nature communications

Article

https://doi.org/10.1038/s41467-025-64718-y

Protein-level batch-effect correction enhances robustness in MS-based proteomics

Received: 25 February 2025

Accepted: 25 September 2025

Published online: 04 November 2025

Check for updates

Qiaochu Chen[1], Zehui Cao[1], Yaqing Liu[1], Naixin Zhang[1,5], Yanming Xie[1], Haonan Chen[1], Yuanbang Mai[1], Shumeng Duan[1], Jiaqi Li[1], Ying Yu[1], Yang Zhao[2], Leming Shi[1,3,4] ✉ & Yuanting Zheng[1,3] ✉

Batch effects, defined as unwanted technical variations caused by differences in labs, pipelines, or batches, are notorious in MS-based proteomics data, wherein protein quantities are inferred from precursor- and peptide-level intensities. However, the optimal stage for batch-effect correction remains elusive and crucial. Leveraging real-world multi-batch data from the Quartet protein reference materials and simulated data, we benchmark batch-effect correction at precursor, peptide, and protein levels combined across two designed scenarios (balanced and confounded), three quantification methods (MaxLFQ, TopPep3, and iBAQ), and seven batch-effect correction algorithms (Combat, Median centering, Ratio, RUV-III-C, Harmony, WaveICA2.0, and NormAE). Our findings reveal that protein-level correction is the most robust strategy, and the quantification process interacts with batch-effect correction algorithms. Furthermore, we extend our analysis to large-scale data from 1431 plasma samples of type 2 diabetes patients in Phase 3 clinical trials, demonstrating the superior prediction performance of the MaxLFQ-Ratio combination. These findings support that batch-effect correction at the protein level enhances multi-batch data integration in large proteomics cohort studies.

Batch effects refer to unwanted variations resulting from technical factors that disturb the biological signals of interest[1–3]. Although the liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) system enables profiling of thousands of samples in large-scale proteomics studies[4–10], the long-term period—lasting for several days, months, or even years—of data generation usually involves multiple reagent or running batches, platform or instrument types, operators, as well as collaborating labs[11,12]. Therefore, the complexity of experimental and analytical procedures in MS-based large-scale proteomics data may lead to batch effects confounded with various factors of interest, thus challenging the reproducibility and reliability of proteomics studies[11]. Meanwhile, MS-based proteomics utilizes the bottom-up strategy[13], in which protein-expression quantities are inferred by a specified quantification method (QM)[14–16] from the extracted ion current (XIC) intensities[17] of multiple peptides assigned to a protein group, usually around three-to-ten peptides per protein group. In addition, precursors are defined as peptides with specific charge states or modifications, and are identified earlier than peptides and proteins. Therefore, it is crucial to consider at which data level to correct batch effects. That is, whether batch-effect correction should be performed earlier (precursor or peptide-level) or later than protein quantification (protein-level) before conducting the intended

[1]State Key Laboratory of Genetics and Development of Complex Phenotypes, Human Phenome Institute and School of Life Sciences, Fudan University, Shanghai, China. [2]National Institute of Metrology, Beijing, China. [3]International Human Phenome Institute (Shanghai), Shanghai, China. [4]Cancer Institute, Shanghai Cancer Center, Fudan University, Shanghai, China. [5]Present address: Department of Public Health and Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ✉e-mail: lemingshi@fudan.edu.cn; zhengyuanting@fudan.edu.cn

explorations on biological properties and functions at the aggregated protein level.

Many batch-effect correction algorithms (BECAs)[12,18–26] have been proposed for removing batch effects. For example, normalization based on means or medians within each batch has been widely used in proteomics data preprocessing[27]. ComBat was first used to modify the mean shift of microarray data across batches by an empirical Bayesian method, and can be applied to proteomics data[28]. RUV-III-C employs a linear regression model to estimate and remove unwanted variation in raw intensities, thereby correcting for batch effects in proteomics data[12]. Ratio of intensities of study samples divided by those of concurrently profiled universal reference materials on a feature-by-feature basis improves cross-batch integration for multi-omics studies[1,22,26,29–32]. Based on principal component analysis (PCA), Harmony[33–35] iteratively clusters cells by similarity and calculates a cluster-specific correction factor to remove batch effects in single-cell RNA sequencing (scRNA-seq) data; it can also be extended to multi-omics data[32]. In large-scale proteomics studies, Čuklina et al. suggested fitting a Locally Estimated Scatterplot Smoothing (LOESS) curve within each batch to fix the injection order-specific MS signal drifts before correcting the discrete batch effect[23]. Recently, BECAs developed for large-scale metabolomics data may generalize to MS-based proteomics. WaveICA[36] (or WaveICA2.0[37]) extracts and removes batch effects by multi-scale decomposition with the time trend of injection orders. Another deep learning-based BECA named NormAE[38] corrects non-linear batch-effect factors learned from neural networks. Above all, the choice of a specific BECA from the overwhelming possibilities is challenging for the end user as different studies may provide conflicting recommendations due to study constraints[1].

Benchmarking studies have provided objective insights in the selection of BECAs for data quality improvement, performing at either the uniformed peptide or protein data level without yet considering the impact of protein quantification methods from the peptide or precursor level[32,39–42]. Callister et al. concluded that linear regression performed well among four BECAs at the peptide level in three high-throughput proteomics datasets[40]. Kultima et al. benchmarked 10 BECA methods with three peptide-level datasets and recommended the use of linear regression to fix experimental order-specific variations[41]. Chawade et al. introduced Normalyzer, a tool for the selection of BECAs adapted to a specific dataset, compatible with multi-type quantitative omics data matrices[42]. Välikangas et al. expanded performance assessment from intra-group variations to inter-group protein differential expressions, emphasizing the effectiveness of linear regression and LOESS methods at the protein level[39]. Yu et al. comprehensively assessed seven BECAs on protein-level proteomics datasets based on the Quartet (中华家系1号) reference materials[32], and found ratio-based scaling to be a universally effective BECA, especially when batch effects are confounded with biological groups of interest[22,26,30,31]. Arend et al. benchmarked 17 normalization methods and 2 batch-effect correction methods at the aggregated protein level, providing an evaluation tool, PRONE for the community[43].

These benchmarking analyses of batch-effect correction in proteomics have broadened our knowledge in the selection of BECAs and also the performance assessment metrics, bridging the gap of reproducibility to some extent. However, the inconsistency of batch-effect correction at the data (i.e., peptide or protein) levels still challenges the harmonization of MS-based proteomics studies. Currently, there is a lack of comprehensive benchmarking studies on batch-effect correction at the precursor, peptide, and protein levels, as well as the adaptation for both multi-lab and large-scale multi-batch scenarios.

Here, we utilize the simulated dataset as well as the real-world multi-lab datasets derived from the Quartet protein reference materials[22] as two typical benchmarking datasets for performance assessments on three (i.e., precursor-, peptide-, and protein-level) batch-effect correction strategies. We evaluated the data matrices at the final aggregated protein level by using both feature-based and sample-based metrics under two designed scenarios (balanced and confounded), combined with three QMs (MaxLFQ, TopPep3, and iBAQ) and seven BECAs (Combat, Median centering, Ratio, RUV-III-C, Harmony, WaveICA2.0, and NormAE). Finally, we also tested the performance of batch-effect removal on a case study with a large-scale proteomics dataset from a type 2 diabetes (T2D) cohort. Our study provides a comprehensive benchmarking of batch-effect correction strategies, revealing better performances of late-data-level workflows in MS-based proteomics.

Results

Overview of the study design

The multi-omics Quartet Project provides the community with multi-batch datasets[1,22,26,29–32,44–49] generated from four grouped reference materials (D5, D6, F7, and M8). For proteomics, each Quartet dataset consists of 12 MS runs from triplicates of tryptic proteins in the fixed injection order. To test over-corrections and false discoveries, we also generated a simulated data matrix with built-in truth comprising triplicates of three biological groups distributed to three batch groups, respectively. Considering that batch effects are usually confounded with biological factors of interest in real-world studies, especially for studies without full randomization or confounded by nature for follow-up analyses, we designed two scenarios, wherein either the known sample groups are balanced (termed Quartet-B, Simulated-B) or confounded with (termed Quartet-C, Simulated-C) among batches. In this study, we leveraged both the simulated dataset and six Quartet datasets as the benchmarks for the following workflows to compare three batch-effect correction strategies, i.e., precursor-, peptide-, and protein-level corrections, in removing unwanted variations while retaining robust biological signals (Fig. 1a).

We focused on the comparison of protein-quantity matrices preprocessed by batch-effect corrections at different data levels, while considering other potential impact factors that may alter the effectiveness of removing the batch effects. Depending on whether or not batch-effect correction occurs and the timing of that correction, our workflow was constructed on three preprocessing strategies: batch-effect corrections at (1) precursor, (2) peptide, and (3) protein levels, with corresponding quantifications that aggregate data from early levels into late levels. The preprocessing strategies were integrated with three QMs (MaxLFQ, TopPep3, and iBAQ) and seven BECAs (Combat, Median centering, Ratio, RUV-III-C, Harmony, WaveICA2.0, and NormAE) on the benchmarking datasets. NormAE requires the input of mass-to-charge ratio (m/z) and retention time (RT) for each feature; it is theoretically available only at the precursor level in the Quartet datasets. For the main benchmarking tests using the other six BECAs, we generated 57 (54 corrected and 3 uncorrected) data matrices at the aggregated protein level, 39 (36 corrected and 3 uncorrected) data matrices at the aggregated peptide level, and 7 (6 corrected and 1 uncorrected) data matrices at the raw precursor level for each design (Quartet-B, Quartet-C, Simulated-B and Simulated-C), respectively (Fig. 1b).

We evaluated the performance of protein profiles from two perspectives, feature-based and sample-based (Fig. 1c). For feature-based quality assessment, we calculated the coefficient of variation (CV) within all technical replicates across different batches for each feature. The simulated data matrix was composed of known feature expression patterns between every two sample groups, thus enabling the assessment of identified DEPs in each data matrix by the Matthews correlation coefficient (MCC) and the Pearson correlation coefficient (RC). For sample-based quality assessment, the signal-to-noise ratio (SNR) can evaluate the resolution in differentiating known Quartet sample groups based on PCA. Sample-based metrics also include the quantified contributions by biological or batch factors through principal variance component analysis (PVCA).

Furthermore, we employed a proteomics dataset (under the ChiHOPE project, https://www.biosino.org/node/project/detail/OEP00002924) comprising 1,431 plasma samples from type 2 diabetes (T2D) patients as a case study. This aimed to demonstrate the effectiveness of batch-effect correction in large-scale scenarios (subsequently termed ChiHOPE) (Fig. 1d). In this large-scale study, three types of QC samples were profiled alongside the study samples for batch-effect monitoring: 16 plasma samples from a healthy male (P10), 16 plasma samples from a healthy female (P11), and 32 pooled plasma samples from the mixture of all study samples (PM). Finally, we evaluated the prediction performances of both categorical variables (i.e., sex) and continuous variables (i.e., Age) after batch-effect correction.

In summary, we conducted a comprehensive benchmarking study, considering multiple scenarios, BECAs, QMs, and various assessment metrics on the optimal timing for proteomics batch-effect correction.

Batch effects existed in uncorrected data matrices at all data levels

To demonstrate the common occurrence of batch effects and the necessity of batch-effect correction, we used feature-based and sample-based metrics to assess batch-related variations in uncorrected data matrices at the precursor, peptide, and quantified protein levels (Fig. 2; Supplementary Fig. 1).

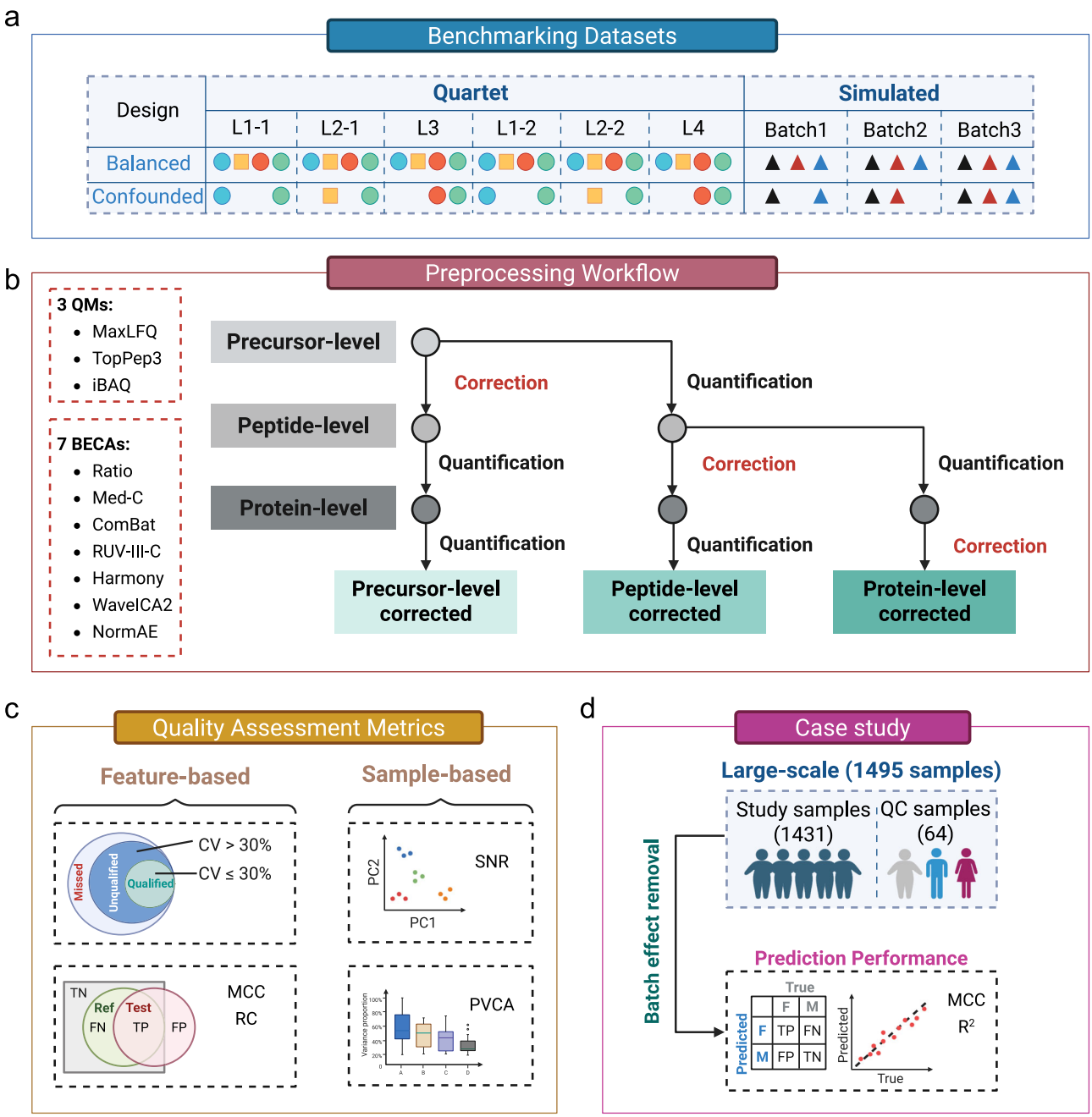


Fig. 1 | Overview of the study design. a Two benchmarking datasets: Quartet and simulated datasets; each point represents for the triplicates of the sample group; there are a total of 72, 36, 27, 21 samples in Quartet-balanced (Quartet-B), Quartet-confounded (Quartet-C), Simulated-balanced (Simulated-B), and Simulated-confounded (Simulated-C) scenarios. **b** three batch-effect correction strategies combined with scenarios, quantitation methods (QMs) and batch-effect correction algorithms (BECAs). **c** feature-based and sample-based quality assessment metrics. **d** A case study: the large-scale T2D-cohort datasets and the metrics for prediction performance assessments. Created in BioRender. Zheng, Y. (2025) https://BioRender.com/089aon4.

https://doi.org/10.1038/s41467-025-64718-y

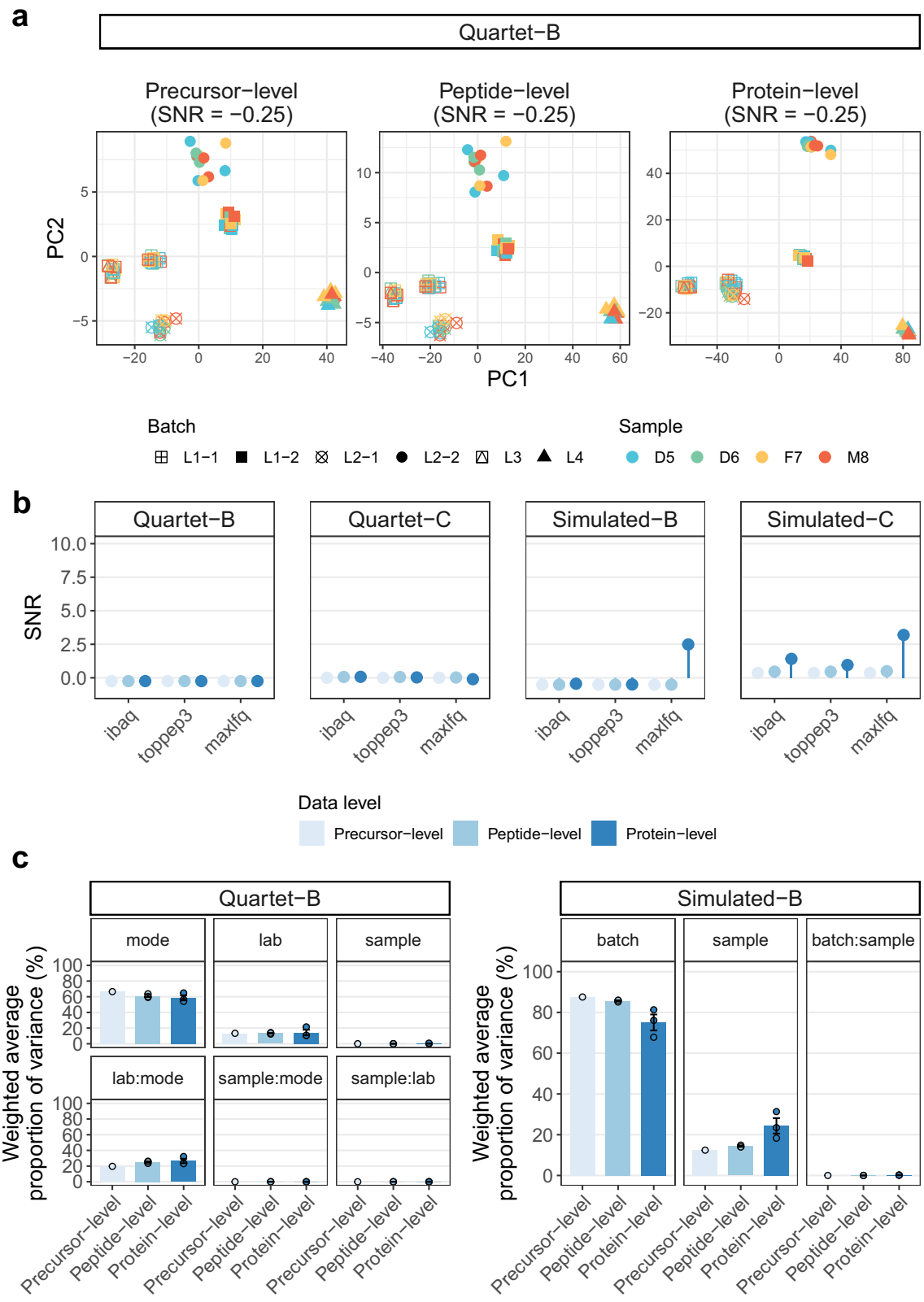


Fig. 2 | Batch effects existed in uncorrected data matrices at all data levels. **a** scatter plots of principal component analysis (PCA) results of the uncorrected precursor-, peptide-, and protein-level data matrices after MaxLFQ protein quantification; shaped by the raw batch; colored by the sample type. **b** bar plots of signal-to-noise ratio (SNR) performance in uncorrected data matrices across all three data levels under four designed scenarios; colored by the data level; faceted by the scenario. **c** bar plots of principal variance component analysis (PVCA) performance in uncorrected data matrices across all three data levels under Quartet-balanced (Quartet-B) (left) and Simulated-balanced (Simulated-B) (right) scenarios; presented as mean values ± s.d. colored by the data level; faceted by the scenario. All data points are plotted individually. For each bar plot: Precursor-level ($n = 1$), Peptide-level ($n = 3$), Protein-level ($n = 3$). All statistics are based on analysis replicates (by quantitation methods). Source data are provided as a Source Data file.

Without batch-effect correction, poor quantity consistency was observed within the same sample group across different batches (Supplementary Fig. 1a). Specifically, mean CV increased from the precursor level (Quartet-B mean ± sd: 0.70 ± 0.45; Quartet-C mean ± sd: 0.53 ± 0.41) to the protein level (Quartet-B mean ± sd: 1.38 ± 0.80; Quartet-C mean ± sd: 0.95 ± 0.66) in Quartet data matrices (Supplementary Fig. 1a). As expected, simulated data CV demonstrated a stepwise increase with introduction of a biological factor and random noise (WithBiol), followed by a scaling factor (WithScal), and finally a batch factor (WithBatch) added to the initial gamma-distributed matrix (Supplementary Fig. 1b). In contrast to Quartet's complex design, simulated dataset exhibited CV decrease from precursor to protein level (Supplementary Fig. 1c). Fold changes for known features were calculated at each level in simulated data, and features with fold changes opposite to expected values or with negligible differences were classified as false discoveries. Most uncorrected data matrices affected by batch effects yielded near-zero MCC values (Supplementary Fig. 1d, e). Notably, only MaxLFQ extracted biological information, improving MCC from −0.04 to 0.41 (Supplementary Fig. 1e).

Using SNR and PVCA metrics based on PCA, we elucidated how batch effects dominated data variability. The step-by-step simulation demonstrated how technical factors progressively outweighed biological signals (Supplementary Fig. 2). Before correction, samples clustered by batch rather than biological group at all three data levels (Fig. 2a; Supplementary Fig. 3). Thus, uncorrected data matrices could not distinguish biological groups, exhibiting near-zero or negative SNR values, while MaxLFQ improved SNR from −0.51 (Simulated-B) and 0.36 (Simulated-C) at the precursor level to 2.49 (Simulated-B) and 3.19 (Simulated-C) at the protein level (Fig. 2b). PVCA revealed the dominant variance contributor. Given Quartet's complex batch sources, data acquisition mode, and laboratory emerged as dominant factors over biological factors (Fig. 2c; Supplementary Fig. 4a). In simulated data, batch factor contribution decreased from precursor to protein levels, while sample factor contribution increased (Supplementary Fig. 4b).

Protein-level correction exhibited robust reproducibility and reliability

We then conducted quality assessment of the profiles using feature-based metrics, focusing on intra-sample-group reproducibility and inter-sample-group reliability of quantified proteins across different batches. From precursor- to protein-level corrected data, we observed a consistent decrease in the overall CVs in all designed scenarios. Specifically, in the Quartet-B scenario, the CV median dropped from 1.34 (uncorrected) to 0.40 (protein-corrected) after MaxLFQ quantification (Fig. 3a), and similarly when quantified by iBAQ or TopPep3. This finding suggested that batch-effect removal at higher data levels improved reproducibility within the same sample groups. Most BECAs exhibited a similar pattern except WaveICA2, which demonstrated the least effective CV performance. The BECA Ratio, when corrected at the protein level, yielded the lowest overall CV (mean ± sd: 0.31 ± 0.32) (Supplementary Fig. 5a).

In simulated scenarios, we compared the fold changes (Group2/Group1 and Group3/Group1) after correction to true values and calculated the MCC metric for each corrected combination. Among the three QMs, iBAQ and TopPep3 did not improve MCC performance, whereas BECAs combined with MaxLFQ displayed a significant increase in MCC, particularly at the protein level under balanced design. Notably, MaxLFQ-quantified and peptide-level corrected data matrices achieved the highest MCC (mean ± sd: 0.56 ± 0.20) under confounded design (Supplementary Fig. 5b). Most BECAs could effectively enhance MCC values at the protein level following MaxLFQ quantification, with RUV-III-C demonstrating the best performance across all three data levels (Fig. 3b). Additionally, under confounded designs, Ratio and ComBat also obtained high MCC values (0.68 and 0.67, respectively) at the peptide level. Scatter plots further elucidated the accuracy of corrected values, indicating that RUV-III-C introduced the least over-correction variability (Fig. 2c).

Nevertheless, batch-effect correction at the protein level was robust to most BECAs; the accuracy of expression patterns could be optimized through a specific combination of QM and BECA.

MaxLFQ unmasked biological signals in protein-level batch-effect correction

The signal-to-noise ratio (SNR), a metric defined in the multi-omics Quartet Project, was employed to assess both the ability to discriminate among different sample groups (specifically D5, D6, F7, and M8) and the technical variations within the same biological group. For sample-based performance assessment, we extended the application of SNR from Quartet to simulated scenarios[50]. Our findings revealed that protein-level corrected data matrices achieved the highest overall SNR values compared to peptide- or precursor-level corrected data across all designed scenarios. This improvement was especially striking under balanced designs (medians: from −0.02 to 12.32 in Quartet-B; from 3.77 to 4.91 in Simulated-B) (Fig. 4a). Furthermore, RUV-III-C, ComBat, and Ratio applied at the protein level yielded high SNR values (mean ± sd: 18.14 ± 8.38, 12.07 ± 6.16, and 11.27 ± 6.74, respectively) (Supplementary Fig. 6a), and MaxLFQ outperformed the other two QMs in SNR when combined with protein-level corrections (Supplementary Fig. 6b). Next, we displayed detailed SNR values for every combination under each design (Fig. 4b; Supplementary Fig. 7). Most BECAs applied to MaxLFQ-quantified protein-level data exhibited high SNRs, especially under balanced designs (Fig. 4b; Supplementary Fig. 7b). Consistently, MaxLFQ-quantified protein-level corrections using RUV-III-C, ComBat, and Ratio achieved higher SNRs than other BECAs under all designed scenarios. The robustness of MaxLFQ in protein-level batch-effect removal was confirmed by the fact that 80% of combinations of QMs and BECAs (20 out of 25) reached the highest SNR values (Fig. 4b; Supplementary Fig. 7). In PCA plots, most BECAs at MaxLFQ-quantified protein-level data enabled samples to cluster by true biological groups (Supplementary Fig. 8).

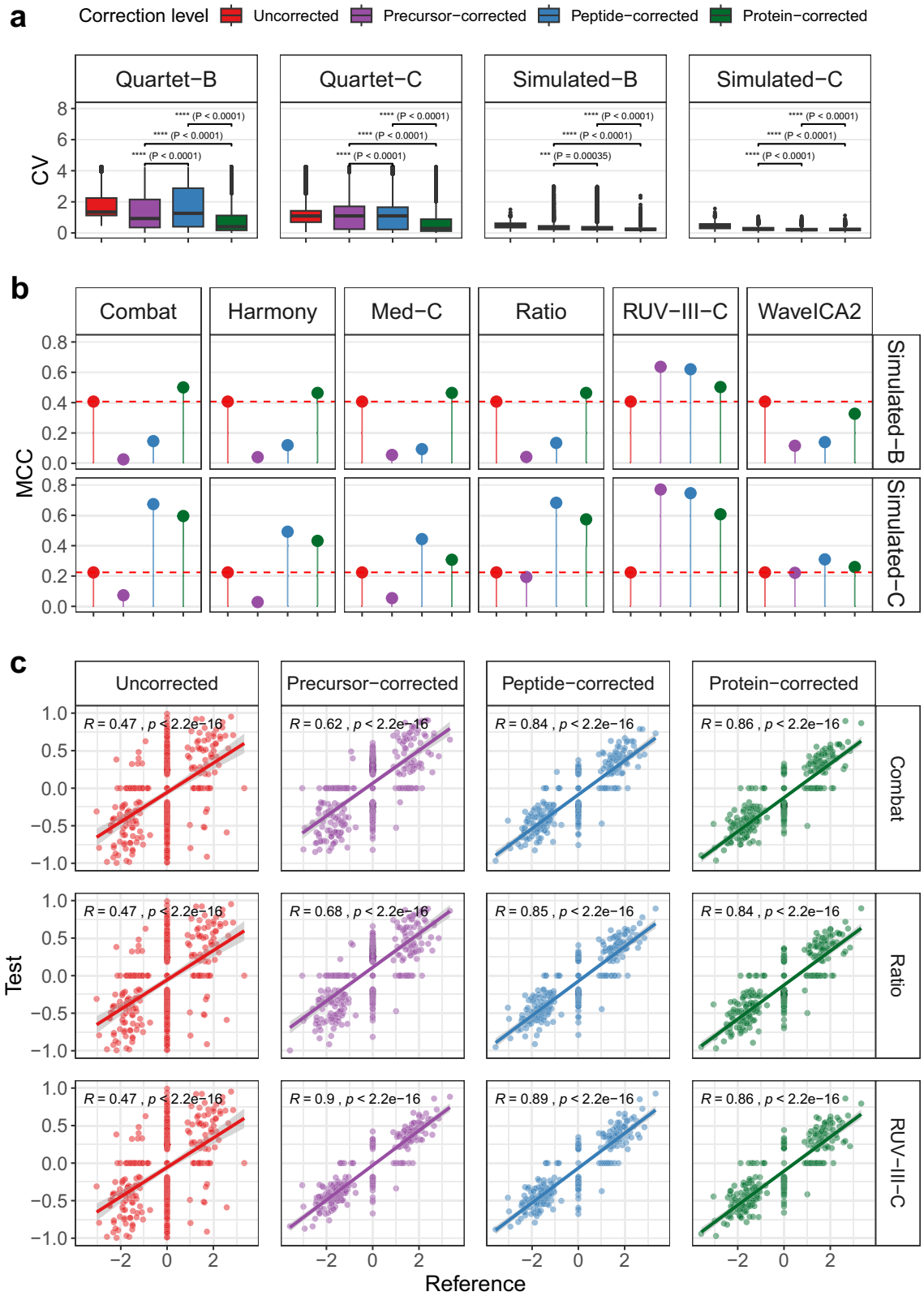
PVCA was used to quantify the contribution of biological and technical factors to data variability. As shown, technical factors were effectively reduced by batch-effect correction at the protein level, alongside increasing effects from the sample factor. Under confounded scenarios, however, batch-sample interactions were slightly enhanced, albeit remaining relatively low (Supplementary Fig. 9). Summary PVCA results under the simulated confounded scenario illustrated the robustness of MaxLFQ when combined with various correction levels, QMs, and BECAs. Consistent with the previous findings, RUV-III-C, ComBat, and Ratio at the MaxLFQ-quantified level ranked highest among all BECAs (Fig. 4c).

Overall, MaxLFQ-quantified protein-level batch-effect corrections surpassed other combinations in sample-based metrics.

Quantification interacted with correction from precursor to protein levels

Feature-based and sample-based quality assessments suggested that correction at the precursor- or peptide-level did not effectively remove batch effects in most combinations. Therefore, we traced back to precursor and the peptide data levels and reassessed all metrics at both the corrected (post-correction) and aggregated (post-quantification) data levels (Fig. 5). At all three data levels, the correction step effectively reduced overall CVs under balanced or confounded designs (Supplementary Fig. 10a), while this reduction could be neutralized or reversed after quantification, particularly under balanced designs (Fig. 5a). This reversal was also observed for SNR performance, which increased after correction but decreased after quantification (Supplementary Fig. 10b; Fig. 5b).

https://doi.org/10.1038/s41467-025-64718-y



Strikingly, examination of FN, FP, TN, and TP counts revealed that correction alone did not significantly increase true discoveries or decrease false discoveries, especially under balanced designs (Supplementary Fig. 10c). In contrast, quantification increased TN and TP counts stepwise from precursor to protein levels, most pronounced at aggregated levels. Conversely, false discoveries showed an ascending trend in FNs with higher aggregation (Fig. 5c). In confounded scenarios, peptide-level corrections yielded the highest true discoveries and lowest false discoveries after to protein-level aggregation, consistent with previous observations (Fig. 3b). PVCA results further suggested quantification amplified batch effects (Fig. 3c), despite reduction at corrected levels (Supplementary Fig. 10d).

In summary, while correction effectively mitigates batch effects, quantification plays a critical, complex role in shaping final data

Fig. 3 | Protein-level correction exhibited robust reproducibility and reliability. **a** box plots of the coefficient of variation (CV) of proteins across different designed scenarios; the P values were calculated using unpaired two-tailed student's *t* tests, $^{****}P < 0.0001$, $^{**}P < 0.01$, $^{*}P < 0.05$; exact *P* values rounded to two significant figures (when not less than 0.0001) are provided above each comparison; colored by the correction level; faceted by the scenario; horizontal lines indicate the median; box boundaries indicate the interquartile range (IQR); whiskers represent values within 1.5× IQR of the first and third quartiles; data points beyond the end of the whiskers are plotted individually. Detailed n statistics are provided (Supplementary Table 2). **b** bar plots of the Matthews correlation coefficient (MCC) values in simulated data; colored by the correction level; faceted by the BECA (by column) and the scenario (by row); data points are plotted individually; the dashed lines represent the Matthews correlation coefficient (MCC) value in the uncorrected data matrix. **c** scatter plots of fold changes (FCs) in the tested (uncorrected or corrected) data matrices and the preset true FCs; the axes are $\log_2$ scaled; the solid lines represent fitted curves from linear regression along with the Pearson correlation coefficient (R) and the corresponding P value rounded to two significant figures; 95% confidence interval (grey ribbon) are displayed. All statistics are based on independent biological replicates (by sample groups) and analysis replicates (by BECAs) before or after MaxLFQ quantification. Source data are provided as a Source Data file.

quality, particularly influencing true discoveries and potentially reintroducing batch-related variability.

Ratio provided the best prediction performance in large-scale data

Finally, we tested protein-level batch-effect correction on large-scale ChiHOPE proteomics data (Fig. 6). The batch effects contained multi-sourced technical noise, manifesting as both injection order-specific MS signal drift and discrete whole-batch shifts. Across all samples, peptide intensities showed discontinuities at instrument cleaning timepoints (dividing the samples into three batches) and exhibited decaying trends with injection order within batches (Supplementary Fig. 11). QC samples (PM, P10, and P11) revealed similar drift patterns to study samples. Discrete batch effects were confirmed by PCA (Supplementary Fig. 12a, b). Substantial batch effects obscured biological differences of interest, emphasizing the need for effective preprocessing before downstream analysis. Based on previous conclusions, we tested BECAs with/without LOESS correction at MaxLFQ-quantified protein level, confirming fundamental batch-effect removal with LOESS. This was supported by true-label clustering in UMAP (using PCs explaining 80% variance) with minimal residual batch effects (Supplementary Fig. 12).

To investigate whether data after protein-level batch-effect corrections allow better prediction of biological groups, we tested all protein profiles from the ChiHOPE cohort using Random Forest. Categorical sex groups and continuous age values served as classification and regression endpoints, respectively. And the negative control endpoints were set by randomly shuffling the sex or the age corresponding to the samples. We divided each protein profile into training or validation set according to the inclusion time of the total 750 subjects in the clinical trials and performed 5-fold cross validation within the training set (757 samples from 394 subjects), then we locked the final evaluation results on the validation set (674 samples from 356 subjects). We calculated MCC and R^2 values to evaluate the performance of the sex and the age prediction, respectively, in the external validation set. As expected, the negative control data matrix was unpredictable with the MCC or R^2 near zero, while the uncorrected and LOESS-corrected alone data matrix remained weak prediction performances with MCC lower than the 0.3 threshold (Fig. 6a). All protein-level batch-effect corrections improved prediction performance (Fig. 6a; 6c), demonstrating the robustness of MaxLFQ quantification. ComBat and Median centering increased the number of FP features in sex prediction (Fig. 6b). For more complex endpoints like age, all corrected data matrices achieved slightly higher R^2 values than uncorrected/LOESS-alone data (Fig. 6c, d). Among BECAs, Ratio preserved maximal data variability for prediction, achieving highest MCC (0.41) and R^2 (0.19). Furthermore, BECAs after LOESS (LOESS-dependent) contributed to significantly higher MCC performances of the corrected data matrices (Supplementary Fig. 13).

We recommend protein-level batch-effect correction in MS-based proteomics with MaxLFQ quantification and Ratio BECA to maximize biological signal extraction while minimizing technical noise.

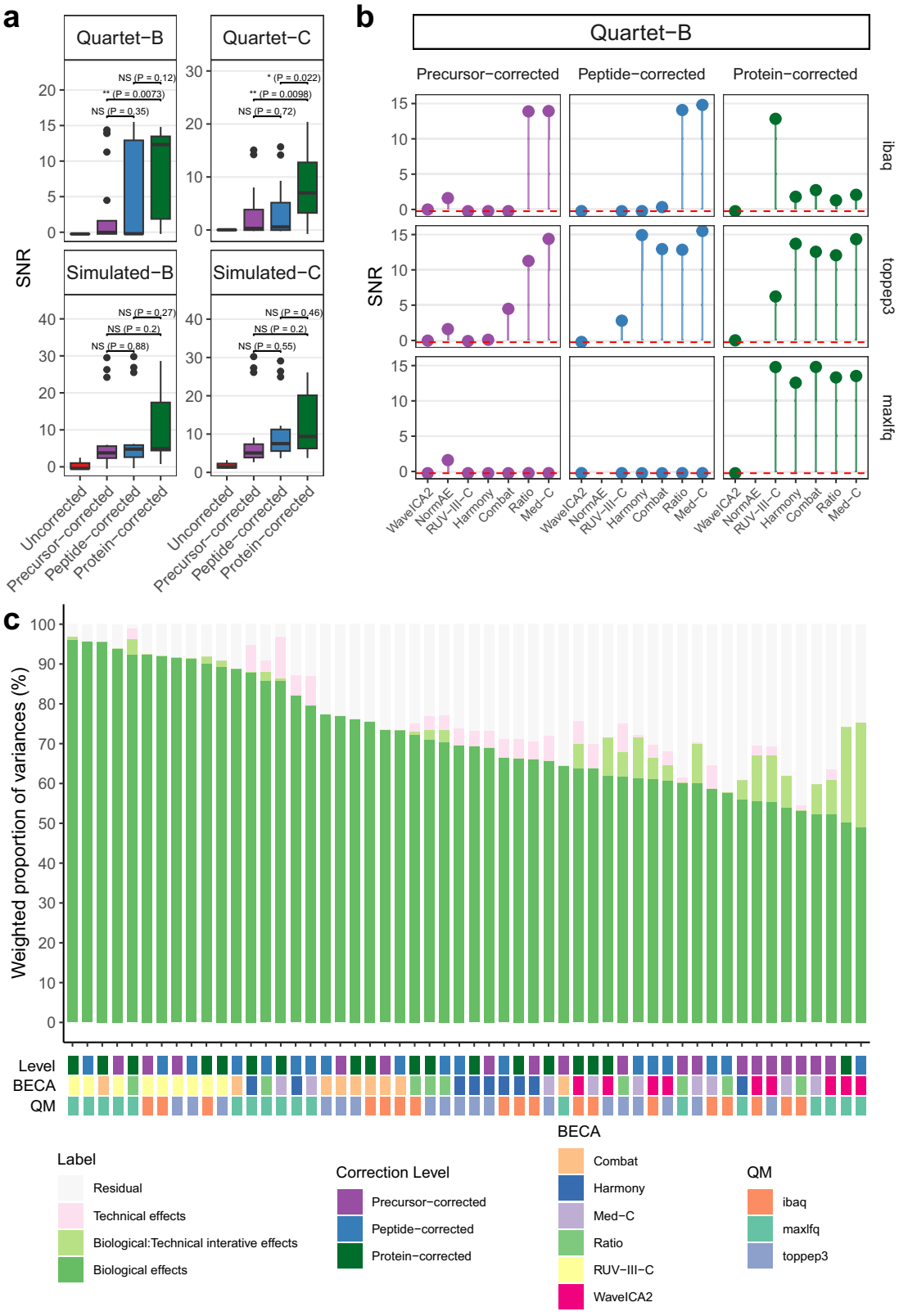
Discussion

To provide evidence on the impact of the timing of protein quantification relative to batch-effect correction, we performed a comprehensive benchmarking study using both the Quartet multi-batch datasets and the simulated datasets with ground truth and tested the combinations between three PQMs (MaxLFQ, iBAQ, and TopPep3) and seven BECAs (Ratio, Median centering, RUV-III-C, ComBat, Harmony, WaveICA2.0, and NormAE) on two designed scenarios (balanced and confounded). We embedded both feature-based (CV and MCC) and sample-based (SNR and PVCA) QC metrics in our quality assessment. Before batch-effect correction, data matrices at all three levels showed overwhelming batch effects, which confused biological signals with technical noise. Our findings show that most protein-level batch-effect corrections effectively remove unwanted technical variations, retaining the reliable true discoveries and maximizing sample characteristics. Performance across multiple combinations demonstrated that the MaxLFQ-quantified protein-level correction strategy is more robust than precursor- or peptide-level corrections under most situations, even though protein quantities are inferred from precursor and peptide intensities. The extension to large-scale data confirmed improved prediction capability after protein-level batch-effect correction, demonstrating the MaxLFQ-Ratio combination achieved the best performance for both categorical (sex) and continuous (age) endpoints, suggesting broad applicability across proteomics datasets.

Multi-dimensional metrics provide a comprehensive understanding of batch-effect correction performance across factors. We used the SNR metric to assess both the discrimination of samples and the robustness of replicates. This PCA-rooted metric, quantifies expected grouping information, enabling generalized comparability across diverse datasets[22,26,29–32]. SNR extends from Quartet and simulated datasets to large-scale datasets with accompanying injections of known QC samples alongside study samples. Additionally, MCC assessed fold-change accuracy after correction, providing a feature-based evaluation perspective. While 80% of batch-effect corrections yielded higher SNR values at the protein level with MaxLFQ quantification, RUV-III-C uniquely removed batch effects across all three data levels in simulated data incorporating true negative control features. Meanwhile, LOESS enhanced MS signal drift removal in large-scale data, consistent with previous research[23]. These results suggest that each BECA's underlying assumptions contribute to scenario-specific performances. Large-scale data revealed that blindly pursuing batch-effect correction inadvertently diminishes true biological variability, as Ratio achieved the best prediction performance despite residual batch effects. The superior MCC performance of ratio-based BECAs was also reported in MAQC-II microarray studies[50].

The quantification process affects batch-effect correction performance. Previous studies have corrected batch effects at the peptide level[7,8,23,51], whereas others correct at the protein level[32,52–54]. Čuklina et al. proposed peptide-level correction before protein quantification to avoid altering final protein abundances[23]. However, batch effects are reportedly more severe at the peptide level than at the protein level[55], suggesting reduction after protein quantification, a finding supported by our results (Fig. 2b, c). Phua et al. also found no superiority in

https://doi.org/10.1038/s41467-025-64718-y



peptide-level correction using guided PCA evaluation[32,56]. Unlike previous benchmarking studies focused on single-level correction without integrating quantification[39–41], we explored optimal correction levels and demonstrated interactions between QMs and BECAs across data levels. Our results suggest that quantification increases true features and decreases false positives (FPs), improving overall MCC despite increased false negatives (FNs). This may stem from inherent averaging: a protein group assembles multiple peptides, and a peptide matches precursors with distinct charges and modifications. Aggregating precursor/peptide data to the protein level improves overall data stability by smoothing the variability of individual peptides. However, when correction is performed at precursor/peptide levels, residual artifacts may be amplified during quantification (Fig. 5d), diminishing correction effectiveness (Supplementary Fig. 10d).

Fig. 4 | MaxLFQ unmasked biological signals in protein-level batch-effect correction. a box plots of the signal-to-noise ratio (SNR) values across different designed scenarios; faceted by the scenario; the P values were calculated using unpaired two-tailed student's T tests, $^{****}P < 0.0001$, $^{**}P < 0.01$, $^{*}P < 0.05$; exact P values rounded to two significant figures (when not less than 0.0001) are provided above each comparison; horizontal lines indicate the median; box boundaries indicate the interquartile range (IQR); whiskers represent values within 1.5× IQR of the first and third quartiles; data points beyond the end of the whiskers are plotted individually; For each box plot: Uncorrected (quantification method (QM) $n = 3$; total data point $n = 3$), Precursor-corrected (batch-effect correction algorithm (BECA) $n = 6$, QM $n = 3$; total data point $n = 18$; except Quaret-balanced (Quartet-B) scenario where BECA $n = 7$, QM $n = 3$; total data point $n = 21$), Peptide-corrected (BECA $n = 6$, QM $n = 3$; total data point $n = 18$), Protein-corrected (BECA $n = 6$, QM $n = 3$; total data point $n = 18$). **b** bar plots of the SNR values in the Quartet-B scenario; colored by the correction level; faceted by the correction level (by column) and the quantification method (by row); data points are plotted individually; the dashed lines represent the SNR value in the uncorrected data matrix. **c** bar plot of principal variance component analysis (PVCA) results of each correction level-BECA-QM combination; the bars are colored by the label of random effects; the cells lying on the x-axis are colored by the detailed combination. All statistics are based on independent biological replicates (by sample groups) and analysis replicates (by BECAs and by QMs). Source data are provided as a Source Data file.

Therefore, quantification before batch-effect removal facilitates more accurate evaluation and correction, enhancing downstream analysis.

Reference materials or QC samples enable determination of batch effect causes[32] and evaluation of correction against known biological groups. In benchmarking, Ratio and RUV-III-C (using Quartet/QC samples) performed well. As RUV-III-C requires negative control features as input--generally unavailable in real-world large-scale cohort studies--we identified inputs using two-way ANOVA across batches. While RUV-III-C learns batch-effect patterns from negative controls, Ratio gently corrects batch effects and maximizes prediction capability by taking accompanying QC samples as a global reference across various batches[16].

This study has several limitations. First, we combined technical factors as overall batch effects without detailed exploration of specific data acquisition modes (e.g., DDA/DIA) or platforms. For example, untargeted (DDA and DIA) and targeted strategies may exhibit distinct patterns in batch-effect correction and quantification. Second, protein-level correction is limited when BECAs require early-stage information (e.g., NormAE needs precursor-level m/z and RT); further exploration is needed at levels earlier than precursors (e.g., raw spectra). Third, we assessed prediction performance using random forest for sex/age endpoints; testing additional endpoints with more deep learning models would strengthen batch-effect correction evaluation. Finally, it remains unclear whether we would obtain the least batch effect and the most biological information of interest if we correct batch effects at each data level and then aggregate together.

In summary, this study provides practical guidance for choosing effective batch-effect correction strategies for MS-based proteomics data. In large-scale proteomics studies, multi-sourced batch effects often confound biological factors of interest. Thus, effective removal requires comprehensive consideration of protein quantification method and BECA combinations. Our findings demonstrate that protein-level correction is generally effective, and MaxLFQ-Ratio achieves optimal performance in both correction and prediction tasks.

Methods

Datasets

Quartet multi-lab datasets. We used six raw datasets generated from the Quartet reference materials as the multi-lab benchmarking datasets[22,29], which are available at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the iProX partner repository[57,58] with the dataset identifier PXD045065. Notably, Quartet Human Proteome Peptide Reference Materials have been approved as the First Class of National Reference Materials (GBW 09908–GBW 09911), by State Administration for Market Regulation (SAMR) of China. All MS raw files were preprocessed by MaxQuant (v2.1.3.0) with the default parameters to obtain the peptide-to-protein identifications (searched against Uniprot human protein database updated on 2022.08.08) and the peptide intensities. DAT1 to DAT3 were generated in the DDA strategy and were searched directly. DAT4 to DAT6 were generated in a DDA-library-based DIA strategy, i.e., the spectrum libraries were constructed through the data generated by the pre-fractioned samples in each batch by the DDA strategy, then the DIA MS files were analyzed by MaxDIA embedded within the MaxQuant software.

In the Balanced design, each Quartet dataset was generated by a triplicate of four types of samples (D5, D6, F7, and M8), thereby composing the combined data matrix with balanced design, i.e, the biological groups are balanced within each batch and between every two batches. In order to design a data matrix in the confounded scenario, we selected one type of sample in each data set: from DAT1 to DAT6, corresponding to D5, F7, M8, D5, F7, and M8, respectively. In this case, the biological signals were thoroughly confounded with the batch factor. Each dataset contains triplicates of D6 to enable the use of the ratio-based batch-effect correction method.

Simulated datasets. We generated simulated data following the methodology described by Hui et al.[59], and the detailed parameters in this study were provided in Supplementary Methods. For samples and batches, the dataset contained three batches (Batch1, Batch2, and Batch3), each comprising triplicates of three sample groups (Group1, Group2, and Group3), totaling 27 samples in the balanced design. In the confounded design, we excluded Group2 from Batch1 and Group3 from Batch2, resulting in sample groups confounded with the batch factor. Each dataset included triplicates of Group1 to enable ratio-based batch-effect correction and MCC evaluation based on known fold changes. For features, we first determined feature proportions impacted by batch, sample group, and interaction effects using two-way ANOVA on Quartet datasets. We then simulated 4950 precursors matched to 3000 peptides and 300 proteins: 200 affected by both sample group ("biological") and batch ("technical") factors, 96 by batch alone, and 4 "housekeeping" proteins.

Large-scale ChiHOPE dataset. We took the plasma proteomics data in the Chiglitazar perturbed Human multi-Omics ProfilE (ChiHOPE) project as a case study. The publicly available ChiHOPE resource encompasses multiple omics data types for 835 patients with T2D. Our proteomics study utilized a subset of this larger cohort, comprising 750 patients (274 females and 476 males, aged from 23 to 70 years) who had plasma samples available for proteomic profiling. From these patients, we analyzed a total of 1431 samples from 1419 unique plasma specimens (12 technical duplicates) collected at baseline and 24 weeks. Three types of quality control samples (PM, P10, and P11) were injected in each batch along with study samples. Specifically, PM was mixed plasma samples from the study samples. P10 and P11 were collected from a healthy man and a healthy woman, respectively. All Ethical approvals were obtained from the Ethical Committees of the study centers (CMAP[60], $n = 26$; CMAS[61], $n = 33$). All procedures which performed in the study involving human participants were in accordance with the ethical standards of the institutional and/or) National Research Committee, and with the Declaration of Helsinki and its later amendments or comparable ethical standards. All participants provided written informed consent.

The samples were measured by data-independent acquisition (DIA) mode in an EASY-nLC 1200 ultra-high-pressure system coupled to a Q Exactive HF-x. All 1,431 peptide samples were injected with two

https://doi.org/10.1038/s41467-025-64718-y

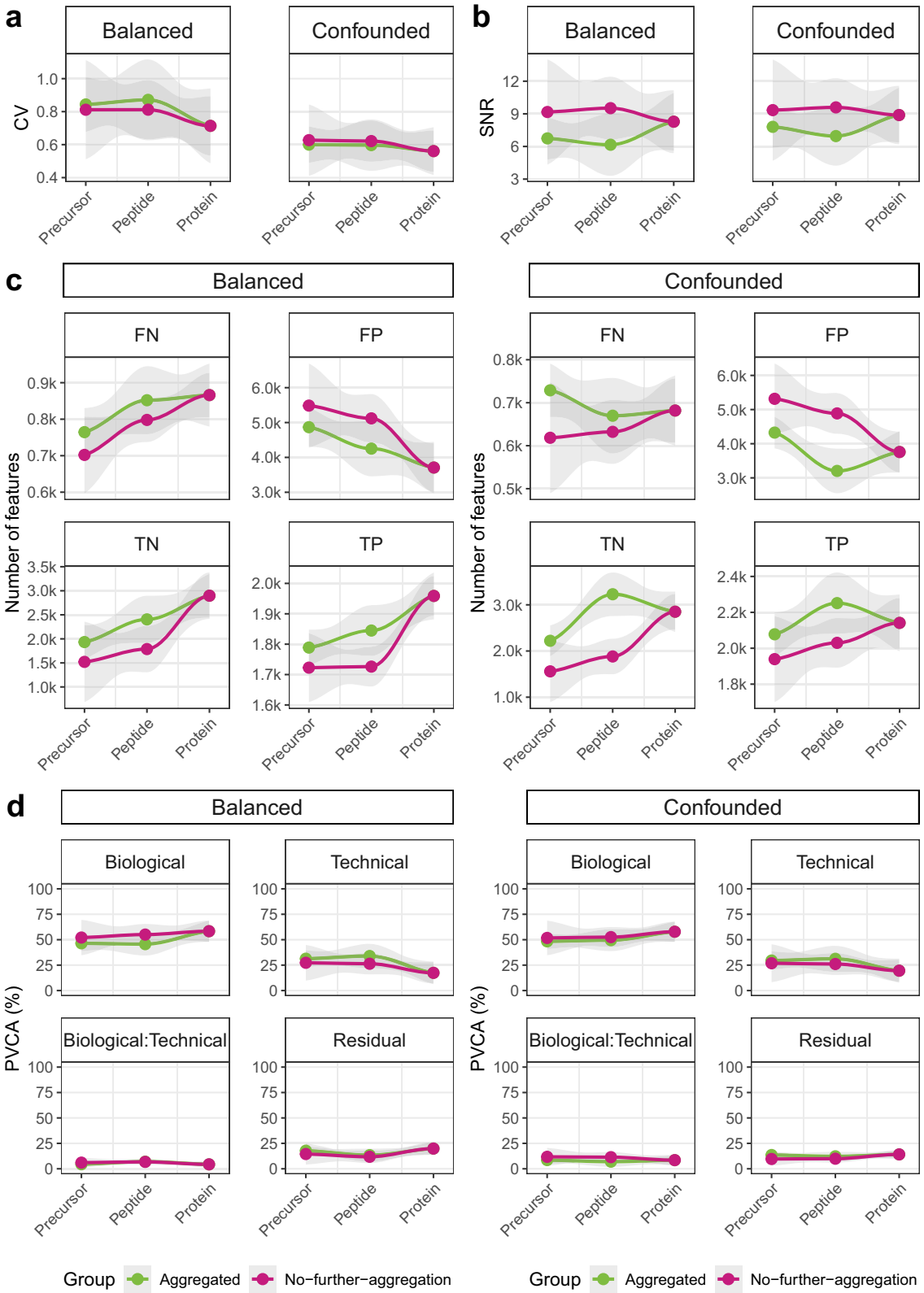


Fig. 5 | Quantification interacted with correction from precursor to protein levels. Plots of the coefficient of variation (CV) (**a**) and signal-to-noise ratio (SNR) (**b**) performances of corrected data matrices across different data levels. **c** plots of the number of false negatives (FNs), false positives (FPs), true negatives (TNs), and true positives (TPs) of corrected data matrices across different data levels. **d** plots of the principal variance component analysis (PVCA) results of corrected data matrices across different data levels. All plots are colored by before (pink) or after (light green) the aggregated quantification step. All plots are fitted with loess curves, with grey ribbon shaded areas indicating the standard error (se) in 95% confidence interval; the data points in the center area represents for mean values individually. All statistics are based on independent biological replicates (by sample groups) and analysis replicates (by batch-effect correction algorithm (BECAs) and by quantification methods (QMs)). Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-025-64718-y

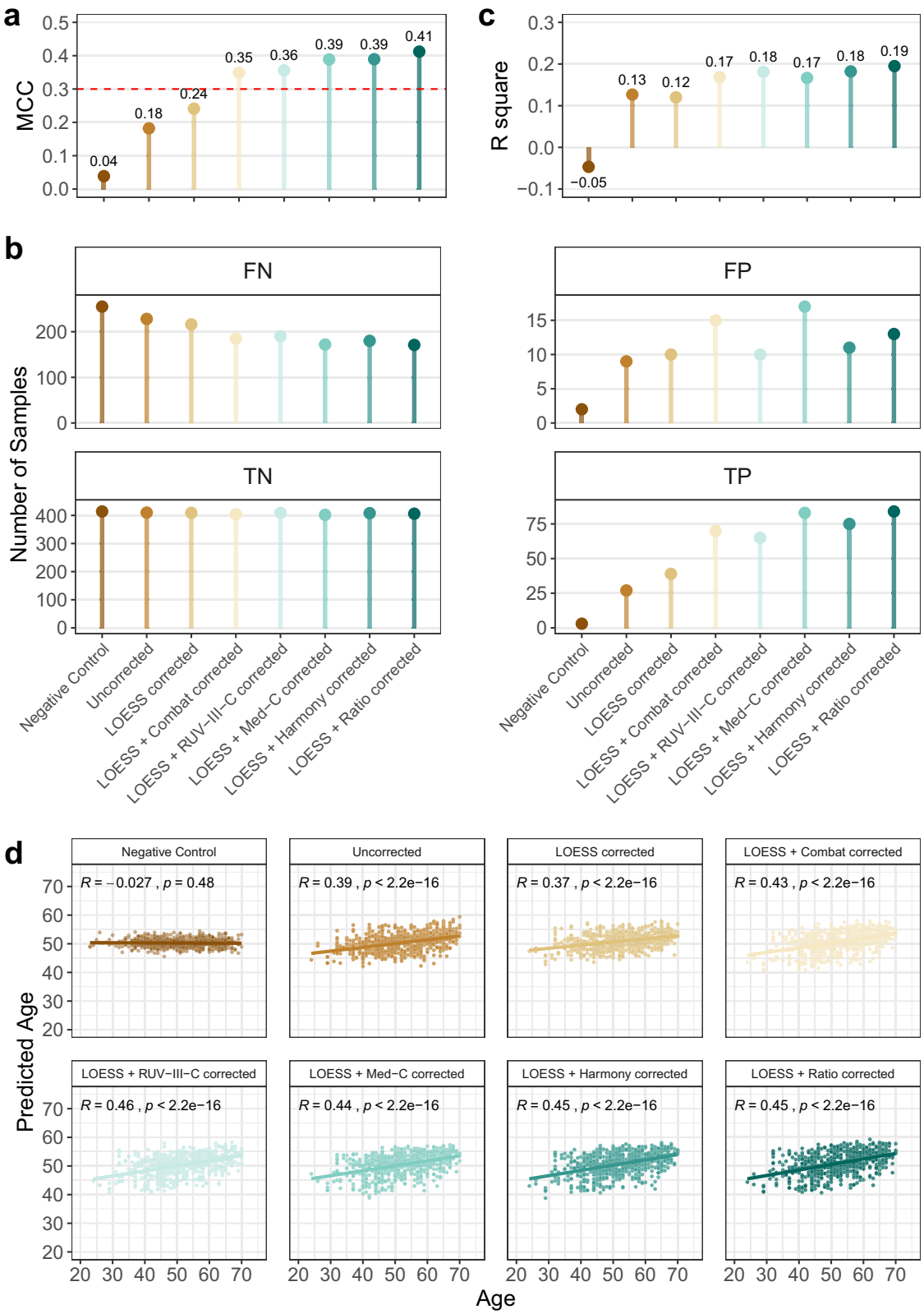


Fig. 6 | Ratio provided the best prediction performance in the large-scale data. The prediction performance on the sex classification. **a** bar plot of the Matthews correlation coefficient (MCC); the dashed lines represent the threshold of MCC value at 0.3; data points are plotted individually. **b** bar plots of the number of false negatives (FNs), false positives (FPs), true negatives (TNs), and true positives (TPs); data points are plotted individually. The prediction performance on the age regression. **c** bar plot of the R square values; data points are plotted individually. **d** scatter plots of predicted age to the true age; the solid lines represent fitted curves from linear regression along with the Pearson correlation coefficient (R) and the corresponding P value rounded to two significant figures; 95% confidence interval (grey ribbon) are displayed. All evaluations are on the validation set after the 5-fold cross validation within the training set. All plots were colored by the label of data matrix. Source data are provided as a Source Data file.

PM samples, one P10 sample, and one P11 sample in each 96-well plate, and 1495 runs of MS files were generated. The MS files were searched against UniProt human protein database (updated on 2019.12.17) using FragPipe (v12.1) with MSFragger (2.2)[62]. DIA data were analyzed using DIA-NN (v1.7.0)[62,63]. The default settings were used for DIA-NN. The identified precursors were quantified by the average of chromatographic fragment ion peak areas across all reference spectra libraries. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (https://proteomecentral.proteomexchange.org) via the iProX partner repository[57,58] with the dataset identifier PXD068273.

Protein quantification

The analysis was carried out using functions implemented in R (version 4.4.1) and R Studio (2024.09.0 + 375).

MaxLFQ. MaxLFQ[64] is a protein quantification method by delayed normalization and maximal peptide ratio extraction. It retains peptides with at least two pair-wise ratios and optimizes protein quantities with the least overall variation. In this paper, we used the R package diann (v1.0.1)[63] to allow for the separate MaxLFQ-based protein quantification. This method is applied to the Quartet multi-lab datasets and the external T2D dataset.

IBAQ. The protein quantification method iBAQ sums up all peptide intensities per sample for each protein group, despite whether the intensities are missing across samples[14]. This method is applied to the Quartet multi-lab datasets.

TopPep3. TopPep3 is based on the observations that the two most intense transitions of the three best flying peptides per protein generated optimal results[15]. This method is applied to the Quartet multi-lab datasets by averaging the three most intense peptides per protein.

Batch-effect correction

For Balanced and Confounded designs, we directly applied seven BECAs to the raw data matrix. For the large-scale dataset, we first fitted a LOESS curve for the log-transformed intensity values of each protein, and estimated the impact of the injection order on the MS signal drift. The estimated impact was subtracted from the intensity values of each feature. Then the data matrix was passed to the typical batch-effect correction workflow. The analysis was carried out using functions implemented in R (v4.4.1) and R Studio (2024.09.0 + 375).

ComBat. ComBat is a commonly used batch-effect correction method in quantitative omics. It is based on a Bayesian framework to optimize the mean and the variance for batch-effect correction[20]. The ComBat function in the sva (v3.52.0) package was applied to the normalized data across batches after log transformation and per-sample normalization of total intensities.

Median centering. In this paper, median centering is the approach that scales the medians per batch to the same, and is mild for data matrices with a high proportion of missing values. This method is applied to the Quartet multi-lab datasets after log transformation and per-sample normalization of total intensities, and to the external T2D dataset after per-feature per-batch LOESS fixing.

Ratio. In the Quartet project, a ratio-based scaling is recommended to enable the comparability across multiple batches at the relative quantitation level[22,26,29–32]. In this paper, the data matrix corrected by the ratio method is achieved by subtracting the means of D6 samples (or Group1 samples in simulated datasets, or PM samples in ChiHOPE study) in each batch after log transformation and per-sample normalization of total intensities.

RUV-III-C. RUV-III-C is a batch-effect correction method that removes unwanted variation from complete intensities per batch in the data matrix[12]. It requires sample label information and negative control features for the estimation and removal of the unwanted variation. In the Quartet datasets, we performed two-way ANOVA analysis at the precursor, peptide, or protein level, and selected quantified precursors, peptides or protein groups with no significance across all samples or an interactive effect between the sample and the batch, but significant across the batches. All ANOVA results can be reproduced by the scripts. In the simulated datasets, 96 batch-effect-only proteins and their matched precursors and peptides were input to RUV-III-C as negative controls. The RUVIII_C function in the RUVIIIC (v1.0.19) package was applied for batch-effect correction.

Harmony. Harmony is an iterative algorithm for robust, scalable, and flexible integration of single-cell datasets, effectively removing batch effects. It projects cells into a shared low-dimensional embedding (e.g., PCA space), grouping cells by biological state rather than by technical variations. Batch-effect correction was applied using the HarmonyMatrix function in the harmony (v1.2.3) package.

WaveICA2.0. WaveICA 2.0 is an improved version of the WaveICA method specifically designed to remove both inter-batch and intra-batch effects in untargeted metabolomics data without requiring explicit batch information. The WaveICA_2.0 function in the WaveICA2.0 (v0.1.0) package was applied for batch-effect correction.

NormAE. NormAE is a deep learning-based method for batch-effect correction in untargeted metabolomics data. It leverages the power of autoencoders to learn and remove technical noise while preserving underlying biological signals. This method was used for batch-effect correction following the instructions at https://github.com/luyiyun/NormAE.

Quality assessment metrics

We used four quality assessment metrics, classified into feature-based and sample-based categories, to evaluate the reproducibility, distinguishability, and reliability of the data matrix. All quality assessments were performed at the quantified protein level. The analysis was carried out using functions implemented in R (version 4.4.1) and R Studio (2024.09.0 + 375).

Coefficient of Variation (CV). The coefficient of variation, defined as the ratio of standard deviation to the mean, was calculated per feature within the same sample groups in each design, i.e., 18 replicates from six batches for each of D5, F7, and M8 in the Quartet balanced design.

Mathew's Correlation Coefficient (MCC) and Relative Correlation (RC). In simulated datasets, we utilized the known fold change of each feature to calculate the MCC value. The positive and negative features were defined by whether their fold change (FC) values were out of (positive) or between the range of 1/1.2-1.2 (negative). If the FC values were in the same direction as the known values (Group2/Group1 and Group3/Group2), then the corresponding feature is labeled as a true feature; otherwise, as a false discovery. With the number of TPs, TNs, FPs, and FNs, we can calculate the Mathew's Correlation Coefficient (MCC) values by the following formula (1):

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

Relative correlation was defined as the Pearson correlation coefficient between the tested and true FC values, computed as follows:

$$RC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

where n is the total number of features, x_i is the $\log_2$ transformed fold change of the i^{th} feature in the test data matrix. y_i is the $\log_2$-transformed fold change of the reference value during data simulation.

Signal-to-Noise Ratio (SNR). The Quartet project defines a metric named signal-to-noise ratio (SNR) based on the known sample labels[22,26,30–32]. The distance is determined as the weighted Euclidean distance between any two samples x and y in the space formed by the first and second principal components after principal components analysis (PCA):

$$\mathrm{dist}(x,y) = \left(\mathbf{W_1}\left(\mathbf{PC}_{1,x} - \mathbf{PC}_{1,y}\right)^2 + \mathbf{W_2}\left(\mathbf{PC}_{2,x} - \mathbf{PC}_{2,y}\right)^2\right)^{\frac{1}{2}} \quad (3)$$

Then, Signal was defined as the root mean square (Root Mean Square, RMS) of the distance between two samples with different biological labels:

$$\mathrm{Signal} = \left(\frac{\sum_{u\neq v}^{K}\sum_{i=1}^{N_u}\sum_{j=1}^{N_v}\mathrm{dist}\left(u_i, v_j\right)^2}{C_N^2 - \sum_{k=1}^{K}C_{N_k}^2}\right)^{\frac{1}{2}} \quad (4)$$

where u_i and v_j represent the replicate i with the sample group u and the replicate j within the sample group v. N_u and N_v are the total number of replicates in the sample group u and the sample group v. N is the total number of samples.

Noise was defined as the RMS of the distance between two replicate samples with the same biological label:

$$\mathrm{Noise} = \left(\frac{\sum_{u=1}^{K}\sum_{i\neq j}^{N_u}\mathrm{dist}\left(u_i, u_j\right)^2}{\sum_{u=1}^{K}C_{N_u}^2}\right)^{\frac{1}{2}} \quad (5)$$

Finally, the SNR is defined as the ten-fold common logarithm of the squared ratio of Signal to Noise:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\mathrm{Signal}}{\mathrm{Noise}}\right)^2 \quad (6)$$

Prediction model

For the large-scale dataset, we performed the sex classification and the age regression tasks using a Random Forest model implemented in the scikit-learn library (v1.7.0) within Python (v3.13.5). The total of 750 subjects were divided into a training set (394 subjects) and a validation set (356 subjects) based on their inclusion time in the clinical trials. A 5-fold cross validation was conducted within the training set (757 injected samples), and the final performance was evaluated on the validation set (674 injected samples).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The statistical data generated in this study have been deposited in the Figshare database at https://doi.org/10.6084/m9.figshare.29567366.v2. The processed Quartet and simulated proteomics data are available in the Figshare database at https://doi.org/10.6084/m9.figshare.29567333.v2. The processed ChiHOPE proteomics data are available at https://doi.org/10.6084/m9.figshare.30028336. All performance assessment results generated in this study are provided in the Source Data file. The raw Quartet and ChiHOPE proteomics data used in this study are available in a ProteomeXchange partner database under accession code PXD045065 and accession code PXD068273. Source data are provided with this paper.

Code availability

The code used to perform the analyses and generate results in this study is publicly available, and has been deposited in a GitHub repository at https://github.com/QiaochuChen/proteomics-batch-effect-correction-benchmarking, under MIT license. The specific version of the code associated with this publication is archived in Zenodo and is accessible via https://doi.org/10.5281/zenodo.17032531[65].

References

1. Yu, Y., Mai, Y., Zheng, Y. & Shi, L. Assessing and mitigating batch effects in large-scale omics studies. *Genome Biol.* **25**, 254 (2024).
2. Goh, W. W. B., Yong, C. H. & Wong, L. Are batch effects still relevant in the age of big data?. *Trends Biotechnol.* **40**, 1029–1040 (2022).
3. Gregori, J. et al. Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. *J. Proteom.* **75**, 3938–3951 (2012).
4. Ge, S. et al. A proteomic landscape of diffuse-type gastric cancer. *Nat. Commun.* **9**, 1012 (2018).
5. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).
6. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
7. Williams, E. G. et al. Multiomic profiling of the liver across diets and age in a diverse mouse population. *Cell Syst.* **13**, 43–57.e6 (2022).
8. Keele, G. R. et al. Regulation of protein abundance in genetically diverse mouse populations. *Cell Genomics* **1**, 100003 (2021).
9. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786 (2015).
10. Zhang, H. et al. Integrated Proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**, 755–765 (2016).
11. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35**, 498–507 (2017).
12. Poulos, R. C. et al. Strategies to enable large-scale proteomics for reproducible research. *Nat. Commun.* **11**, 3793 (2020).
13. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteom.* **73**, 2092–2123 (2010).
14. Schwanhüusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
15. Ludwig, C., Claassen, M., Schmidt, A. & Aebersold, R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry*. *Mol. Cell. Proteom.* **11**, M111.013987 (2012).
16. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ*. *Mol. Cell. Proteom.* **13**, 2513–2526 (2014).
17. Rozanova, S. et al. Quantitative Mass Spectrometry-Based Proteomics: An Overview. in *Methods in Molecular Biology* vol. 2228 85–116 (Humana Press Inc., 2021).
18. Leek, J. T. Svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
19. Galitzine, C. et al. Nonlinear regression improves accuracy of characterization of multiplexed mass spectrometric assays. *Mol. Cell. Proteom.* **17**, 913–921 (2018).

20. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
21. Sundararaman, N. et al. BIRCH: An automated workflow for evaluation, correction, and visualization of batch effect in bottom-up mass spectrometry-based proteomics data. *J. Proteome Res* **22**, 471–481 (2023).
22. Zheng, Y. et al. Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials. *Nat. Biotechnol.* **42**, 1133–1149 (2024).
23. Čuklina, J. et al. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol. Syst. Biol.* **17**, e10240 (2021).
24. Dammer, E. B., Seyfried, N. T. & Johnson, E. C. B. Batch correction and harmonization of –Omics datasets with a tunable median polish of ratio. *Front. Syst. Biol.* **3** (2023).
25. Carry, P. M. et al. Propensity scores as a novel method to guide sample allocation and minimize batch effects during the design of high throughput experiments. *BMC Bioinforma.* **24**, 86 (2023).
26. Tian, S. et al. Quartet protein reference materials and datasets for multi-platform assessment of label-free proteomics. *Genome Biol.* **24**, 202 (2023).
27. O'Rourke, M. B. et al. What is normalization? The strategies employed in top-down and bottom-up proteome analysis workflows. *Proteomes* **7**, (2019).
28. Lee, A. H. et al. Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. *Nat. Commun.* **10**, 1092 (2019).
29. Yang, J. et al. The Quartet Data Portal: integration of community-wide resources for multiomics quality control. *Genome Biol.* **24**, 245 (2023).
30. Yu, Y. et al. Quartet RNA reference materials improve the quality of transcriptomic data through ratio-based profiling. *Nat. Biotechnol.* **42**, 1118–1132 (2024).
31. Zhang, N. et al. Quartet metabolite reference materials for inter-laboratory proficiency test and data integration of metabolomics profiling. *Genome Biol.* **25**, 34 (2024).
32. Yu, Y. et al. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol.* **24**, 201 (2023).
33. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
34. Chen, W. et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. *Nat. Biotechnol.* **39**, 1103–1114 (2021).
35. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
36. Deng, K. et al. WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Anal. Chim. Acta* **1061**, 60–69 (2019).
37. Deng, K. et al. WaveICA 2.0: a novel batch effect removal method for untargeted metabolomics data without using batch information. *Metabolomics* **17**, 87 (2021).
38. Rong, Z. et al. NormAE: Deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data. *Anal. Chem.* **92**, 5082–5090 (2020).
39. Välikangas, T., Suomi, T. & Elo, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief. Bioinform* **19**, 1–11 (2018).
40. Callister, S. J. et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res* **5**, 277–286 (2006).
41. Kultima, K. et al. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol. Cell. Proteom.* **8**, 2285–2295 (2009).
42. Chawade, A., Alexandersson, E. & Levander, F. Normalyzer: A tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res* **13**, 3114–3120 (2014).
43. Arend, L. et al. Systematic evaluation of normalization approaches in tandem mass tag and label-free protein quantification data using PRONE. *Brief. Bioinform.* **26**, bbaf201 (2025).
44. Ren, L. et al. Quartet DNA reference materials and datasets for comprehensively evaluating germline variant calling performance. *Genome Biol.* **24**, 270 (2023).
45. Jia, P. et al. Haplotype-resolved assemblies and variant benchmark of a Chinese Quartet. *Genome Biol.* **24**, 277 (2023).
46. Wang, S. et al. De novo and somatic structural variant discovery with SVision-pro. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02190-7. (2024)
47. Wang, D. et al. A real-world multi-center RNA-seq benchmarking study using the Quartet and MAQC reference materials. *Nat. Commun.* **15**, 6167 (2024).
48. Chen, Q. et al. Plasma-free blood as a potential alternative to whole blood for transcriptomic analysis. *Phenomics* **4**, 109–124 (2024).
49. Ren, L., Shi, L. & Zheng, Y. Reference Materials for Improving Reliability of Multiomics Profiling. *Phenomics* **4**, 487–521 (2024).
50. Luo, J. et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenom. J.* **10**, 278–291 (2010).
51. Messner, C. B. et al. Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell Syst.* **11**, 11–24.e4 (2020).
52. Huang, Z. et al. Brain proteomic analysis implicates actin filament processes and injury response in resilience to Alzheimer's disease. *Nat. Commun.* **14**, 2747 (2023).
53. Zhu, T. et al. BatchServer: A web server for batch effect evaluation, visualization, and correction. *J. Proteome Res* **20**, 1079–1086 (2021).
54. Voß, H. et al. HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nat. Commun.* **13**, 3523 (2022).
55. Graw, S. et al. proteiNorm - A user-friendly tool for normalization and analysis of TMT and label-free protein quantification. *ACS Omega* **5**, 25625–25633 (2020).
56. Phua, S.-X., Lim, K.-P. & Goh, W. W.-B. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Comput Struct. Biotechnol. J.* **20**, 4369–4375 (2022).
57. Chen, T. et al. iProX in 2021: Connecting proteomics data sharing with big data. *Nucleic Acids Res* **50**, D1522–D1527 (2022).
58. Ma, J. et al. Iprox: An integrated proteome resource. *Nucleic Acids Res* **47**, D1211–D1217 (2019).
59. Hui, H. W. H., Chan, W. X. & Goh, W. W. Bin. Assessing the impact of batch effect associated missing values on downstream analysis in high-throughput biomedical data. *Brief. Bioinform.* **26**, bbaf168 (2025).
60. Ji, L. et al. Efficacy and safety of chiglitazar, a novel peroxisome proliferator-activated receptor pan-agonist, in patients with type 2 diabetes: a randomized, double-blind, placebo-controlled, phase 3 trial (CMAP). *Sci. Bull.* **66**, 1571–1580 (2021).
61. Jia, W. et al. Chiglitazar monotherapy with sitagliptin as an active comparator in patients with type 2 diabetes: a randomized, double-blind, phase 3 trial (CMAS). *Sci. Bull.* **66**, 1581–1590 (2021).
62. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).
63. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).

64. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, Termed MaxLFQ*. *Mol. Cell. Proteomics*. **13**, 2513–2526 (2014).
65. Chen, Q. Protein-level batch-effect correction enhances robustness in MS-based proteomics. https://github.com/QiaochuChen/proteomics-batch-effect-correction-benchmarking, https://doi.org/10.5281/zenodo.17032531 (2025).

Acknowledgements

This study was supported in part by the National Key R&D Project of China (2022YFF0608404 to Y. Zheng and 2023YFC3402501 to L.S.), the National Natural Science Foundation of China (T2425013 to Y. Zheng, 32370701 to L.S., 32470692 to Y. Zheng, and 32170657 to L.S.), the Natural Science Foundation of Shanghai (24JS2840100 to Y. Zheng), Shanghai Municipal Science and Technology Major Project (2023SHZDZX02 to L.S.), and the 111 Project (B13016 to L.S.). We also thank CFFF (Computing for the Future at Fudan) and the Human Phenome Data Center of Fudan University for computing support.

Author contributions

Y. Zheng and L.S. conceived and designed the study. Q.C. performed data preprocessing, assessment analyses, and visualizations. Y.X. performed the model predictions. Q.C. and N.Z. drafted the manuscript. Q.C. and Y. Zhao organized and deposited all raw and preprocessed data. All authors, including Z.C., Y.L., H.C., Y.M., S.D., J.L., Y.Y., and Y. Zhao, contributed to the critical review and revision of the manuscript. The final version was reviewed and approved by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64718-y.

Correspondence and requests for materials should be addressed to Leming Shi or Yuanting Zheng.

Peer review information *Nature Communications* thanks Wilson Goh, Hui Peng, and the other anonymous reviewer for their contribution to the peer review of this work. A peer review file is available.