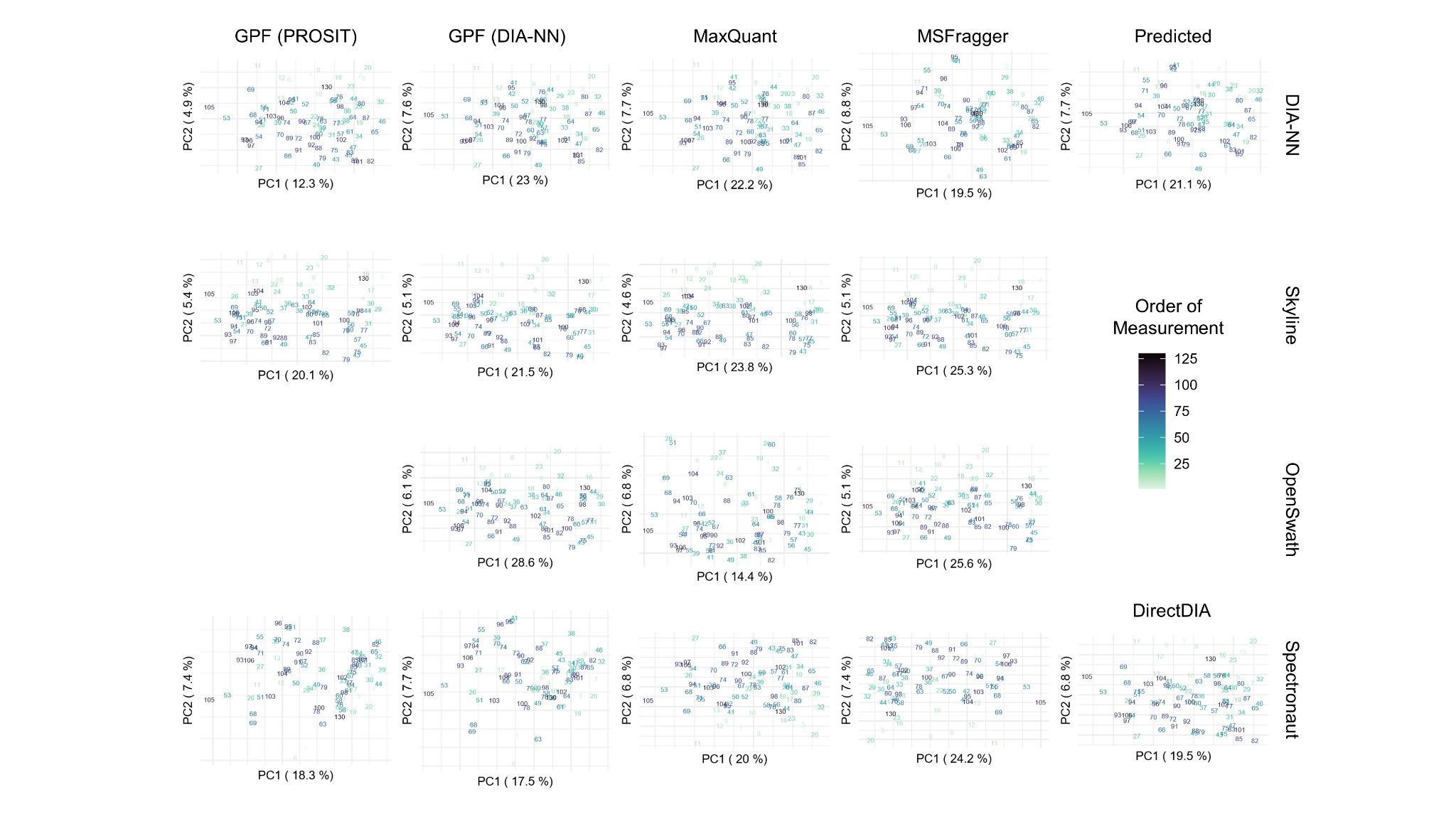
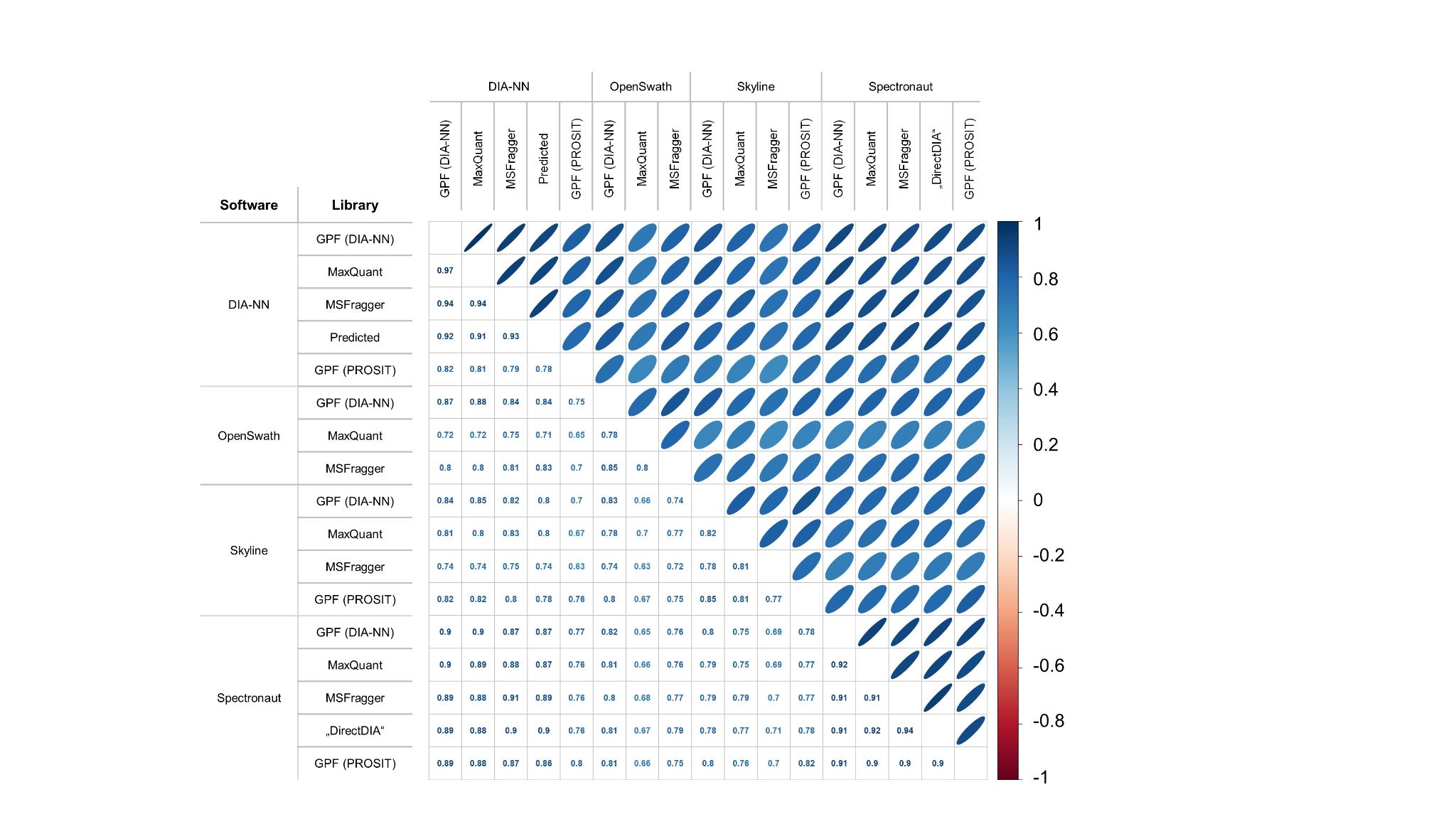
# Supplementary Information



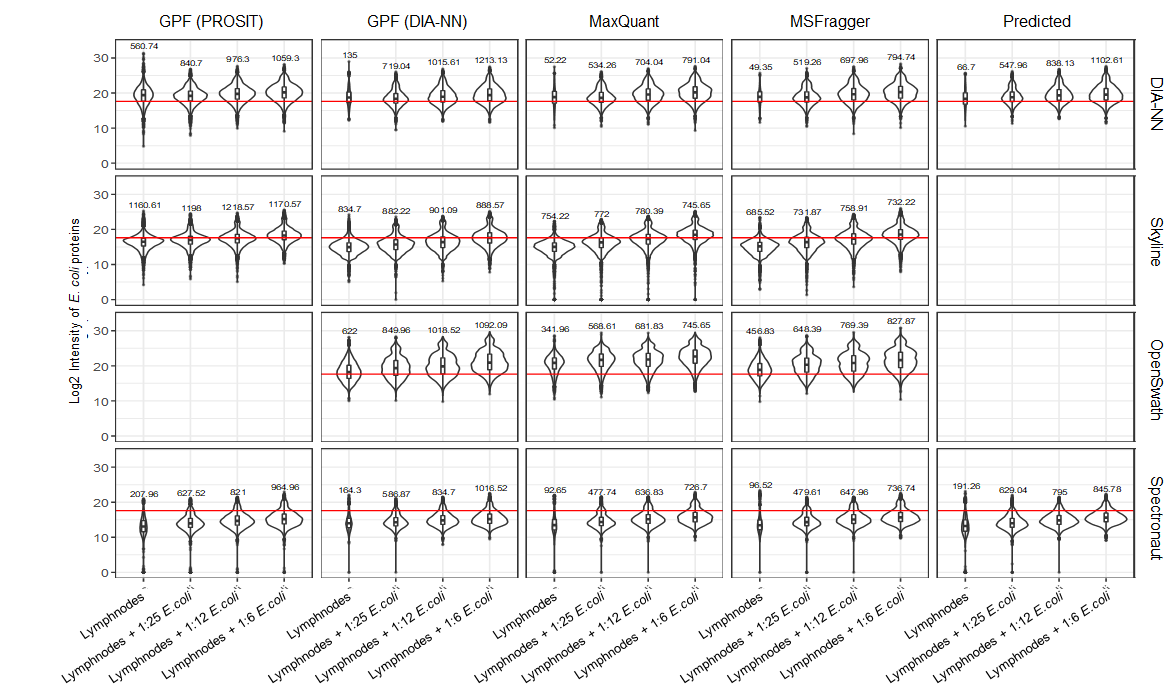
Supplementary Figure S1. No Batch Effects Are Observed in this Benchmark Dataset.

NIPALS (Nonlinear Iterative Partial Least Squares) principal component analysis (PCA) was performed based on protein abundance of DIA analysis workflows following quantile normalization. Chronological order of measurement was in an ascending manner. Sample 28, which belongs to spike-in condition 6:1, is not included in this plot as it represents an outlier due to a high degree of missing values. NIPALS was used, as it can directly be applied to data with missing values.



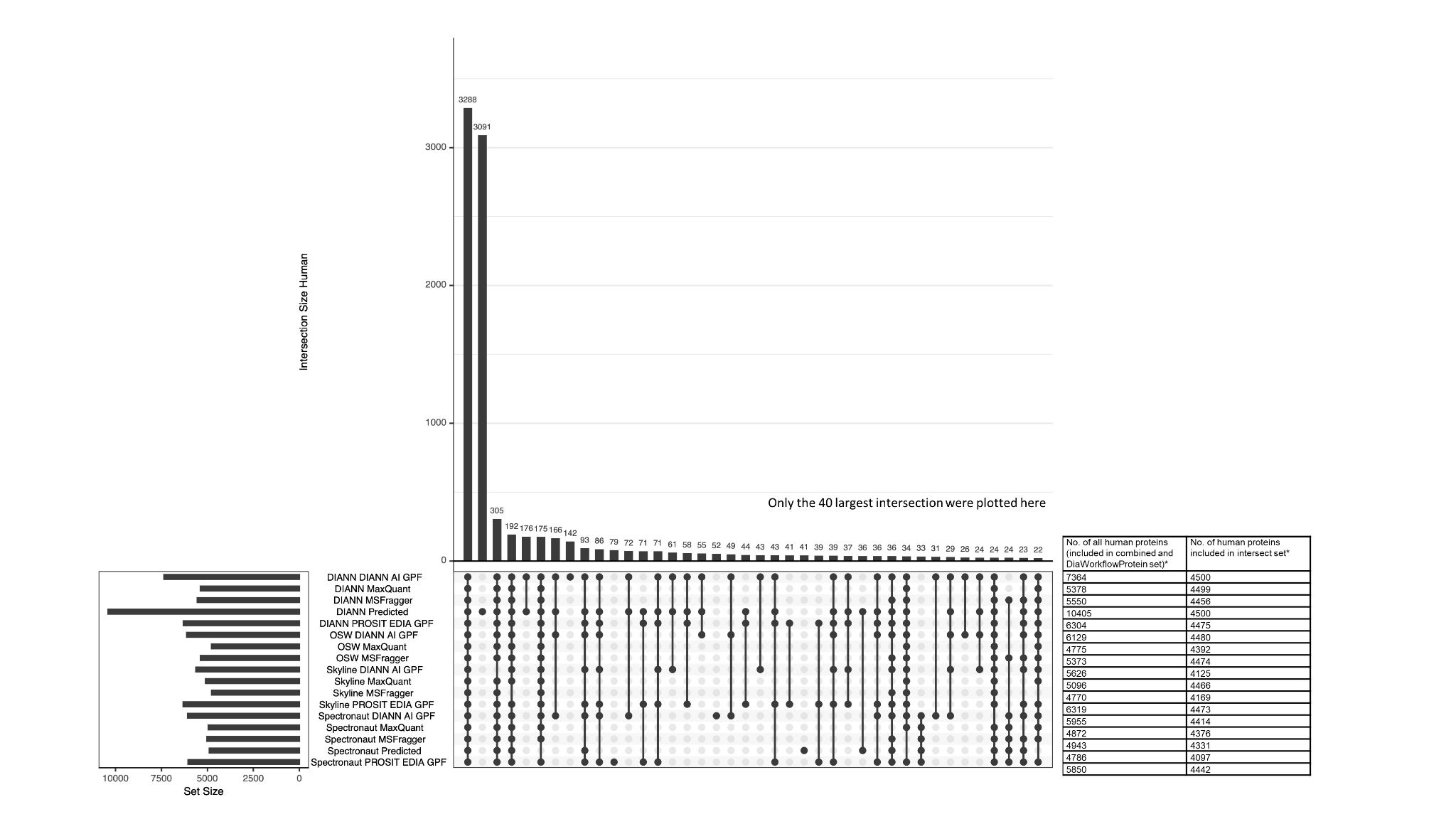
Supplementary Figure S2. Correlation of Protein Intensities Mainly Depends on Employed DIA Analysis Software

Pearson correlation between log2 protein intensities of all DIA analysis workflows (using all complete pairs of observations). The calculated correlation is based on the ​​3966 proteins common to all DIA analysis workflows. Color gradient and elliptic shape indicate correlation (circles / white represent lower correlation (0) as compared to ellipses / blue (1) ).



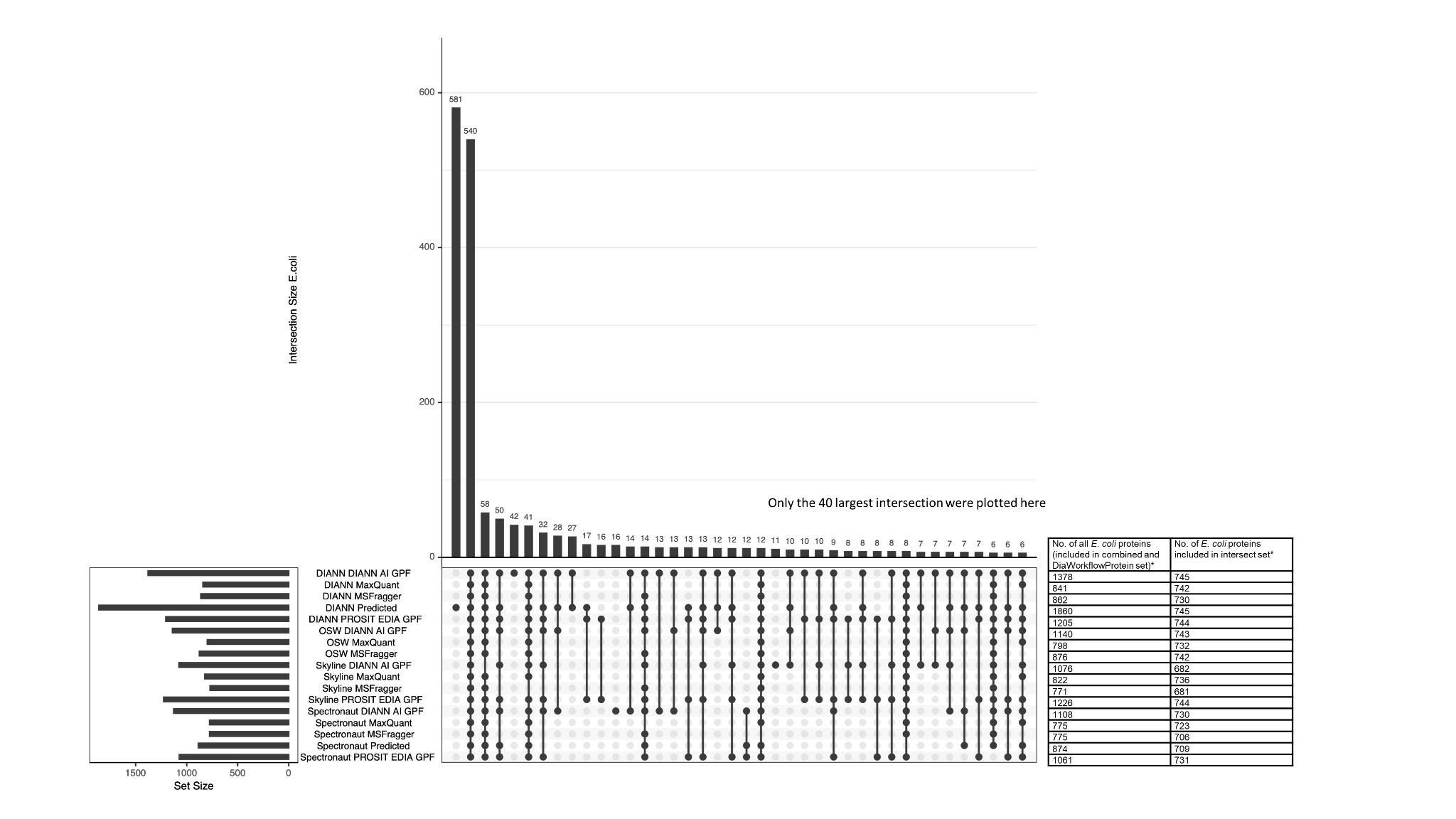
Supplementary Figure S3. *E.coli* proteins not being present in a biological sample are reported as missing by DIA-NN and Spectronaut, while Skyline and OpenSwath assume small intensity values for these many missing proteins.

Log2 protein abundance distribution of *E. coli* proteins separated by the four spike-in conditions. The overall median is indicated by the red line. The average number of identified *E. coli* proteins per sample within each DIA analysis workflow is displayed above each violin plot.



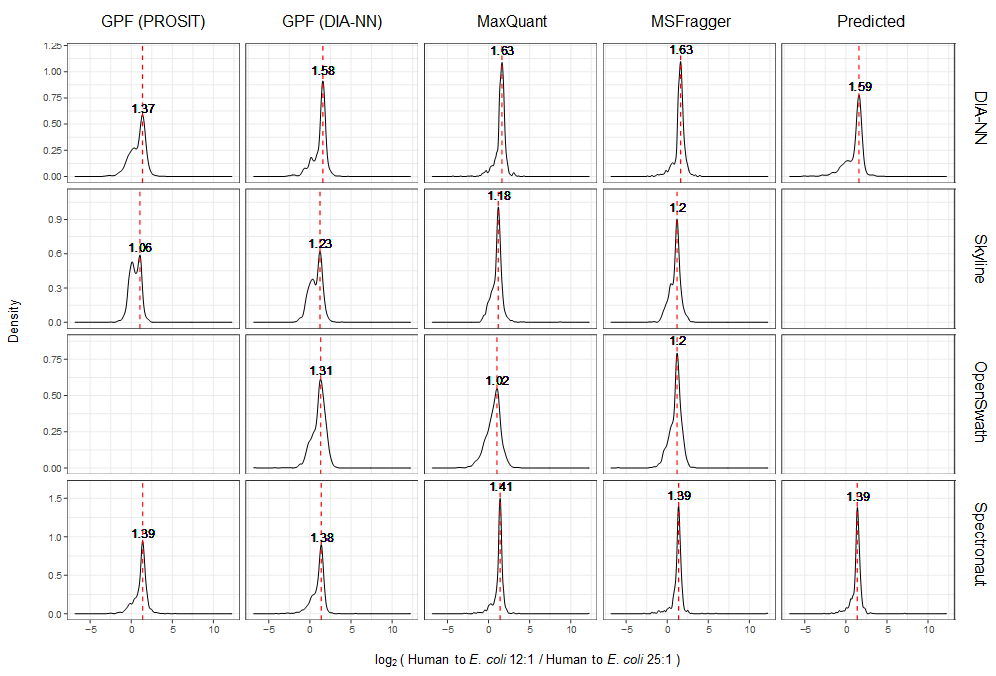
Supplementary Figure S4: Most Human Protein Identifications Are Shared Between All DIA Workflows.

Overlap of human proteins found in the DIA analysis workflows depicted via an upset plot (left) and number of proteins used in the ‘DIA workflow’/’Combined’ (left table column) and ‘Intersection’ (right table column) reference protein list (right). Only proteins appearing in one of the spike-in conditions 12:1 or 25:1 were included for this figure and only the 40 largest intersections are displayed in the upset plot. If DIA workflow results contained protein identifiers composed of more than one protein, those proteins are counted as individual entities. In cases where the protein identifier contains more than one protein, proteins contained in the bootstrap datasets are included in the analysis if one of the proteins the protein identifier is composed of matches an entry in the Intersection protein set. Unlike for the upset plot, in the table to the right protein identifiers containing multiple protein names are counted as only one protein.



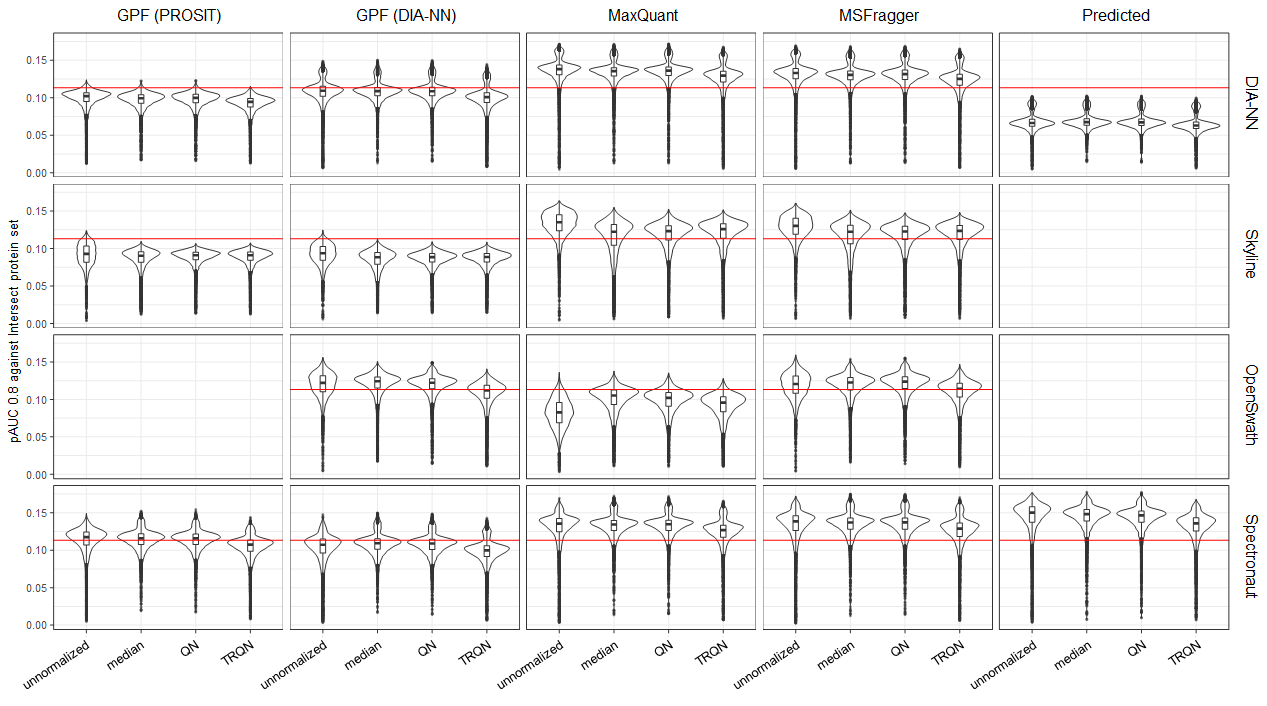
Supplementary Figure S5: Most *E. coli* Protein Identifications Are Unique to ‘DIANN Predicted’.

See legend of Supplementary Figure S4, but instead of human proteins, *E. coli* proteins are summarized.



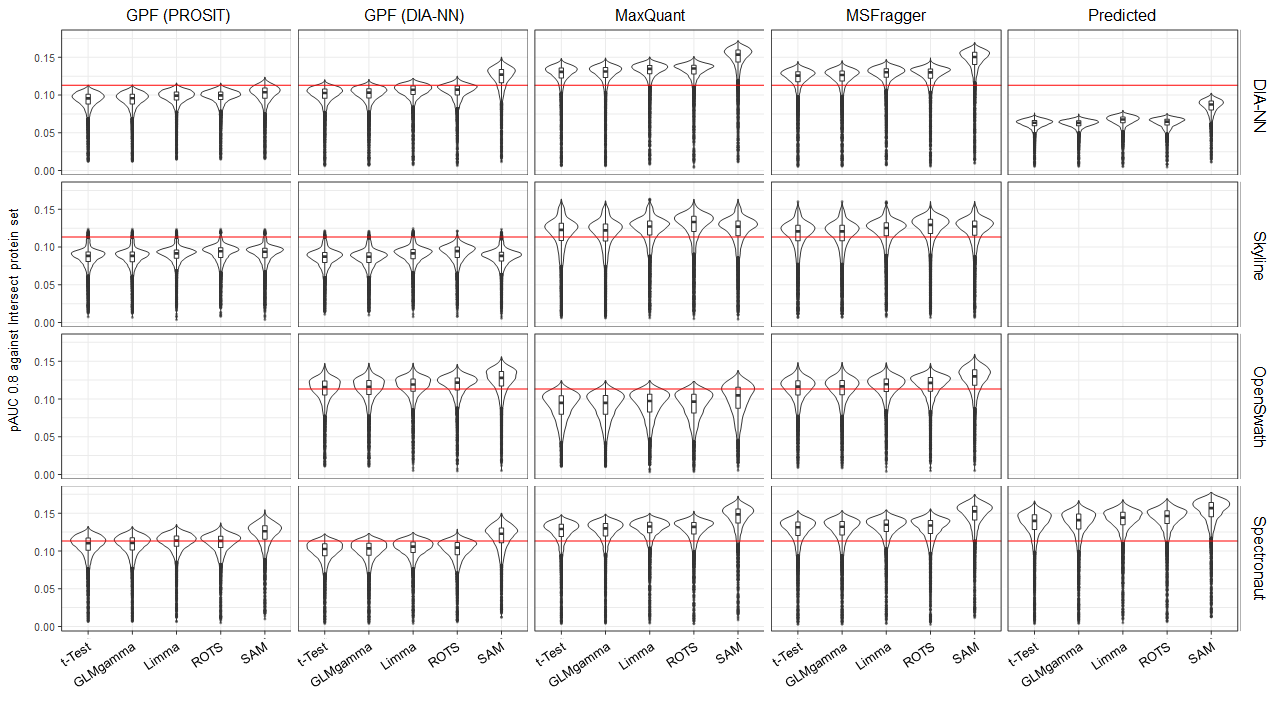
Supplementary Figure S6: Accuracy of Fold Change Mainly Depends on Employed DIA Analysis Software.

*E. coli* log2 fold changes observed between the spike-in conditions human to *E. coli* 12:1 and 25:1. The theoretical log2 fold change is 1.11. The dotted red line indicates the mode of the respective distribution.



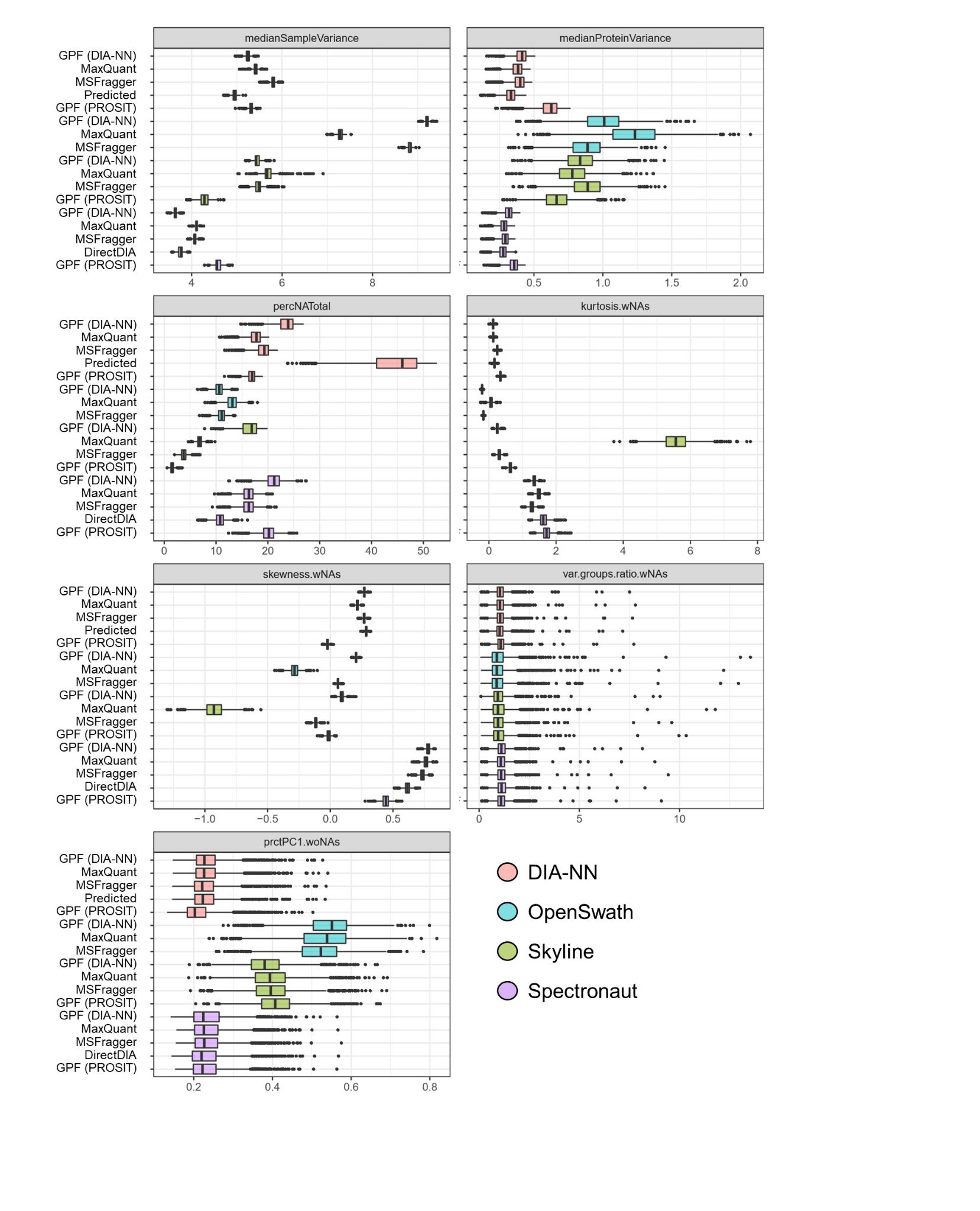
Supplementary Figure S7: Performing no normalization on our benchmark dataset resulted in best prediction performance for most DIA workflows.

Comparison of normalization options for no sparsity reduction (NoSR). pAUC was calculated based on the DIA workflow reference protein list. The red line indicates the overall median.



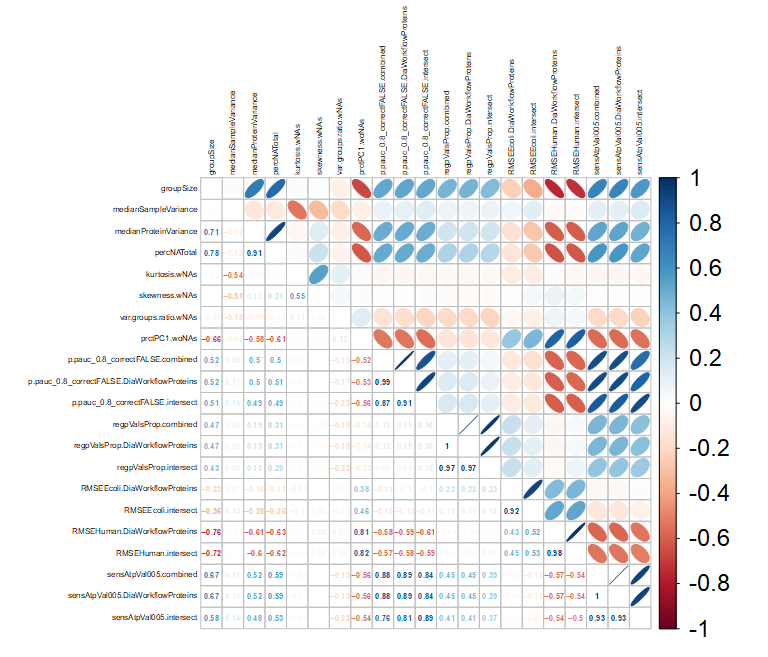
Supplementary Figure 8: The statistical tests SAM and ROTS perform best in detecting differentially abundant proteins.

Comparison of statistical tests for no sparsity reduction (NoSR). pAUC was calculated based on the DIAworkflow reference protein list. The red line indicates the overall median.



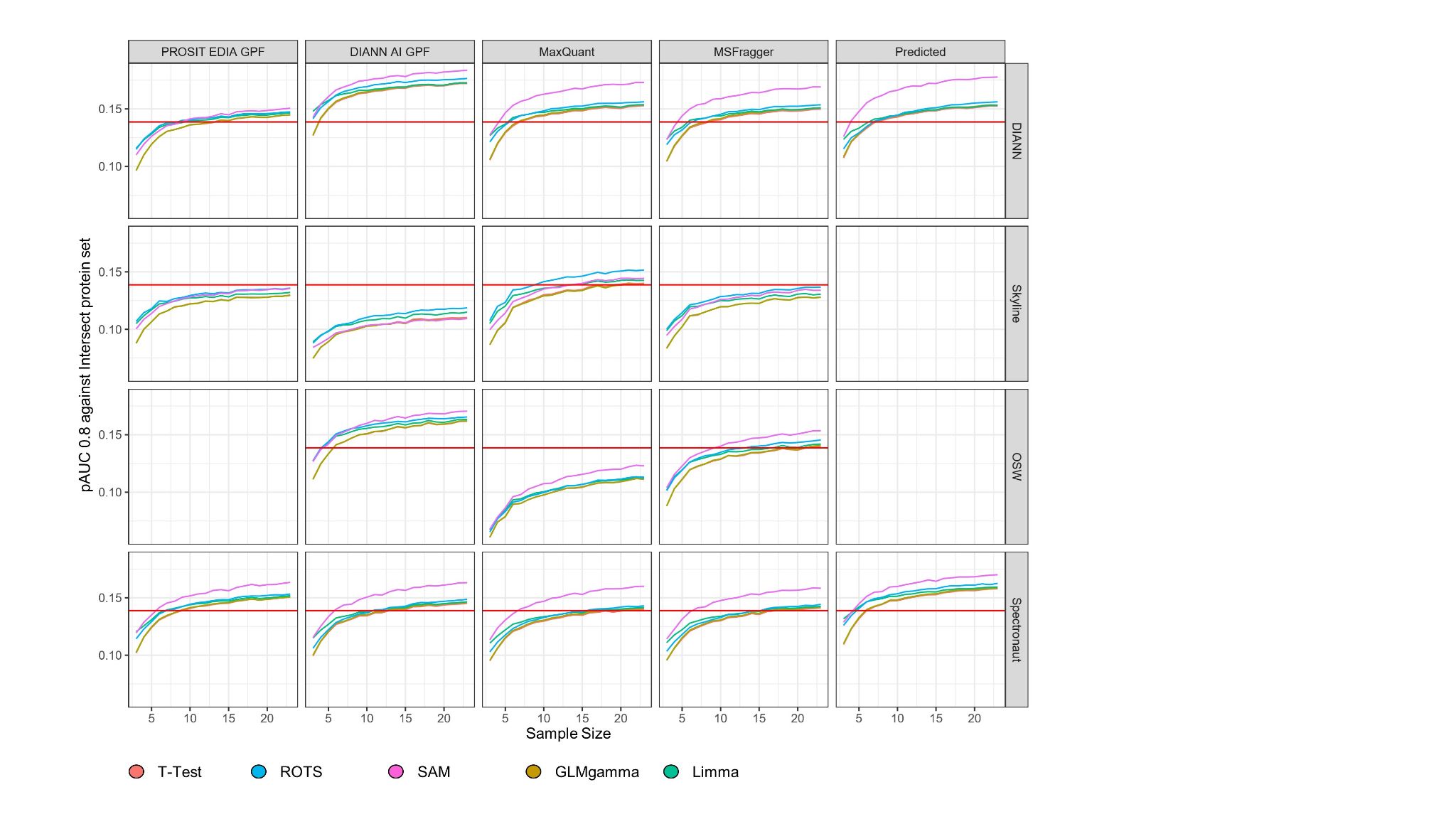
Supplementary Figure S9 OpenSwath shows the highest variance in protein intensities (together with Skyline) and sample intensities, as well as the highest percentage of the contribution of the first principal component to the total variance.

Terminology of data characteristics extensions (.wNAs and .woNAs indicate if proteins were included or excluded if they contained missing values): medianSampleVariance = median of the variances of the samples, medianProteinVariance = median of the variances of the proteins, percNATotal = percentage of missing values, kurtosis.wNAs = kurtosis, skewness.wNAs = skewness, var.groups.ratio.wNAs = median of the ratio of the protein variances of two groups (here: the two spike-in conditions 25:1 and 12:1), prcPC1.woNAs = percentage of the contribution of the first principal component to the total variance.



Supplementary Figure S10 Sample variance, kurtosis, skewness and the ratio of variances between two spike-in conditions have little influence on the performance of statistical tests.

Pearson correlations between data characteristics and performance measures of bootstrap datasets, (using all complete pairs of observations). The correlations are shown for the DIA analysis workflow consisting of DIA-NN in combination with the *in silico* predicted GPF-refined (DIA-NN) spectral library. The data characteristics were log2-transformed prior to the correlation analysis. For terminology of data characteristics see Supplementary Figure S9. Terminology of performance measures (extensions .combined, .DiaWorkflowProteins, .intersect refer to the reference protein list based on which the performance measure has been calculated): p.pauc\_0.8\_correctFALSE = partial area under the curve (pAUC), regValsProp = estimated proportion of regulated proteins based on p-value distribution (1-π0), RMSEEcoli = root-mean-square error (RMSE) of all *E. coli* proteins, RMSEHuman = RMSE of all human proteins, sensAtpVal005 = sensitivity calculated by predicting all proteins with p-value < 0.05 to be differentially abundant.



Supplementary Figure S11: SAM performs best for all software suites, except for Skyline for which ROTS is superior. limma performs well for small sample sizes.  
Prediction performance of different statistical tests in pAUC over different sample sizes. The lines, which are color-coded by statistical test, represent the median over the respective sample size.