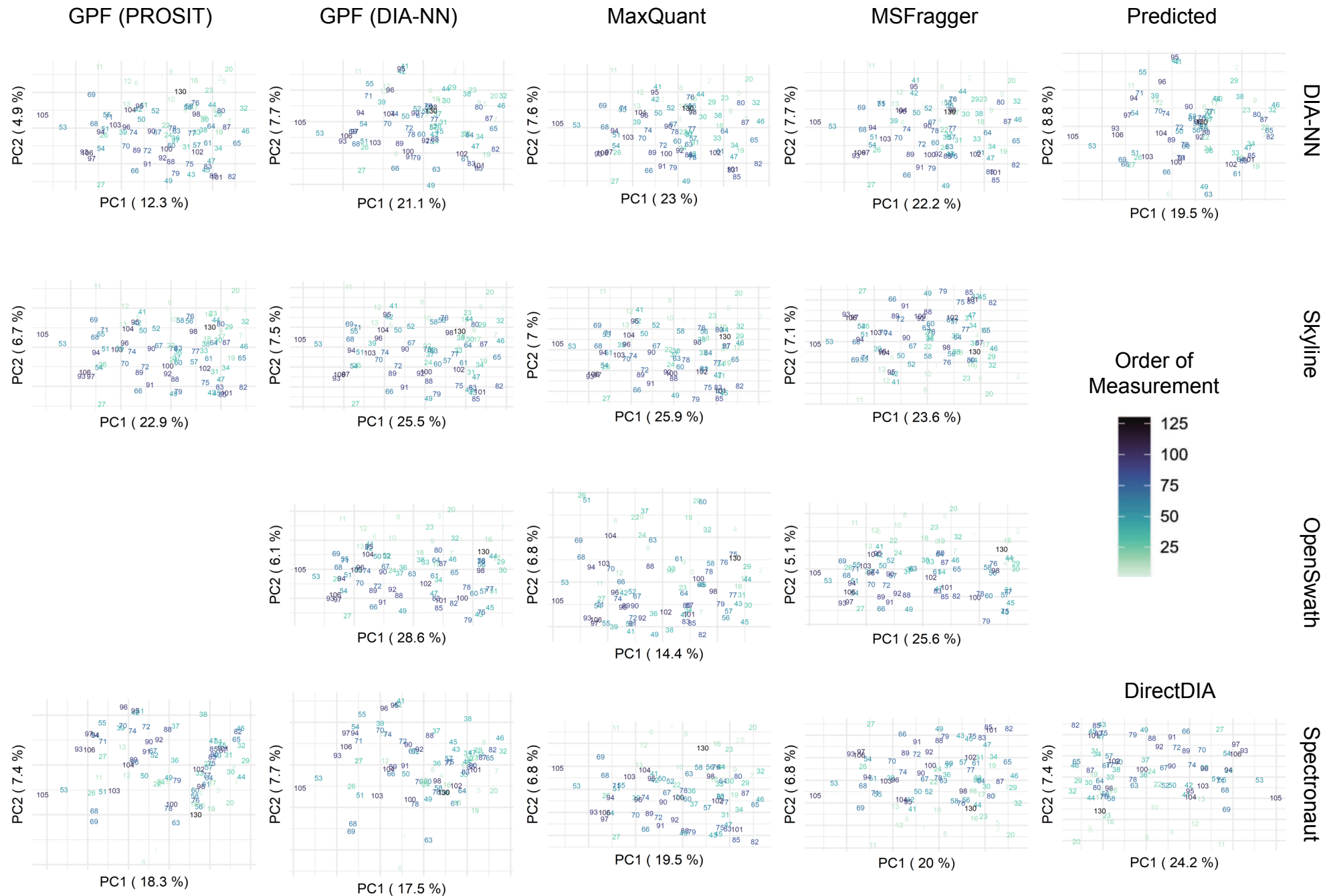


Supplementary Information:

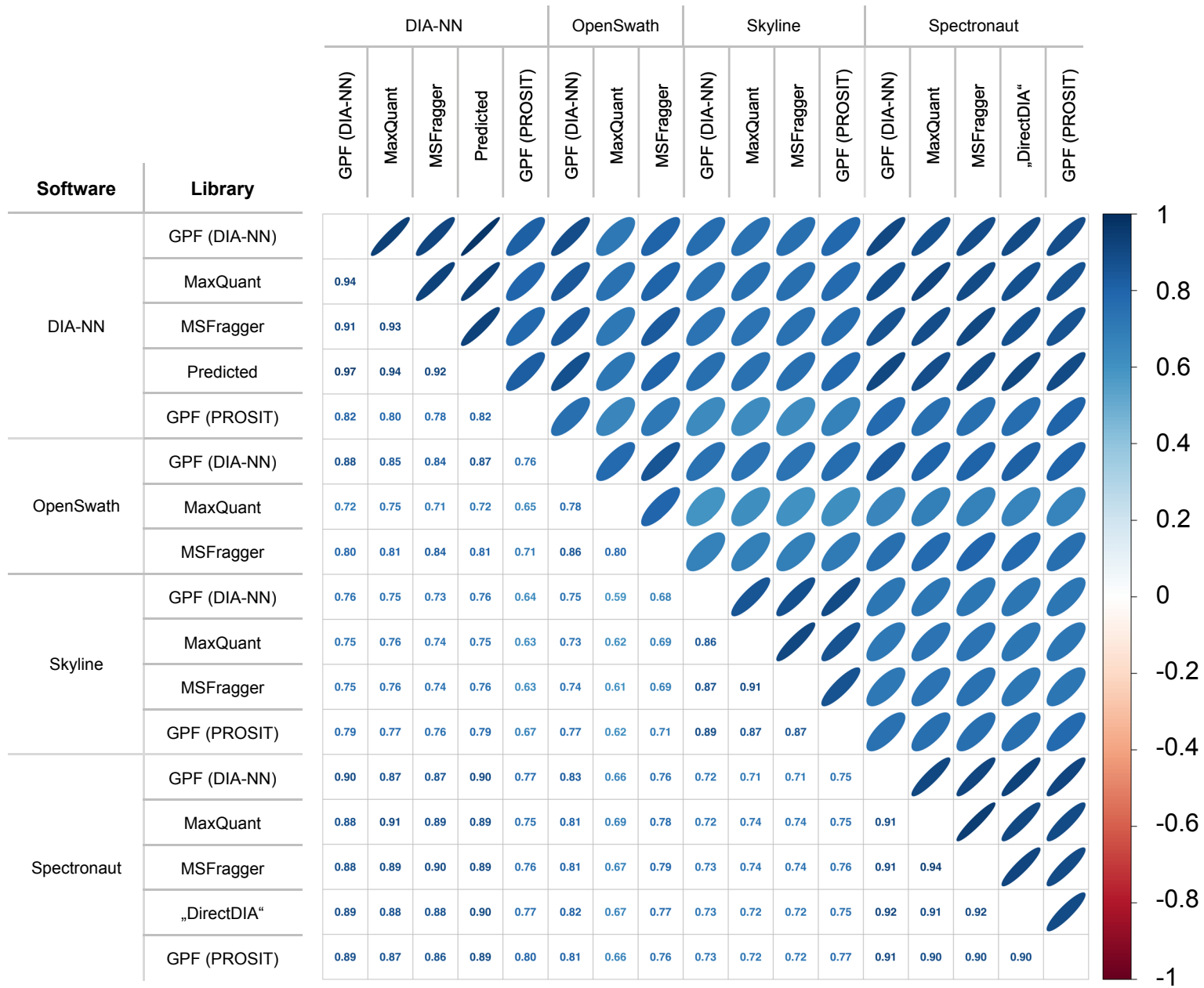
**Benchmarking of Analysis Strategies  
for Data-Independent Acquisition  
Proteomics Using a Large-Scale  
Dataset Comprising Inter-Patient  
Heterogeneity**

O. Schilling et al



## Supplementary Figure 1. No Batch Effects Are Observed in this Benchmark Dataset.

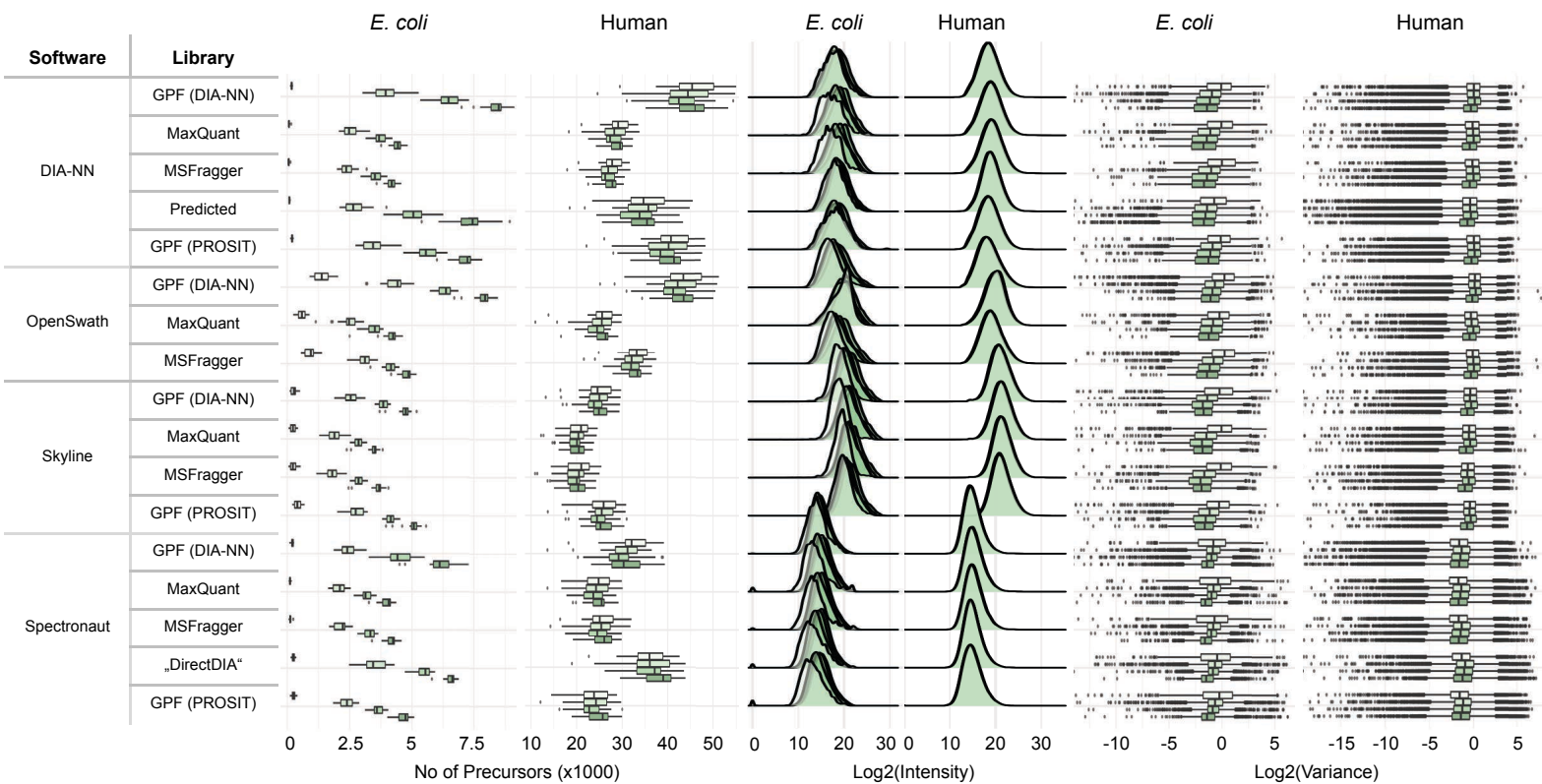
NIPALS (Nonlinear Iterative Partial Least Squares) principal component analysis (PCA) was performed based on protein abundance of DIA analysis workflows following quantile normalization. Chronological order of measurement was in an ascending manner, indicated by the increasing darkness of label color. Sample 28, which belongs to spike-in condition 1:6, 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. Source data are provided as a Source Data file.





## Supplementary Figure 2. Correlation of Protein Intensities Mainly Depends on Employed DIA Analysis Software

Pearson correlation between log<sub>2</sub> 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. Pearson correlation ranges from -1 (perfect negative correlation, red) to 1 (perfect positive correlation, blue). This is also indicated by the elliptic shape (round indicates lower correlation as opposed to a more narrow shape indicating a higher (positive) correlation). Source data are provided as a Source Data file.



○ Human only

● *E. coli* to Human 1:25

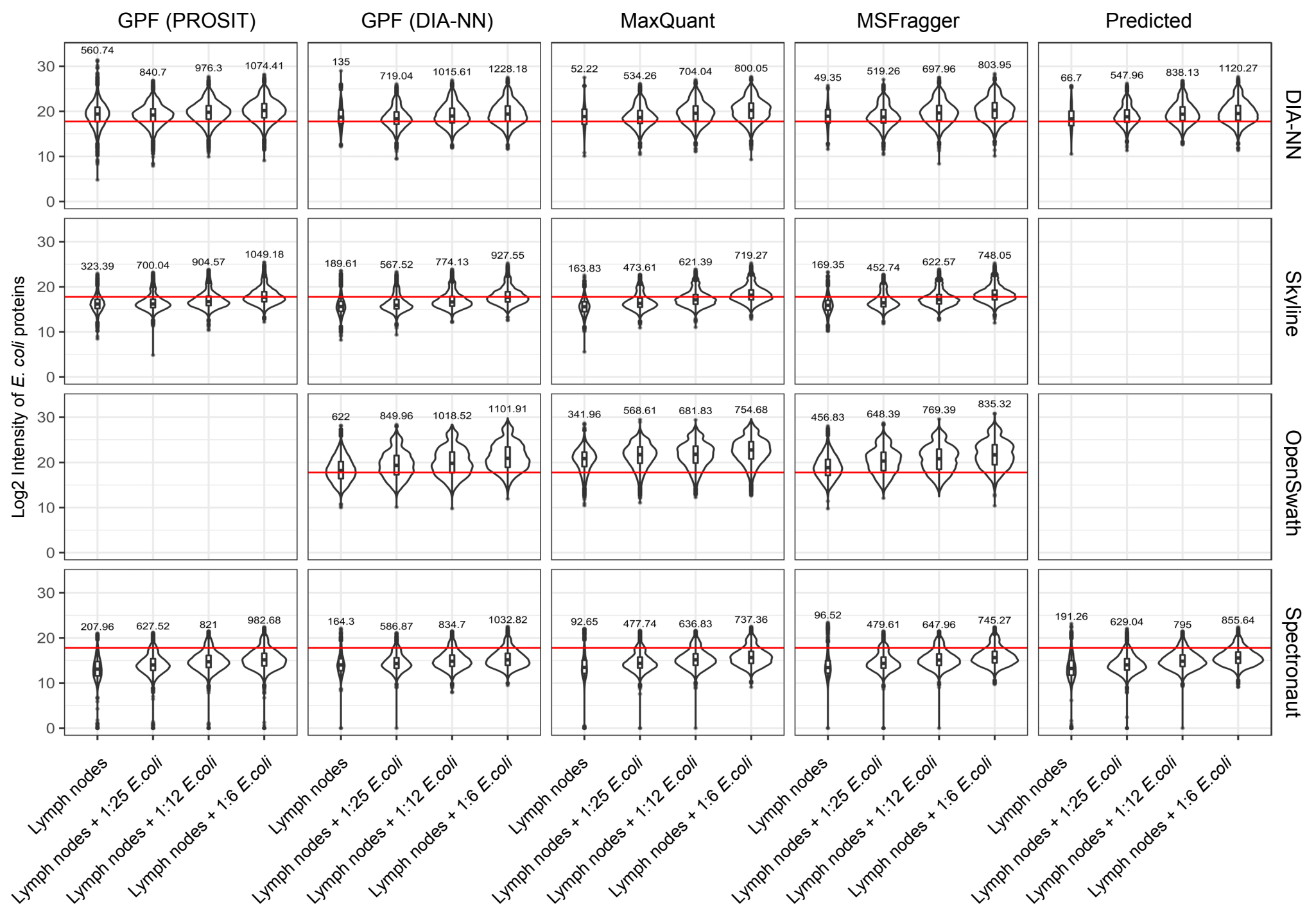
● *E. coli* to Human 1:12

● *E. coli* to Human 1:6

Supplementary Figure 3: Precursor number, distribution and variance for each DIA analysis workflow separated by species and color-coded by spike-in condition.

Left: Number of all identified and quantified proteins in all 92 samples. Center: Log<sub>2</sub> intensity distributions. Intensities smaller than 0 were excluded. Right: Log<sub>2</sub> variance. Log<sub>2</sub> variance values smaller than -12 were excluded from this plot.

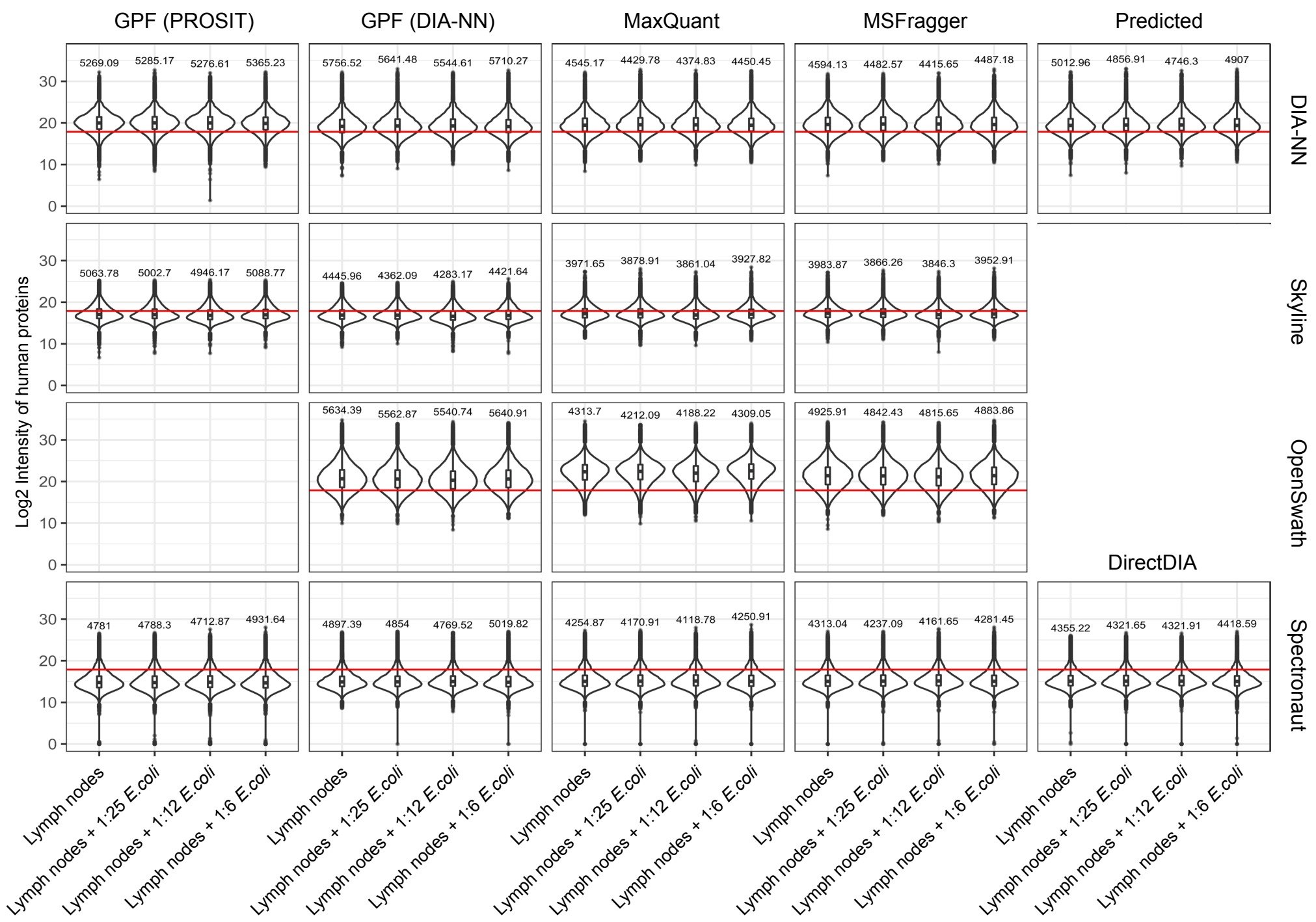
For spike-condition 1:6 data of n=22 biologically independent samples have been used and for each of the other spike-in conditions data of n=23 biologically independent samples have been used. The boxplots show median (center line), interquartile range (IQR, extending from the first to the third quartile) (box), and  $1.5 * IQR$  (whiskers). Source data are provided as a Source Data file.



Supplementary Figure 4: *E. coli* proteins not being present in a biological sample are reported as missing by DIA-NN, Skyline, and Spectronaut, while OpenSwath assumes small intensity values for these missing proteins.

Log<sub>2</sub> 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.

For spike-condition 1:6 data of n=22 biologically independent samples have been used and for each of the other spike-in conditions data of n=23 biologically independent samples have been used. The boxplots show median (center line), interquartile range (IQR, extending from the first to the third quartile) (box), and  $1.5 * IQR$  (whiskers). Source data are provided as a Source Data file.



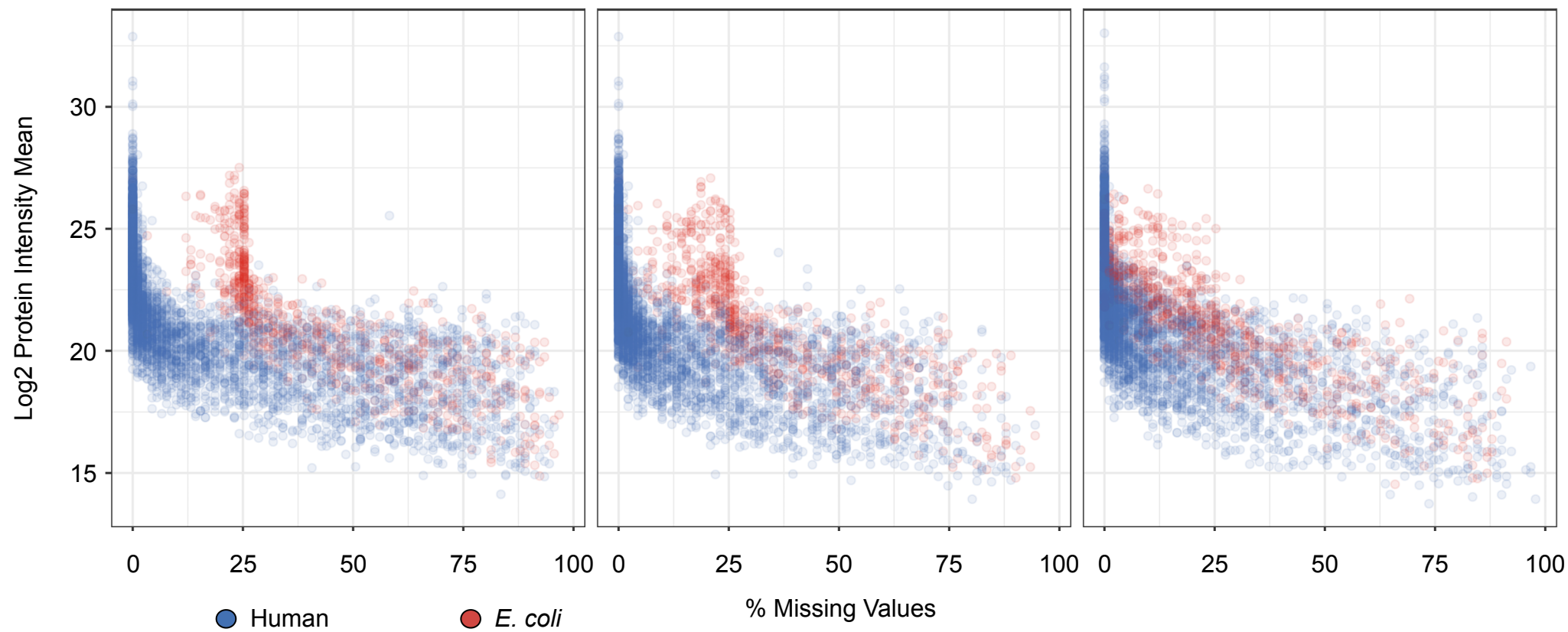
Supplementary Figure 5: Human proteins are identified and quantified equally across all spike-in conditions.

Log<sub>2</sub> protein abundance distribution of human proteins separated by the four spike-in conditions. The overall median is indicated by the red line. The average number of identified human proteins per sample within each DIA analysis workflow is displayed above each violin plot. For spike-condition 1:6 data of n=22 biologically independent samples have been used and for each of the other spike-in conditions data of n=23 biologically independent samples have been used. The boxplots show median (center line), interquartile range (IQR, extending from the first to the third quartile) (box), and 1.5 \* IQR (whiskers). Source data are provided as a Source Data file.

TRIC filtered 1 % FDR

TRIC filtered 5 % FDR

PyProphet filtered 1 % FDR

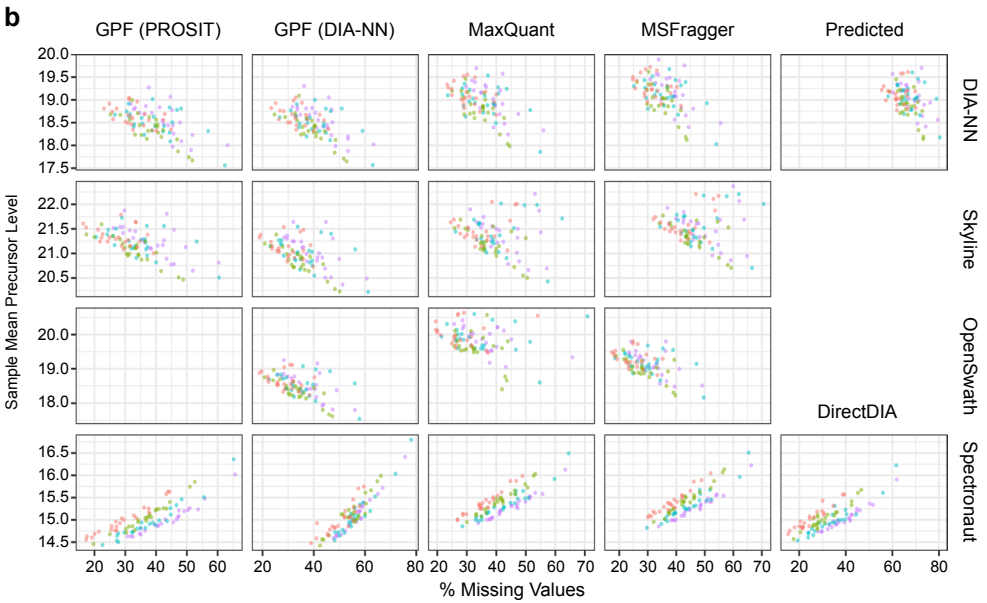
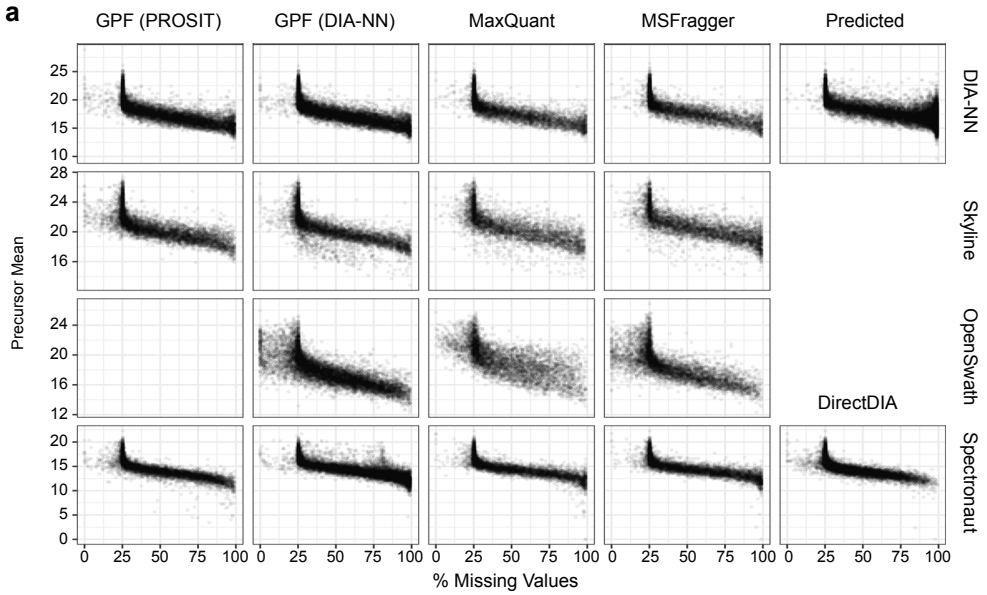




Supplementary Figure 6. Using TRIC OpenSwath shows a similar missingness behavior

to the other DIA software suites, e.g. for example, the accumulation at 25% missing values, which originates from the human-only samples.

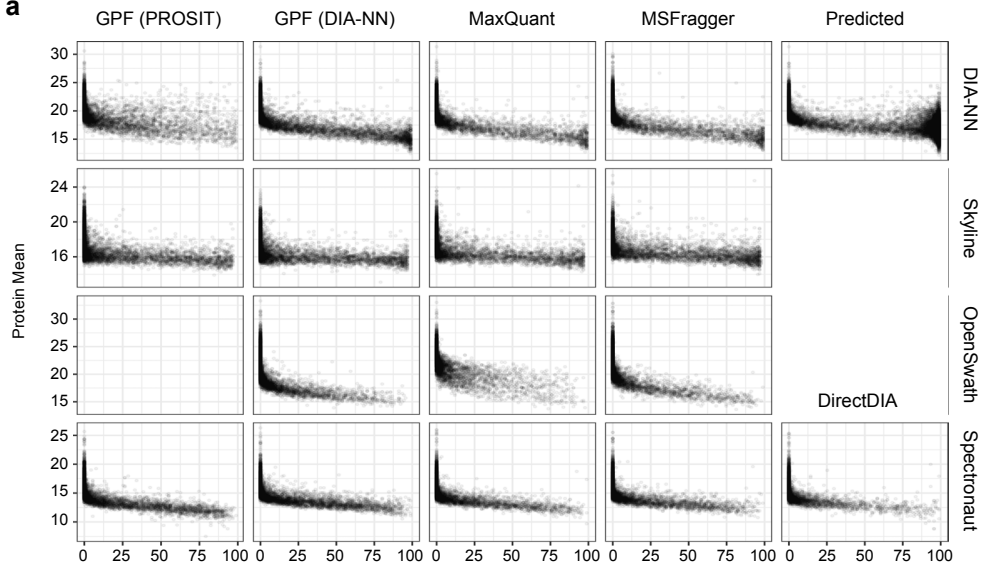
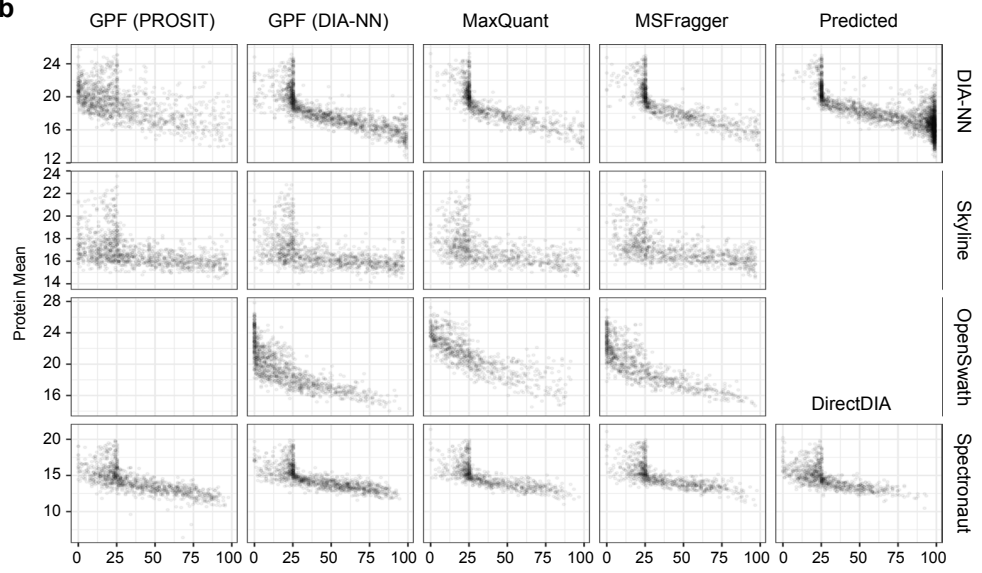
FDR filtering applied to the OpenSwath protein data of the 92 lymph node samples using TRIC with an FDR of 1% (left), TRIC with an FDR of 5% (center), and PyProphet with an FDR of 1 % (right). PyProphet is used in our study. Human and *E. coli* proteins are depicted in blue and red, respectively. Source data are provided as a Source Data file.



- Lymph nodes + 1:6 *E. coli*
- Lymph nodes + 1:12 *E. coli*
- Lymph nodes + 1:25 *E. coli*
- Lymph nodes

## Supplementary Figure 7. Missing Value Characteristics and Correlations on the Precursor Level.

a) Unlike on the protein level (Figure 3A) OpenSwath handles precursors of small intensity similarly to the other DIA software suites. Means of log<sub>2</sub> intensities of identified *E. coli* precursors plotted against the percentage of missing values in the respective precursor. *E. coli* precursors are not physically present in 25% of samples. b) The correlation between the missingness within samples and the sample mean over all human and *E. coli* precursors of these samples varies with the employed DIA software. Unlike on the protein level DIA-NN shows a negative correlation between missingness and sample mean. Sample means of precursor intensities are plotted against the percentage of missing values in the respective sample. Source data are provided as a Source Data file.

**a****b**

Supplementary Figure 8. Figure 3A separated by species into a) human and b) *E. coli* proteins.

Means of log<sub>2</sub> intensities of identified proteins plotted against the percentage of missing values in the respective protein. *E. coli* proteins are not physically present in 25% of samples. Source data are provided as a Source Data file.

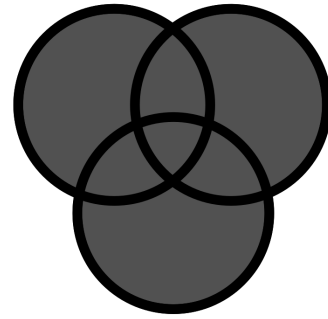
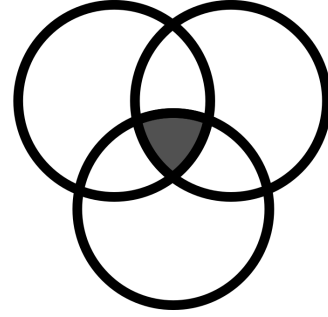
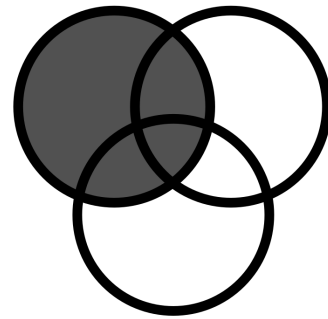
True Positive Rate (TPR) = # TP/ # *E. coli* proteins

False Positive Rate (FPR) = # FP/ # Human proteins

Problem:

**No ground truth** about number of existing total and regulated proteins in the biological samples known

→ # *E. coli* proteins and # human proteins can refer to different protein lists:

Protein list	All proteins present in...	Identification quality	Quantification quality	Theoretical maximum of TPR
 <p><b>Combined</b> (<i>E. coli</i> 2125, human: 11533)</p>	≥ 1 DIA analysis workflow	TPI, FPI ↑ TNI, FNI ↓	↓	DIA workflow-dependent: # identified / # combined
 <p><b>Intersection</b> (<i>E. coli</i> 740, Human: 4512)</p>	≥ 80% DIA analysis workflows	TNI, FPI ↓ TNI, FNI ↑	↑	1
 <p><b>DIAWorkflow</b></p>	DIA workflow-dependent	DIA workflow-dependent	DIA workflow-dependent	1

## Supplementary Figure 9. Summary of the Reference Protein Lists.

### Abbreviations:

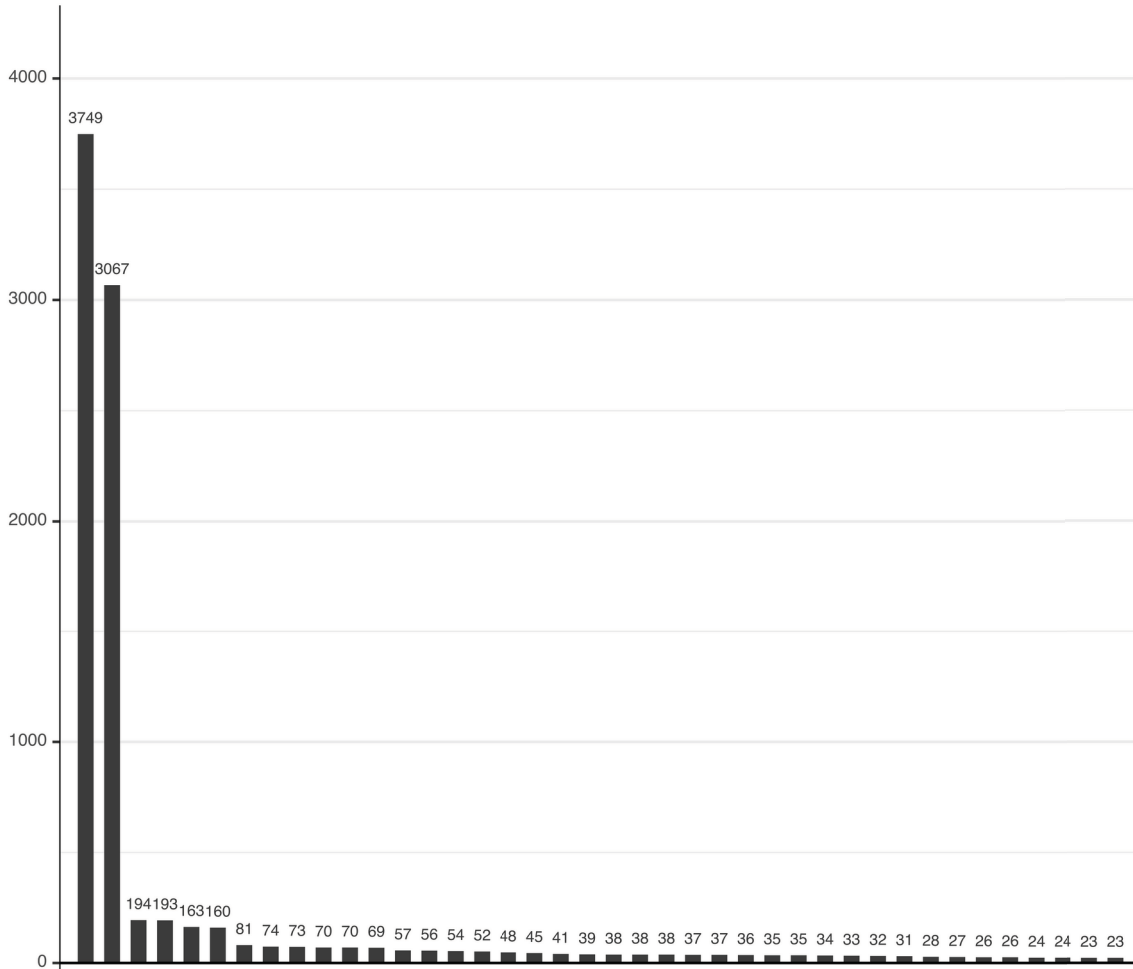
TPI = true positive identifications,

FPI = false positive identifications,

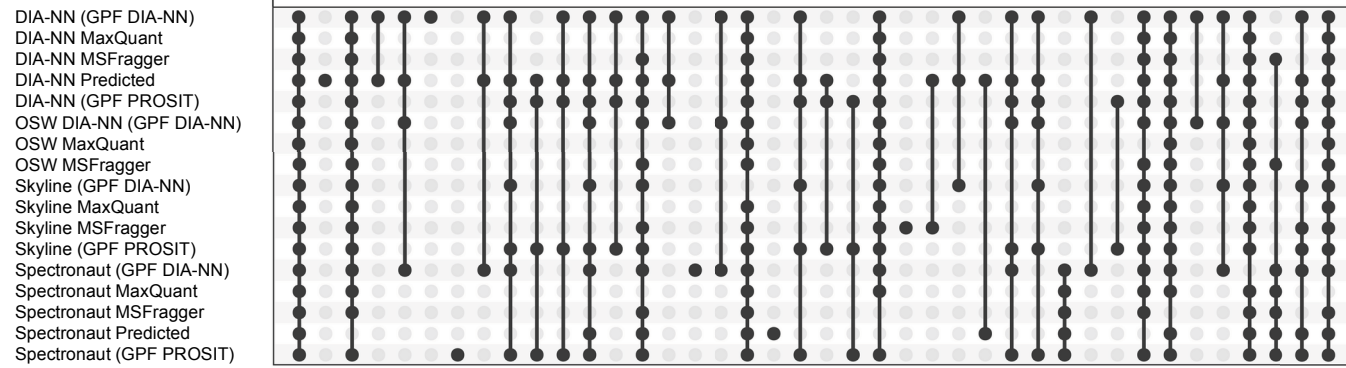
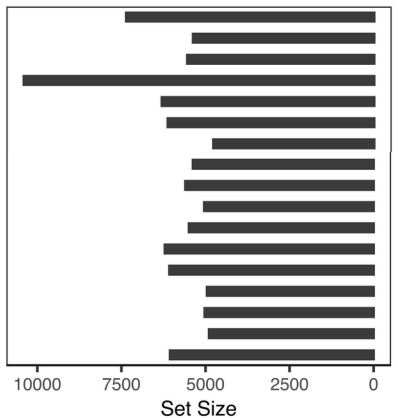
TNI = true negative identifications,

FNI = false negative identifications

Intersection Size Human



No. of all human proteins (included in combined and DiaWorkflowProtein set)      No. of human proteins included in intersect set

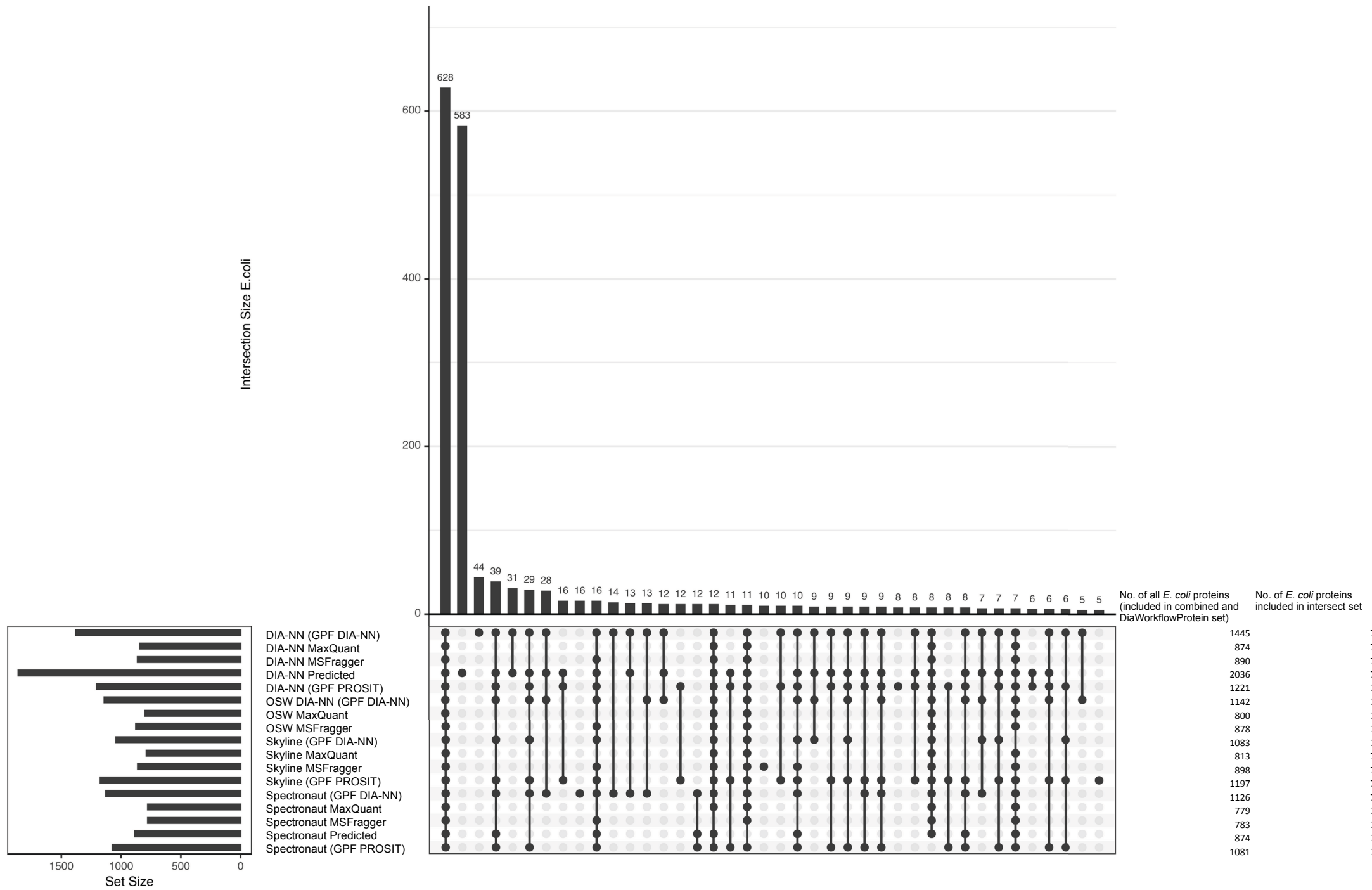


7468	4513
5440	4513
5637	4474
10763	4513
6320	4483
6130	4492
4777	4403
5375	4480
5611	4407
5052	4476
5515	4468
6214	4478
5959	4424
4878	4398
4947	4341
4793	4099
5852	4451



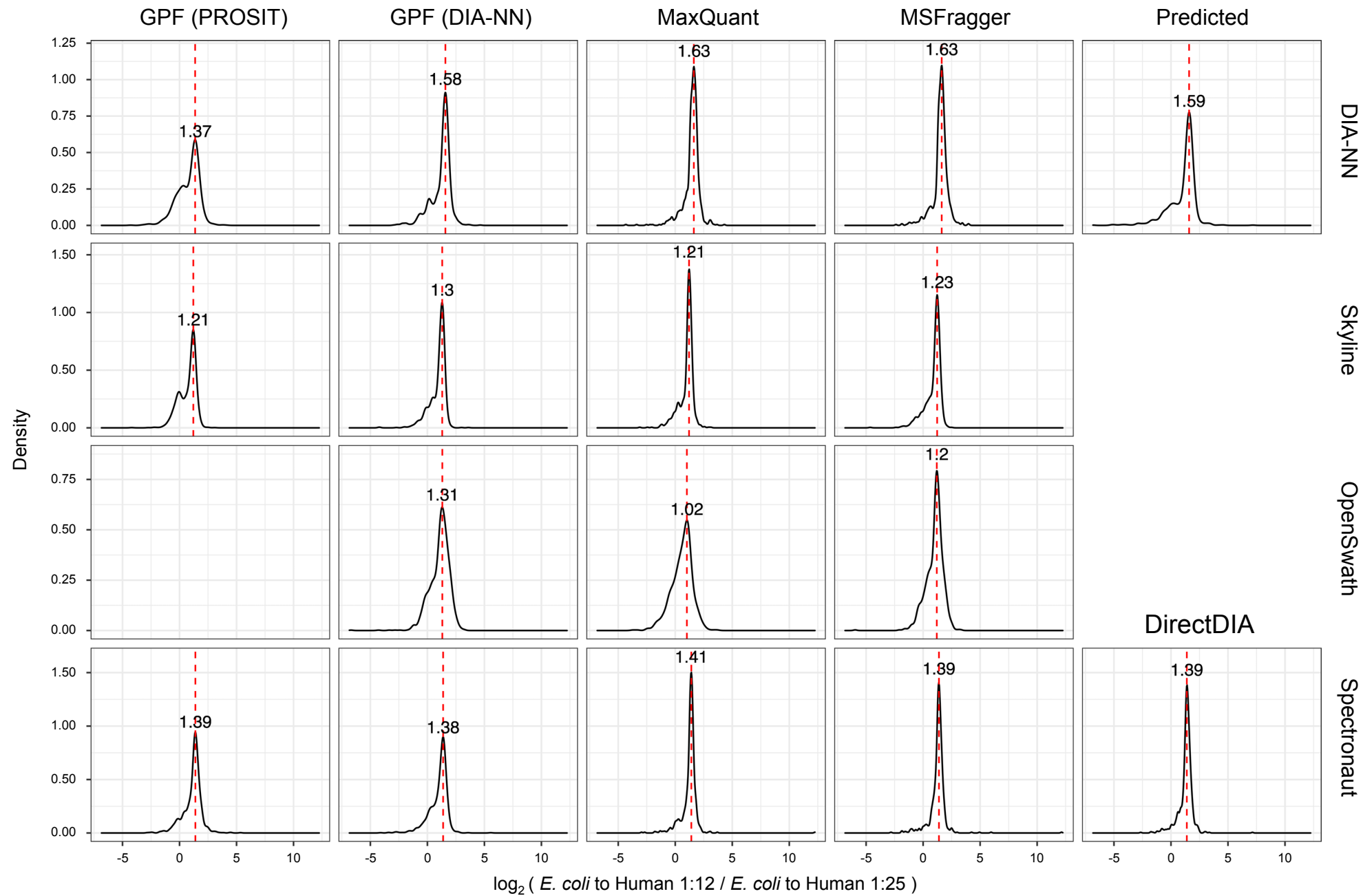
Supplementary Figure 10: Most Human Protein Identifications Are Shared Between All DIA Workflows with 'DIA-NN Predicted' representing the DIA Analysis Workflow capturing the highest Number of Unique Human Proteins.

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 1:12 or 1:25 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 list. Unlike for the upset plot, in the table to the right protein identifiers containing multiple protein names are counted as only one protein. Source data are provided as a Source Data file.



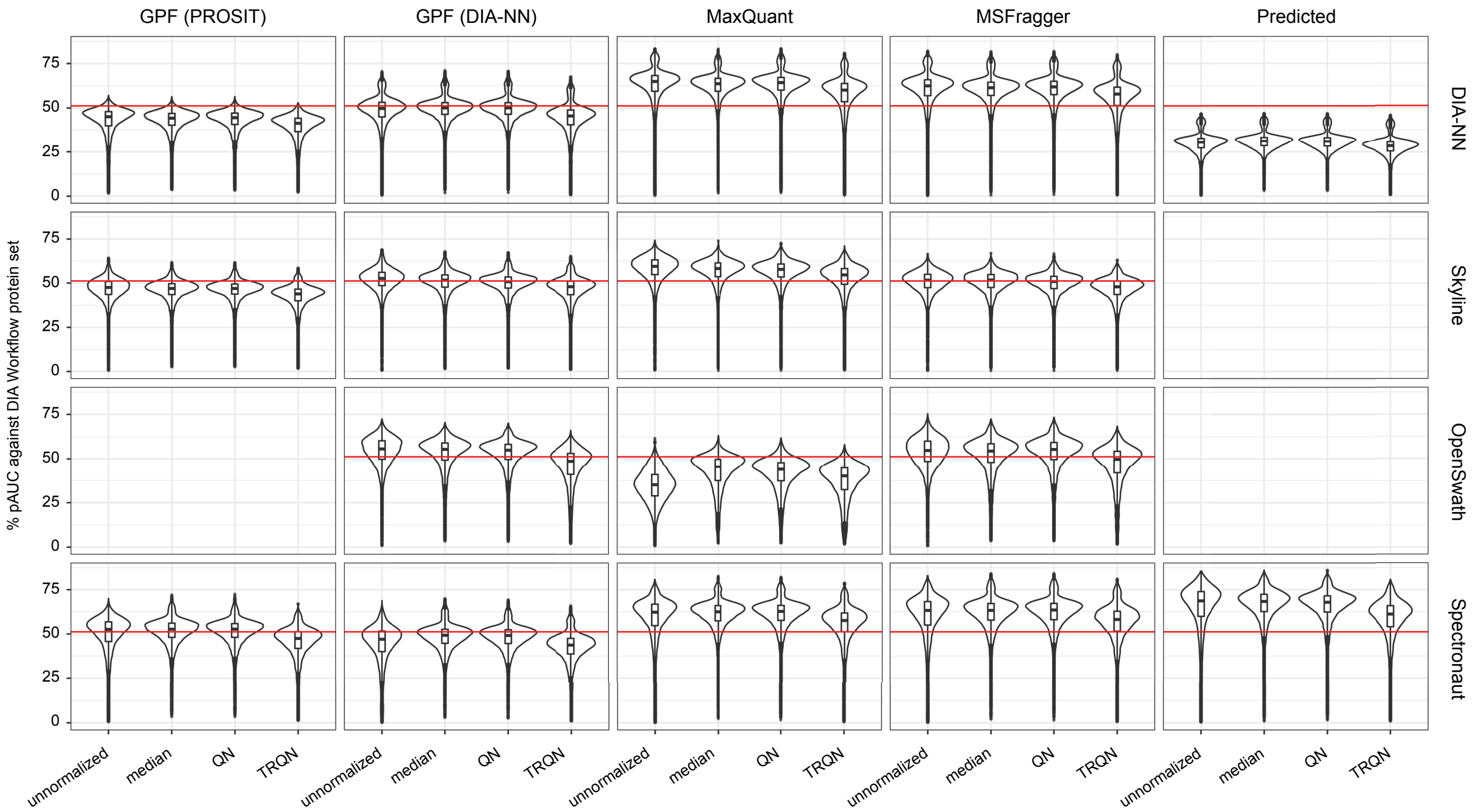
Supplementary Figure 11: Most *E. coli* Protein Identifications Are Shared Between All DIA Workflows, with 'DIA-NN Predicted' representing the DIA Analysis Workflow capturing the highest Number of Unique *E. coli* Proteins.

See legend of Supplementary Figure 10, but instead of human proteins *E. coli* proteins are summarized. Source data are provided as a Source Data file.



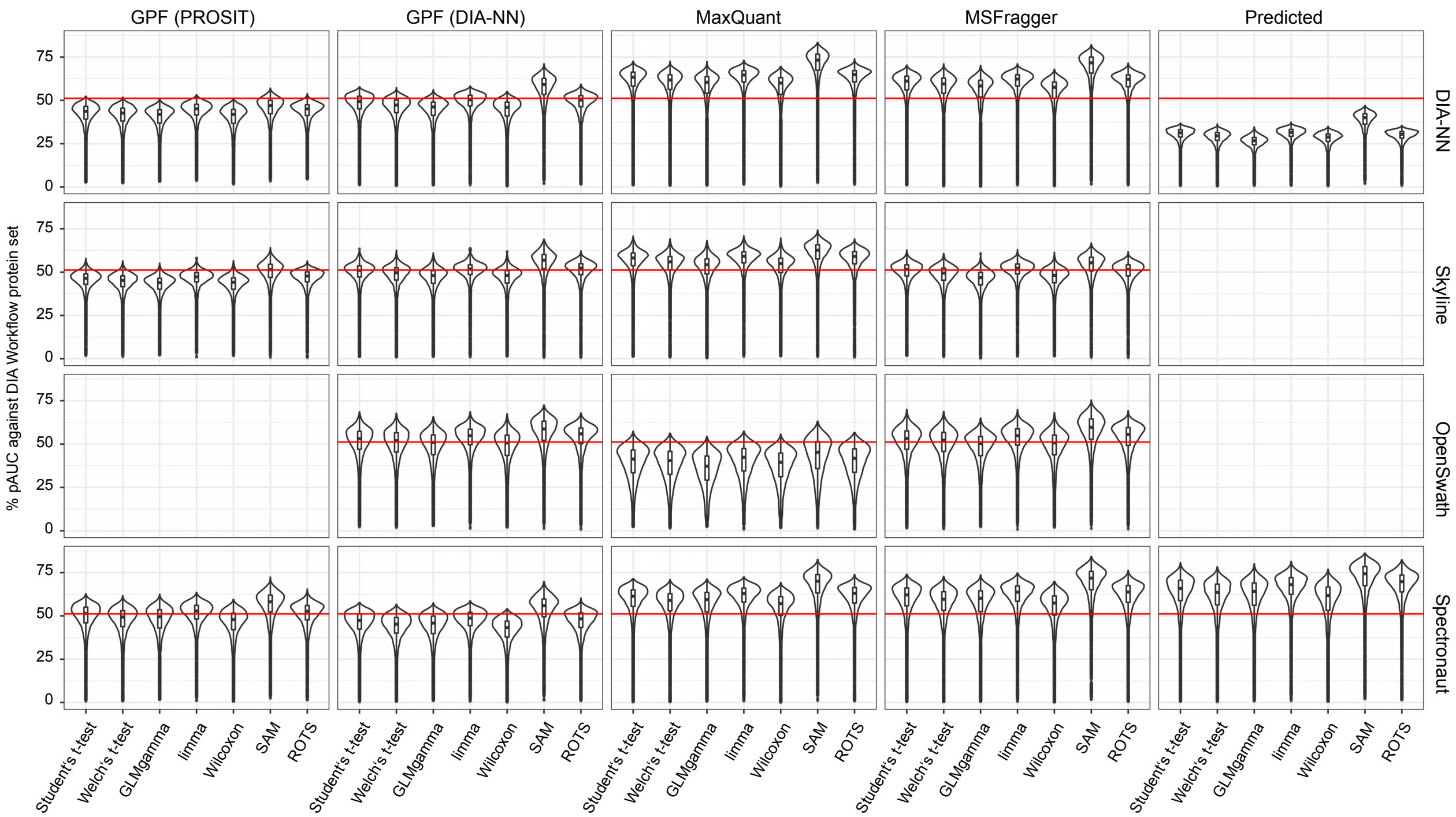
## Supplementary Figure 12: Accuracy of Fold Change Mainly Depends on Employed DIA Analysis Software.

*E. coli* log<sub>2</sub> fold changes observed between the spike-in conditions *E. coli* to human 1:12 and 1:25. The theoretical log<sub>2</sub> fold change is 1.11. The dotted red line indicates the mode of the respective distribution. The occurrence of a second smaller peak around 1 likely implies human proteins being incorrectly annotated as *E. coli* proteins. For each spike-in condition data of n=23 biologically independent samples have been used. Source data are provided as a Source Data file.



Supplementary Figure 13: Performing no normalization on our benchmark dataset resulted in best prediction performance for most DIA workflows.

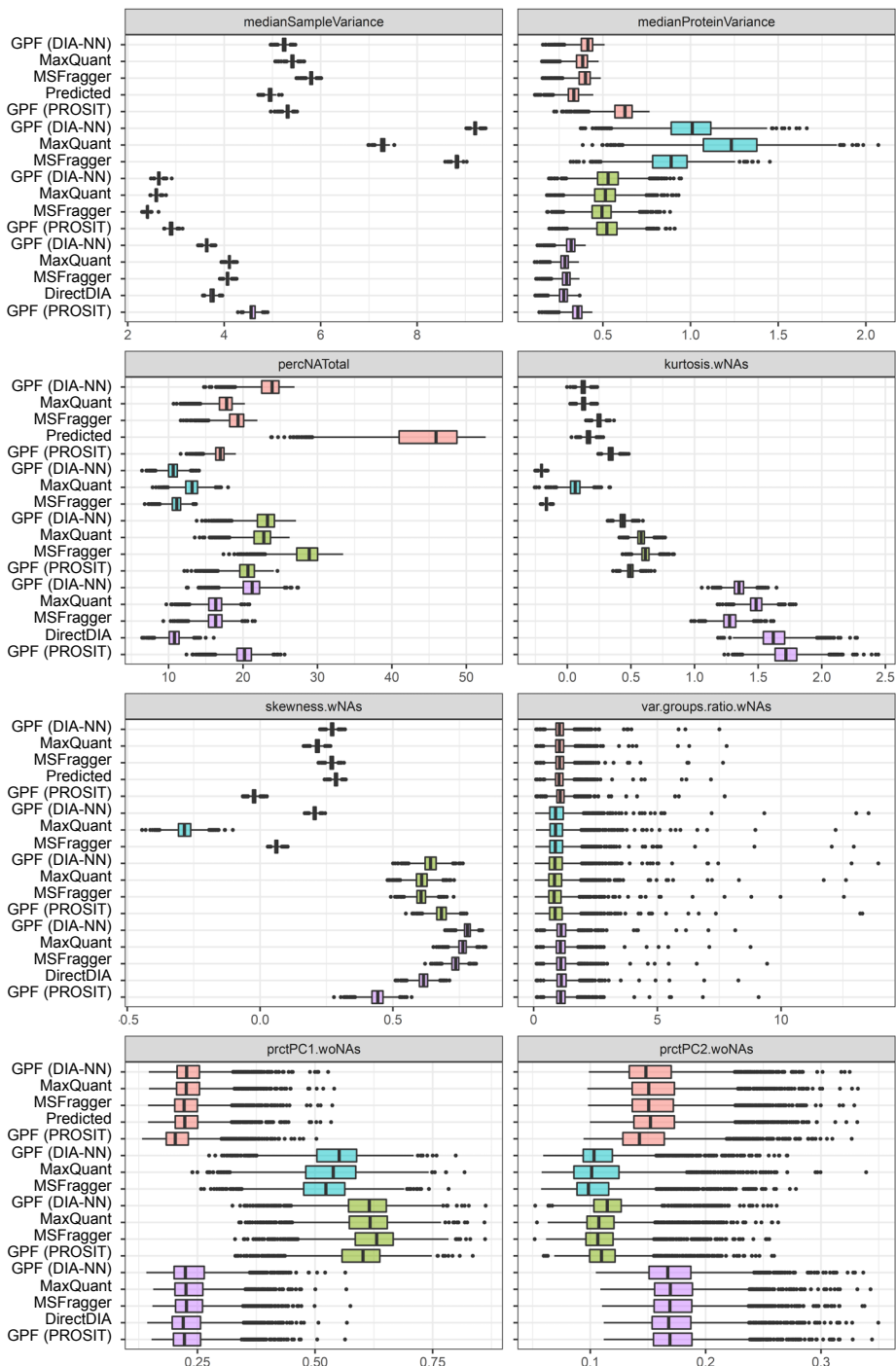
Comparison of normalization options for no sparsity reduction (NoSR). pAUC was calculated based on the 'DIA workflow' reference protein list. The red line indicates the overall median. Each subplot is based on  $n=2100$  bootstrap datasets which have been generated by drawing with replacement from data of  $n=23$  biologically independent samples of spike-in conditions 1:12 and 1:25, respectively. The sample size of these bootstrap datasets ranged from 3 to 23 samples, which due to drawing with replacement can appear multiple times. For each subplot that comes to a total of  $n=2100 \times 3$  sparsity reductions  $\times 7$  statistical tests = 44100 data points per normalization setting. The boxplots show median (center line), interquartile range (IQR, extending from the first to the third quartile) (box), and  $1.5 \times$  IQR (whiskers). Source data are provided as a Source Data file.





Supplementary Figure 14: The statistical test SAM performs best in detecting differentially abundant proteins.

Comparison of statistical tests for no sparsity reduction (NoSR). pAUC was calculated based on the DIA workflow reference protein list. The red line indicates the overall median. All seven statistical tests were two-sided and not adjusted for multiple testing. Each subplot is based on  $n=2100$  bootstrap datasets which have been generated by drawing with replacement from data of  $n=23$  biologically independent samples of spike-in conditions 1:12 and 1:25, respectively. The sample size of these bootstrap datasets ranged from 3 to 23 samples, which due to drawing with replacement can appear multiple times. For each subplot that comes to a total of  $n=2100 \times 3$  sparsity reductions  $\times 4$  normalizations = 25200 data points per statistical test setting. The boxplots show median (center line), interquartile range (IQR, extending from the first to the third quartile) (box), and  $1.5 \times \text{IQR}$  (whiskers). Source data are provided as a Source Data file.



● DIA-NN

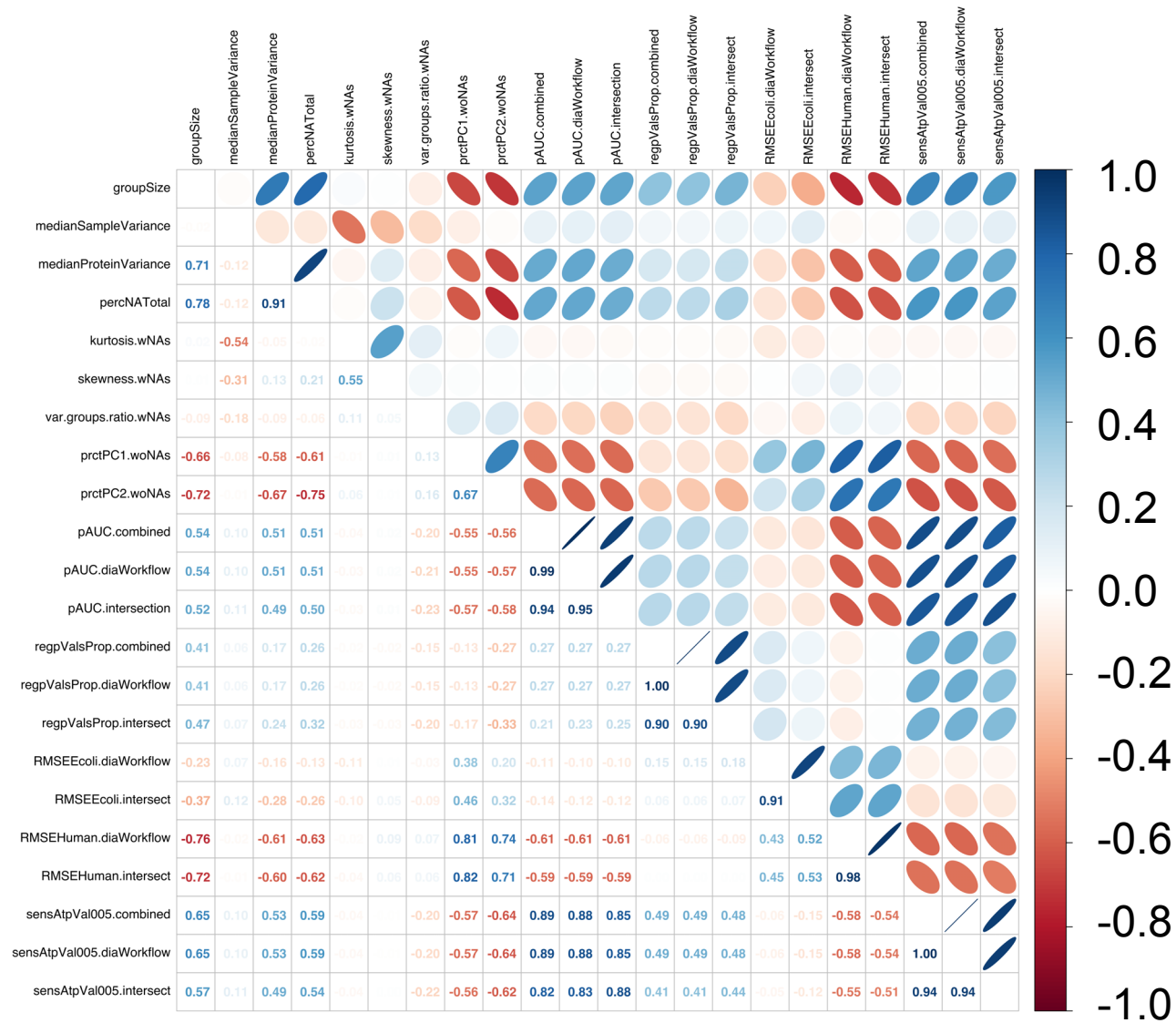
● Skyline

● OpenSwath

● Spectronaut

Supplementary Figure 15: 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 and second principal component to the total variance.

Distribution of selected data characteristics for all bootstrap datasets of the different DIA analysis workflows. 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 1:25 and 1:12), prcPC1.woNAs = percentage of the contribution of the first principal component to the total variance, prcPC2.woNAs = percentage of the contribution of the second principal component to the total variance. Each of the eight subplots is based on n=2100 bootstrap datasets which have been generated by drawing with replacement from data of n=23 biologically independent samples of spike-in conditions 1:12 and 1:25, respectively. The sample size of these bootstrap datasets ranged from 3 to 23 samples, which due to drawing with replacement can appear multiple times. The boxplots show median (center line), interquartile range (IQR, extending from the first to the third quartile) (box), and  $1.5 * IQR$  (whiskers). Source data are provided as a Source Data file.



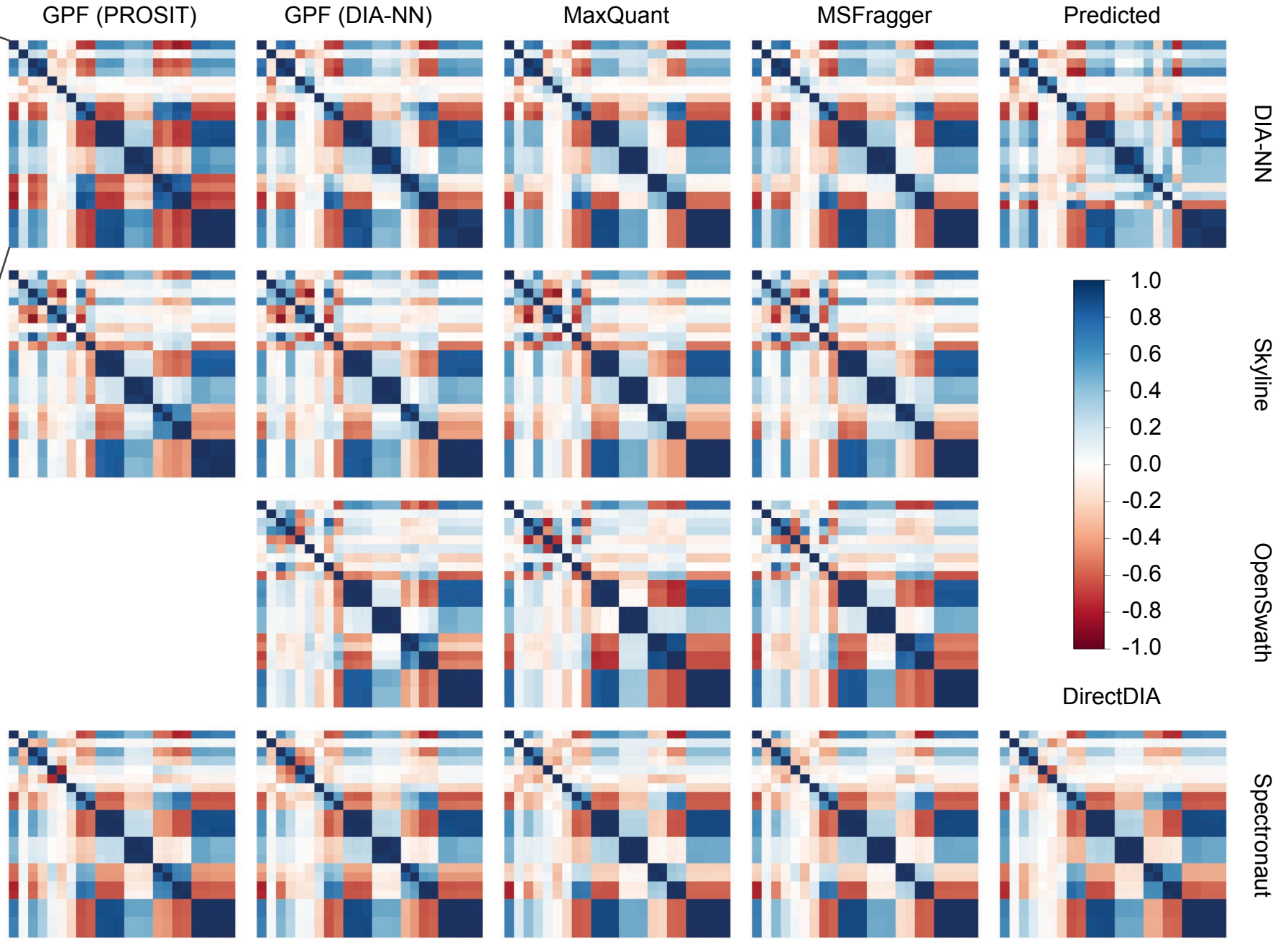
Supplementary Figure 16: Sample variance, kurtosis, skewness and the ratio of variances between two spike-in conditions have little influence on the performance of statistical tests.

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 that used DIA-NN in combination with the *in silico* predicted GPF-refined (DIA-NN) spectral library. The data characteristics were log<sub>2</sub>-transformed prior to the correlation analysis. For terminology of data characteristics see Supplementary Figure 15. Terminology of performance measures (extensions .combined, .DiaWorkflow, .intersect/.intersection refer to the reference protein list based on which the performance measure has been calculated): pAUC= partial area under the curve (pAUC), regValsProp = estimated proportion of regulated proteins based on p-value distribution ( $1-\pi_0$ ), RMSE<sub>Ecoli</sub> = root-mean-square error (RMSE) of all *E. coli* proteins, RMSE<sub>Human</sub> = RMSE of all human proteins, sensAtpVal005 = sensitivity calculated by predicting all proteins with p-value < 0.05 to be differentially abundant. Pearson correlation ranges from -1 (perfect negative correlation, red) to 1 (perfect positive correlation, blue)

The analysis is based on n=2100 bootstrap datasets which have been generated by drawing with replacement from data of n=23 biologically independent samples of spike-in conditions 1:12 and 1:25, respectively. The sample size of these bootstrap datasets ranged from 3 to 23 samples, which due to drawing with replacement can appear multiple times.

That results in a total of n=2100\*3 sparsity reductions\*4 normalizations\*7 statistical tests=176400 cases, for which both prediction performance as well as information on data characteristics are available. Source data are provided as a Source Data file.

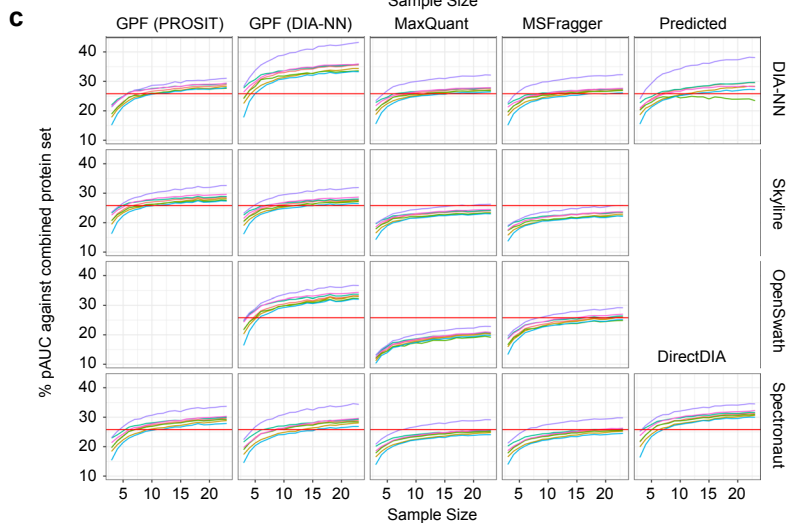
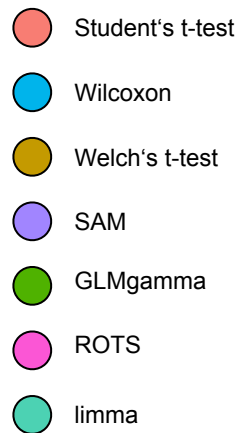
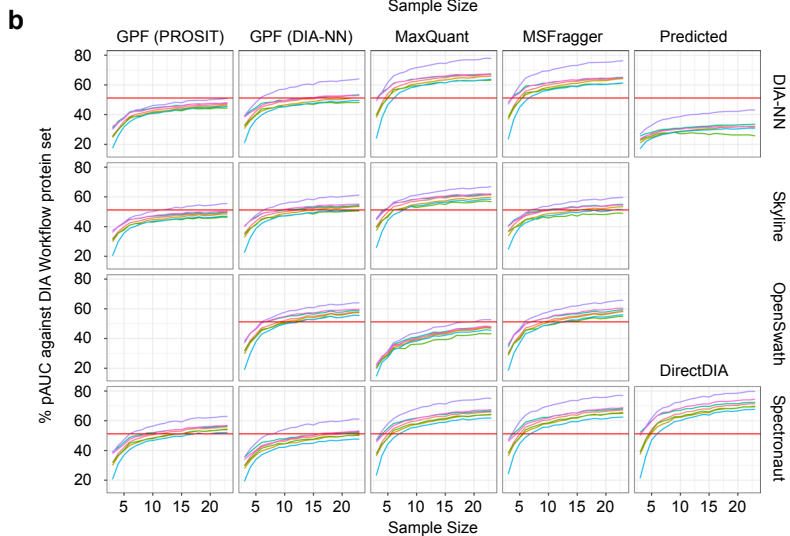
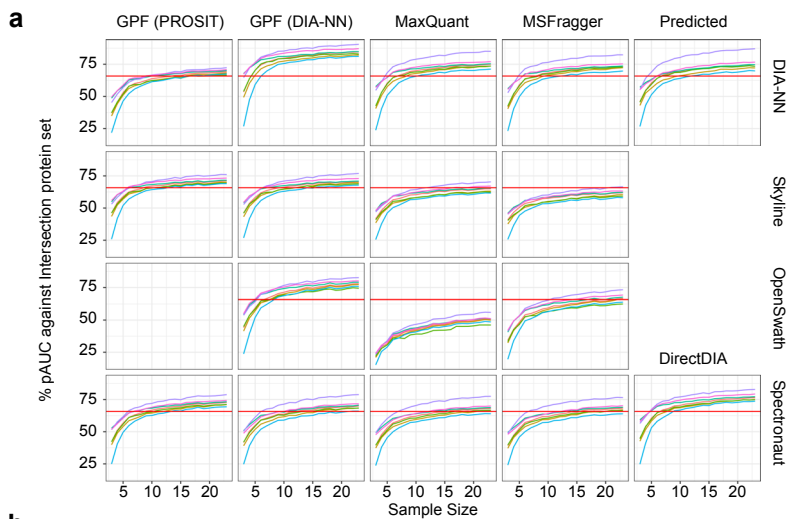
groupSize  
 medianSampleVariance  
 medianProteinVariance  
 percNATotal  
 kurtosis.wNAs  
 skewness.wNAs  
 var.groups.ratio.wNAs  
 prctPC1.wNAs  
 prctPC2.wNAs  
 pAUC.combined  
 pAUC.diaWorkflow  
 pAUC.intersection  
 regpValsProp.combined  
 regpValsProp.diaWorkflow  
 regpValsProp.intersect  
 RMSEEcoli.diaWorkflow  
 RMSEEcoli.intersect  
 RMSEHuman.diaWorkflow  
 RMSEHuman.intersect  
 sensAtpVal005.combined  
 sensAtpVal005.diaWorkflow  
 sensAtpVal005.intersect



DirectDIA

## Supplementary Figure 17 The correlation between the data characteristics varies with DIA analysis workflow

Pearson correlations between data characteristics and performance measures of bootstrap datasets (using all complete pairs of observations) for all 17 DIA analysis workflows. Pearson correlation ranges from -1 (perfect negative correlation, red) to 1 (perfect positive correlation, blue). The terminology of the data characteristics and performance measures is analogous to Supplementary Figure 16. The analysis is based on  $n=2100$  bootstrap datasets which have been generated by drawing with replacement from data of  $n=23$  biologically independent samples of spike-in conditions 1:12 and 1:25, respectively. The sample size of these bootstrap datasets ranged from 3 to 23 samples, which due to drawing with replacement can appear multiple times. That comes to a total of  $n=2100 \times 3$  sparsity reductions  $\times 4$  normalizations  $\times 7$  statistical tests = 176400 cases, for which both prediction performance as well as information on data characteristics are available. Source data are provided as a Source Data file.

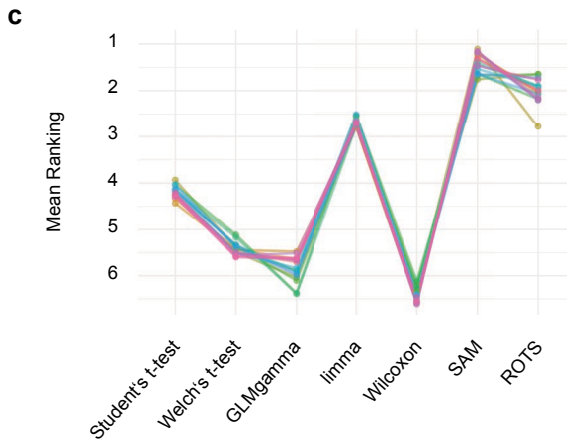
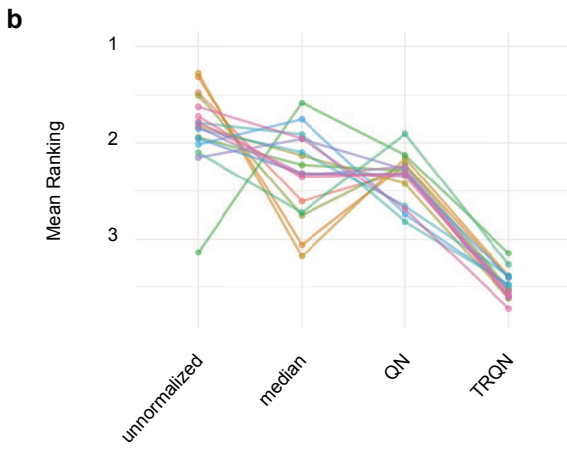
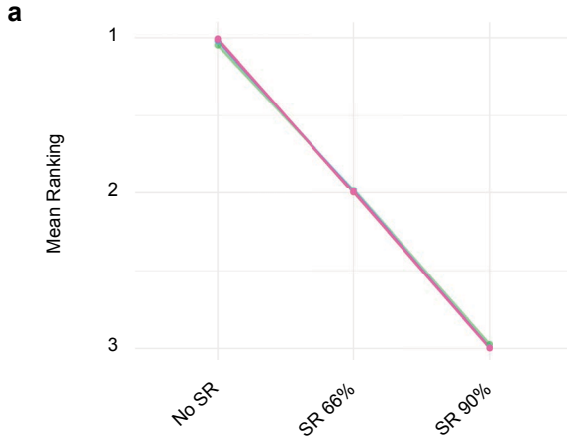




Supplementary Figure 18: SAM performs best for all software suites. limma performs well for small sample sizes.

Prediction performance of different statistical tests in pAUC over different sample sizes for protein reference lists a) Intersection, b) DiaWorkflow, and c) Combined. The lines, which are color-coded by statistical test, represent the median over the respective sample size. All three reference protein lists are similar in terms of the order of best performing statistical tests. The red line indicates the overall median. All seven statistical tests were two-sided and not adjusted for multiple testing.

Each subplot is based on  $n=2100$  bootstrap datasets which have been generated by drawing with replacement from data of  $n=23$  biologically independent samples of spike-in conditions 1:12 and 1:25, respectively. The sample size of these bootstrap datasets ranged from 3 to 23 samples, which due to drawing with replacement can appear multiple times. Each data point the lines are build upon represents the median of  $n=(2100/21)*3$  sparsity reductions\*4 normalizations=1200 data points. Source data are provided as a Source Data file.

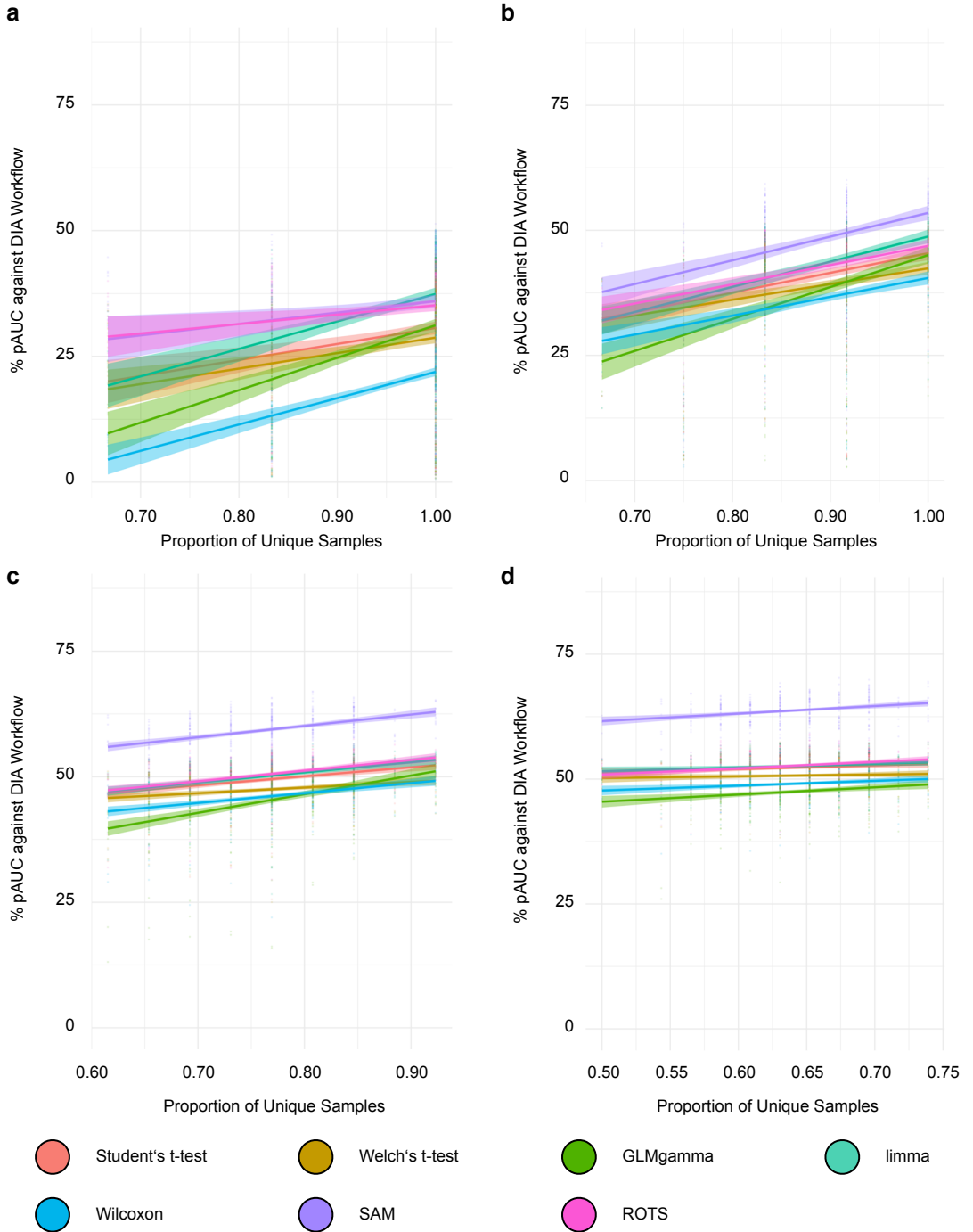


Software	Library	Color
DIA-NN	GPF (DIA-NN)	Red
	MaxQuant	Orange
	MSFragger	Yellow-Orange
	Predicted	Yellow-Green
OpenSwath	GPF (PROSIT)	Green
	GPF (DIA-NN)	Light Green
	MaxQuant	Light Green
Skyline	MSFragger	Teal
	GPF (DIA-NN)	Teal
	MaxQuant	Light Blue
Spectronaut	MSFragger	Blue
	GPF (PROSIT)	Blue
	GPF (DIA-NN)	Purple
	MaxQuant	Purple
„DirectDIA“	MSFragger	Pink
	GPF (PROSIT)	Pink
	GPF (DIA-NN)	Pink

Supplementary Figure 19: Ranked performance of DIA analysis workflows for different choices of a) sparsity reduction, b) normalization, and c) statistical tests.

Specifically, e.g. for statistical tests, for each bootstrap dataset-DIA workflow-sparsity reduction-normalization-statistical test combination the rank of each statistical test was calculated based on the pAUC. Then the mean rank over all bootstrap dataset-DIA workflow - statistical test combinations was calculated, and, subsequently, the results were averaged over all bootstrap datasets. This has been done for all three reference protein lists and all three results were averaged such that, finally, each statistical test got a rank assigned for each DIA analysis workflow. The lower the rank, the better the performance.

All seven statistical tests were two-sided and not adjusted for multiple testing. Source data are provided as a Source Data file.



Supplementary Figure 20. All investigated statistical tests show a comparable decrease in prediction performance for an increasing number of duplicate samples in the bootstrap datasets due to bootstrapping with replacement.

This is shown for sample sizes a) 3, b) 6, c) 13, and d) 23 at the example of DIA analysis workflow DIA-NN GPF (DIA-NN) when no sparsity reduction was applied. For each sample size linear regression lines with a 95% confidence interval are depicted for each statistical test. All seven statistical tests were two-sided and not adjusted for multiple testing.

For the comparison of performances it is most important that the bias through duplicated samples is comparable between the different statistical tests, which is, indeed, the case here.

Source data are provided as a Source Data file.