

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

Corresponding Author: Professor Yuanting Zheng

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

What are the noteworthy results?

- Via extensive simulations and tests involving multiple datasets, balance and imbalance, classification and regression tasks, the authors reported that about 70% of the time, protein-based batch correction is superior to early peptide-based batch correction.

Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.

- The authors cite the latest works in the field. And update their analysis to also include machine learning task performances, which is increasingly important in bioinformatics. The findings, if correct, are fundamental, and can change/alter behaviors and analysis workflows, representing a best practice in the field. The idea of early correction has been purported in several works, including Goh et al, and Cuklina et al. This work is possibly, the first to investigate and test this extensively.

Does the work support the conclusions and claims, or is additional evidence needed?

- The work is impressive. There are extensive tests across a wide variety of scenarios. However, I am concerned if the findings, based on a limited set of data, are sufficiently representative of all data in the proteomics field. Besides DDA and DIA platforms, I think we need more granularity on the types of platforms being tested and evaluated. perhaps there may be interesting findings here. for example, in DDA, the constituent peptides representing each protein may be different across samples/runs and each peptide carries different information on the batch effect. this may be less of a problem for DIA runs or TDA runs even. I think we need to be more specific and give more insight on the nature of the platform and acquisition paradigm, instead of arriving at more generalized statistic. I believe there are differences there.

I am also curious if the manifestation of batch effect, in this particular dataset, is also representative of all batch effects. I think there needs to be more control over the magnitude, nature, and distribution of batch effects in the data, to arrive at a more nuanced and meaningful understanding.

The authors need to also incorporate other metrics in their estimation of the batch factor in their data. Namely, I think PVCA, gPCA, etc, should also be incorporated to better estimate the component of batch information. Besides using the PCA plots, which can be more subjective in interpretation, I suggest the authors also consider using the side-by-side boxplots (<https://pmc.ncbi.nlm.nih.gov/articles/PMC5374662/>) to help review the PC-based association of batch effects, with concomitant effect sizes.

I also think the authors should weigh their observations on their dataset, on potential mechanism and explanation as to why 70% of the time, protein correction is superior. Saying it just is, can be dangerous and biased, as this may only apply to the nature of the datasets + batch effects, and cannot generalize elsewhere. It is more satisfying to find out exactly under what circumstances is this so. is it because, the information aggregation approaches tend to amplify or estimate the batch factor more accurately? if so, then a closer look at how the protein assembly methods combine this information is important. and

we have to show via proof-by-simulation or mathematical proof that this is indeed the case, and that the improved results is because of signal amplification, making it easier to reliably estimate and remove the batch factor.

in this regard, there should also be explicit insights into the BECAs themselves. The authors did show that methods like RUV are worse than others. Again, I think this may be a function of the nature of the batch effects/class effects in the data. Each BECA is designed with different batch assumptions in mind, so I think we need to first understand the batch factor, and batch distributions, before we make a generalized comment that a particular is inferior. In other benchmarks and studies, RUV has been shown to be a top performer. So i think what is reported in this study, has to be tempered and fair. Moreover, this is not a benchmark study, but rather to demonstrate that early correction at the peptide level may be counter productive.

Finally, I think when it comes to early correction, it is possible to go even earlier than peptides. it is possible to go into the spectra realm, or do adjustments based on summaries or representations of spectra. I think this aspect should be answered so that it doesn't just because a binary comparison between peptides and proteins. The proteins are assembled from the peptides, but the peptides are inferred from the spectra, and this is bulk of the real data that comes from MS. The question I have is... how much batch/class information do we lose if we do not fully utilize spectra. some reports claim that only 30% of spectra are matched to peptide, and the remainder are lost. your paper will be much more valuable, if we can look into information maximization concerns. is batch estimation using all spectra, prior to peptide and protein inference, more valuable?

Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?

- see above

Is the methodology sound? Does the work meet the expected standards in your field?

-see above

Is there enough detail provided in the methods for the work to be reproduced?

-the method is well described and should be reproducible.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

What are the noteworthy results?

The study benchmarks multiple combinations of batch-effect correction algorithms (BECAs) and protein quantification methods (PQMs) to determine whether correction at the peptide or protein level yields better outcomes. The most important reported result is that protein-level correction outperformed peptide-level correction in 72.2% of tested combinations across several metrics. However, this emphasis on average performance hides what should be the most relevant finding: what is the best possible combination?

For example, in Figure 4a, the highest signal-to-noise ratio (SNR) is actually achieved when correcting at the peptide level using the Ratio method and MaxLFQ, contradicting the main conclusion. Similarly, Figure 4b–e lacks clarity on what the barplots and boxplots represent. If they are based on the 12 combinations of PQMs and BECAs, this aggregation makes protein-level correction appear more consistent, but hides the fact that peptide-level correction yielded the single best score. In the context of choosing methods for a real experiment, researchers are likely to care about the best-performing pipeline, not the most consistent across all combinations.

Will the work be of significance to the field and related fields? How does it compare to the established literature?

The work is relevant and useful to the proteomics field, where batch effects are a major concern. It addresses an important question that has been raised before: whether correction should happen at the peptide or protein level (e.g., Phua et al., PMC9411064). Until now, there hasn't been a systematic benchmark comparing these options across multiple methods and datasets. This study attempts to fill that gap and provides empirical support for a widely adopted, but under-justified, practice: correcting at the protein level.

However, the current presentation weakens the impact. Highlighting average scores across methods dilutes the value of identifying top-performing pipelines, which is ultimately what matters in applied settings.

The study would be of more significance if more methods were benchmarked. For example, waveica (<https://doi.org/10.1016/j.aca.2019.02.010>), harmony (doi: 10.1038/s41592-019-0619-0) or deep learning based BECAs (e.g. NormAE: <https://doi.org/10.1021/acs.analchem.9b05460>, BERNN: DOI: 10.1038/s41467-024-48177-5) .

Does the work support the conclusions and claims, or is additional evidence needed?

Not entirely. The main claim, that protein-level correction is superior because it performs better on average, is misleading. It ignores the fact that the best performing combination in at least one major metric (SNR) is achieved at the peptide level. This undercuts the narrative that protein-level correction should be favored overall. The authors should revise their framing: instead of focusing on averages, they should highlight scenarios where each level of correction is preferable. The results are of interest, but the conclusions and claims could benefit from being revised.

Are there any flaws in the data analysis, interpretation, and conclusions? Do these prohibit publication or require revision?

Yes, the paper requires revision. The biggest issue is the interpretation of the results, especially the reliance on average scores to claim protein-level superiority. The work would be much more useful if it focused on identifying the best-performing method combinations and under what conditions they occur. It does validate that protein-level correction corrects batch effect effectively in more combinations of scenarios, especially using PQM other than MaxLFQ.

Another issue is the incomplete codebase. While a GitHub repository is provided, it is not usable in its current state: paths are hardcoded, essential folders are missing, and there's no way to obtain or reproduce the raw data used in the study. Harmony was implemented in the code but excluded from the benchmark without explanation.

Is the methodology sound? Does the work meet the expected standards in your field?

The benchmark design is solid overall. It includes a diverse set of real and reference datasets, and a reasonable variety of BECAs and PQMs. That said, there are a few methodological weaknesses that need to be addressed. In the large-scale dataset (1431 samples from 836 individuals), it's unclear how the authors handled data splitting. If samples from the same individual appear in both training and test sets, that would introduce data leakage and inflate performance metrics like MCC and R^2 . This possibility needs to be ruled out.

Moreover, the use of R^2 for sex prediction is questionable, this is a classification task, and R^2 is not appropriate. MCC is the correct choice here and should be used consistently. Including clustering metrics such as ARI or AMI would also strengthen the evaluation.

Is there enough detail provided in the methods for the work to be reproduced?

Not yet. The methods section is generally well written, but the code and supplementary material fall short of reproducibility standards. The GitHub repository lacks proper documentation, instructions, and example data. Paths in the scripts are hardcoded, necessary folders must be manually created, and there's no mapping between scripts and the figures they generate. A minimal working example and clear installation/setup guide (except for a list of packages to install) are missing. Without these, the work cannot be reliably reproduced by others, and the code is not a usable community resource in its current form.

Other comments

Large-scale multi-batch dataset

It's unclear how 1431 samples were collected from 836 individuals. If multiple samples were taken from the same individuals, these must be kept in the same data split during classification to avoid data leakage in R^2 or MCC calculations (Figure 5).

The paper mentions a LOESS fit in the "Batch effect correction" section, but it is only specified to be used on the T2D dataset in median centering batch effect correction.

Figures – General Comments

Many figures lack context: it could be clearer that the mean and standard deviation are values from all possible scenarios possible. For example, it seems the boxplots in fig 4b are probably the 12 values of all combinations of PQM + BECA, but it is not clear. This is confusing in many figures.

Figure 1

In panel a), make it clearer that 72 and 1495 are the number of samples in the study.

Figure 2

PCA plots (b-c) are difficult to interpret visually. Either improve readability or consider alternatives.

Figure 4

Panels c, e: BRInfo should indicate clearly that lower values are better.

Panels d–e: ANOVA is used, but it doesn't test for specific differences between peptide- and protein-level corrections.

Figure 5

Panel a): R^2 is nearly perfectly negative, how is this explained? Such a result is unexpected and suggests a strong negative correlation.

R^2 is not appropriate for classification tasks. MCC should be used instead, especially since the Methods say R^2 was only used for age prediction.

Figure 6

Panel a): Radar plots need a legend indicating which color is peptide vs. protein.

Panel b): The 72% result is shown as 0.72. Consider reformatting to avoid confusion.

Panel c): Several columns are empty. While it's helpful to summarize all scenarios, the figure is hard to interpret. Clarify the rationale for the "total scores". Consider adding a subsection to explain this in the "Quality assessment metrics" section.

Metrics like BRInfo and missed proteins are likely inverted to be on the same scale as other metrics, but this is not explained.

Why aren't unqualified proteins included in the scoring?

How is MCC scaled? Was the minimum value used to map the range to [0,10]? Some MCC values appear close to 0, but that doesn't match the expected [-1, 1] range. This needs clarification.

(Remarks on code availability)

The code is present, but no results could be reproduced because all paths are hardcoded and no data is available. The authors should add instructions on how to get the data to reproduce at least the results from the publicly available data. It should be easy to reproduce the results without having to manually modify the code to change the paths where the data are placed.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have answered my queries satisfactorily. I have no further comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

A README file with enough instructions for installing and running the application is available, the code can be easily executed after installing the required R packages.

Reviewer #3

(Remarks to the Author)

The authors made great improvements to the manuscript and added important results as suggested. Some minor changes could improve it further.

1- Some bar plots might be better represented as boxplots, or at least with error bars added. For example, in figure 6a-b and figure 3b, the results would be more robust if cross-validation was executed and results with the variation obtained presented.

2- I am wondering LOESS improves the BECAs. I think that since the authors included LOESS correction without any BECA, the current version could be enough, but we do not know if the BECAs alone are sufficient or LOESS is essential to get the best results.

3- Some figure legends could be more complete. For example, in figure 2, it could improve readability to include what -B and -C mean, though they are defined in the results.

(Remarks on code availability)

I was able to run the docker, but only the test.R file can be run.

the file .Rprofile prevents me from using the .R files. After removing source("renv/activate.R") I can start using them.

I am on windows and had to remove the options of having multiple cores to be able to use the R files, but without multiple cores it was taking forever, so I did not replicate past step 0.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

What are the noteworthy results?

- Via extensive simulations and tests involving multiple datasets, balance and imbalance, classification and regression tasks, the authors reported that about 70% of the time, protein-based batch correction is superior to early peptide-based batch correction.

Reply:

We thank the reviewer for clearly highlighting our key findings.

Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.

- The authors cite the latest works in the field. And update their analysis to also include machine learning task performances, which is increasingly important in bioinformatics. The findings, if correct, are fundamental, and can change/alter behaviors and analysis workflows, representing a best practice in the field. The idea of early correction has been purported in several works, including Goh et al, and Cuklina et al. This work is possibly, the first to investigate and test this extensively.

Reply:

We thank the reviewer for the insightful comments. As shown in the more comprehensive analysis in our revised manuscript, our findings are correct. We agree that that “*are fundamental, and can change/alter behaviors and analysis workflows, representing a best practice in the field*”.

Does the work support the conclusions and claims, or is additional evidence needed?

- The work is impressive. There are extensive tests across a wide variety of scenarios. However, I am concerned if the findings, based on a limited set of data, are sufficiently representative of all data in the proteomics field. Besides DDA and DIA platforms, I think we need more granularity on the types of platforms being tested and evaluated. perhaps there may be interesting findings here. for example, in DDA, the constituent peptides representing each protein may be different across samples/runs and each peptide carries different information on the batch effect. this may be less of a problem for DIA runs or TDA runs even. I think we need to be more specific and give more insight on the nature of the platform and acquisition paradigm, instead of arriving at more generalized statistic. I believe there are differences there.

Reply:

We appreciate this important point regarding the representativeness of our datasets and the need for more granular insights into different acquisition platforms. We acknowledge that data acquisition mode indeed contributed to technical variations in our Quartet datasets, as described in the revised manuscript (**Fig. 2c; Extended Data Fig. 4a**). While diverse technical factors as well as known biological groups were

involved in the Quartet datasets, we may not encompass all possible variations in proteomics data and batch-effect manifestations. We have added text in Discussion of the revised manuscript on the potential impact of different acquisition platforms (e.g., DDA, DIA, and TDA). Although various technical factors (i.e., mode and lab) were taken as part of the overall batch effects, the primary conclusion on the effectiveness of protein-level batch-effect correction by most BECAs in such complex and multi-sourced scenarios still stands. While a more comprehensive exploration of all platform types is beyond the scope of this study, we highlight this caveat as a crucial area for future research (**Discussion section, page 17, lines 440-444**).

I am also curious if the manifestation of batch effect, in this particular dataset, is also representative of all batch effects. I think there needs to be more control over the magnitude, nature, and distribution of batch effects in the data, to arrive at a more nuanced and meaningful understanding.

Reply:

We thank the reviewer for the comments and suggestions. What makes our study unique from previous benchmarking studies is that we have considered the data level and the quantification aggregation into the process of assessing the final batch-effect-correction performance. This multi-faceted workflow, inherently demanding a benchmarking dataset with representative batch effects, guided our choice of the Quartet datasets.

The leveraged Quartet datasets were all generated by the same batch of peptide mixtures from immortalized B-lymphoblastoid cell lines (LCLs) of the Quartet family. In this case, the sample preparation process remained the same and does not disturb the thereafter injections to the LC-MS instruments, which maximize the characteristics of batch effects existing in mass spectrometry. Moreover, four sample groups were sufficient for balancing between explicit (but multi-sourced) batch effects and complex biological background.

While we kept the Quartet datasets as one type of our benchmarking datasets, we also simulated a dataset and designed the corresponding balanced and confounded scenarios (**Methods section, pages 19-20, lines 485-498**). Performances of all QC metrics demonstrated a stepwise deterioration, with the introduction of the biological factor and random noises (Withbiol), followed by the scaling factor (Withscal), and finally the batch factor (Withbatch) to the initial gamma-distributed data matrix in the revised manuscript (**Extended data Fig. 1b; Extended data Fig. 1d; Extended data Fig. 2**). Consistent with findings from the Quartet datasets, the batch-effect removal at the protein level improved reproducibility within the same groups (**Fig. 3a**) and true biological grouping across all samples in the simulated datasets (**Fig. 4a**). These simulation results confirmed our observations from the Quartet datasets.

The authors need to also incorporate other metrics in their estimation of the batch factor in their data. Namely, I think PVCA, gPCA, etc, should also be incorporated to better estimate the component of batch information. Besides using the PCA plots, which can be more subjective in interpretation, I suggest the authors also consider using the side-by-side boxplots (<https://pmc.ncbi.nlm.nih.gov/articles/PMC5374662/>) to help review the PC-based association of batch effects, with concomitant effect sizes.

Reply:

We thank the reviewer for the suggestions. We have now incorporated the PVCA (Principal Variance Component Analysis) results as an important perspective of sample-based assessment, to quantify the proportion of variance explained by batch and biological factors in our datasets (**Fig. 4; Extended figure 9**). This provides a more objective and quantitative measure beyond PCA plots. We have also included side-by-side boxplots for PC-based association (**Extended figure 8b**), as suggested by the reviewer. PCA scatter plots were presented in extended figures to support the evidence of SNR performance (**Results section, page 11-12, lines 285-296**).

I also think the authors should weigh their observations on their dataset, on potential mechanism and explanation as to why 70% of the time, protein correction is superior. Saying it just is, can be dangerous and biased, as this may only apply to the nature of the datasets + batch effects, and cannot generalize elsewhere. It is more satisfying to find out exactly under what circumstances is this so. is it because, the information aggregation approaches tend to amplify or estimate the batch factor more accurately? if so, then a closer look at how the protein assembly methods combine this information is important. and we have to show via proof-by-simulation or mathematical proof that this is indeed the case, and that the improved results is because of signal amplification, making it easier to reliably estimate and remove the batch factor.

Reply:

We agree with the reviewer that this is a crucial point, and simply stating the observation without a deeper mechanistic explanation limits its generalizability. We have traced back to the precursor and the peptide data levels, and re-assessed all metrics at both the corrected (post-correction step) and aggregated (post-quantification step) levels.

In the Results section of the revised manuscript, we have explored these interactions (**Fig. 5; Extended data Fig. 10**). Through data simulation, we found that true negatives (TNs) and true positives (TPs) progressively increased with higher aggregation (from precursor to protein levels) after quantification. Conversely, false negative (FN) features showed an ascending trend with increasing aggregation levels (**Fig. 3b**). PVCA results further suggested that the quantification step amplified batch effects (**Fig. 3c**), though these effects were dramatically reduced at all corrected levels (**Results section, pages 11-12, lines 300-325**).

In the Discussion section, we have now elaborated on mechanisms explaining protein-level correction's potential superiority. We propose that protein quantification methods aggregate information from multiple peptides and precursors, leading to more robust estimation and amplification of true biological signals while diluting peptide-specific noise. However, residual artifacts from earlier-level corrections may be amplified during quantification, diminishing overall effectiveness. Thus, while BECAs reduce batch effects, subsequent quantification critically influences data quality by increasing FNs and potentially reintroducing batch-related variability (**Discussion section, pages 16-17, lines 409-430**).

While formal mathematical proof remains beyond manuscript's scope, we provide detailed qualitative explanation and propose this for future theoretical investigation.

in this regard, there should also be explicit insights into the BECAs themselves. The authors did show that methods like RUV are worse than others. Again, I think this may be a function of the nature of the batch effects/class effects in the data. Each BECA is designed with different batch assumptions in mind, so I think we need to first understand the batch factor, and batch distributions, before we make a generalized

comment that a particular is inferior. In other benchmarks and studies, RUV has been shown to be a top performer. So I think what is reported in this study, has to be tempered and fair. Moreover, this is not a benchmark study, but rather to demonstrate that early correction at the peptide level may be counter productive.

Reply:

We have clarified that our primary aim is to compare the optimal timing of batch-effect correction (precursors, peptides, and proteins), rather than to provide a definitive ranking of BECAs themselves. Despite this, we have reframed our manuscript and objectively described the performance on each specific metric of different BECAs and quantification methods (QMs).

In the Results section of the revised manuscript, RUV-III-C performed well especially in the simulated datasets (**Fig. 3b**). This is because the input negative control features were preset to be true unrelated to biological signals during data simulation. In real-world datasets (Quartet and ChiHOPE), we have re-defined the input negative control features, which were determined by a two-way analysis of variance (ANOVA) test across four grouped samples (Quartet) or three grouped QC samples (ChiHOPE). In the previous version of our manuscript, we filtered the features with neither significance across all samples nor interactive effect between the sample and the batch, but did not further select those batch-related features. In the present version of the revised manuscript, we have fixed this point and revised the relevant descriptions (**Methods section, page 22, lines 563-569**). The updated results suggested that the reliability of the negative controls contributed to the final batch-effect performance of RUV-III-C. For prediction performance assessment, Ratio yielded the highest MCC and R square values (**Fig. 6**), consistent with conclusions from previous MAQC-II studies on microarray gene-expression data by Luo et al. (PMID: 20676067)

In the revised Discussion section, we have also rephrased that the underlying assumptions of each BECA contributed to their distinct performances under specific scenarios leveraging multi-dimensional QC metrics. We agree that the effectiveness of a BECA can be highly dependent on the specific nature and distribution of batch effects. Furthermore, 80% batch-effect corrections yielded higher SNR values at protein level when combined with MaxLFQ (**Fig. 4b**), suggesting the critical role of quantification method. Finally, we recommended MaxLFQ-Ratio combination as the best practice, and emphasized the necessity of reference materials or accompanying QC samples in large-scale studies (**Discussion section, pages 15-16, lines 390-408**).

Finally, I think when it comes to early correction, it is possible to go even earlier than peptides. it is possible to go into the spectra realm, or do adjustments based on summaries or representations of spectra. I think this aspect should be answered so that it doesn't just because a binary comparison between peptides and proteins. The proteins are assembled from the peptides, but the peptides are inferred from the spectra, and this is bulk of the real data that comes from MS. The question I have is... how much batch/class information do we lose if we do not fully utilize spectra. some reports claim that only 30% of spectra are matched to peptide, and the remainder are lost. your paper will be much more valuable, if we can look into information maximization concerns. is batch estimation using all spectra, prior to peptide and protein inference, more valuable?

Reply:

We really appreciated this fascinating comment. We have expanded our benchmarking studies to the precursor level (earlier than the peptide level), combined with more batch-effect correction algorithms (BECAs) and scenarios, to confirm the robustness of batch-effect correction at the protein-level (comparing to the precursor and peptide levels) (**Results section, pages 6-7, lines 149-166**).

We agree that exploring batch correction at the spectral level is important, as spectra represent the most information of MS data which could be lost and mis-aggregated to precursor, peptide and protein levels. In the Discussion section of the revised manuscript, we have added text on future directions of batch correction on the raw spectra (**Discussion section, page 17, lines 444-447**). While our current study focuses on the data matrix at precursor, peptide and protein levels due to the complexity and computational demands of spectral-level analysis, we will investigate whether batch estimation and correction using all spectra (even unmatched ones) prior to peptide and protein inference could lead to superior results and better information maximization. This would extend our "data level of correction" concept beyond the current scope and highlights future directions.

Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?

- see above

Reply:

We sincerely hope that the reviewer find our responses and the revised manuscript thorough and convincing.

Is the methodology sound? Does the work meet the expected standards in your field?

-see above

Reply:

We sincerely hope that the reviewer find our responses and the revised manuscript thorough and convincing.

Is there enough detail provided in the methods for the work to be reproduced?

-the method is well described and should be reproducible.

Reply:

We thank the reviewer for the comment. We have further enhanced the reproducibility by addressing the specific code availability and data accessibility issues raised by Reviewer #3.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reply:

We great acknowledge and appreciate the valuable comments and suggestions of both reviewers in the co-reviewing process.

Reviewer #3 (Remarks to the Author):

What are the noteworthy results?

The study benchmarks multiple combinations of batch-effect correction algorithms (BECAs) and protein quantification methods (PQMs) to determine whether correction at the peptide or protein level yields better outcomes. The most important reported result is that protein-level correction outperformed peptide-level correction in 72.2% of tested combinations across several metrics. However, this emphasis on average performance hides what should be the most relevant finding: what is the best possible combination?

Reply:

We thank the reviewer for this insightful comment. We agree that researchers may ultimately be interested in identifying the best-performing pipeline for their specific needs. Thus, we have extensively revised our manuscript accordingly. Beyond the general conclusion on the superior performance at protein-level batch-effect correction, we now have also demonstrated the distinct performance of BECAs, QMs, and their combinations. In addition, the quantification process impacts the batch-effect correction performance, and similar findings have also been elucidated by tracing back to the precursor and the peptide data levels with re-assessment by all metrics (**Fig. 5; Extended figure 10**). Finally, our findings on prediction performance across different BECAs suggest that Ratio preserved the most information in the data for prediction tasks, showing both the highest MCC (0.41) and R-square (0.19) values (**Fig. 6**) (**Results section, pages 13-14, lines 343-364**). Together, we recommended Ratio at MaxLFQ-quantified protein level as the best practice for batch-effect correction in MS-based proteomics.

For example, in Figure 4a, the highest signal-to-noise ratio (SNR) is actually achieved when correcting at the peptide level using the Ratio method and MaxLFQ, contradicting the main conclusion. Similarly, Figure 4b–e lacks clarity on what the barplots and boxplots represent. If they are based on the 12 combinations of PQMs and BECAs, this aggregation makes protein-level correction appear more consistent, but hides the fact that peptide-level correction yielded the single best score. In the context of choosing methods for a real experiment, researchers are likely to care about the best-performing pipeline, not the most consistent across all combinations.

Reply:

We thank the reviewer for this comment. In the previous version of our manuscript, Ratio at the peptide level before MaxLFQ quantification did achieve the highest SNR in the Quartet confounded (Quartet-C) scenario. In the revised manuscript, after a more refined and comprehensive analysis that accounts for the impact of quantification methods from precursors to peptides (a crucial aspect highlighted in our revised Methods), the MaxLFQ-quantified protein level combined with Ratio now consistently yields superior SNR performance under the Quartet-C design (**Extended figure 7**). This refinement underscores our finding that the quantification process plays a significant role in batch-effect correction performance, which was not fully explored in the previous version. We also found that MaxLFQ-RUV-III-C was highly effective in removing batch effects across all three levels in simulated datasets (**Fig. 3b**), attributable to the reliability of negative control features.

Additionally, our primary aim is based on a comprehensive evaluation of multiple performance metrics (including, but not limited to, SNR) and across different datasets. Therefore, we have simplified **Figure 4**, and added PVCA analysis as the traditional way of quantifying the biological and the technical effects. The PVCA results showed that post-quantification batch-effect correction retained more proportion of variance explained by the biological factor (the sample) at the final quantified protein data level (**Extended figure 9; Extended table 4**) (**Results section, pages 10-11, lines 261-298**).

Will the work be of significance to the field and related fields? How does it compare to the established literature?

The work is relevant and useful to the proteomics field, where batch effects are a major concern. It addresses an important question that has been raised before: whether correction should happen at the peptide or protein level (e.g., Phua et al., PMC9411064). Until now, there hasn't been a systematic benchmark comparing these options across multiple methods and datasets. This study attempts to fill that gap and provides empirical support for a widely adopted, but under-justified, practice: correcting at the protein level.

Reply:

We sincerely appreciate the reviewer's comment on the importance of our work and for recognizing its significance and utility to the proteomics field.

However, the current presentation weakens the impact. Highlighting average scores across methods dilutes the value of identifying top-performing pipelines, which is ultimately what matters in applied settings.

Reply:

We appreciate the comments and suggestions and agree that identifying top-performing combinations is important. In the revised manuscript, as suggested by the reviewer, we have shifted our focus from presenting average performance scores to considering the distinct performance of specific BECAs, quantification methods (QMs), and their combinations across various metrics and scenarios. We now recommend the MaxLFQ-Ratio combination as the best practice for batch-effect correction in MS-based proteomics, particularly at the protein level.

The study would be of more significance if more methods were benchmarked. For example, waveica (<https://doi.org/10.1016/j.aca.2019.02.010>), harmony (doi: 10.1038/s41592-019-0619-0) or deep learning based BECAs (e.g. NormAE: <https://doi.org/10.1021/acs.analchem.9b05460>, BERNN: DOI: 10.1038/s41467-024-48177-5) .

Reply:

We appreciate the comments and suggestions and agree that benchmarking an even wider array of methods would further enhance the study's significance and provide a more comprehensive overview of the landscape of batch-effect correction methods. As suggested by the reviewer, we have included Harmony, NormAE, and WaveICA, which has greatly enriched our comparative analysis (**Methods section, pages 22-23, lines 572-589**). In the revised manuscript, Harmony did not show superiority on

either batch-effect correction, or classification and regression prediction tasks, compared with Ratio, RUV-III-C and ComBat (**Figures 3-4 and 6**). In addition, we encountered specific challenges that prevented us from fully exploring these methods in our analysis:

- 1) WaveICA: We have already included WaveICA2.0 to the Quartet as well as simulated datasets. However, when we applied WaveICA to our large-scale ChiHOPE proteomics data, we encountered extreme memory demands (e.g., requesting 117382.6 Gb for $\text{dim}(X) = [2369, 1495]$). This exceeded our computational resources, limiting its systematic benchmarking to smaller datasets only.
- 2) NormAE: While integrated, NormAE requires m/z and retention time as input, making it inherently suited for precursor-level correction in metabolomics, not directly comparable to the protein level in proteomics. We've clarified this distinction in the Results and Discussion sections.
- 3) BERNN: We faced persistent errors, likely due to hard-coded elements (i.e., the file path) in its codebase. Thus, we excluded it from our extended benchmarking efforts.

Does the work support the conclusions and claims, or is additional evidence needed?

Not entirely. The main claim, that protein-level correction is superior because it performs better on average, is misleading. It ignores the fact that the best performing combination in at least one major metric (SNR) is achieved at the peptide level. This undercuts the narrative that protein-level correction should be favored overall. The authors should revise their framing: instead of focusing on averages, they should highlight scenarios where each level of correction is preferable. The results are of interest, but the conclusions and claims could benefit from being revised.

Reply:

We thank the reviewer for the insightful comments. We fully agree that focusing on average performance can be misleading and that researchers ultimately seek to identify the best-performing pipeline for their specific needs. In the revised manuscript, we no longer emphasize the average performance (the average quality score integrated by linearly scaling each QC metric to 1 ~ 10) in Figure 6 of the previous version of the manuscript.

In the revised manuscript, our comprehensive benchmarking efforts have been based on both the Quartet and simulated datasets, expanded to three data levels (precursor, peptide, and protein levels) combined with three QMs (MaxLFQ, TopPep3, and iBAQ) and seven BECAs (Combat, Median centering, Ratio, RUV-III-C, Harmony, WaveICA2.0, and NormAE) under two designed scenarios (balanced and confounded).

In the revised Results section, we have confirmed the robustness of batch-effect correction at the protein-level (comparing to the precursor and peptide levels). There is a stepwise decrease of overall CVs from precursor to protein levels (**Fig. 3a**). Most BECAs applied to the protein level can effectively improve MCC values from uncorrected data matrices (**Fig. 3b**). Moreover, 80% batch-effect corrections yielded higher SNR values at protein level when combined with MaxLFQ quantification (**Figs. 4a-b; Extended figure 7**), and the superiority was also supported by PVCA results (**Extended figure 9**). Therefore, the results suggest that quantification prior to batch-effect removal may be a more practical strategy for accurately evaluating and correcting batch effect factors.

In the revised Discussion section, we have discussed the underlying assumptions of each BECA contributing to their distinct performances under specific scenarios, providing researchers with more

actionable guidance for method selection. For example, RUV-III-C performed well in each correction level, especially in simulated datasets (**Fig. 3**), where preset true negative control features (**Methods section, page 22, lines 563-569**) were incorporated as input. We believe that this revision greatly enhances the practical utility of our findings. We sincerely thank the reviewer again for the insightful suggestion.

Are there any flaws in the data analysis, interpretation, and conclusions? Do these prohibit publication or require revision?

Yes, the paper requires revision. The biggest issue is the interpretation of the results, especially the reliance on average scores to claim protein-level superiority. The work would be much more useful if it focused on identifying the best-performing method combinations and under what conditions they occur. It does validate that protein-level correction corrects batch effect effectively in more combinations of scenarios, especially using PQM other than MaxLFQ.

Reply:

We thank the reviewer for this insightful comment. As we stated in our response to the reviewer's earlier comments, we have revised the manuscript extensively and no longer conclude solely on the average quality scores of linearly scaled QC metrics. We have conducted a broader benchmarking study on more data levels, more BECAs, and more datasets. In the revised manuscript, we confirmed the robustness of protein-level batch-effect correction in multi-dimensional quality assessment, and described the specific performance of each QC metric (**Figs. 3-4**). For example, despite that most BECAs yielded higher SNR values at protein level, RUV-III-C uniquely showed power of removing batch effects across all three data levels in the simulated data (**Extended figures 7b-c**), which incorporated preset known negative control features as input (**Fig. 4a**). Furthermore, we demonstrated the best prediction performance of Ratio-MaxLFQ combination, providing a practical workflow for correcting batch effect factors, thus helping the thereafter downstream analysis.

Another issue is the incomplete codebase. While a GitHub repository is provided, it is not usable in its current state: paths are hardcoded, essential folders are missing, and there's no way to obtain or reproduce the raw data used in the study. Harmony was implemented in the code but excluded from the benchmark without explanation.

Reply:

We sincerely apologize for the incompleteness in the codebase, and we have taken immediate action to rectify it:

- 1) All codes for the preprocessing and statistical analyses are available on GitHub at <https://github.com/QiaochuChen/proteomics-batch-effect-correction-benchmarking>. All hardcoded paths have been removed and replaced with relative paths (**Code availability, page 26, lines 666-671**).
- 2) All benchmarking datasets are available at https://figshare.com/articles/dataset/All_corrected_uncorrected_data_matrices_at_precursor_peptide_protein_levels_/29567333?file=56278295 (**Data availability, page 26, lines 656-665**).

- 3) All statistical results and figures are available at https://figshare.com/articles/dataset/All_statistical_results_and_figures_generated_for_the_manuscript/_29567366?file=56278352 (**Data availability, page 26, lines 656-665**).
- 4) Harmony has been included in the current benchmarking study of the revised manuscript.

Is the methodology sound? Does the work meet the expected standards in your field?

The benchmark design is solid overall. It includes a diverse set of real and reference datasets, and a reasonable variety of BECAs and PQMs. That said, there are a few methodological weaknesses that need to be addressed. In the large-scale dataset (1431 samples from 836 individuals), it's unclear how the authors handled data splitting. If samples from the same individual appear in both training and test sets, that would introduce data leakage and inflate performance metrics like MCC and R^2 . This possibility needs to be ruled out.

Reply:

We sincerely apologize for the ambiguity of the description, and we have clarified the number of subjects and samples in the revised Results and Methods sections. The number 836 is the total number of subjects that participated in the clinical trials and agreed to have their biospecimen profiled with multi-omics technologies. One subject only participated plasma collection at baseline but no at other timepoints, thus was excluded from the final cohort (<https://www.biosino.org/node/project/detail/OEP00002924>). Meanwhile, there were a total number of 750 subjects who agreed to participate in the proteomics study, and provided 1431 plasma samples at baseline or week 24. The training set (394 subjects) and the validation set (356 subjects) were divided by subject (not by sample) according to their inclusion time in the clinical trials. All samples from the same subject will be either in the training or test set, but no in both, which prevents data leakage (**Methods section, pages 25, lines 648-655**).

Moreover, the use of R^2 for sex prediction is questionable, this is a classification task, and R^2 is not appropriate. MCC is the correct choice here and should be used consistently. Including clustering metrics such as ARI or AMI would also strengthen the evaluation.

Reply:

We sincerely appreciate the suggestion. We completely agree with the reviewer that R^2 is not suitable for classification tasks. In the revised manuscript, we have corrected this: MCC (Matthews Correlation Coefficient) is now consistently used for the sex classification task, while R^2 is reserved for the age regression task (**Results section, page 13, lines 352-354**). While ARI/AMI offer valuable insights into unsupervised grouping, incorporating a comprehensive clustering analysis would expand the scope beyond our current core objectives and the direct impact on downstream predictive modeling.

Is there enough detail provided in the methods for the work to be reproduced?

Not yet. The methods section is generally well written, but the code and supplementary material fall short of reproducibility standards. The GitHub repository lacks proper documentation, instructions, and example data. Paths in the scripts are hardcoded, necessary folders must be manually created, and there's no mapping between scripts and the figures they generate. A minimal working example and clear

installation/setup guide (except for a list of packages to install) are missing. Without these, the work cannot be reliably reproduced by others, and the code is not a usable community resource in its current form.

Reply:

We sincerely apologize for the ambiguity, and we have re-organized all code scripts. Specifically, we have provided a docker and a README file for reproducing our results. In the revised scripts, all hardcoded file paths have been removed and all directories can be created automatically in each ordered script for saving all results and figures. Each script was renamed for mapping to the six main figures and their relevant extended figures.

Other comments

Large-scale multi-batch dataset

It's unclear how 1431 samples were collected from 836 individuals. If multiple samples were taken from the same individuals, these must be kept in the same data split during classification to avoid data leakage in R^2 or MCC calculations (Figure 5).

Reply:

We sincerely apologize for the ambiguity of the description, and we have clarified the number of subjects and samples in the revised Results and Methods sections. Yes, samples from the same individual were kept in the same data split during classification to avoid data leakage.

The paper mentions a LOESS fit in the “Batch effect correction” section, but it is only specified to be used on the T2D dataset in median centering batch effect correction.

Reply:

We sincerely appreciate the comment. We have clarified that the LOESS fitting process has been introduced to each tested BECAs on the ChiHOPE dataset in the revised Results and Methods sections. The LOESS fitting was proposed by Čuklina et al. (**Reference 23**) to fix the injection order-specific MS signal drifts within each discrete batch with a relatively large number of samples before typical batch-effect correction workflow for large-scale proteomics data. For the Quartet and simulated datasets, the injection order-specific MS signal drifts problem was not observed because only a small number of samples were processed in a distinct batch, thus LOESS was not applied to these datasets. In the revised manuscript, we have demonstrated the benefit of LOESS before applying each BECA for the ChiHOPE dataset (**Extended figures 11-12**) (**Results section, page 13, lines 338-342**).

Figures – General Comments

Many figures lack context: it could be clearer that the mean and standard deviation are values from all possible scenarios possible. For example, it seems the boxplots in fig 4b are probably the 12 values of all

combinations of PQM + BECA, but it is not clear. This is confusing in many figures.

Reply:

We sincerely apologize for the ambiguity of the description. In the revised manuscript, we have added the detailed number of combinations in the expanded benchmarking study (**Results section, page 7, lines 161-166**).

Figure 1

In panel a), make it clearer that 72 and 1495 are the number of samples in the study.

Reply:

We thank you for the suggestion. We have improved **Figure 1** accordingly, and clarified what each point represents for as well as the specific number of samples in the figure legends.

Figure 2

PCA plots (b-c) are difficult to interpret visually. Either improve readability or consider alternatives.

Reply:

We thank you for the suggestion. We have improved **Figure 2**, and added SNR and PVCA results as panels (b-c). The batch-effect diagnosis of the large-scale ChiHOPE dataset has been moved to **Extended figures 11-12**.

Figure 4

Panels c, e: BRinfo should indicate clearly that lower values are better.

Reply:

We thank you for the suggestion. Since BRinfo is not a commonly used metric for batch-effect correction, we have replaced it with PVCA results.

Panels d–e: ANOVA is used, but it doesn't test for specific differences between peptide- and protein-level corrections.

Reply:

We thank you for the suggestion. We have replaced ANOVA by Student's T tests between peptide- and protein-, peptide- and precursor-, and protein- and precursor-level corrections in the revised panel a).

Figure 5

Panel a): R^2 is nearly perfectly negative, how is this explained? Such a result is unexpected and suggests a strong negative correlation.

Reply:

We thank you for the comment. As RUV-III-C requires negative control features as input, we carefully checked the methodology in the previous manuscript, where we have determined the input negative control features by a two-way analysis of variance (ANOVA) test across three grouped QC samples in ChiHOPE dataset. In the previous manuscript, we filtered features with neither significance across all samples nor interactive effect between the sample and the batch, without further selecting those batch-related features. This misidentification of negative controls likely led to the unexpected, nearly perfectly negative R^2 value, as the algorithm might have inadvertently amplified or inverted the batch effect instead of correcting it. In the present revised manuscript, we have fixed this problem and revised the relevant descriptions (**Methods section, page 22, lines 563-569**). The updated results have been moved to **Figure 6** in the revised manuscript, suggesting that the reliability of the negative controls contributed to the final batch-effect performance of RUV-III-C. We thank the reviewer again for catching this discrepancy.

R^2 is not appropriate for classification tasks. MCC should be used instead, especially since the Methods say R^2 was only used for age prediction.

Reply:

We appreciate this important comment. We've fixed this in the revised manuscript: MCC is now used for the sex classification task, while R^2 is reserved for age prediction.

Figure 6

Panel a): Radar plots need a legend indicating which color is peptide vs. protein.

Panel b): The 72% result is shown as 0.72. Consider reformatting to avoid confusion.

Panel c): Several columns are empty. While it's helpful to summarize all scenarios, the figure is hard to interpret. Clarify the rationale for the "total scores". Consider adding a subsection to explain this in the "Quality assessment metrics" section.

Metrics like BRInfo and missed proteins are likely inverted to be on the same scale as other metrics, but this is not explained.

Why aren't unqualified proteins included in the scoring?

How is MCC scaled? Was the minimum value used to map the range to [0,10]? Some MCC values appear close to 0, but that doesn't match the expected [-1, 1] range. This needs clarification.

Reply:

We sincerely appreciate the reviewer for the detailed comments and suggestions for correction regarding **Figure 6**. In the revised manuscript, we have removed the original Figure 6 (which presented average quality scores from linearly normalized metrics) as we agree with the reviewer's previous comment that emphasizing average performance across all scenarios can be misleading and dilute the specific performances in different QC metric. Instead, we have extensively revised our presentation of results:

The content from the original **Figure 5** (the prediction performance of ChiHOPE dataset) has been moved to become the new **Figure 6**. We have introduced a new **Figure 5** (exploring the interactions between quantification and correction) to further elucidate the impact of quantification on batch-effect correction performance, tracing back to precursor and peptide data levels. This restructuring of the whole manuscript allows us to provide a more nuanced and insightful discussion regarding how final batch-effect correction performance is impacted by the quantification and the correction steps across different data levels. Consequently, the specific questions related to the previous Figure 6 are now no longer relevant to the revised manuscript's figures and discussion.

Reviewer #3 (Remarks on code availability):

The code is present, but no results could be reproduced because all paths are hardcoded and no data is available. The authors should add instructions on how to get the data to reproduce at least the results from the publicly available data. It should be easy to reproduce the results without having to manually modify the code to change the paths where the data are placed.

Reply:

We thank the reviewer for this important point regarding code for result reproducibility and data availability. We sincerely apologize for the initial incompleteness. As detailed in our reply to a similar comment above (under “Another issue is the incomplete codebase”), we have already taken immediate action to rectify these issues.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have answered my queries satisfactorily. I have no further comments.

Reply:

We greatly acknowledge and appreciate the reviewer for all constructive suggestions.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #2 (Remarks on code availability):

A README file with enough instructions for installing and running the application is available, the code can be easily executed after installing the required R packages.

Reply:

We greatly acknowledge and appreciate the valuable comments and suggestions of both reviewers in the co-reviewing process.

Reviewer #3 (Remarks to the Author):

The authors made great improvements to the manuscript and added important results as suggested. Some minor changes could improve it further.

Reply:

We greatly acknowledge and appreciate the valuable comments and suggestions of both reviewers in the co-reviewing process.

1- Some bar plots might be better represented as boxplots, or at least with error bars added. For example, in figure 6a-b and figure 3b, the results would be more robust if cross-validation was executed and results with the variation obtained presented.

Reply:

Thank you for your comments. The data points are plotted individually in figure 6a-b and figure 3b, and we have clarified this in the figure legends.

2- I am wondering LOESS improves the BECAs. I think that since the authors included LOESS correction without any BECA, the current version could be enough, but we do not know if the BECAs alone are sufficient or LOESS is essential to get the best results.

Reply:

Thank you for your suggestions. We have added the supplementary figure 13 to demonstrate the benefit from LOESS in the classification task.

3- Some figure legends could be more complete. For example, in figure 2, it could improve readability to include what -B and -C mean, though they are defined in the results.

Reply:

Thank you for your comments. We have provided the full term for each abbreviation in the figure legends.

Reviewer #3 (Remarks on code availability):

I was able to run the docker, but only the test.R file can be run.

the file .Rprofile prevents me from using the .R files. After removing source("renv/activate.R") I can start using them.

I am on windows and had to remove the options of having multiple cores to be able to use the R files, but without multiple cores it was taking forever, so I did not replicate past step 0.

Reply:

We appreciate your patience, and are really sorry for the inconvenience.

Our code leverages multi-core processing for performance optimization. We strongly recommend running this Docker container on a Linux or macOS environment, as the R parallelization functions (mclapply) rely on the fork mechanism, which is not supported natively on Windows.

If you are using Docker Desktop on Windows, you may experience significant performance degradation when running multi-core tasks, as the container's multi-core functionality is limited by the underlying Windows Subsystem for Linux (WSL).

We have added this statement to the README file, and removed the file .Rprofile in the GitHub and Docker.