Supporting Information - 2 - Figures: *Complete Simulation of timsTOF PASEF Raw Datasets with Timsim Enables Precise Benchmarking of False Discovery and Phosphosite Localization Error Rates*

David Teschner[1,2], Zixuan Xiao[3], Tim Maier[1,2], David Gomez-Zepeda[4,5], Mateusz K. Łącki[6], Michał P. Startek[7], Ute Distler[6], Tanja Ziesmann[6], Mathias Wilhelm[3,8], Andreas Hildebrandt[1,2], and Stefan Tenzer[4,5,6]

[1]Institute of Computer Science, Johannes Gutenberg University, 55128 Mainz, Germany
[2]Institute for Quantitative and Computational Biosciences (IQCB), Johannes Gutenberg University, 55128 Mainz, Germany
[3]Computational Mass Spectrometry, School of Life Sciences, Technical University of Munich, Freising 85354, Germany
[4]Helmholtz Institute for Translational Oncology, Mainz, Germany
[5]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[6]University Medical Center, Johannes Gutenberg University, 55131 Mainz, Germany
[7]Department of Mathematics, Informatics, and Mechanics, University of Warsaw, 02-097 Warsaw, Poland
[8]Munich Data Science Institute (MDSI), Technical University of Munich, Garching 85748, Germany

March 3, 2026

1 Supplementary Figures

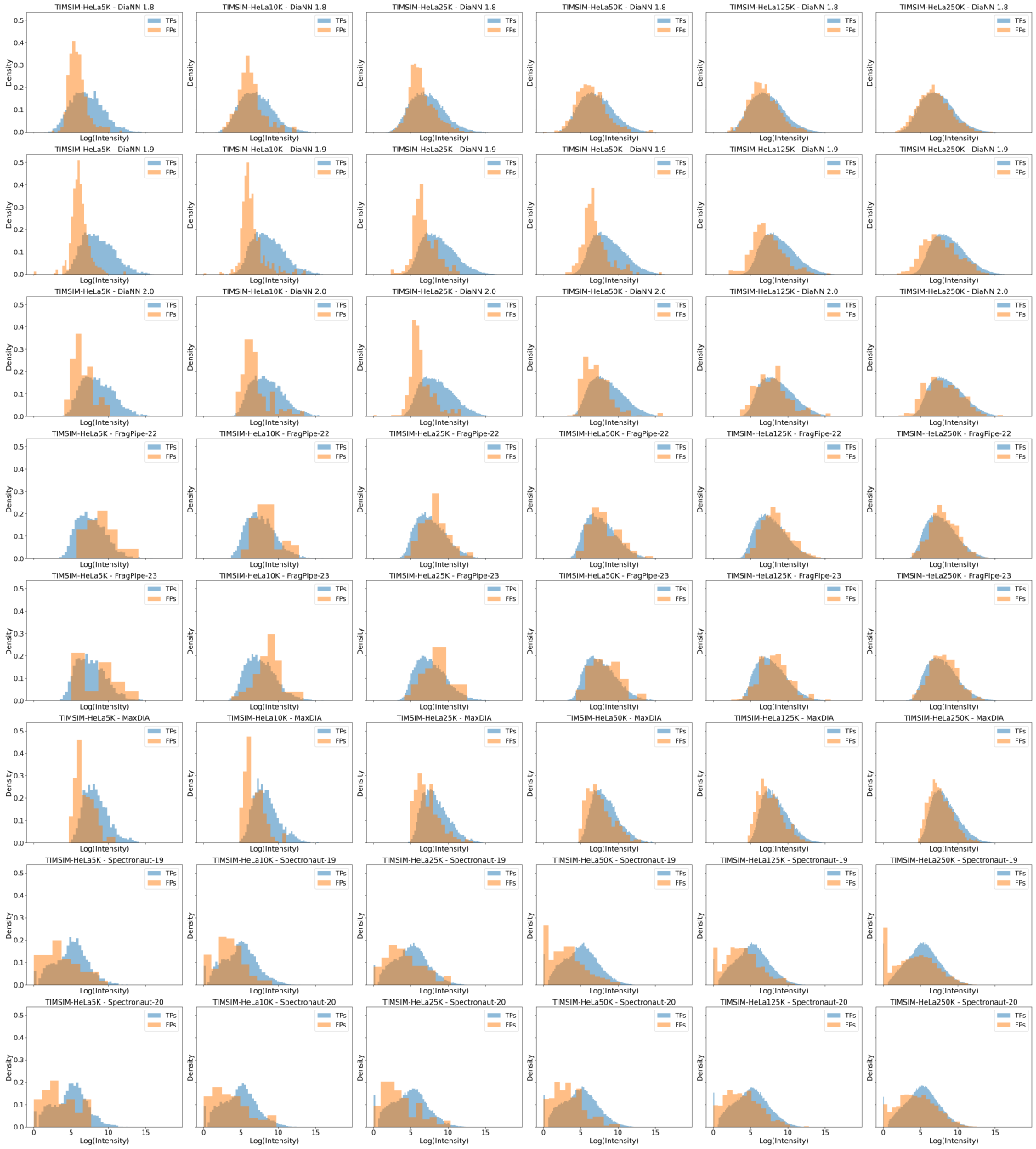

Figure 1: **Intensity distribution of true positives (TPs) and false positives (FPs) across all benchmarked software tools.** With the exception of FragPipe diaTracer, most tools exhibit a tendency to report FPs at lower precursor intensities, particularly at lower sample complexities. This bias diminishes as sample complexity increases. Observe that in a perfect case, the results of raw data analysis should not leave place for learning how to discern real and false discoveries, so the above distributions are expected to be the same.

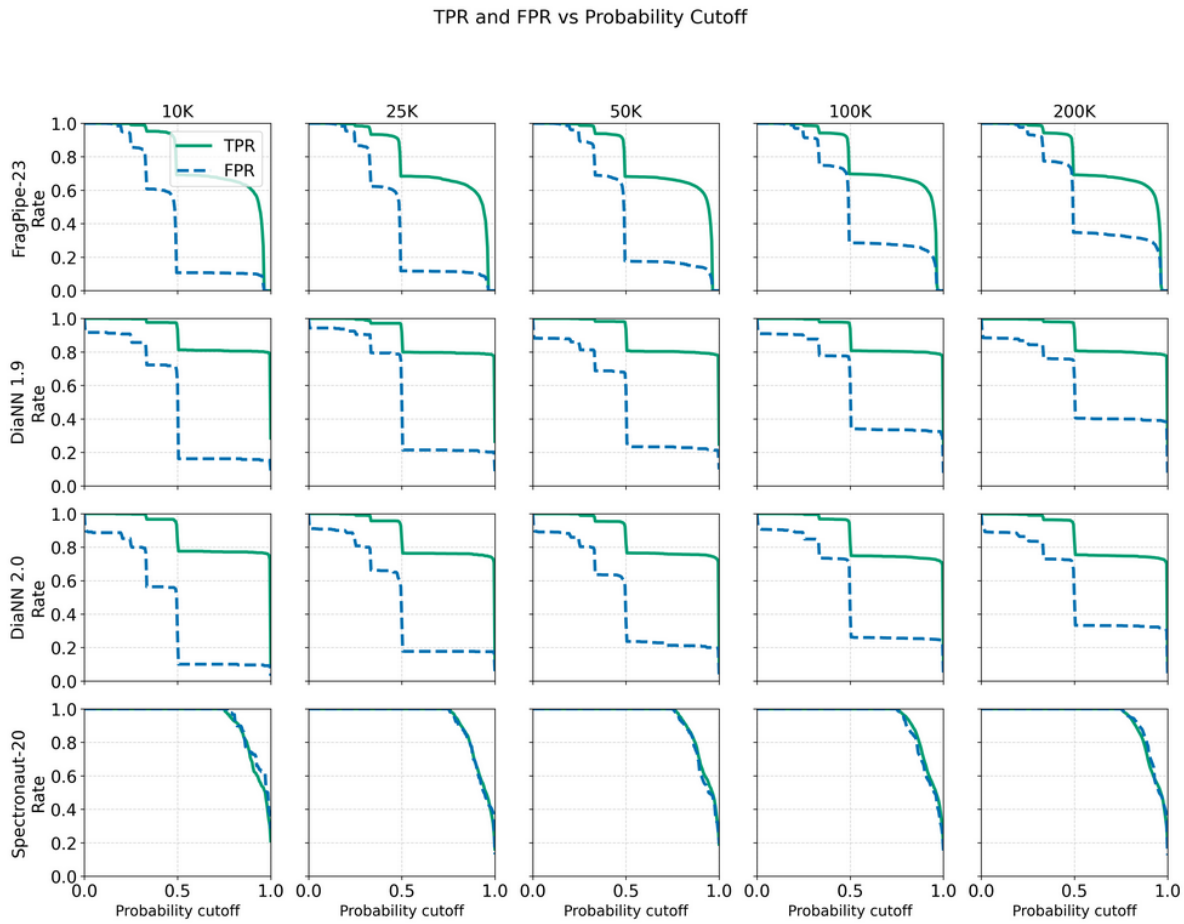

Figure 2: **Separation of true and false phosphosite identifications across probability cutoffs.** For FragPipe v23, DIA-NN v1.9, and DIA-NN v2.0, false positives decline sharply with increasing site localization probability, while true positives decrease more gradually. Notably, the steepest drop in false positives for both tools occurs above a cutoff of 0.5. In contrast, Spectronaut's scoring appears less effective at distinguishing true from false identifications, showing a weaker separation across the probability range.

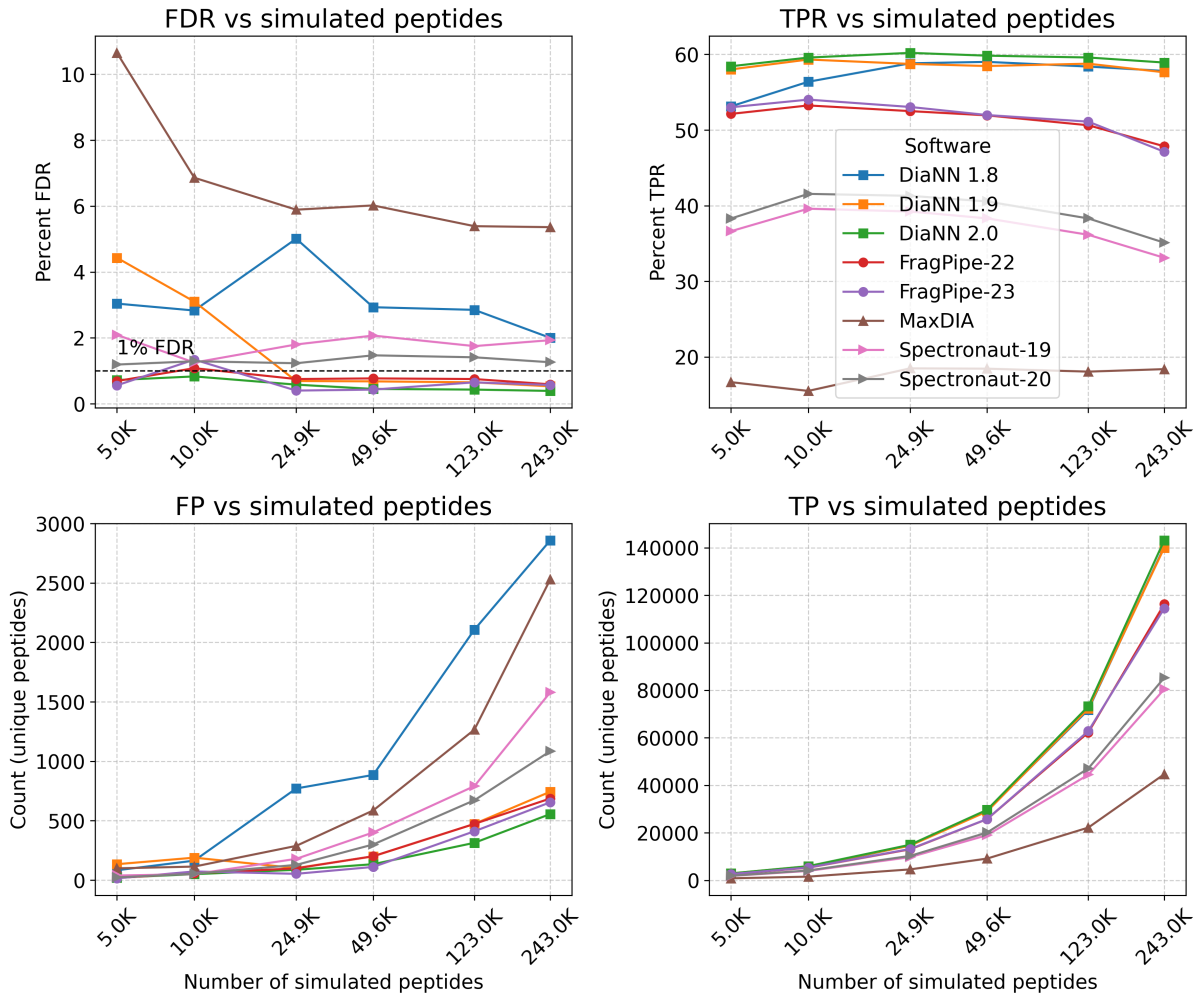

Figure 3: **Benchmarking of software tools for processing Hela dia-PASEF datasets, evaluating false discovery rate (FDR) control and identification performance across varying simulated dataset complexities (peptide level). a)** Relative true FDR as a function of dataset complexity (peptide level). Only FragPipe in diaTracer mode and DIA-NN v2.0 consistently maintain FDR at or below the target threshold of 1%. Other tools either show elevated FDR across all complexities (Spectronaut), only achieve target-level FDR in high-complexity datasets (DIA-NN v1.8, v1.9), or do not reach the target FDR at all (MaxDIA). **b)** True positive rate (TPR) as a function of dataset complexity. DIA-NN shows the highest number of true positives. **c, d)** Absolute numbers of false positives (FPs) and true positives (TPs), respectively, across all tools and dataset complexities.

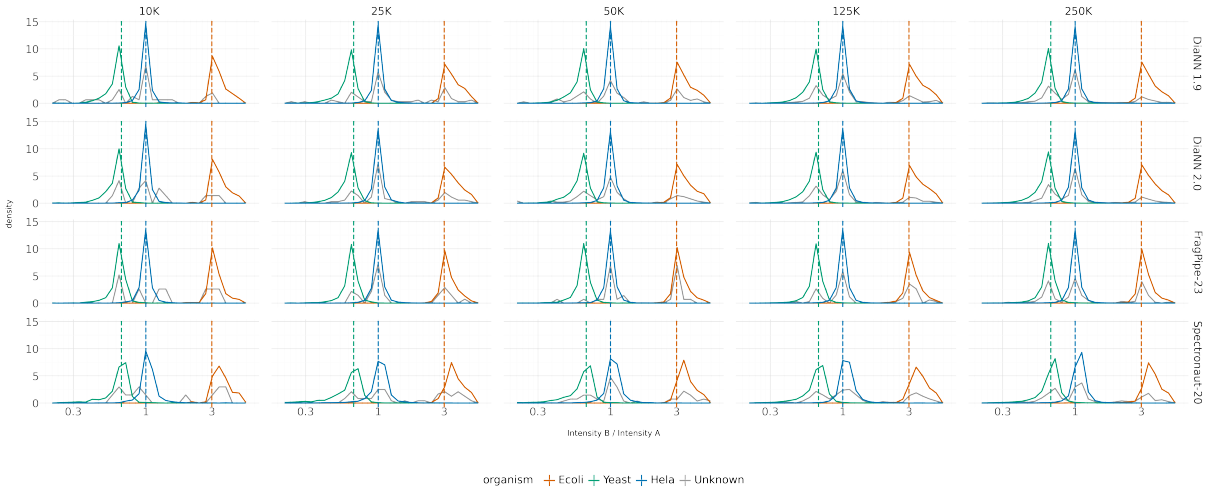

Figure 4: **Label-free quantification benchmarking for dia-PASEF data using a simulated HeLa–Yeast–*E. coli* mixture.** Rows show the relative intensity distributions reported by each software tool, split by organism. False positive identifications, those not present in the simulation, are labeled as unknown. Each column represents a different simulated sample complexity, ranging from 10,000 to 250,000 peptides. The dashed vertical lines indicate the ground-truth fold changes between replicates, colored by organism.

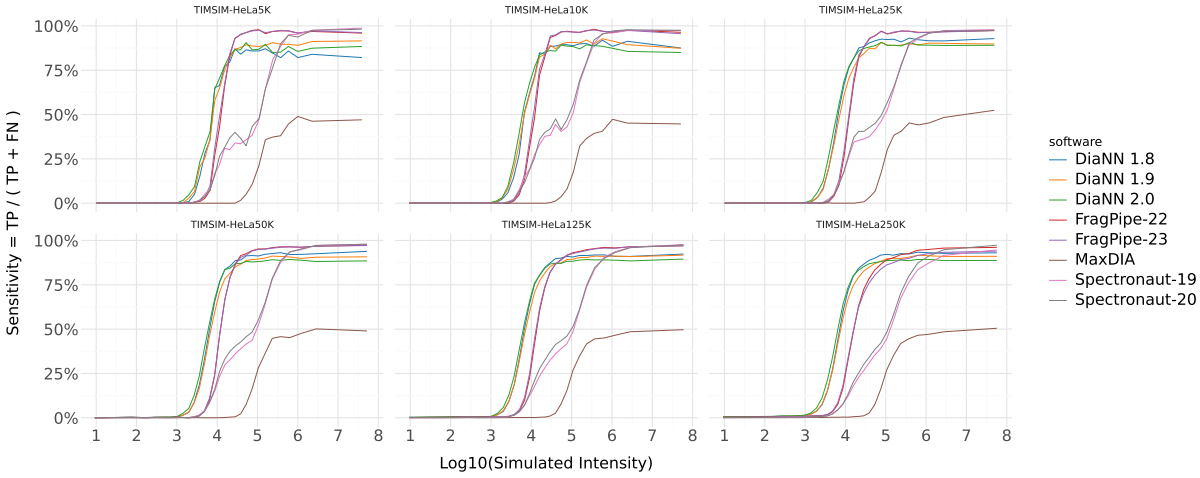

Figure 5: **Sensitivity analysis for different software tools on dia-PASEF Hela complexity ramp.** DIA-NN in all versions turned out most sensitive in this benchmark, picking up IDs with lower simulated intensity before any other software. Interestingly, FragPipe and Spectronaut turned out slightly more sensitive for higher intensity signals, implicating that there might be a tradeoff where along the signal intensity range software reliably picks up IDs. This effect became smaller as simulated dataset complexity increased.

Figure 6: MS/MS-based identification (peptide ions) variation of replicates from different simulation setups for Match-Between Run evaluation.

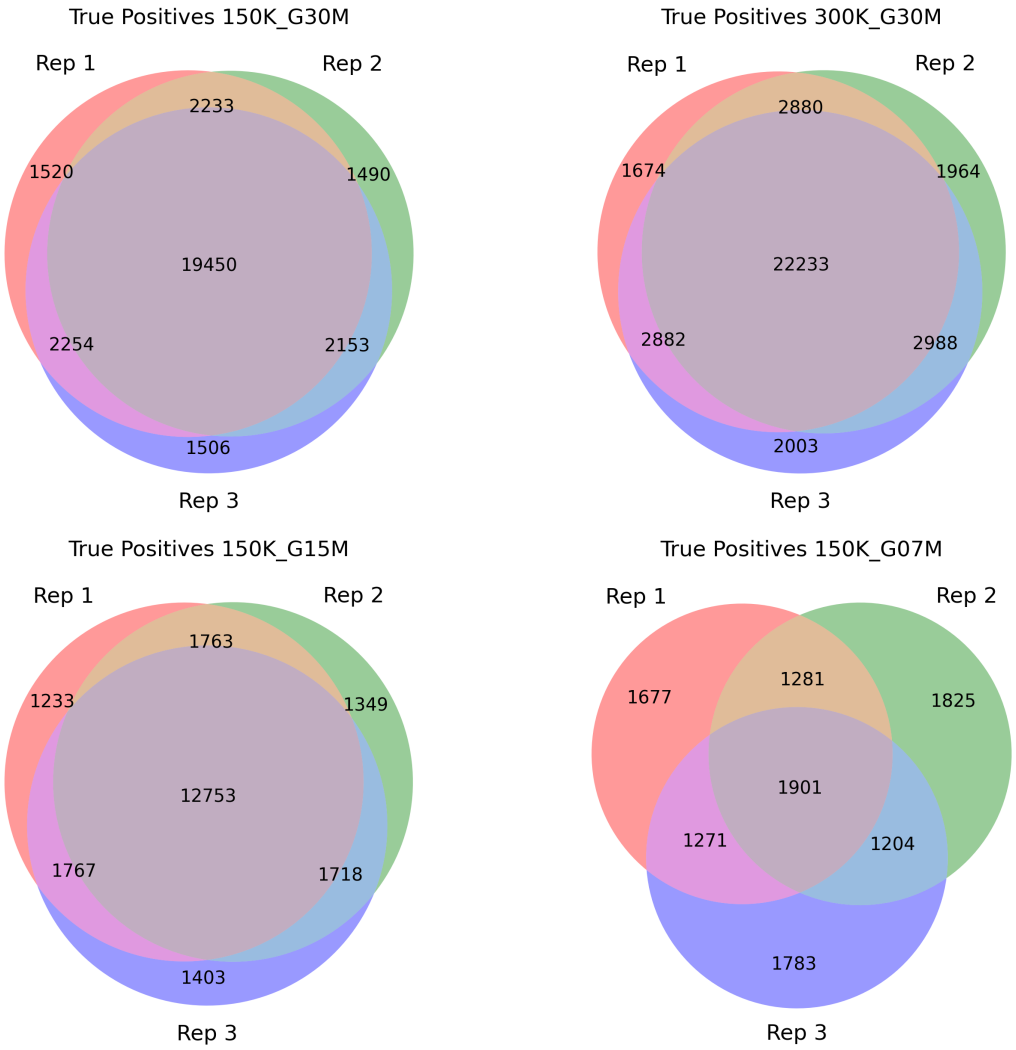

Figure 7: MS/MS-based true positive (TP) identification (peptide ions) variation of replicates from different simulation setups for Match-Between Run evaluation by MaxQuant.

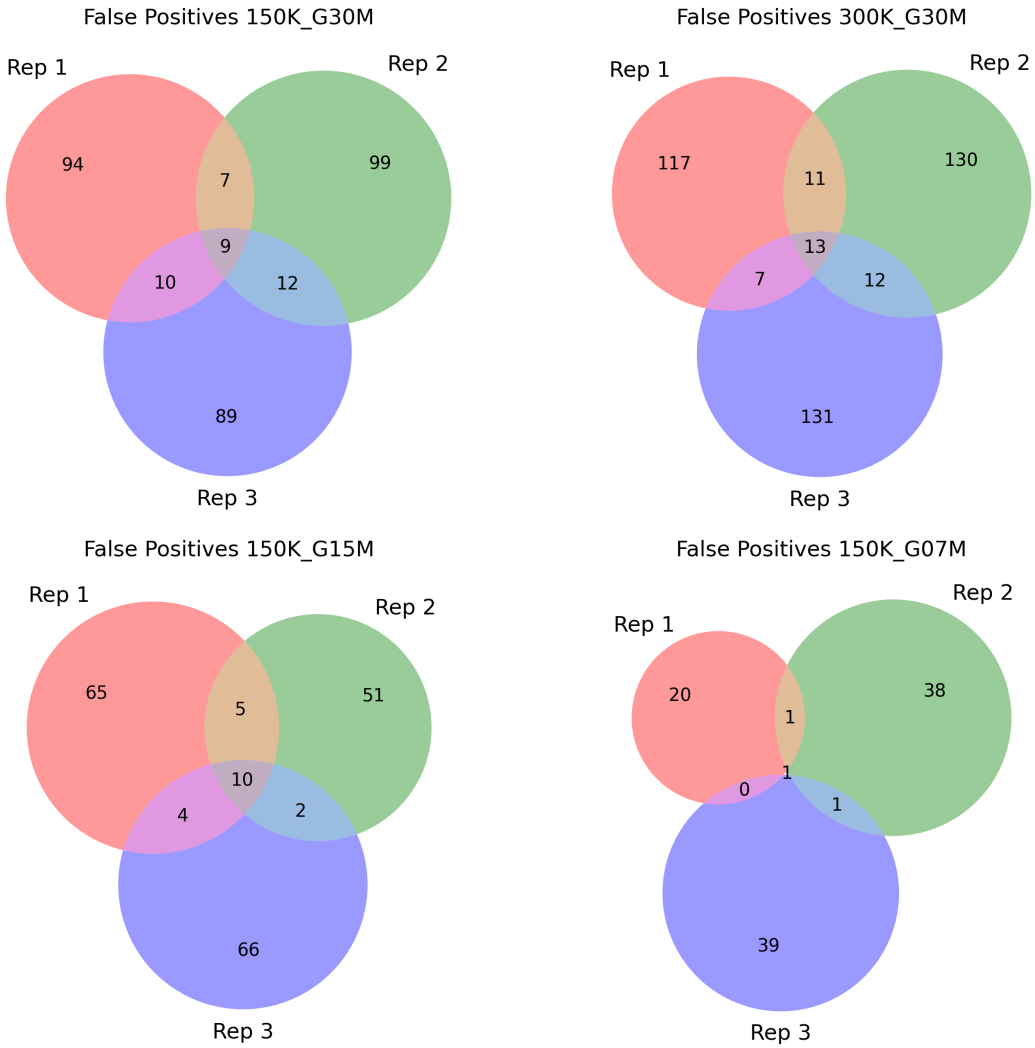

Figure 8: MS/MS-based false positive (FP) identification (peptide ions) variation of replicates from different simulation setups for Match-Between Run evaluation by MaxQuant.

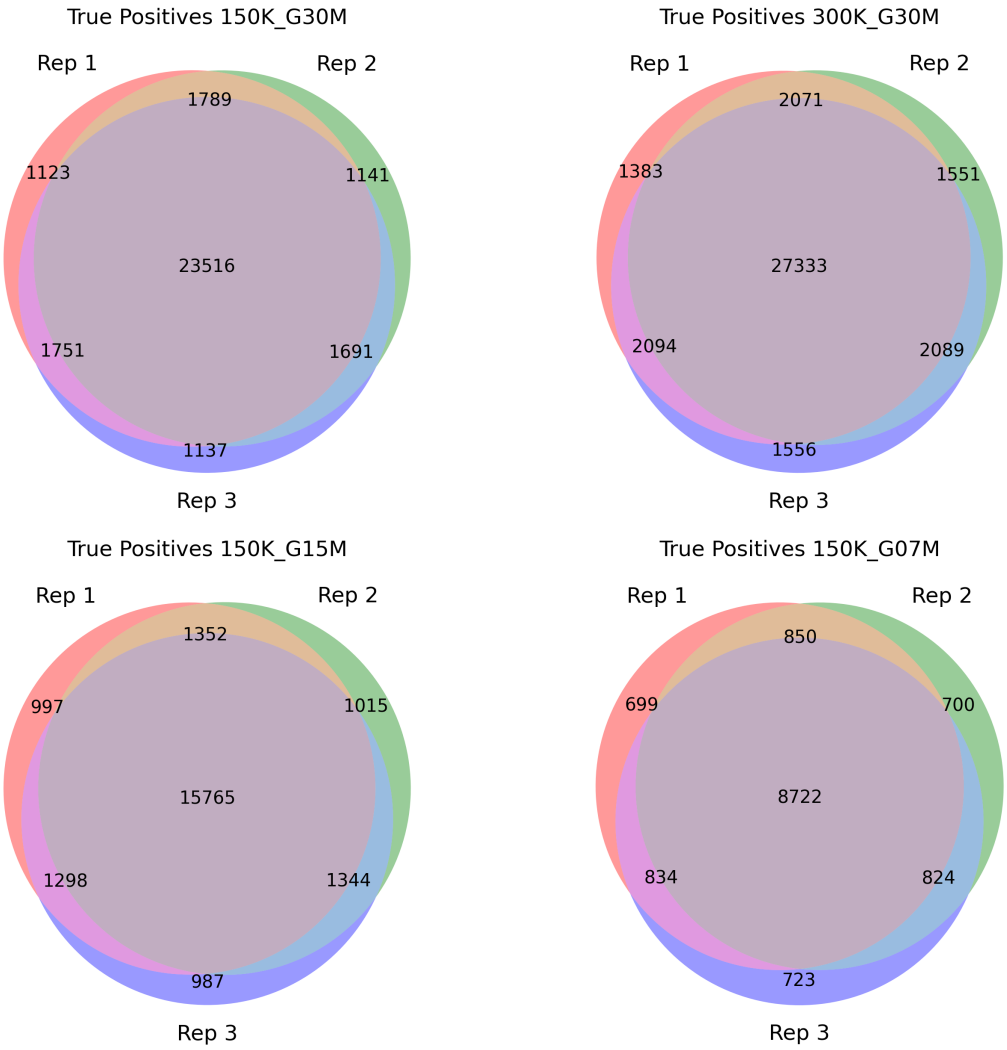

Figure 9: MS/MS-based true positive (TP) identification (peptide ions) variation of replicates from different simulation setups for Match-Between Run evaluation by FragPipe.

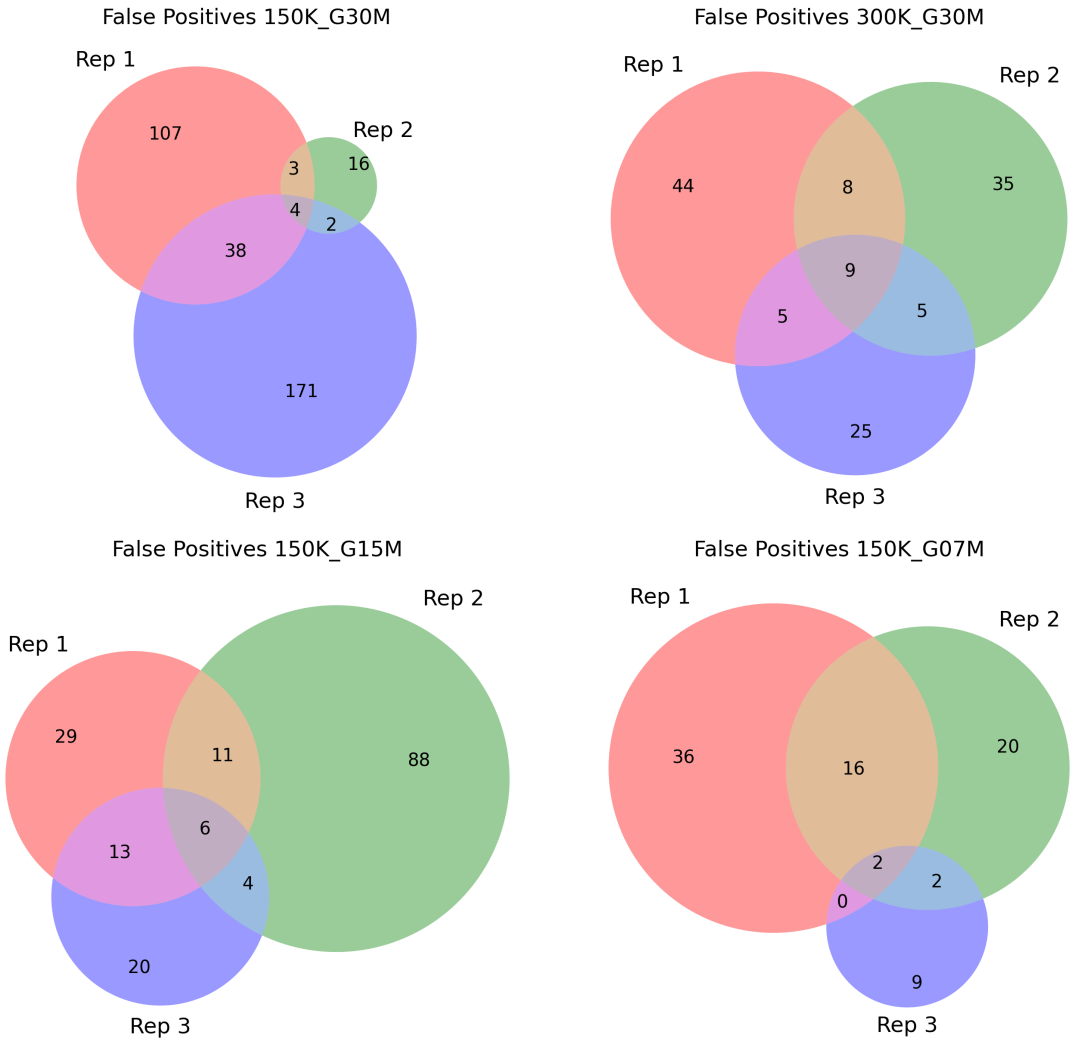

Figure 10: MS/MS-based false positive (FP) identification (peptide ions) variation of replicates from different simulation setups for Match-Between Run evaluation by FragPipe.

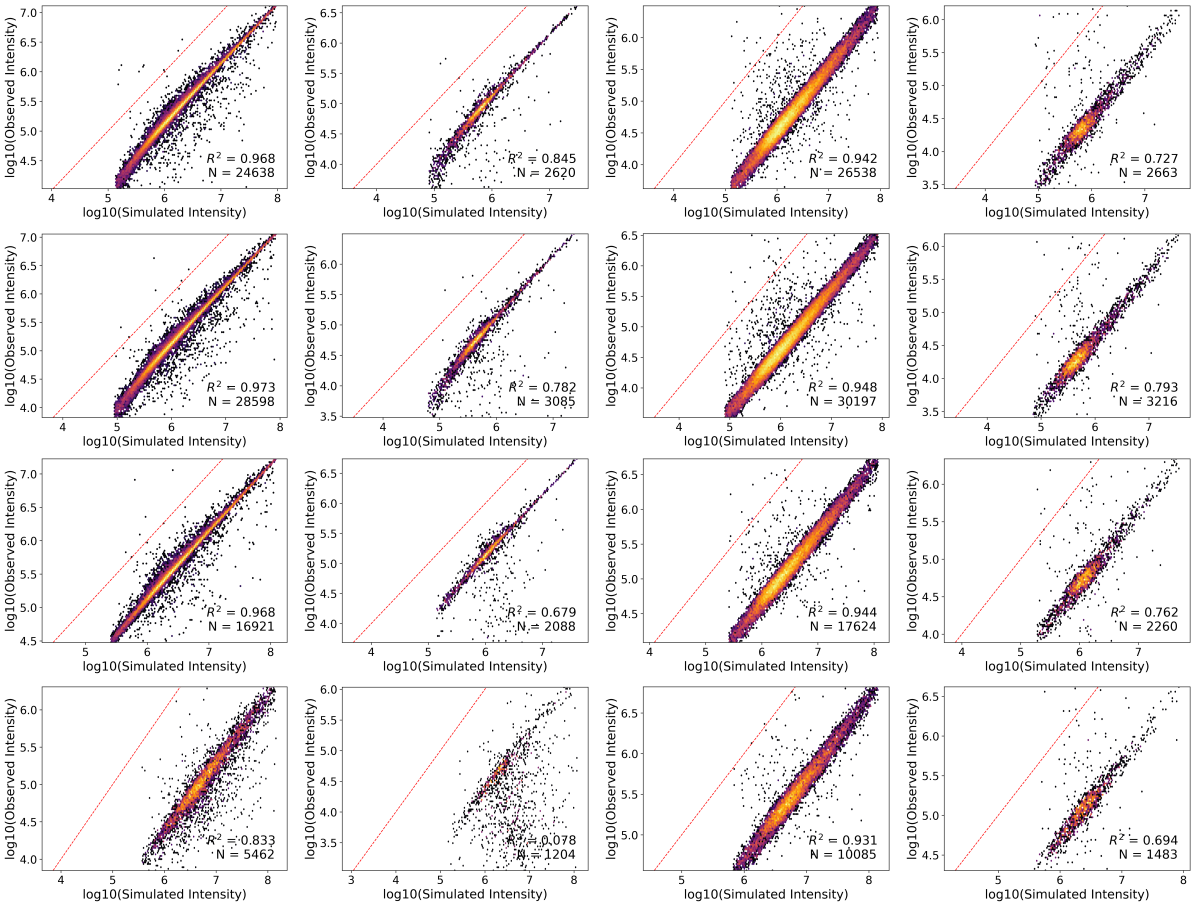

Figure 11: **Quantification accuracy of Match-Between-Run(MBR) results of software tools on 4 different simulation of dda-PASEF HeLa dataset.** Hexbin plots showing simulated vs reported log-intensities per ion (true positive only). Pearson's correlation coefficient (r) and number of datapoints in each subplot (N) are indicated. Each column represents one replicate. **Row one**: 30-minute LC gradient with 150,000 peptides. **Row two**: 30-minute LC gradient with 300,000 peptides. **Row three**: 7.5-minute LC gradient with 150,000 peptides. **Row four**: 15-minute LC gradient with 150,000 peptides. **Column one**: MaxQuant MS/MS identified ions. **Column two**: MaxQuant MBR ions. **Column three**: IonQuant (FragPipe) MS/MS identified ions. **Column four**: IonQuant (FragPipe) MBR ions.