Enhanced protein identification and quantification improve differential expression analysis

Dejan Stepec, Maximilien Burq, Bryan Crampton, Rehan Chinoy, Melissa Miao, Charlie Kim, Peter Cimermancic

Tesorai Inc.

Abstract

Accurate protein identification and quantification are central to extracting meaningful biological insights from mass spectrometry (MS)-based proteomics. However, limitations in peptide detection, data completeness, and quantification precision often constrain the accuracy of detection of differentially expressed proteins (DEPs). In this study, we present an optimized proteomics workflow that enhances both protein identification rates and quantitative accuracy and normalization strategies, thereby improving the sensitivity and reliability of differential expression analysis.

Implementation of the enhanced workflow resulted in a significant increase in protein identifications—up to 25% more unique proteins compared to standard pipelines—and improved quantification consistency across replicates. Downstream differential expression analysis revealed a higher number of confidently detected DEPs, including low-abundance regulatory proteins often missed by conventional methods.

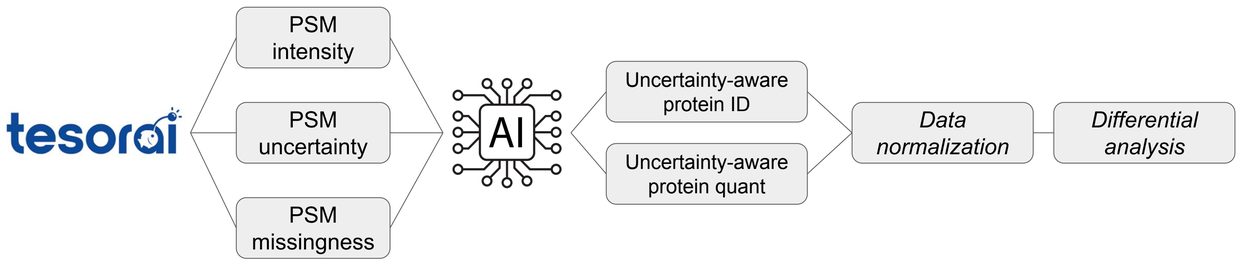

Figure 1: The optimized Tesorai proteomics workflow. An AI model integrates PSM intensity, uncertainty, and missingness data to enhance protein identification and quantification, forming the basis for subsequent data normalization and differential analysis.

Methods

Uncertainty-aware protein quantification

The standard method for protein quantification consists in taking the average, or max across all the peptides mapping to that protein (sometimes accounting for protein length – IBAQ method). This loses important information related to peptide digestibility, ionizability, as well as observed uncertainty on peptide identification and intensity estimation. Our approach instead combines these informations to produce a holistic protein-level quantity estimation.

Uncertainty-aware protein identification

Protein groups are typically identified by taking the maximum up the identification scores of their associated peptides. This ignores information about all the peptides which have scores lower than the maximum. Given wide heterogeneity in peptide ionizability and digestibility, most observed peptides are thus never taken into account. Our approach combines all available information, resulting in more confident protein group identification.

Optimized normalization

Normalization is performed by selecting the sample with the highest data completeness as a reference, and then linearly aligning all other samples to it.

Conclusion

These findings demonstrate that increasing the accuracy and completeness of protein identification and quantification substantially enhances the biological interpretability of proteomics data. The optimized workflow provides a robust framework for researchers seeking to increase the depth and reproducibility of MS-based proteomic analyses, ultimately enabling more sensitive detection of subtle proteomic changes across biological conditions.

Improved quantification

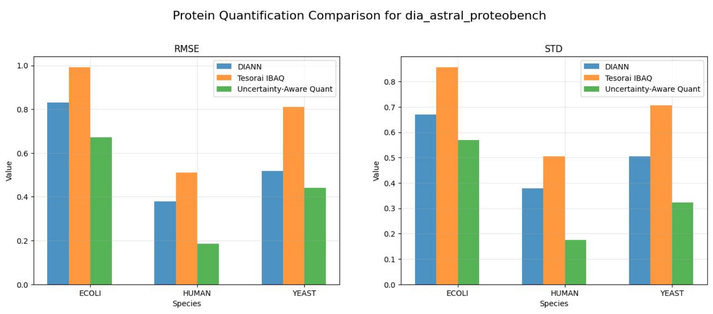

Figure 2: Improved quantification accuracy and precision with uncertainty-aware quantification. Using a 3-mix species DIA Astral (Van Puyvelde 2022) dataset, we benchmark the quantification accuracy by measuring the difference between expected and observed log-2 fold-change intensities. Uncertainty-Aware Quantification method shows consistently lower RMSE and STD than Tesorai iBAQ and DIA-NN, indicating higher accuracy and reproducibility.

Increased identifications

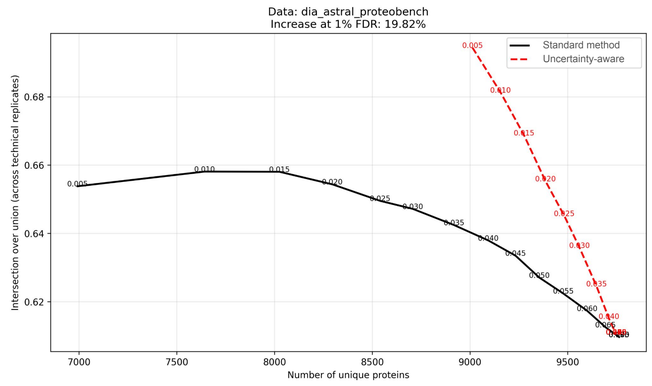

Figure 3: Improved protein identification using uncertainty-aware method. Uncertainty-aware protein identification approach (red, dashed) compared to the standard method (black, solid) on the DIA Astral dataset. The optimized method yields 19.8% more unique proteins at a 1% FDR (0.010). It is also more consistent across replicates, measured by the ratio between the number proteins found in all 3 replicates (intersection) over to the number found in any one of the 3 (union).

Optimized data normalization

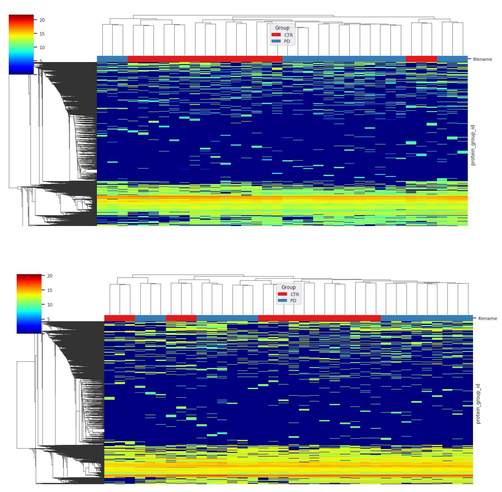

Figure 4: Effect of linear alignment-based normalization on data consistency. The clustered heatmaps display protein intensities (rows) across all samples (columns). The top colorbar indicates Control (CTR, blue) and PD (red).

(Top) Before normalization, prominent vertical banding and sample-driven clustering indicate technical batch effects.

(Bottom) After linear alignment-based normalization, inter-sample variation is visibly reduced, enhancing the underlying biological structure and improving data quality for differential analysis.

End-to-end workflow reduces false positives… … and identifies 2x more differentially expressed proteins in people with Parkinson's Disease

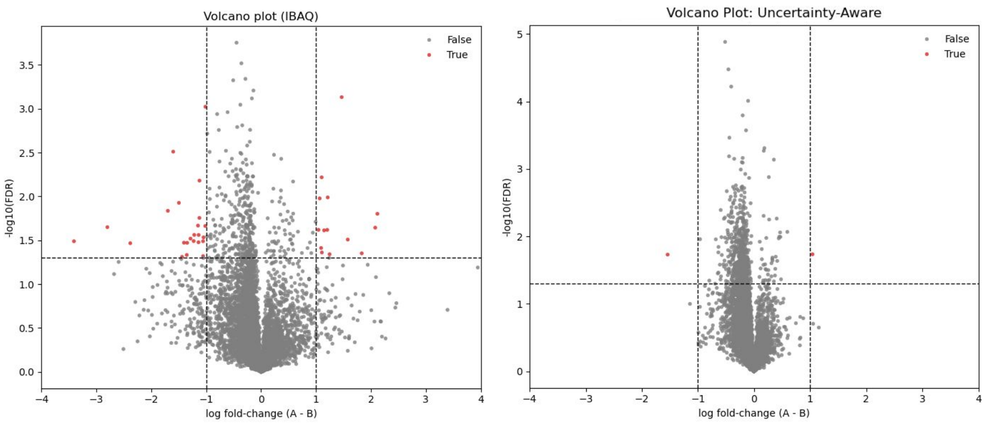

Figure 5: Comparison of differential expression results between traditional iBAQ and the new UAQ approach in a QC dataset (Van Puyvelde 2022). We selected only human proteins, which were mixed in equal amounts across all six files. As no biological difference is expected, all detected DEPs represent false positives. UAQ markedly reduces spurious detections ($p < 0.01$, $|\log_2 FC| > 1$).

Comparative analysis between conventional iBAQ and the Uncertainty-Aware Quantification (UAQ) approach further highlights the benefits of the optimized workflow. In a QC dataset with identical samples, UAQ effectively suppresses false positives, demonstrating superior quantitative stability. Conversely, in a Parkinson's disease cohort, UAQ identified more biologically relevant differential proteins, underscoring its enhanced sensitivity in real biological contexts.

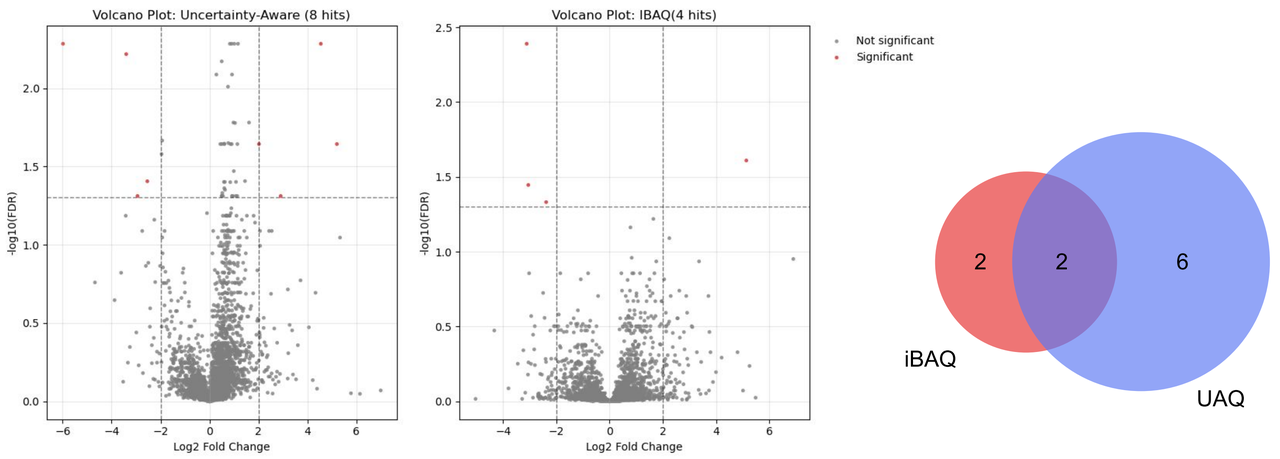

Figure 6: Application to a Parkinson's disease (Zhu et al 2024) dataset reveals that UAQ detects a higher number of biologically relevant DEPs compared to iBAQ ($q < 0.02$, $|\log_2 FC| > 2$), indicating improved sensitivity and accuracy for identifying true proteomic changes.

Genes identified (**red:** known role in Parkinson's Disease, **bold**: known role in Alzheimer's Disease)

- UAQ: **NDRG2**, **APOA4, CTNNA2**, **GRIPAP1**, **LAMB2**, **PMP2**, **MYH11**, PPFIA3
- IBAQ: **NDRG2**, **APOA4**, **MAG**, TMCO1

References:

- Zhu et al 2024 Single-cell transcriptomic and proteomic analysis of Parkinson's disease brains (PXD030142)
- Van Puyvelde 2022 A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics (PXD028735)

www.tesorai.com info@tesorai.com