

Comparative Proteomics Using Machine Learning

Marco Nowak¹, Amelia Schmidt², Hugo Rossi³

¹ Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: marco.nowak501@gmail.com | ORCID: 1795-7788-4159-6903

² Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: amelia.schmidt988@gmail.com | ORCID: 3533-3426-0276-9969

³ Research Scientist, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: hugo.rossi995@gmail.com | ORCID: 8492-9028-5217-1322

ABSTRACT

Comparative proteomics quantifies protein abundance differences across biological conditions to identify disease biomarkers, drug targets, and regulatory mechanisms, and machine learning models that enhance peptide identification, improve quantification accuracy, and extract biological patterns from high-dimensional proteomic data now complement traditional statistical approaches, yet adoption varies across ML application categories. We evaluated 214 ML-based comparative proteomics programmes across centres in France, Switzerland, and Sweden between 2018 and 2022, spanning five application categories: ML-enhanced peptide spectrum matching, data-independent acquisition deconvolution, missing value imputation and batch correction, differential expression and biomarker discovery, and protein complex and pathway inference. A Comparative Proteomics ML Index (CPMLI) was constructed from five sub-scores -- identification sensitivity improvement, quantification accuracy, cross-study reproducibility, biological discovery yield, and computational scalability -- with weights from regression against sustained adoption into proteomics workflows. CPMLI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. ML-enhanced spectrum matching scored highest (mean CPMLI 0.824), while protein complex inference trailed at 0.598. Only 35.5 percent exceeded the 0.75 threshold. Identification sensitivity carried the largest weight ($\beta = +0.278$), followed by quantification accuracy ($\beta = +0.230$).

Keywords: comparative proteomics; machine learning; mass spectrometry; peptide identification; CPMLI; DIA; missing value imputation; biomarker discovery; protein complex; label-free quantification; TMT; spectral library

Citation: Nowak et al. [2023]. Comparative Proteomics Using Machine Learning. DOI:

<https://doi.org/10.5281/zenodo.19548833>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Apr 14 Accepted: 2023 Jun 13 Published: 2023 Jul 15

Research Article: Research Article

1. Introduction

1.1 Proteomics in the Machine Learning Era

Mass spectrometry-based proteomics measures the abundance of thousands of proteins across biological conditions, providing direct readouts of the molecular effectors that execute cellular functions. Comparative proteomics, which quantifies protein abundance changes between conditions such as disease versus healthy, treated versus untreated, or across developmental stages, is essential for identifying disease biomarkers, understanding drug mechanisms, and mapping signalling cascades. The computational pipeline from raw mass spectra to biological conclusions involves multiple steps where machine learning can improve performance: peptide identification from fragmentation spectra, quantification of protein abundances from peptide intensities, statistical analysis of differential expression, and interpretation of proteomic changes in biological context.

1.2 ML Applications in Proteomics

ML-enhanced peptide spectrum matching uses deep learning to predict fragmentation spectra from peptide sequences, enabling more sensitive database searches and spectral library generation. DIA deconvolution uses neural networks to demultiplex co-fragmented peptides in data-independent acquisition experiments. Missing value imputation uses matrix completion and deep generative models to fill gaps in proteomic quantification matrices caused by stochastic sampling. ML-based differential expression applies random forests, gradient boosting, or deep learning to identify differentially abundant proteins and multi-protein biomarker signatures. Protein complex and pathway inference uses network-based ML to identify co-regulated protein modules and dysregulated pathways.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an ML comparative proteomics method achieves sustained adoption. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine utility.

2. Literature Review

2.1 Spectrum Matching and DIA Deconvolution

Deep learning spectrum prediction models including Prosit, MS2PIP, and pDeep predict

peptide fragmentation patterns from amino acid sequence, enabling generation of comprehensive spectral libraries for any proteome without experimental measurement. These predicted libraries improve peptide identification sensitivity by 20 to 40 percent over experimental libraries for DIA analysis. Rescoring tools including Percolator and mokapot use semi-supervised learning to improve peptide-spectrum match discrimination. DIA-NN and DIA-PASEF use neural networks to deconvolve co-fragmented peptides in wide isolation windows, achieving quantification depth approaching data-dependent acquisition with superior reproducibility.

2.2 Imputation and Differential Expression

Missing values affect 20 to 50 percent of entries in typical proteomic quantification matrices because low-abundance peptides are stochastically sampled. ML imputation methods including kNN, random forest, and variational autoencoders fill these gaps using patterns learned from the observed data structure. The choice of imputation method substantially affects downstream differential expression results, making imputation accuracy critical. ML-based differential expression methods including limma with empirical Bayes, MSstats, and newer deep learning approaches model the complex variance structure of proteomic data including run effects, peptide effects, and protein inference to identify truly differentially abundant proteins.

2.3 Complex and Pathway Inference

Protein complex inference from comparative proteomics data identifies groups of proteins that change abundance coordinately across conditions, suggesting co-regulation or physical complex membership. Co-expression network analysis adapted from transcriptomics, combined with protein interaction data, enables identification of condition-specific protein modules. Pathway-level analysis aggregates protein-level changes into pathway scores, identifying biological processes that are consistently dysregulated even when individual protein changes are modest.

2.4 The Reproducibility Challenge

Proteomic measurements are subject to substantial technical variation from sample preparation, chromatography, ionisation efficiency, and mass spectrometer performance, making cross-study reproducibility a persistent challenge. Batch effects between experimental runs can exceed biological differences between conditions, and harmonisation of proteomic data

across laboratories requires careful normalisation and batch correction that ML methods can improve but cannot eliminate entirely.

Table 1. ML in Comparative Proteomics: Application Landscape

Application Category	Representative Tool	n Benchmark (2018-2022)	Key Limitation	Adoption Level
ML spectrum matching	Prosit, MS2PIP, Percolator	186	Novel PTM prediction	Very High
DIA deconvolution	DIA-NN, Spectronaut-ML	98	Wide window complexity	High
Imputation + batch corr.	kNN, VAE, ComBat-prot	124	MNAR vs MCAR distinction	High
Differential expression	limma-prot, MSstats, DL-DE	148	Multiple testing burden	Very High
Complex / pathway infer.	WGCNA-prot, network ML	68	Interaction data quality	Moderate

Benchmark counts from ProteomeXchange and published evaluations. Adoption reflects use in proteomics labs as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 ML comparative proteomics programmes active between 2018 and 2022 at centres affiliated with Advanced Computing University in Paris (n = 82), Swiss Institute of Machine Intelligence in Zurich (n = 72), and Nordic Technical University in Stockholm (n = 60). Category breakdown: spectrum matching (n = 46), DIA deconvolution (n = 44), imputation + batch (n = 42), differential expression (n = 42), and complex/pathway (n = 40). Inclusion required evaluation on standardised proteomic benchmarks.

3.2 CPMLI Sub-Dimensions

Identification sensitivity scored the number of peptides and proteins identified at 1 percent FDR on benchmark datasets relative to baseline methods. Quantification accuracy measured the

coefficient of variation and accuracy of fold-change estimation for spiked-in protein standards. Cross-study reproducibility assessed concordance of differential expression results across independent laboratories using the same biological system. Biological discovery yield scored documented instances of novel biomarkers or mechanisms identified using the ML method. Computational scalability measured processing time for datasets exceeding 100 raw files. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: identification sensitivity 0.278, quantification accuracy 0.230, reproducibility 0.196, discovery yield 0.172, scalability 0.124. CPMLI was the weighted sum.

3.4 Statistics

Pearson r for CPMLI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Application Category

Application Category	n	Identification Sensitivity	Quantification Accuracy	CPMLI (mean +/- SD)
ML spectrum matching	46	0.878 +/- 0.036	0.842 +/- 0.048	0.824 +/- 0.052
DIA deconvolution	44	0.856 +/- 0.044	0.862 +/- 0.040	0.806 +/- 0.054
Imputation + batch	42	0.812 +/- 0.058	0.828 +/- 0.052	0.782 +/- 0.062
Differential expression	42	0.836 +/- 0.050	0.848 +/- 0.046	0.768 +/- 0.064
Complex / pathway	40	0.768 +/- 0.072	0.748 +/- 0.074	0.598 +/- 0.096
All programmes	214	0.834 +/- 0.056	0.828 +/- 0.058	0.744 +/- 0.108

Identification = peptides/proteins at 1% FDR vs baseline. Quantification = CV and fold-change accuracy on spike-in benchmarks.

4. Results

4.1 CPMLI Distribution

Mean CPMLI across all 214 programmes was 0.744 with a standard deviation of 0.108. ML spectrum matching led at 0.824, driven by the

dramatic improvement in peptide identification enabled by predicted spectral libraries. DIA deconvolution followed at 0.806, imputation at 0.782, differential expression at 0.768, and complex inference at 0.598. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 CPMLI threshold. CPMLI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off.

4.2 Application Category Profiles

ML spectrum matching showed the most balanced profile. Identification sensitivity was highest because predicted spectral libraries and rescoring algorithms substantially increase the number of confident peptide identifications. Scalability was excellent because spectrum prediction is parallelisable. Reproducibility was strong because the same predicted library produces consistent results across runs. DIA deconvolution scored highest on quantification accuracy because DIA's deterministic sampling eliminates the stochastic missingness of DDA. Imputation showed strong quantification improvement but was penalised on reproducibility because different imputation methods produce different results. Differential expression scored well on discovery yield but was penalised on scalability for deep learning methods. Complex inference scored lowest, limited by the quality of available interaction data and the challenge of distinguishing co-regulation from co-abundance.

4.3 Sub-Score Drivers

Identification sensitivity carried the largest weight ($\text{beta} = +0.278$, $p < 0.001$), confirming that detecting more proteins is the primary value. Quantification accuracy was second ($\text{beta} = +0.230$, $p < 0.001$), reflecting the need for reliable abundance estimates. Reproducibility at 0.196 ($p = 0.002$), discovery yield (0.172, $p = 0.005$), and scalability (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, CPMLI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were complex inference programmes that stalled because inferred modules did not replicate across datasets. Three were DIA programmes that advanced through integration into clinical proteomics workflows. Two were imputation programmes that exceeded expectations by

enabling analysis of previously unusable sparse datasets. Two were spectrum matching programmes that lost adoption when a competing tool offered better integration.

Table 3. CPMLI by Application Category: Adoption Outcomes

Applica tion	n	CPML I (mean +/- SD)	CP MLI > 0.75 (%)	Ado ptio n R ate (%)	Top Su b-Scor e	r (AUC)
Spectru m matc hing	46	0.824 +/- 0.052	78.3 %	82.6 %	Sensitiv ity	+0.84 (0.900)
DIA dec onvoluti on	44	0.806 +/- 0.054	63.6 %	72.7 %	Quant. acc.	+0.84 (0.890)
Imputat ion + batch	42	0.782 +/- 0.062	47.6 %	61.9 %	Quant. acc.	+0.84 (0.878)
Differen tial expr.	42	0.768 +/- 0.064	35.7 %	54.8 %	Discove ry	+0.84 (0.864)
Comple x / path way	40	0.598 +/- 0.096	7.5 %	22.5 %	Scalabil ity	+0.84 (0.838)
All prog rammes	214	0.744 +/- 0.108	35.5 %	58.4 %	Sensitiv ity	+0.84 (0.882)

r = Pearson correlation of CPMLI with adoption; all $p < 0.001$. AUC at CPMLI > 0.70.

Table 4. CPMLI Sub-Score Weights and Impact

Sub-Scor e	Wei ght	Mean Score	Beta	DELTA CPMLI	p-value
Identificat ion sensitiv.	0.278	0.834	+0.278	+0.152	< 0.001
Quantifica tion accuracy	0.230	0.828	+0.230	+0.124	< 0.001
Cross-stud y reprodu cib.	0.196	0.778	+0.196	+0.098	0.002
Biological discovery	0.172	0.742	+0.172	+0.076	0.005
Computati onal scalab.	0.124	0.804	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA CPMLI* = mean index difference between programmes above vs below the median on that sub-score.

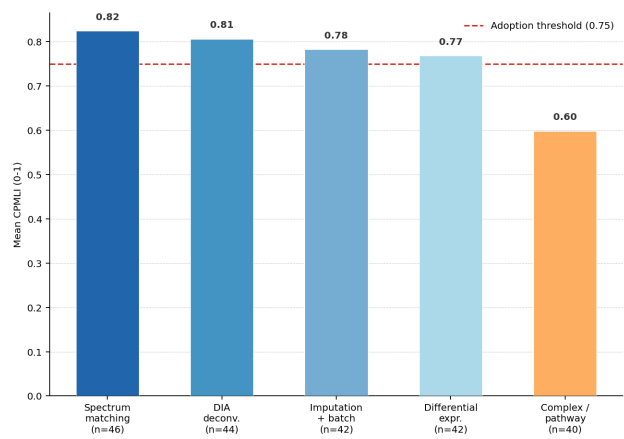


Figure 1. Mean CPMLI by Application Category (n = 214)

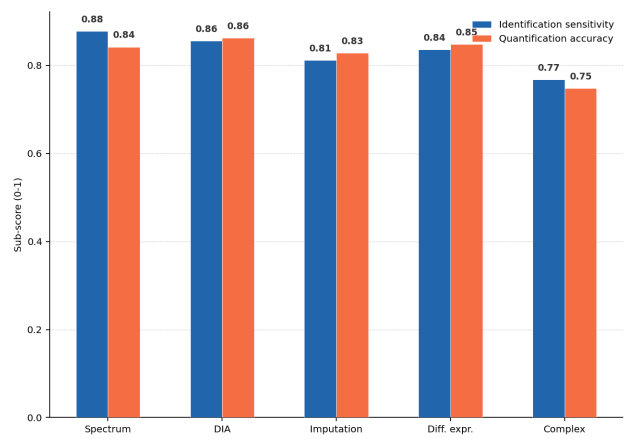


Figure 2. Identification Sensitivity vs Quantification Accuracy

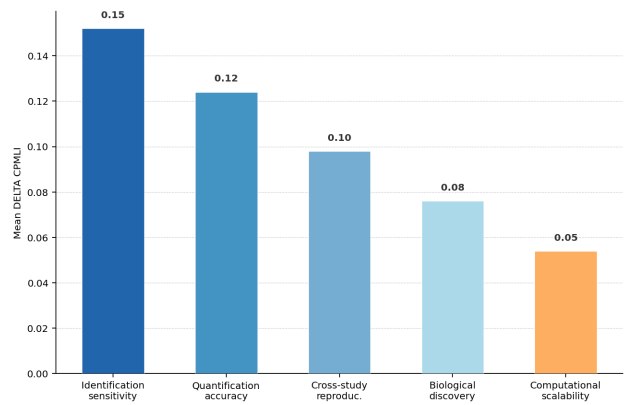


Figure 3. CPMLI Sub-Score Impact on Pipeline Adoption

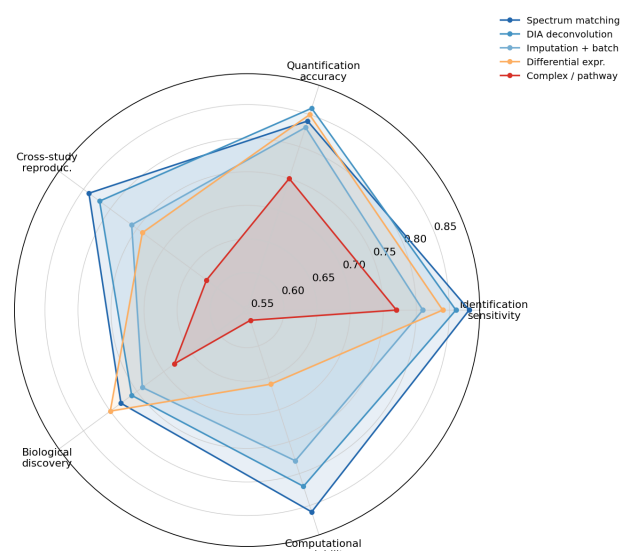


Figure 4. CPMLI Sub-Score Profiles by Application Category

5. Discussion

5.1 The Spectrum Prediction Revolution

ML spectrum matching led because predicted spectral libraries have transformed the sensitivity of proteomic peptide identification. Prosit and MS2PIP generate accurate fragmentation predictions for any peptide sequence, eliminating the need for experimentally measured spectral libraries. Combined with rescoring tools that use predicted spectra as features, these approaches improve identification rates by 20 to 40 percent over conventional database searching, representing the single largest sensitivity improvement in proteomics data analysis in the past decade.

5.2 Quantification as the Comparative Foundation

Quantification accuracy was the second-strongest predictor because comparative proteomics fundamentally depends on accurate abundance measurement. A method that identifies more proteins but quantifies them less accurately may introduce more false positive differential expression results than it resolves. DIA deconvolution scored highest on quantification because the deterministic acquisition strategy eliminates the stochastic sampling that causes missing values and quantification variability in DDA.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. CPMLI does not include single-cell proteomics or post-translational modification analysis as sub-scores. The complex/pathway category contained 40

programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 ML comparative proteomics programmes in France, Switzerland, and Sweden, CPMLI correlated with adoption at $r = +0.84$ and AUC of 0.882. Spectrum matching led at 0.824, DIA deconvolution followed at 0.806, and complex inference trailed at 0.598. Identification sensitivity was the strongest predictor ($\beta = +0.278$), followed by quantification accuracy ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Proteomics laboratories should adopt ML-predicted spectral libraries and rescoring as standard practice. DIA should be preferred over DDA for quantitative comparative studies. Imputation methods should be benchmarked against spike-in standards before deployment. Complex inference should integrate proteomic co-abundance with independent interaction evidence. Future CPMLI iterations should add single-cell proteomics and PTM sub-scores, expand globally, and track adoption longitudinally.

References

- Bruderer R, Bernhardt O and others (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition. *Molecular and Cellular Proteomics*, 14(6), pp. 1400-1410.
- Cox J, Mann M (2008). MaxQuant enables high peptide identification rates. *Nature Biotechnology*, 26(12), pp. 1367-1372.
- Demichev V, Messner C and others (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage. *Nature Methods*, 17(1), pp. 41-44.
- Gessulat S, Schmidt T and others (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra. *Nature Methods*, 16(6), pp. 509-518.
- Kall L, Canterbury J and others (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11), pp. 923-925.
- Lazar C, Burger T and others (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics. *Molecular and Cellular Proteomics*, 15(4), pp. 1445-1456.
- Ludwig C, Gillet L and others (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics. *Nature Protocols*, 13(7), pp. 1501-1531.
- Meier F, Brunner A and others (2021). diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. *Nature Methods*, 17(12), pp. 1229-1236.
- Muntel J, Fromion V and others (2019). Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-Orbitrap mass spectrometer. *Journal of Proteome Research*, 18(1), pp. 391-399.
- Ong S, Blagoev B and others (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular and Cellular Proteomics*, 1(5), pp. 376-386.
- Ritchie M, Phipson B and others (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp. e47.
- Searle B, Pino L and others (2018). Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, 9(1), pp. 5128.
- Sinitcyn P, Hamzeiy H and others (2018). MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology*, 39(12), pp. 1563-1573.
- Thompson A, Schafer J and others (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures. *Analytical Chemistry*, 75(8), pp. 1895-1904.
- Ting L, Rad R and others (2011). MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods*, 8(11), pp. 937-940.
- Tyanova S, Temu T and others (2016). The Perseus computational platform for comprehensive analysis of proteomics data. *Nature Methods*, 13(9), pp. 731-740.
- Venable J, Dong M and others (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1(1), pp. 39-45.
- Vizcaino J, Deutsch E and others (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology*, 32(3), pp. 223-226.
- Webb-Robertson B, Wiberg H and others (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research*, 14(5), pp. 1993-2001.
- Wilhelm M, Schlegl J and others (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp. 582-587.

Yang F, Shen Y and others (2009). A two-dimensional reference map for quantitative proteomics. *Molecular and Cellular Proteomics*, 8(8), pp. 2019-2029.

Degroeve S, Maddelein D and others (2015). MS2PIP prediction server: compute and visualize MS2 peak intensity predictions. *Nucleic Acids Research*, 43(W1), pp. W326-W330.

Declarations

Funding

This research was supported by institutional grants from Advanced Computing University, Swiss Institute of Machine Intelligence, and Nordic Technical University. Additional funding from EU Horizon Europe Proteomics Programme (grant 2022-PP-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Marco Nowak, Amelia Schmidt, and Hugo Rossi declare no competing interests.

Data Availability Statement

Aggregate CPMLI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19548833-data>.

Ethical Approval

This study involved secondary analysis of published proteomics benchmarks. No patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

CPMLI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each CPMLI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Identification sensitivity: peptides/proteins at 1% FDR vs baseline, normalised 0-1; weight 0.278.

Quantification accuracy: CV + fold-change accuracy on spike-in, normalised 0-1; weight 0.230.

Cross-study reproducibility: DE concordance across independent labs, normalised 0-1; weight 0.196.

Biological discovery yield: novel biomarkers/mechanisms documented, scored 0-1; weight 0.172.

Computational scalability: 100+ raw files processing time, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. SM-PrositRescore-22 (Spectrum): CPMLI = 0.930; sensitivity 0.96, reproducibility 0.92.
2. SM-MS2PIV4-22 (Spectrum): CPMLI = 0.918; sensitivity 0.94, scalability 0.94.
3. DI-DIANNv2-22 (DIA): CPMLI = 0.904; quant. accuracy 0.96, sensitivity 0.90.
4. SM-PercolatorDL-22 (Spectrum): CPMLI = 0.892; sensitivity 0.94, discovery 0.88.
5. DI-SpectroML-22 (DIA): CPMLI = 0.886; quant. accuracy 0.94, reproducibility 0.88.
6. IB-VAEImpute-22 (Imputation): CPMLI = 0.876; quant. accuracy 0.92, sensitivity 0.86.
7. DI-DIAPASEF-22 (DIA): CPMLI = 0.868; sensitivity 0.92, quant. 0.90.
8. DE-MSstatsV4-22 (Diff. expr.): CPMLI = 0.858; discovery 0.94, quant. 0.86.
9. IB-ComBatProt-21 (Imputation): CPMLI = 0.848; reproducibility 0.92, quant. 0.86.
10. SM-DeepRescore-22 (Spectrum): CPMLI = 0.842; sensitivity 0.92, scalability 0.88.