chan, and W. W. B. Goh, "Assessing the Impact of Batch Effect Associated Missing Values on Downstream Analysis in High-Throughput Biomedical Data," *Briefings in Bioinformatics* 26, no. 2 (2025): bbaf168, https://doi.org/10.1093/bib/bbaf168.

15. H. W. H. Hui, W. Kong, H. Peng, and W. W. B. Goh, "The Importance of Batch Sensitization in Missing Value Imputation," *Scientific Reports* 13, no. 1 (2023): 3003, https://doi.org/10.1038/s41598-023-30084-2.

16. H. Mischak, A. Vlahou, and J. P. A. Ioannidis, "Technical Aspects and Inter-Laboratory Variability in Native Peptide Profiling: The CE-MS Experience," *Clinical Biochemistry* 46, no. 6 (2013): 432–443, https://doi.org/10.1016/j.clinbiochem.2012.09.025.

17. A. Latosinska, J. Siwy, H. Mischak, and M. Frantzi, "Peptidomics and Proteomics Based on CE-MS as a Robust Tool in Clinical Application: The Past, the Present, and the Future," *Electrophoresis* 40 (2019): 2294–2308, https://doi.org/10.1002/elps.201900091.

18. J. Jantos-Siwy, E. Schiffer, K. Brand, et al., "Quantitative Urinary Proteome Analysis for Biomarker Evaluation in Chronic Kidney Disease," *Journal of Proteome Research* 8, no. 1 (2009): 268–281, https://doi.org/10.1021/pr800401m.

19. L. Harris, W. E. Fondrie, S. Oh, and W. S. Noble, "Evaluating Proteomics Imputation Methods With Improved Criteria," *Journal of Proteome Research* 22, no. 11 (2023): 3427–3438, https://doi.org/10.1021/acs.jproteome.3c00205.

20. S. Tyanova, T. Temu, P. Sinitcyn, et al., "The Perseus Computational Platform for Comprehensive Analysis of (prote)Omics Data," *Nature Methods* 13, no. 9 (2016): 731–740, https://doi.org/10.1038/nmeth.3901.

21. *Statistical Analysis With Missing Data*, 2nd Edition (Wiley, 2025).

22. S. Zhang, "Nearest Neighbor Selection for Iteratively kNN Imputation," *Journal of Systems and Software* 85, no. 11 (2012): 2541–2552, https://doi.org/10.1016/j.jss.2012.05.073.

23. A. Kowarik and M. Templ, "Imputation With the R Package VIM," *Journal of Statistical Software* 74 (2016): 1–16, https://doi.org/10.18637/jss.v074.i07.

24. ComBat: Adjust for batch effects using an empirical Bayes framework in sva: Surrogate Variable Analysis. (n.d.). Retrieved 1 May 2025, from https://rdrr.io/bioc/sva/man/ComBat.html.

25. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva Package for Removing Batch Effects and Other Unwanted Variation in High-throughput Experiments," *Bioinformatics* 28, no. 6 (2012): 882–883, https://doi.org/10.1093/bioinformatics/bts034.

26. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. (n.d.). Portal de Periódicos da CAPES. Retrieved 30 April 2025, from https://www.periodicos.capes.gov.br/index.php/acervo/buscador.html?task=detalhes&id=W2794521141.

27. Batch Effects in Single-cell RNA-sequencing Data Are Corrected by Matching Mutual Nearest Neighbors | *Nature Biotechnology* (n.d.). Retrieved 29 January 2026, from https://www.nature.com/articles/nbt.4091.

28. Jaccard Similarity—An overview | ScienceDirect Topics. (n.d.). Retrieved 5 May 2025, from https://www.sciencedirect.com/topics/computer-science/jaccard-similarity.

29. I. Belczacka, A. Latosinska, J. Siwy, et al., "Urinary CE-MS Peptide Marker Pattern for Detection of Solid Tumors," *Scientific Reports* 8, no. 1 (2018): 5227, https://doi.org/10.1038/s41598-018-23585-y.

30. J. Siwy, W. Mullen, I. Golovko, J. Franke, and P. Zürbig, "Human Urinary Peptide Database for Multiple Disease Biomarker Discovery," *PROTEOMICS—Clinical Applications* 5, no. 5–6 (2011): 367–374, https://doi.org/10.1002/prca.201000155.

31. D. M. Good, P. Zürbig, À. Argilés, et al., "Naturally Occurring Human Urinary Peptides for Use in Diagnosis of Chronic Kidney Disease," *Molecular & Cellular Proteomics: MCP* 9, no. 11 (2010): 2424–2437, https://doi.org/10.1074/mcp.M110.001917.

32. M. Li and G. K. Smyth, "Missing Values Are Informative in Label-free Shotgun Proteomics Data: Estimating the Detection Probability Curve," *bioRxiv* (2022): 2022–2007, https://doi.org/10.1101/2022.07.02.498573.

33. X. Yu, X. Xu, J. Zhang, and X. Li, "Batch Alignment of Single-cell Transcriptomics Data Using Deep Metric Learning," *Nature Communications* 14, no. 1 (2023): 960, https://doi.org/10.1038/s41467-023-36635-5.

34. H. Webel, L. Niu, A. B. Nielsen, et al., "Imputation of Label-free Quantitative Mass Spectrometry-based Proteomics Data Using Self-supervised Deep Learning," *Nature Communications* 15, no. 1 (2024): 5405, https://doi.org/10.1038/s41467-024-48711-5.

35. M. Medo, D. M. Aebersold, and M. Medová, "ProtRank: Bypassing the Imputation of Missing Values in Differential Expression Analysis of Proteomic Data," *BMC Bioinformatics [Electronic Resource]* 20, no. 1 (2019): 563, https://doi.org/10.1186/s12859-019-3144-3.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File 1: pmic70111-sup-0001-FigureS1.pdf.
Supporting File 2: pmic70111-sup-0002-FigureS2.pdf.
Supporting File 3: pmic70111-sup-0003-FigureS3.pdf.
Supporting File 4: pmic70111-sup-0004-FigureS4.pdf.
Supporting File 5: pmic70111-sup-0005-FigureS5.pdf.
Supporting File 6: pmic70111-sup-0006-FigureS6.pdf.
Supporting File 7: pmic70111-sup-0007-FigureS7.pdf.
Supporting File 8: pmic70111-sup-0008-FigureS8.pdf.
Supporting File 9: pmic70111-sup-0009-FigureS9.pdf.
Supporting File 10: pmic70111-sup-0010-FigureS10.pdf.
Supporting File 11: pmic70111-sup-0011-FigureS11.pdf.
Supporting File 12: pmic70111-sup-0012-FigureS12.pdf.
Supporting File 13: pmic70111-sup-0013-FigureS13.pdf.
Supporting File 14: pmic70111-sup-0014-FigureS14.pdf.
Supporting File 15: pmic70111-sup-0015-TableS1.pdf.
Supporting File 16: pmic70111-sup-0016-TableS2.pdf.
Supporting File 17: pmic70111-sup-0017-TableS3.pdf.
Supporting File 18: pmic70111-sup-0018-TableS4.pdf.
Supporting File 19: pmic70111-sup-0019-TableS5.pdf.
Supporting File 20: pmic70111-sup-0020-TableS6.pdf.
Supporting File 21: pmic70111-sup-0021-TableS7.pdf.