Northeastern
Khoury College of Computer and Information Sciences

Integration of longitudinal quality metrics enhances differential analysis in noisy large-scale Mass Spectrometry(MS)-based proteomics experiments

Devon Kohler[1,2]; Eralp Dogu[3]; Manuel Magana[4]; Mrittika Bhattacharya[4];

Ozge Karayel[4]; Veronica G Anania[4]; Olga Vitek[1,2]

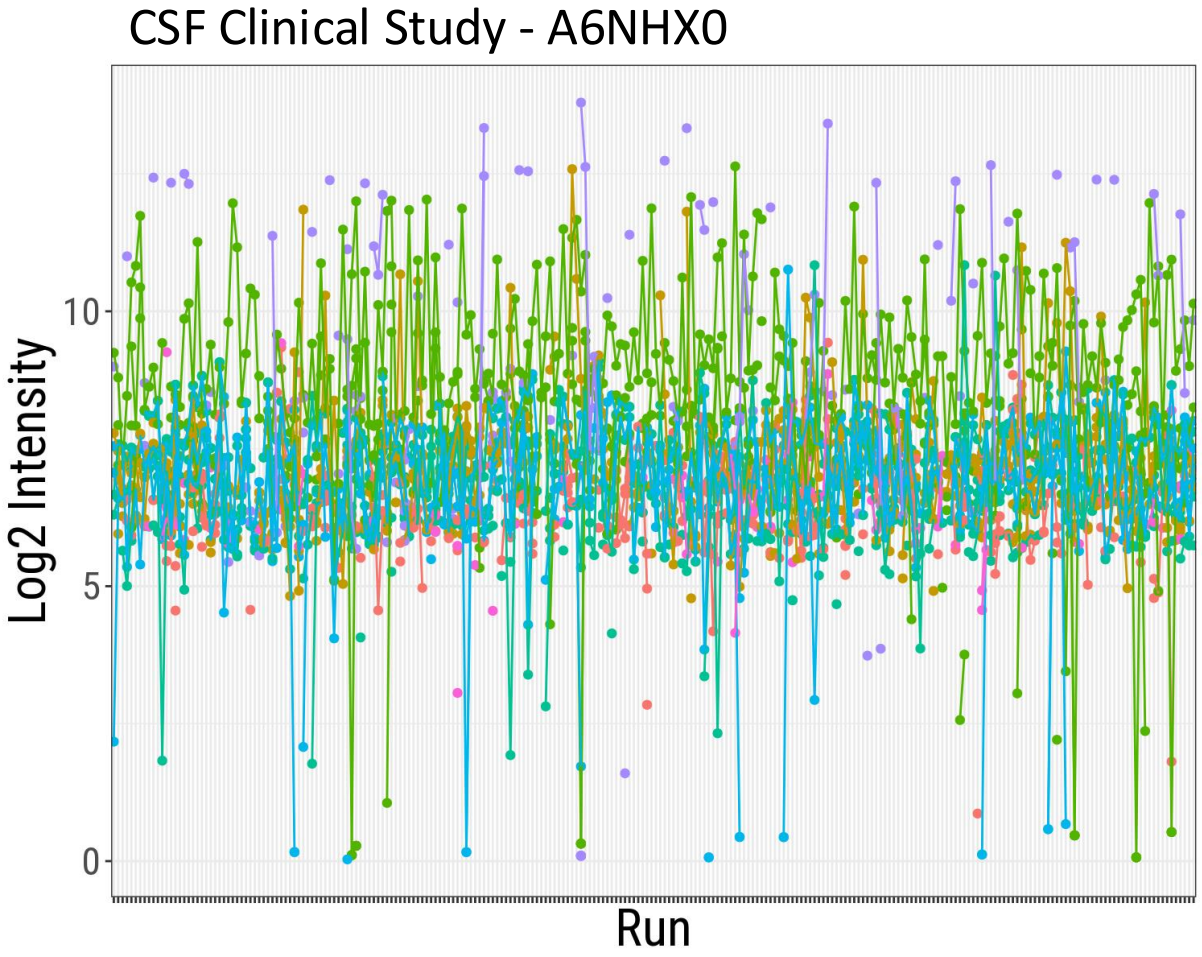
1. Northeastern University, Boston, MA;
2. Barnett Institute for Chemical and Biological Analysis, Boston, MA;
3. Mugla Sitki Kocman University, Köstekli, Turkey;
4. Genentech Inc., South San Francisco, CA;

Conflict of Interest

Veronica G. Anania, Manuel Magana, Ozge Karayel and Mrittika Bhattacharya are employees of Genentech

Maintaining data quality becomes harder as experiments increase in scale and complexity

- Poor quantitative values persist even when leveraging advanced tools
- Existing statistical solutions largely target intensity-based corrections
- We propose a method that leverages spectral peak quality metrics to enhance differential analysis

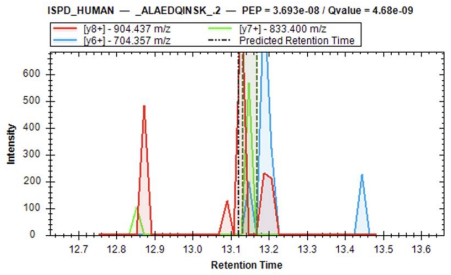


Outline

- Problem statement
- Background
  - Existing differential analysis methods
  - Informative quality metrics from spectral processing tools
- Incorporating quality metrics into differential analysis
- Case study and benchmarking

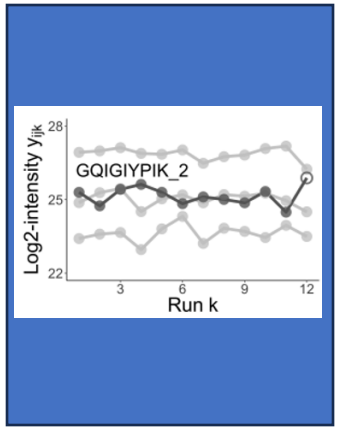
Standard summarization-based differential analysis workflow

Identified and Quantified Peaks

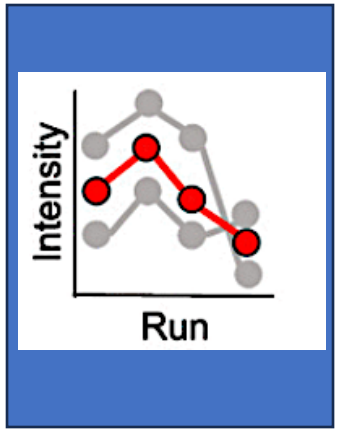


Data filtering

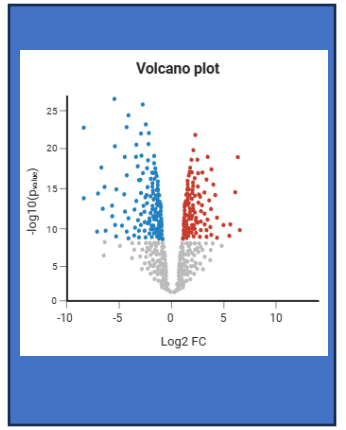
Feature selection



Protein-level summarization

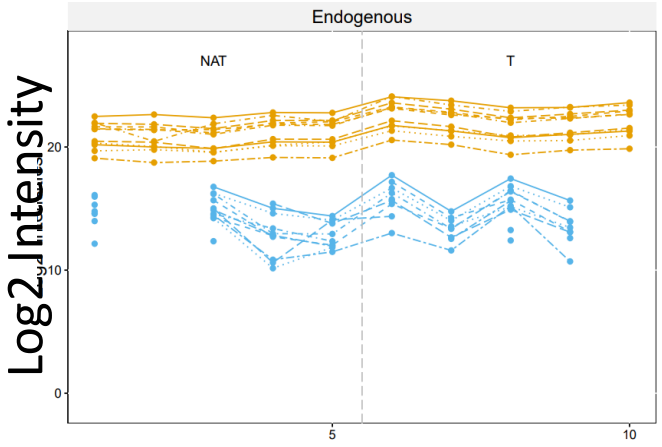


Differential analysis



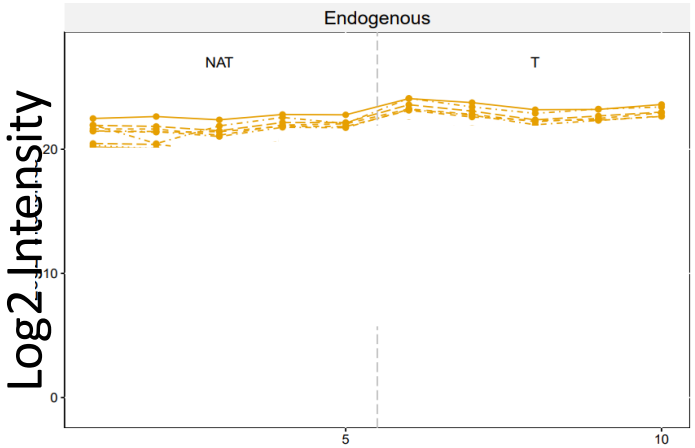
Feature selection removes fragments which adversely affect summarization

All



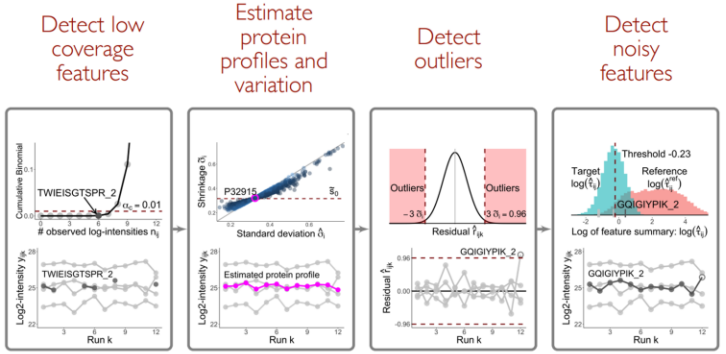
MS Runs

Top-N selection



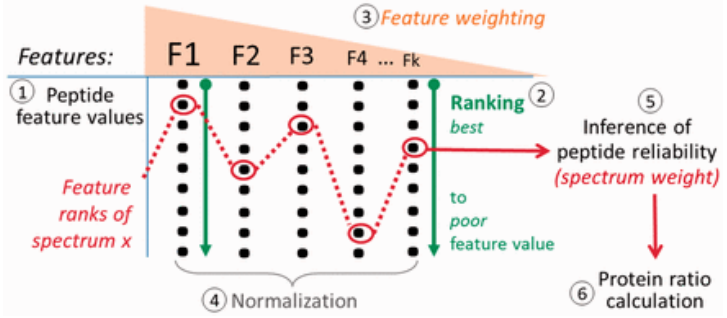
MS Runs

Best features



Tsai *et al*. Molecular & Cellular Proteomics, 19 (6), 944 – 959. (2020).

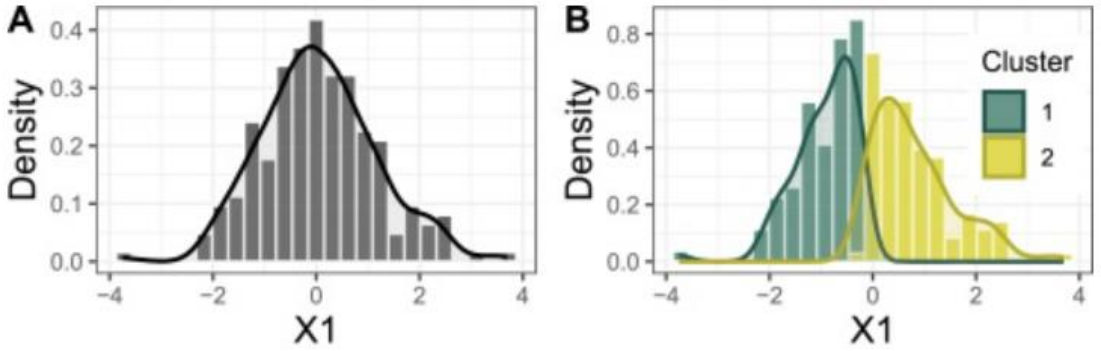
iPQF



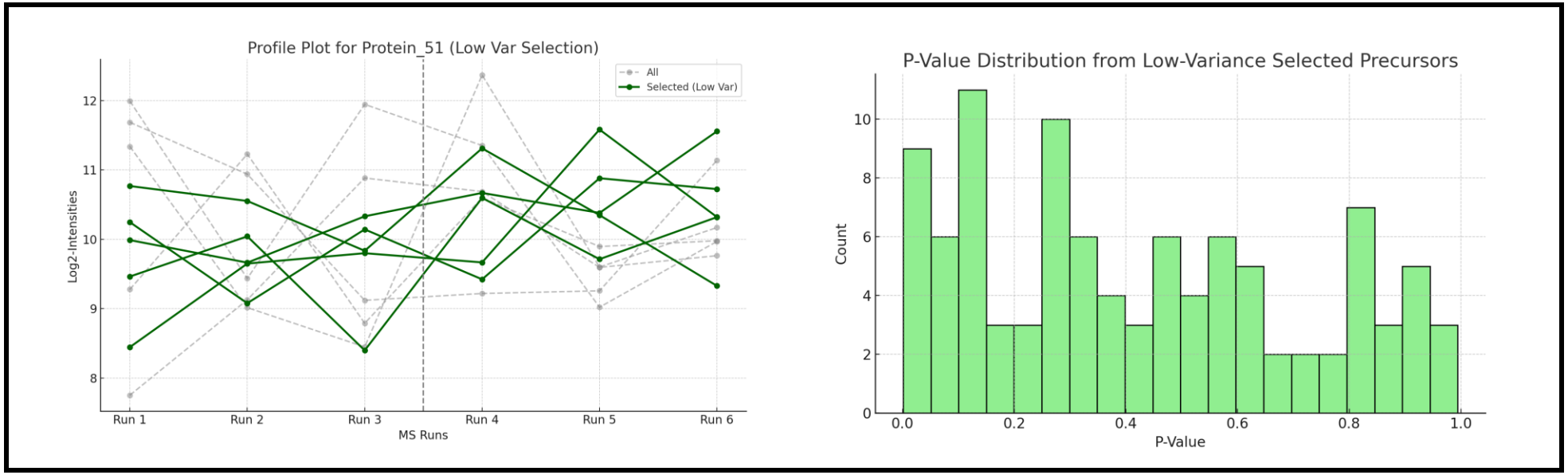
Fischer *et al*. Bioinformatics, 32(7), 1040–1047. (2016).

Intensity-based feature selection can create a double dipping problem

- Double dipping can lead to false positives



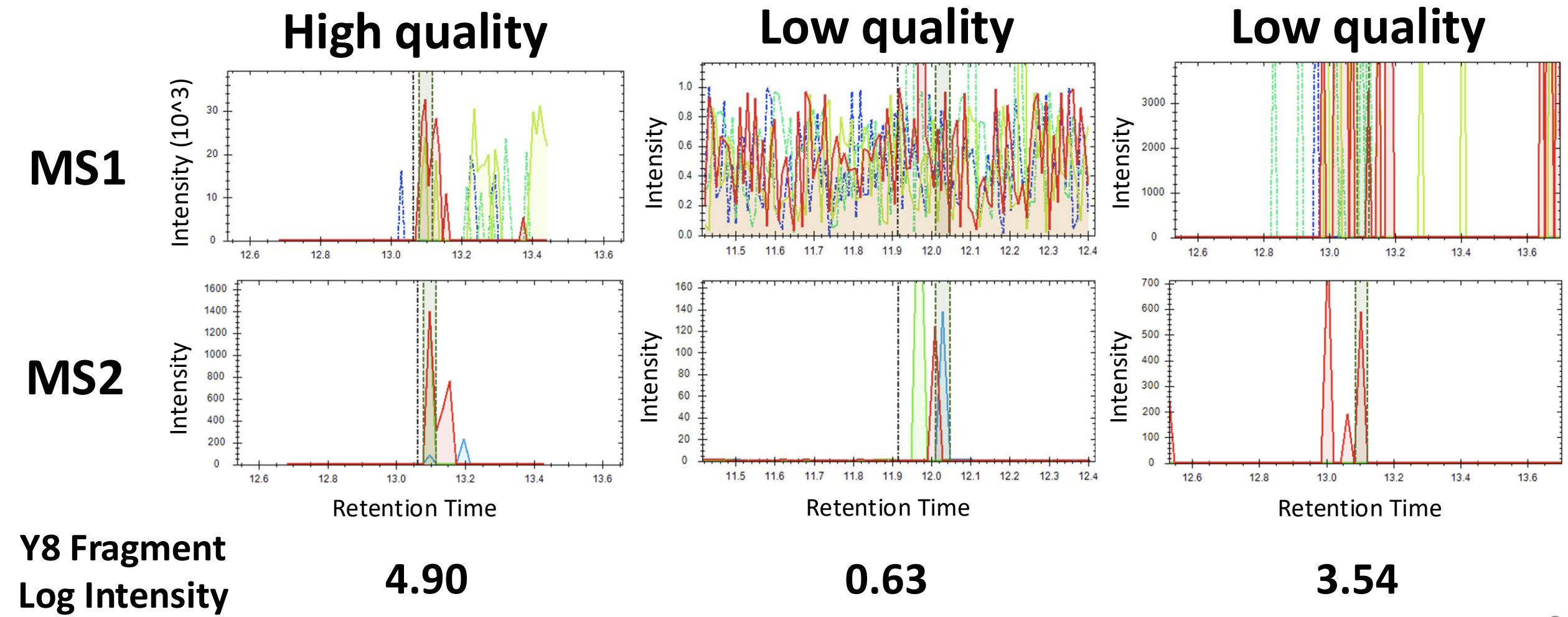
Hivert et al. Computational Statistics & Data Analysis. Vol 193. (2024)



Outline

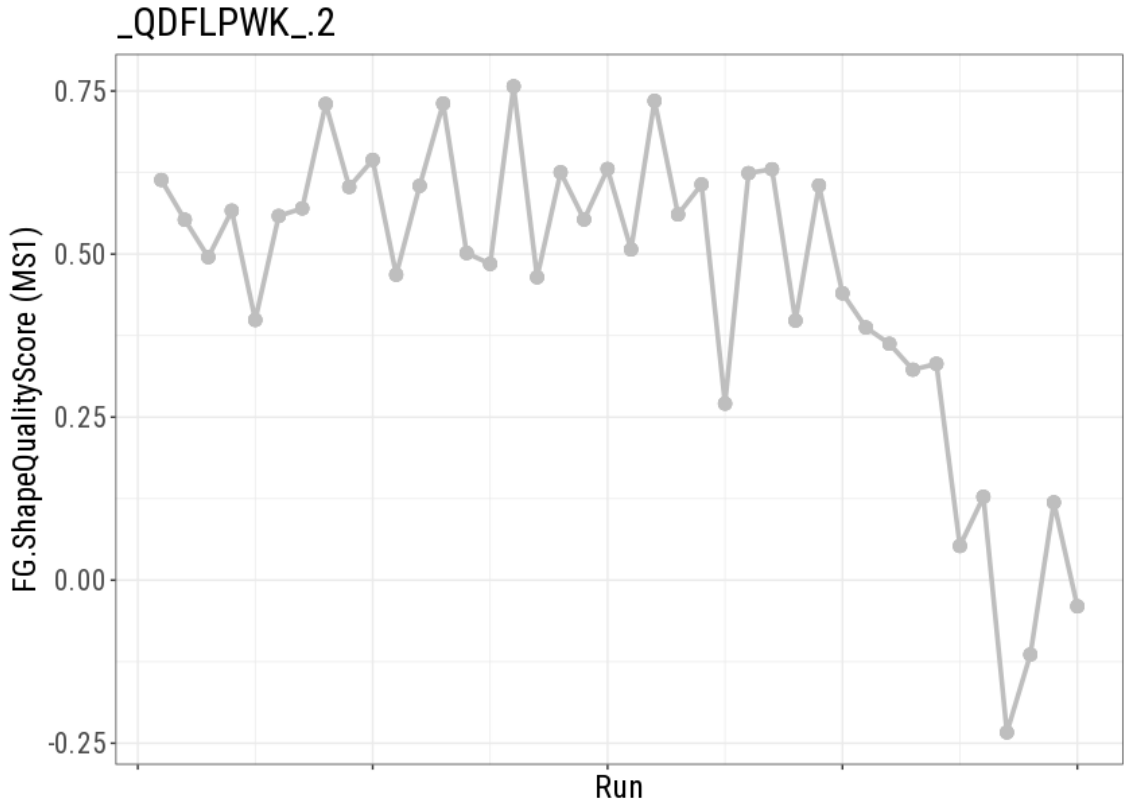
- Problem statement
- Background
  - Existing differential analysis methods
  - Informative quality metrics from spectral processing tools
- Incorporating quality metrics into differential analysis
- Case study and benchmarking

Spectral processing tools provide metrics which are informative of the quantification accuracy



Longitudinal context provides additional insight from quality metrics

- Including time of collection as another dimension can help identify instrumental trends
- Temporal aspect can reveal drift, degradation, or batch effects
- Can correct for instrument performance on a precursor level (as opposed to experiment-wide)

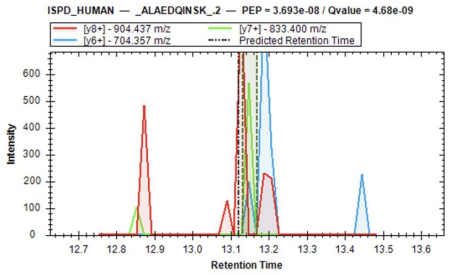


Outline

- Problem statement
- Background
- Incorporating quality metrics into differential analysis
- Case study and benchmarking

Replace feature selection with quality metric weighting

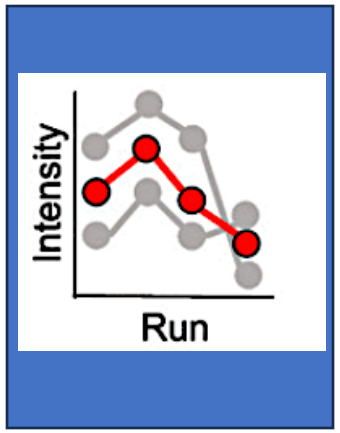
Identified and Quantified Peaks



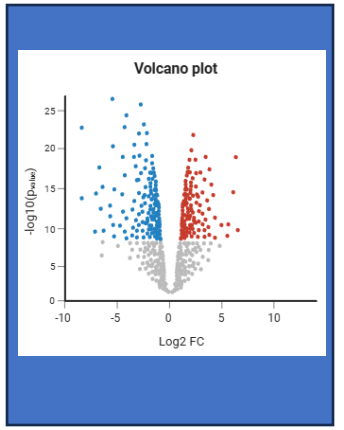
Data filtering

Quality metric weighting

Protein-level summarization

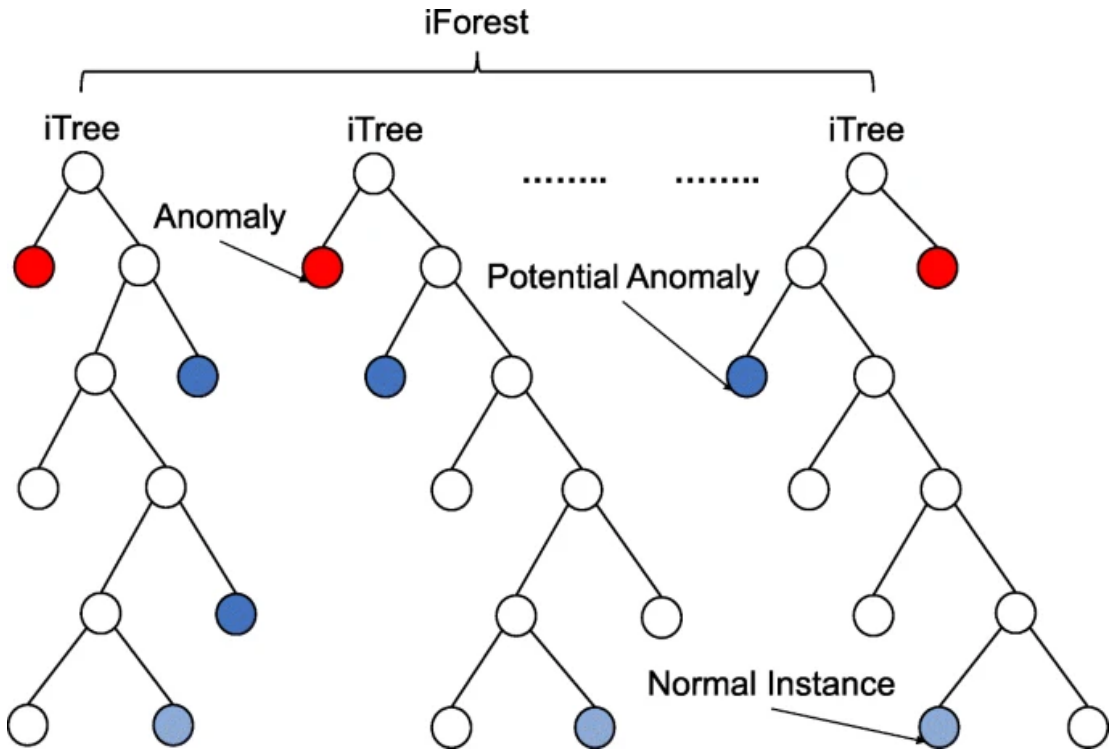


Differential analysis



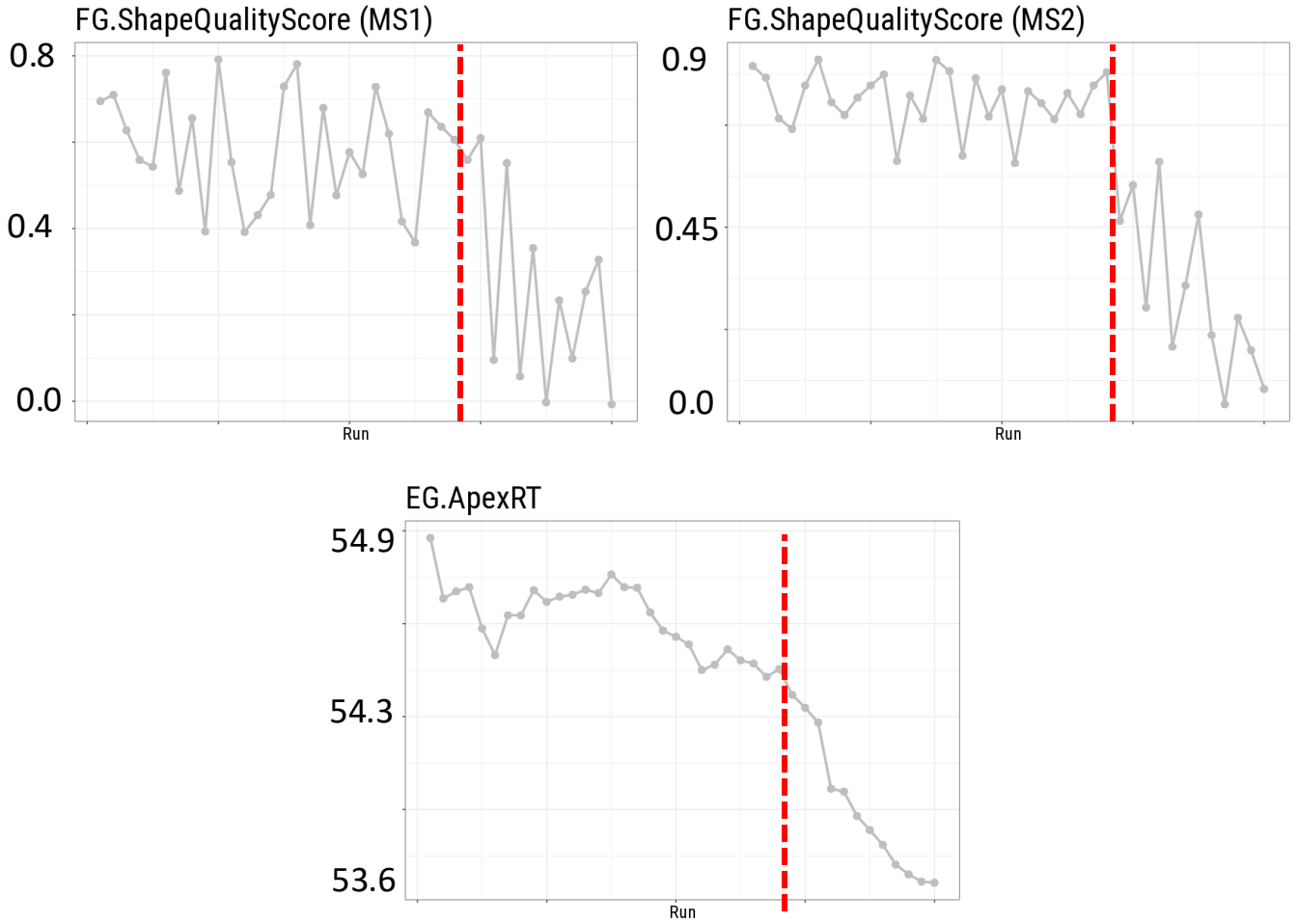
Isolation forest translates quality metrics into informative weights

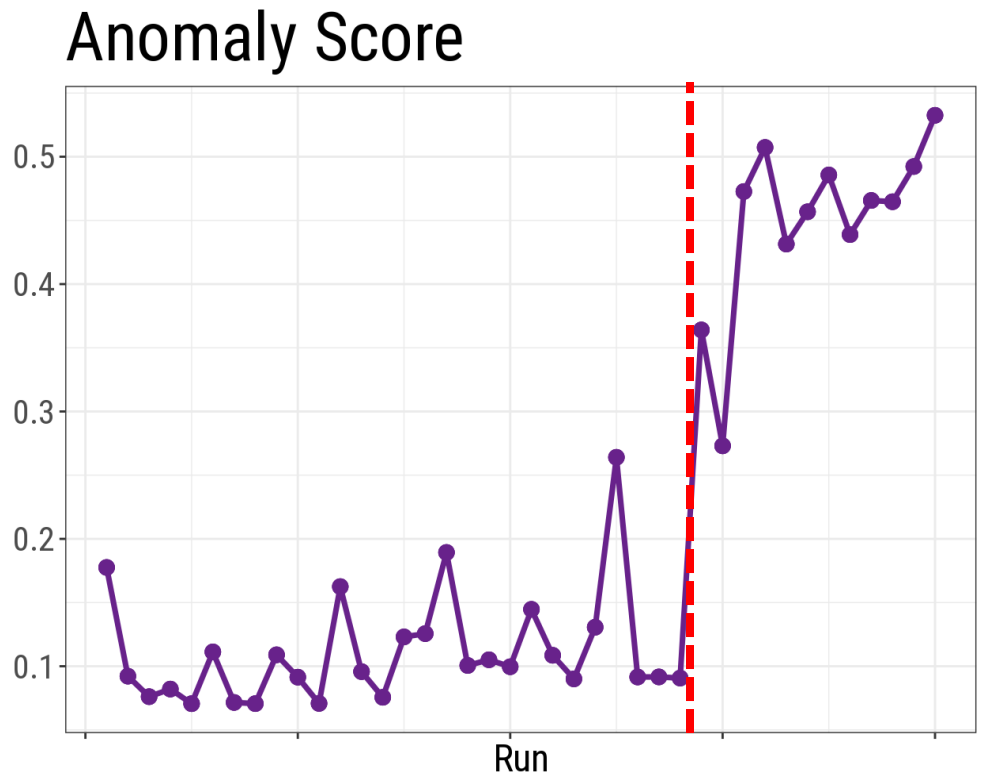
- Unsupervised anomaly detection algorithm
- No labels required and can automatically adapt to new data
- Highly anomalous values are treated as poor quality
- Incorporate longitudinal features via feature engineering



Regaya *et al*. Multimed Tools Appl. 80, 28161–28177 (2021).

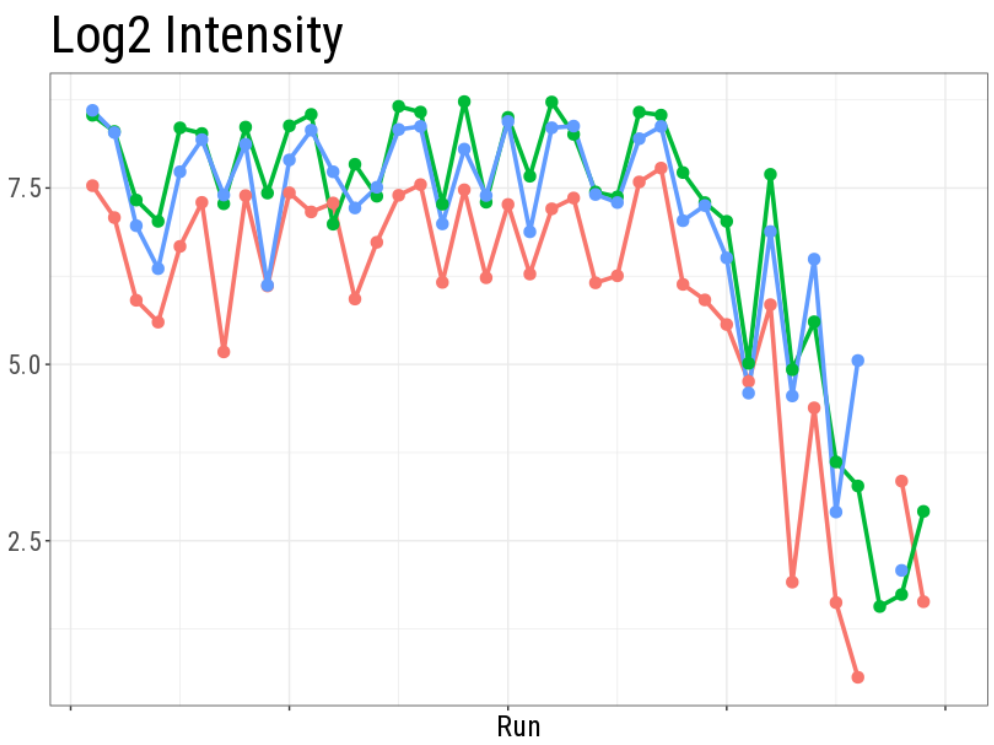
Anomaly scores integrate quality metrics

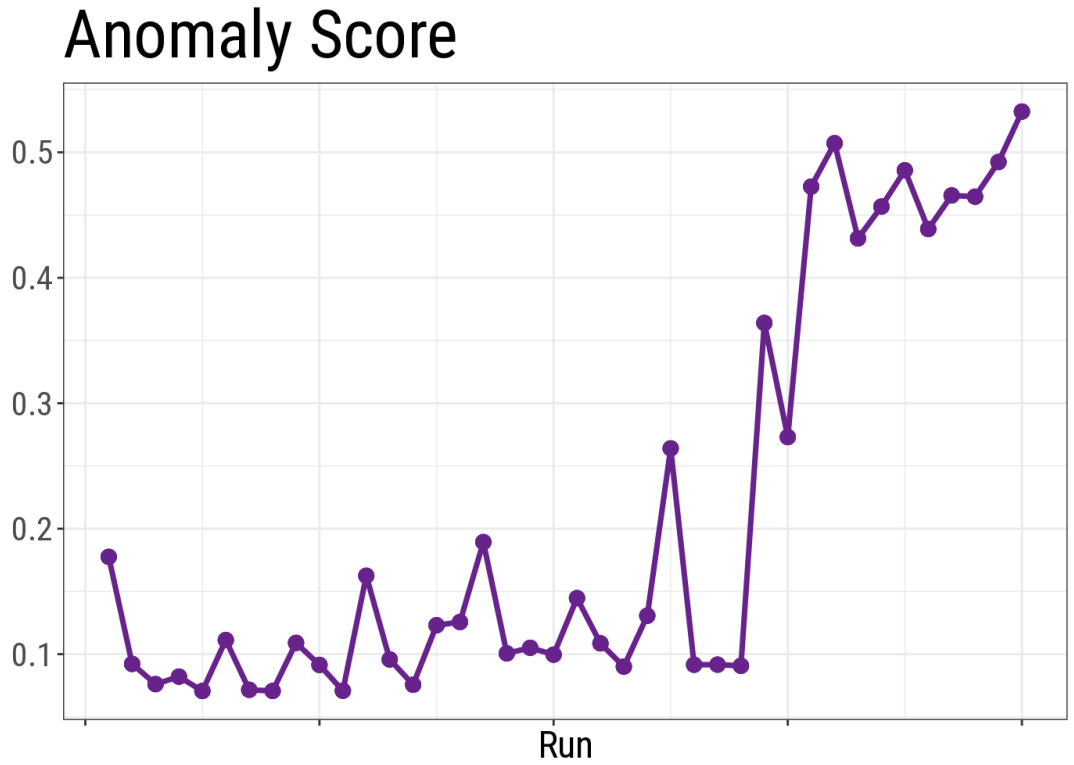




K562 benchmark – ISPD – ALAEDIQINSK 2

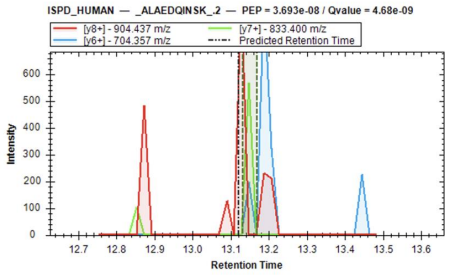
Anomaly model broadly correlates with intensities without ever seeing them





Incorporate quality weights into summarization and differential analysis

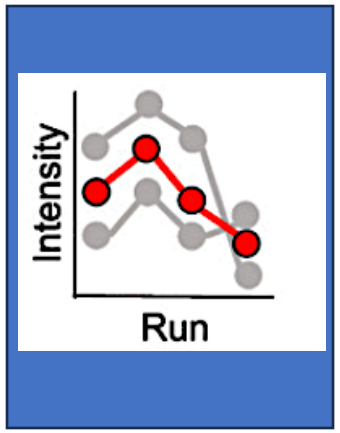
Identified and Quantified Peaks



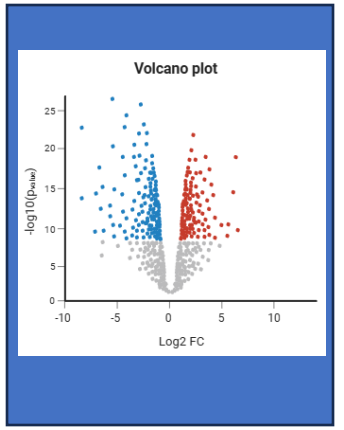
Data filtering

Peak quality weighting

Protein-level summarization



Differential analysis



Protein summarization using weighted least squares with anomaly scores as weights

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl}$$

where $\sum_{i=1}^{I} Run_i = 0 \ and \sum_{j=1}^{J} Feature_j = 0$

$$\epsilon_{ijkl} \sim N(0, \sigma^2)$$

Define weight w_i as the reciprocal of the σ_i^2 in maximum likelihood

$$w_{ijkl} = \frac{1}{a_{ijkl}}$$

Define weighted loss function

$$\sum_{n=1}^{ijkl} w_n \left(y_n - \mu + Run_{ijk} + Feature_l\right)^2$$

Outline

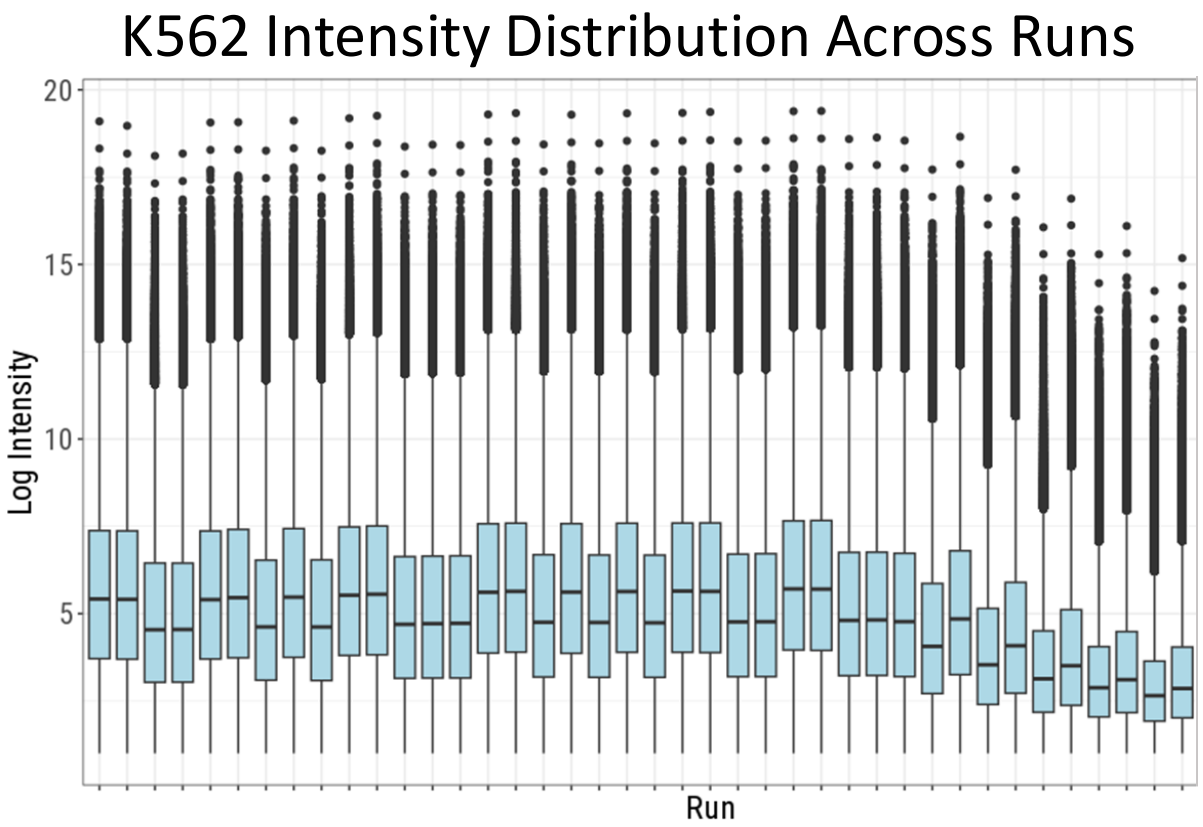
- Problem statement
- Background
- Incorporating quality metrics into differential analysis
- Case study and benchmarking

Benchmarking strategy

- Experimental data
  - K562 + CSF benchmark experiments
  - Biological mixture data
  - Real world clinical study
- Comparison methods
  - Base MSstats
  - msqrob2
  - MaxLFQ + limma
  - DEqMS

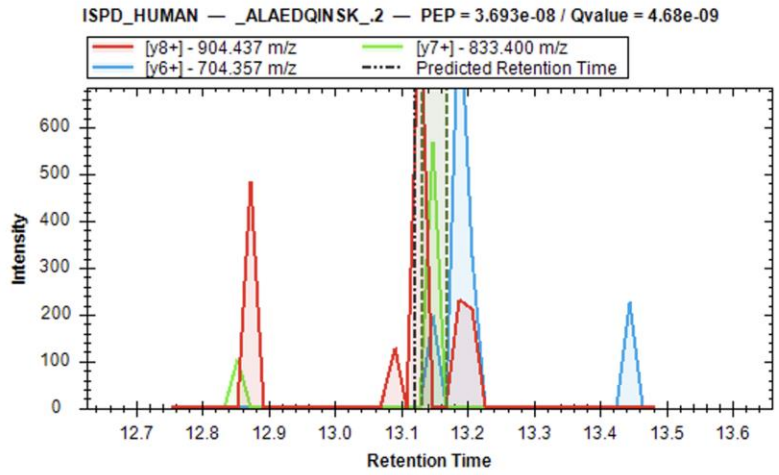
Design of K562 and CSF experiments

- Two experiments using K562 cell lines and CSF samples
- Two conditions with one $\log_2$ fold change difference
- First 30 runs show consistent, high-quality measurements
- Last 10 runs drift to lower intensities

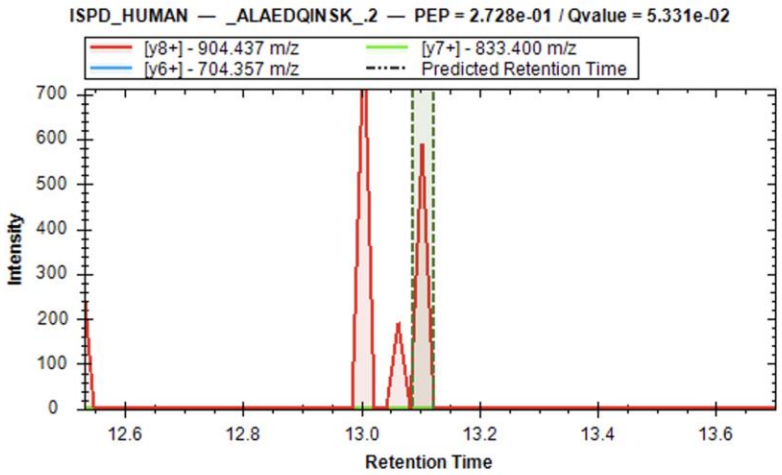


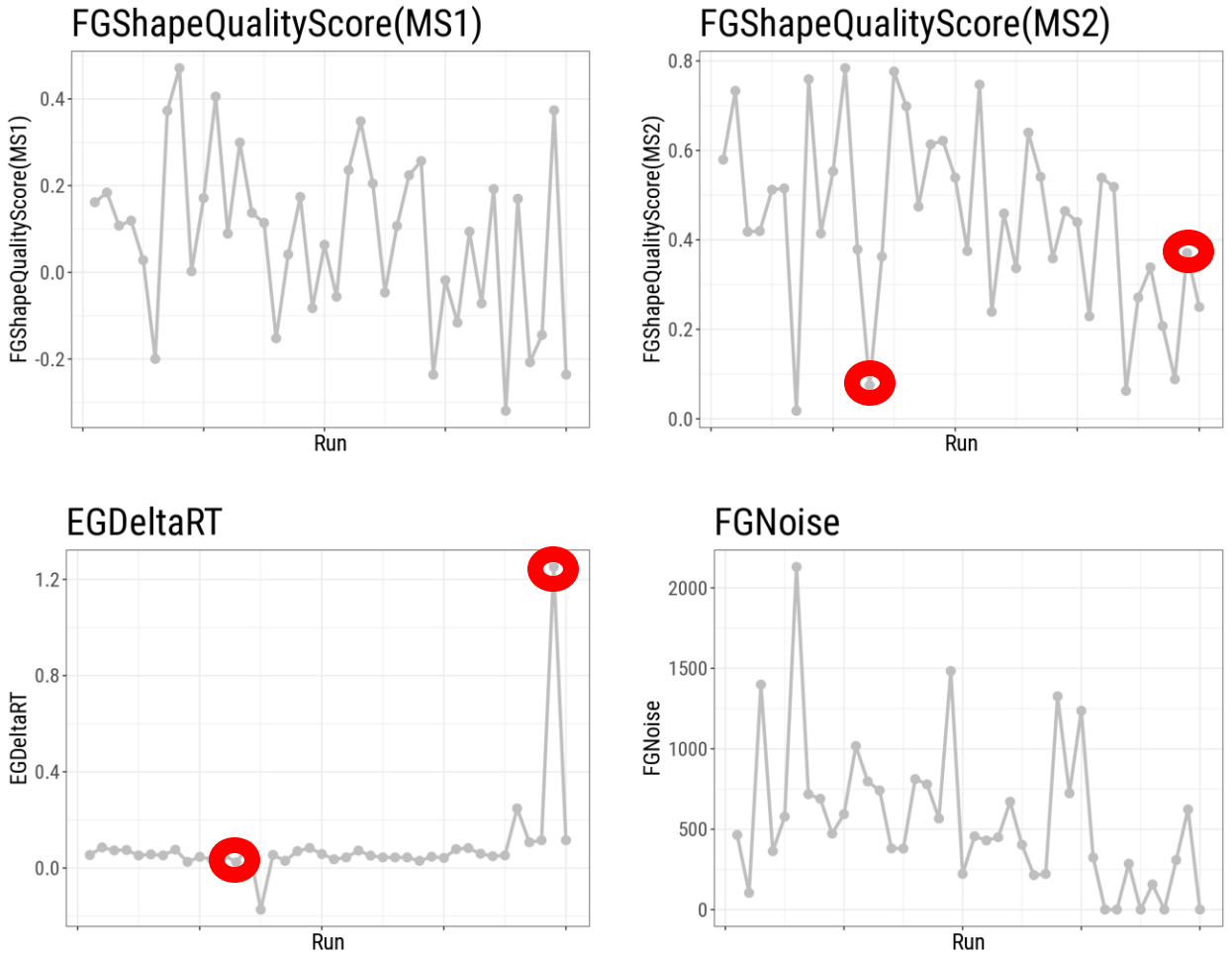
K562 experiment case study – A4D126

Run 13

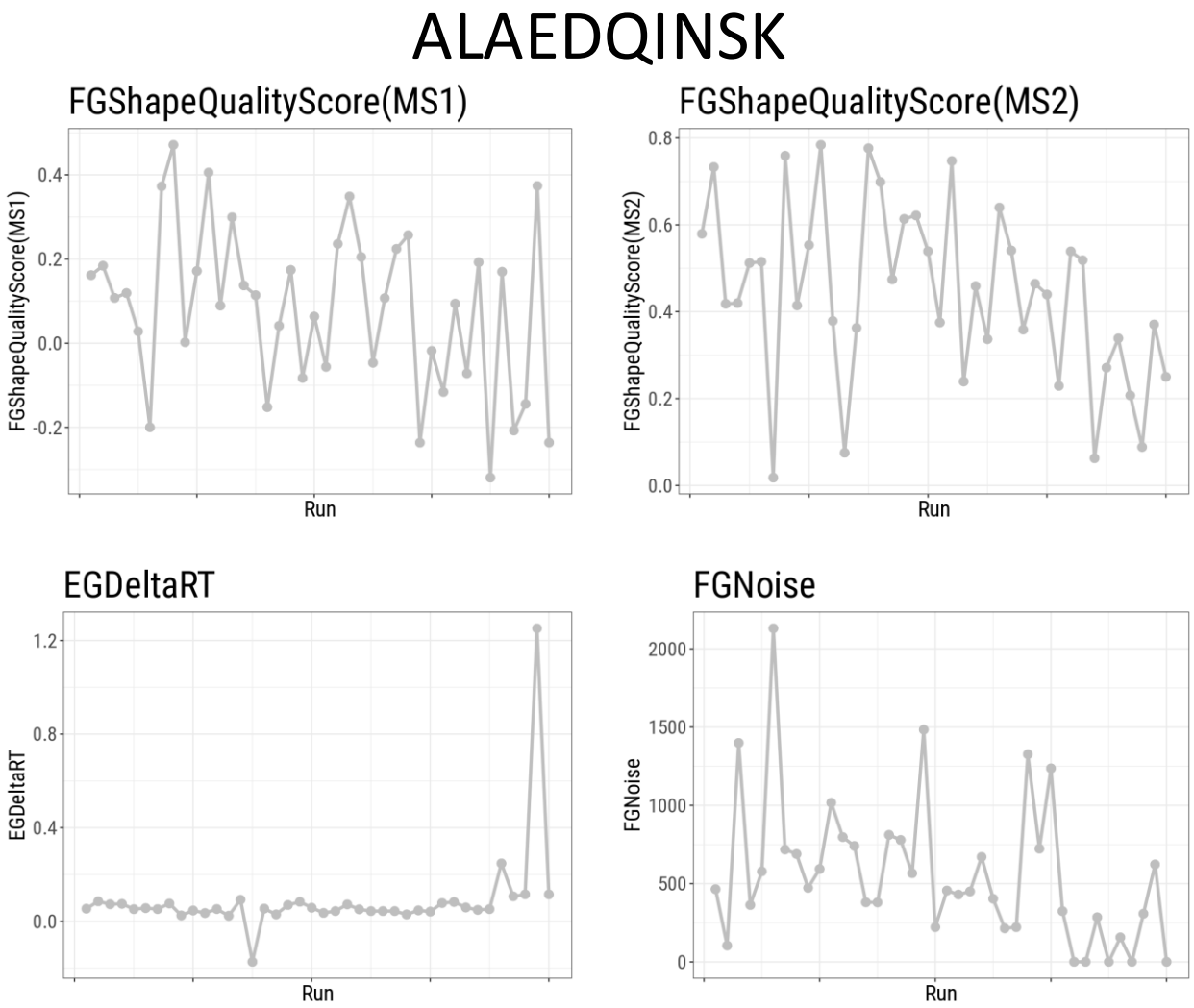


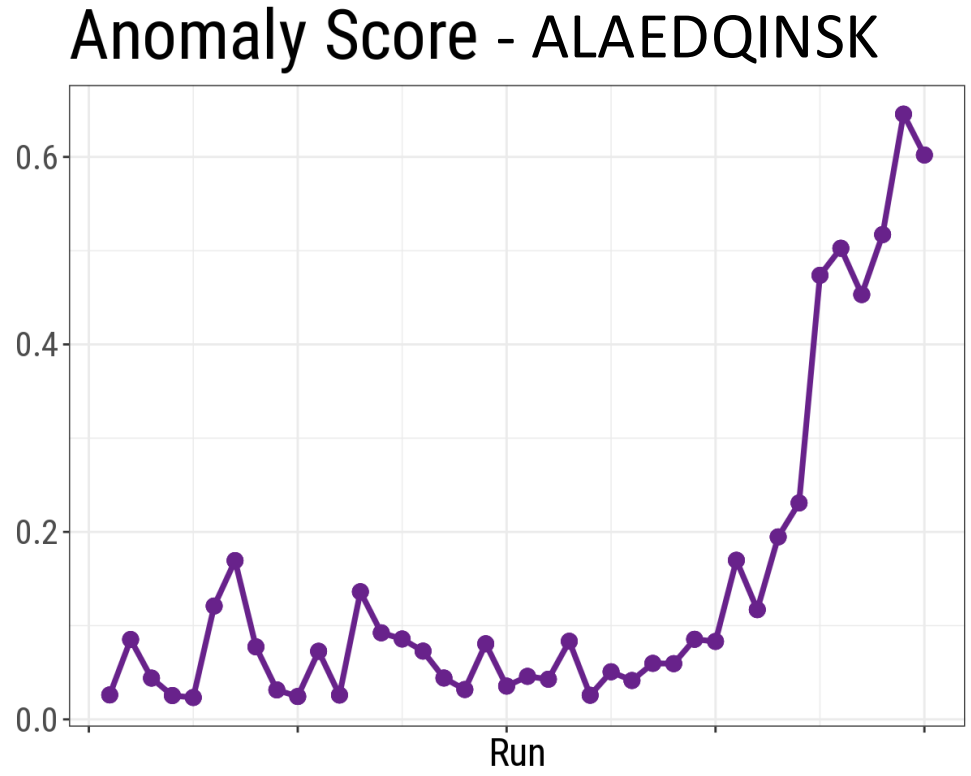
Run 39



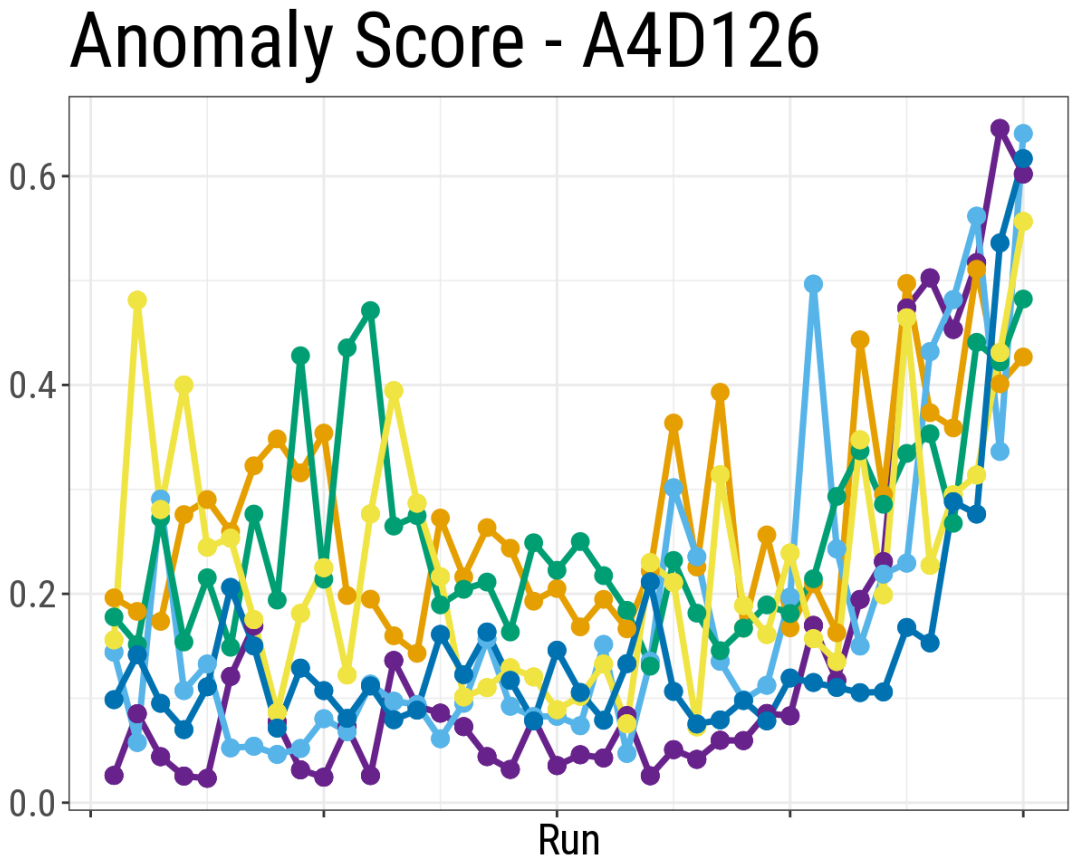


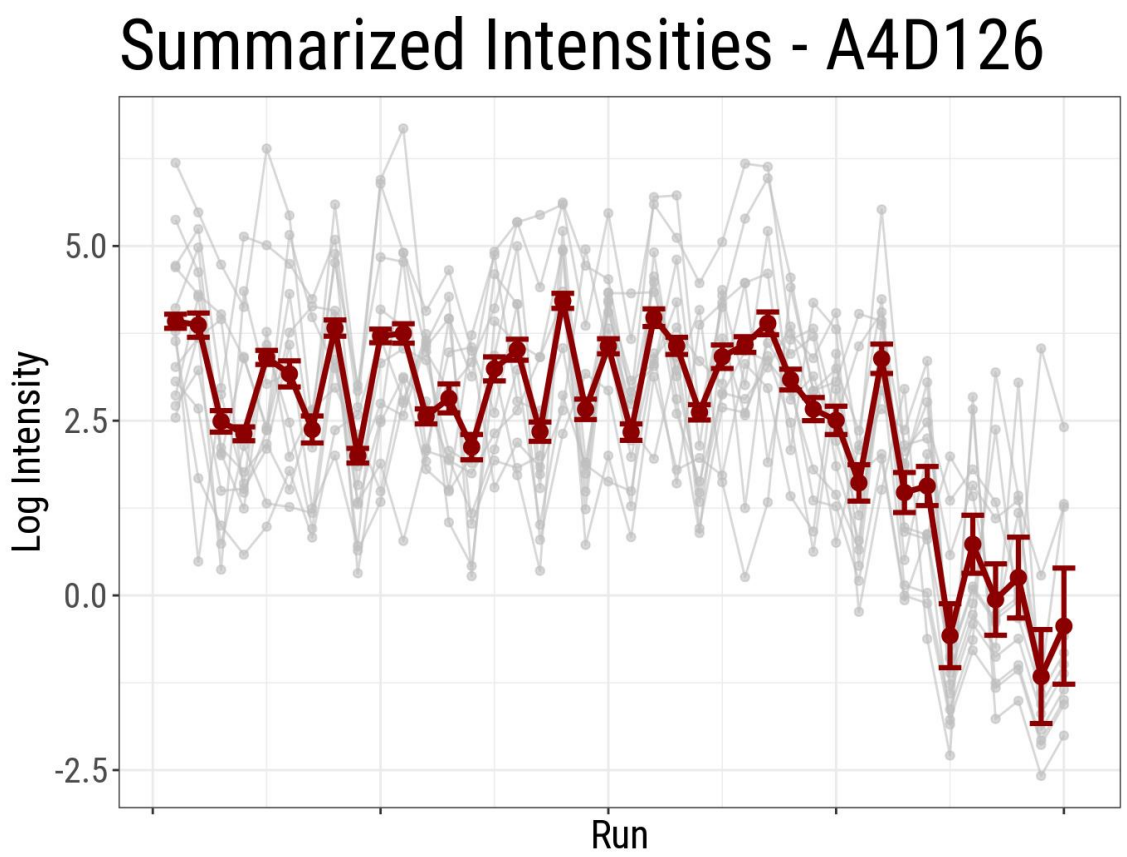
Isolation forest transforms quality metrics into anomaly scores





Anomaly scores calculated across all precursors and used in weighted summarization





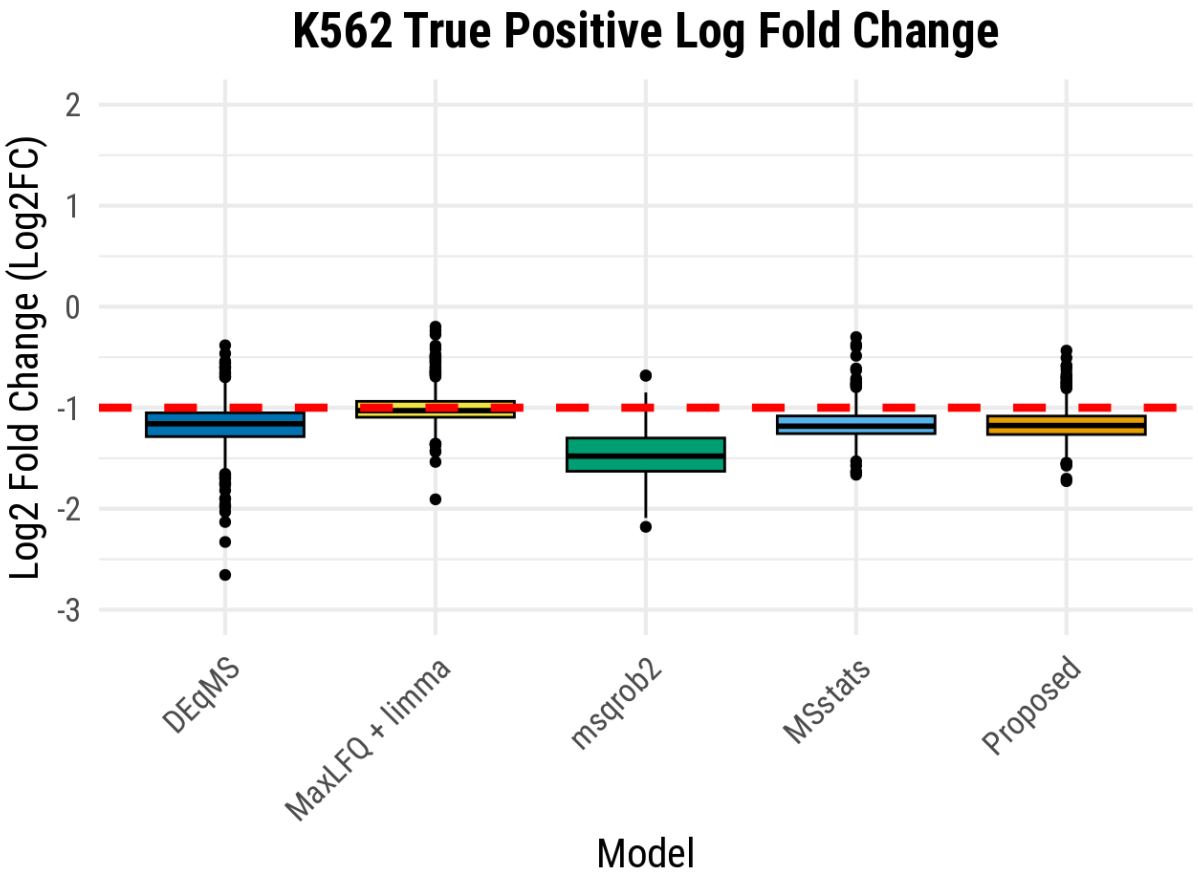
Quality weighted differential analysis reduces standard error

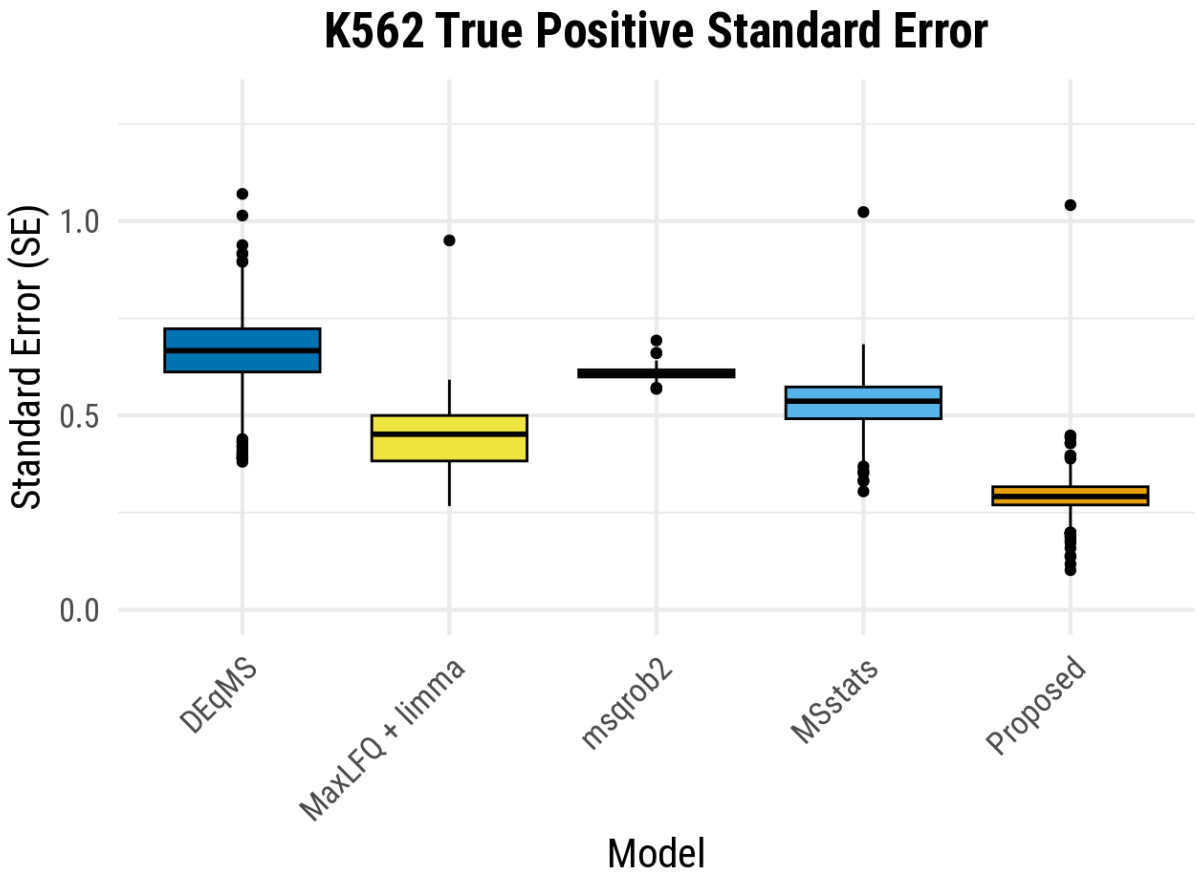
K562 Benchmark - A4D126

Model	Log Fold Change*	Standard Error	Adj P-value
Proposed	-1.17	.25	.002
MSstats	-1.16	.45	.680

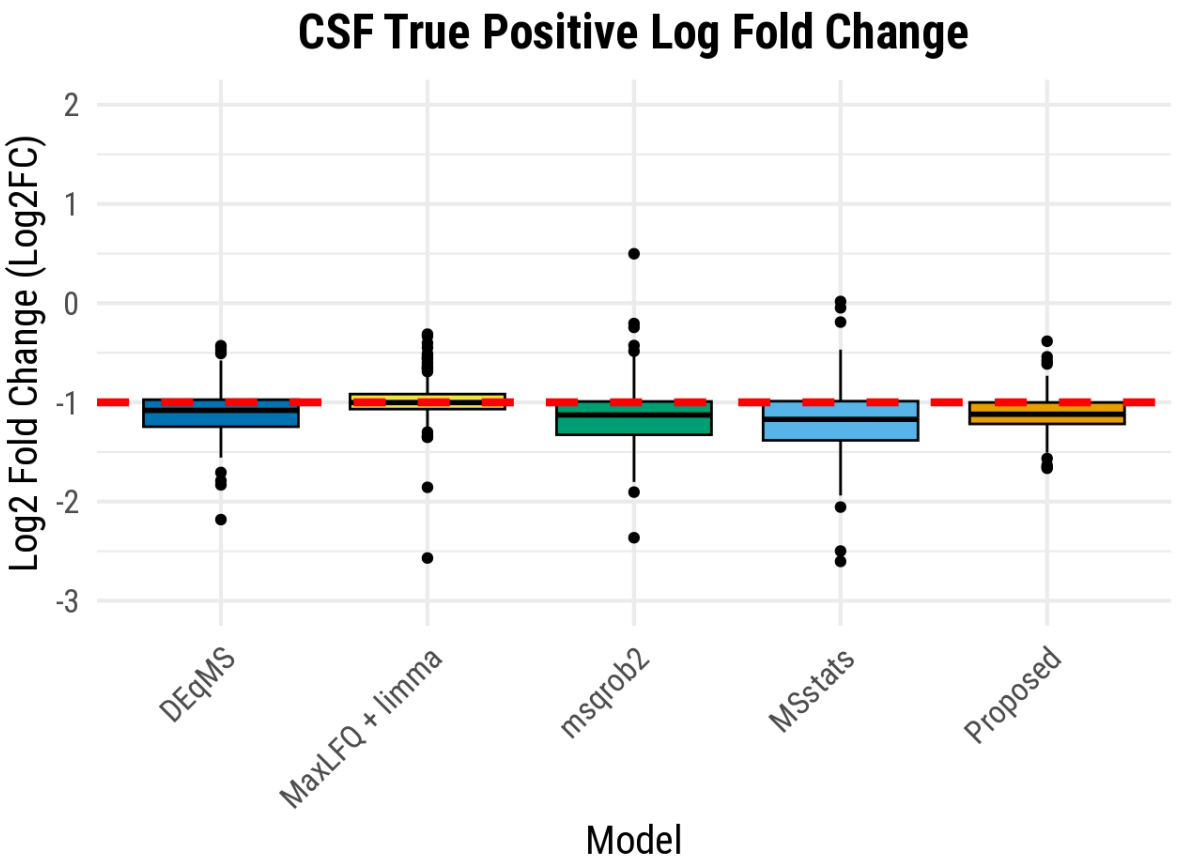
*** True $\log_2$ fold change = -1**

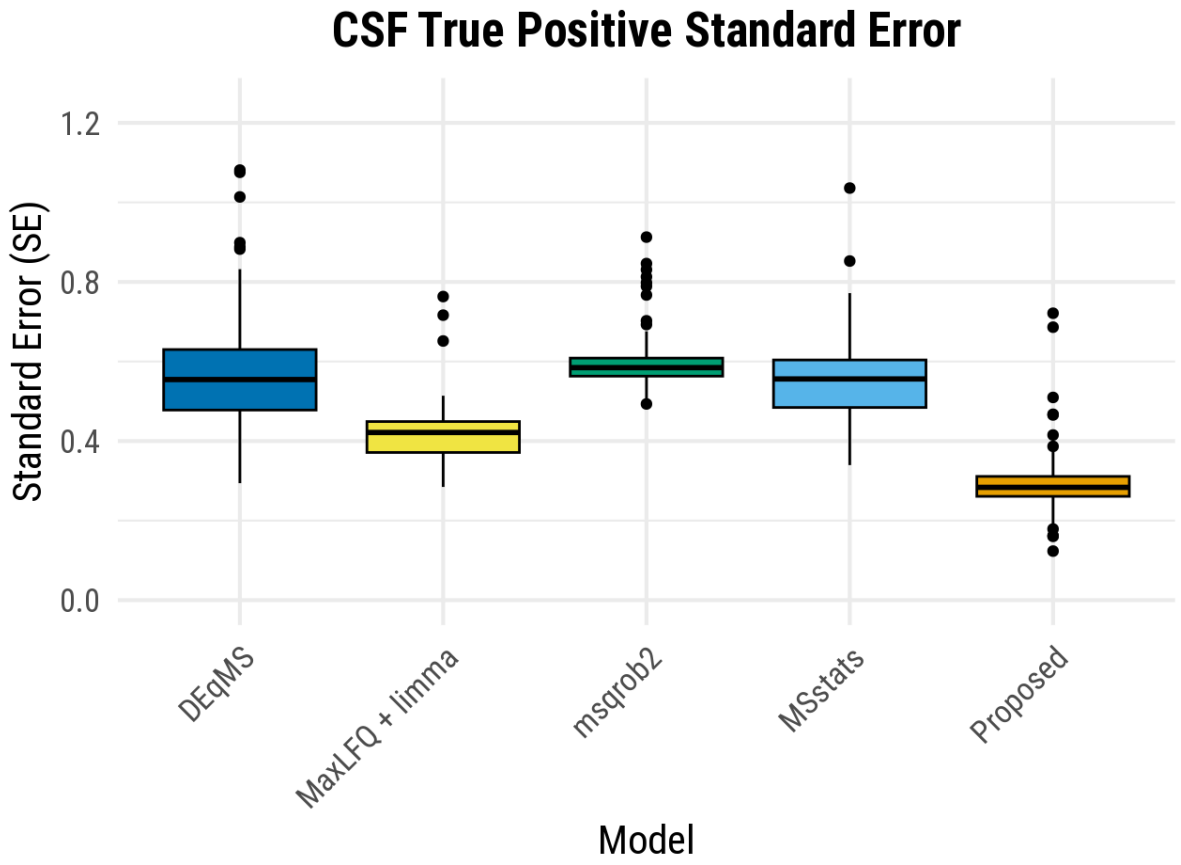
K562 Benchmark





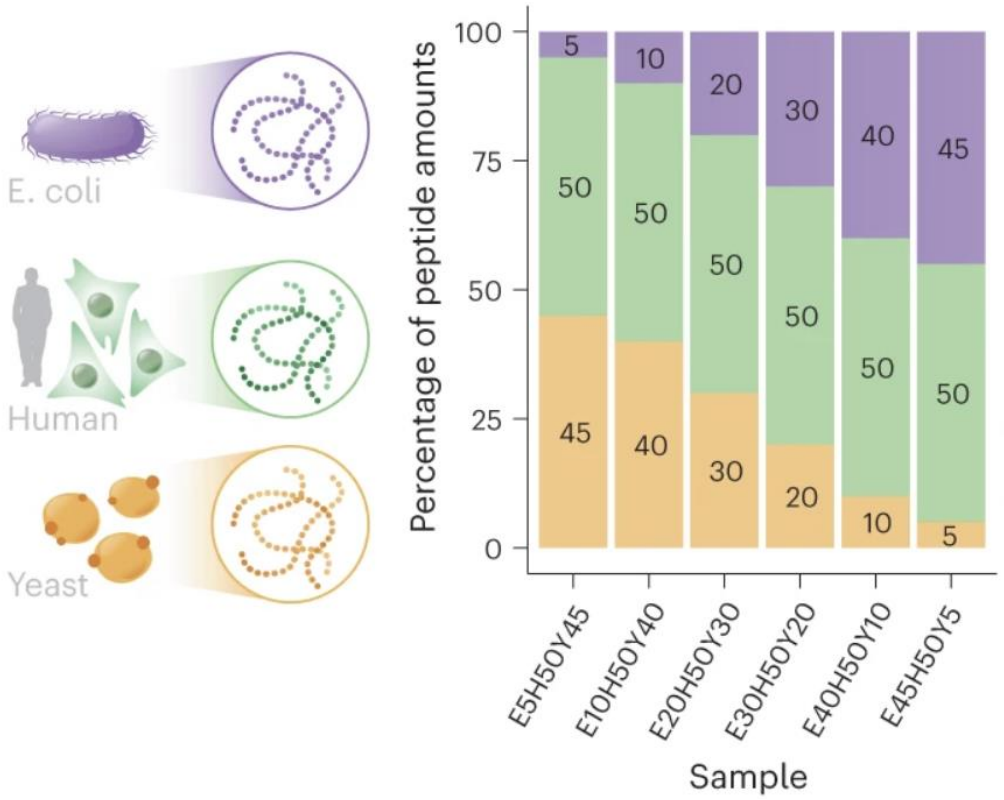
CSF Benchmark





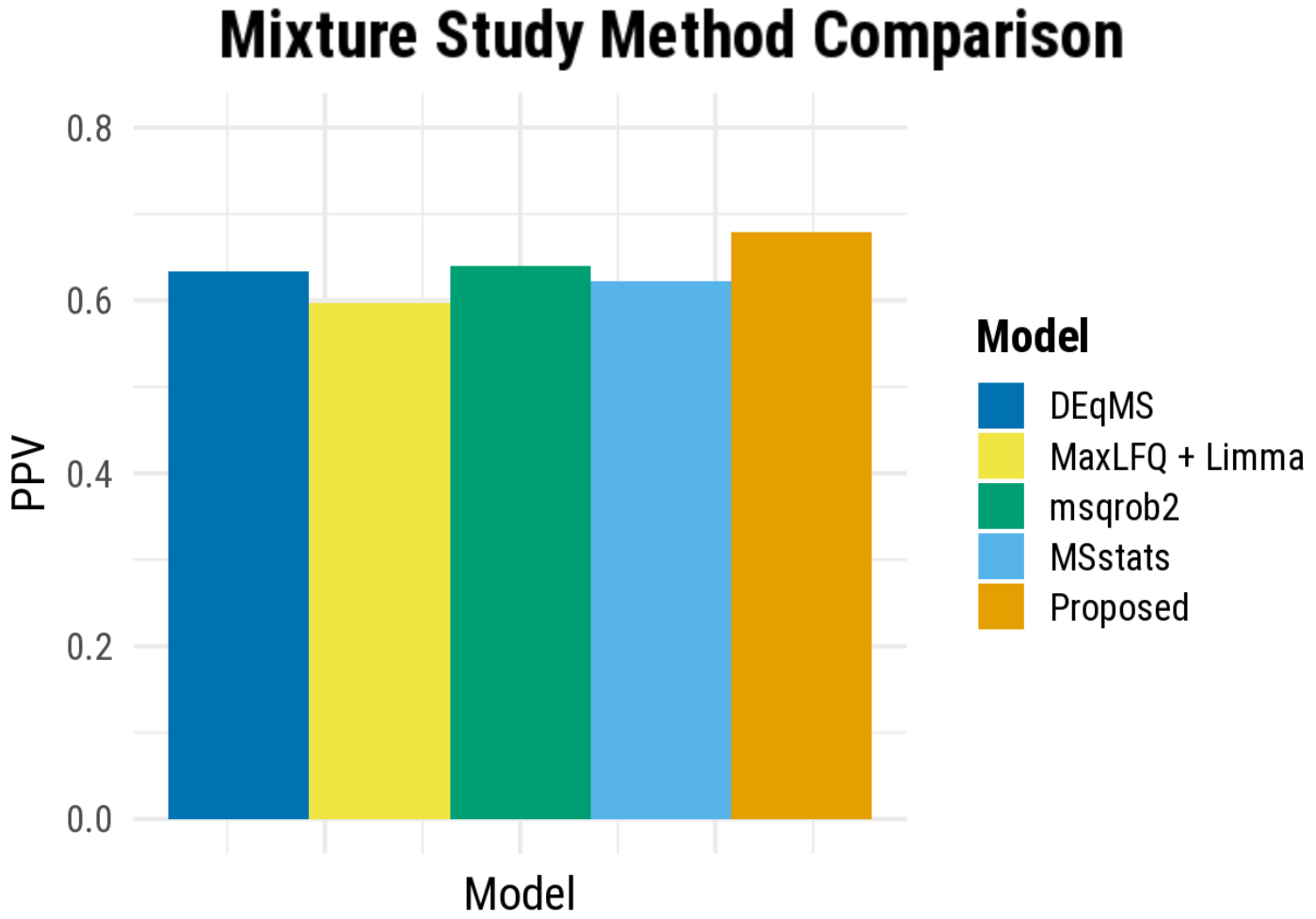
Mixture data experimental design

- Human, Yeast, and E. coli mixed at 6 different concentrations
- Measured on an Orbitrap Astral and acquired with DIA
- ~12,000 proteins measured across all organisms



Guzman *et al*. Nat Biotechnol 42, 1855–1866 (2024).

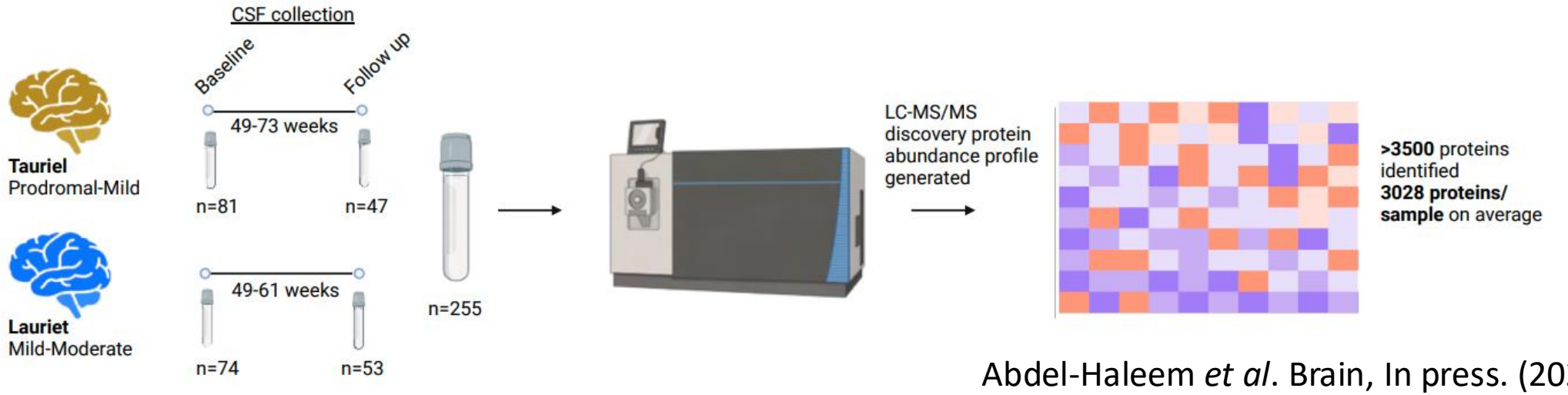
The proposed approach maintained performance even without many low quality quantifications



Best controlled FDR while providing similar power

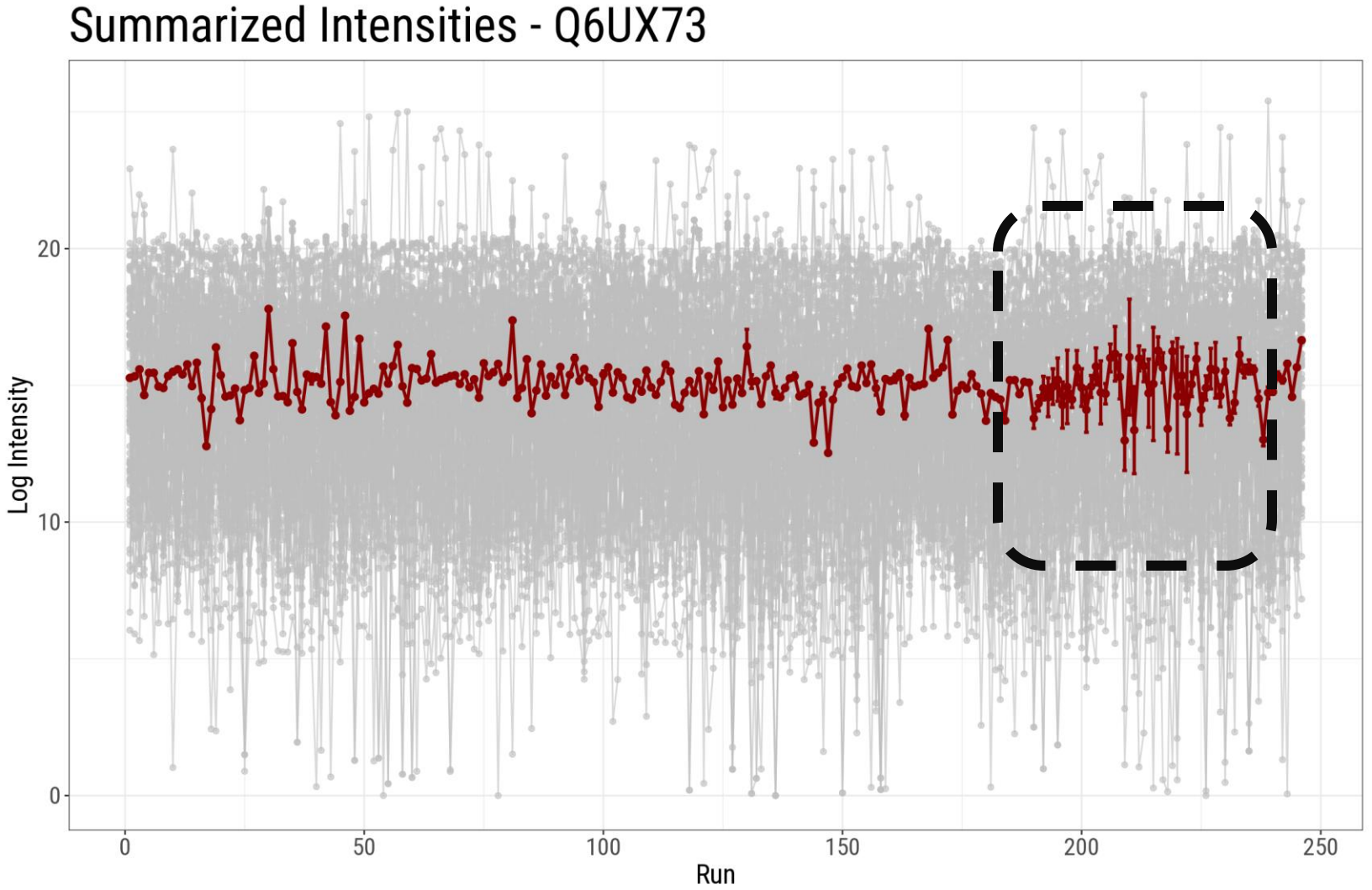
CSF analysis of semorinemab Ph2 trials in Alzheimer's disease

- Large cerebrospinal fluid clinical proteomics dataset studying Alzheimer's disease
- More than 250 CSF samples
- ~3500 proteins measured (random 1000 protein subset for analysis)
- Acquired with DIA



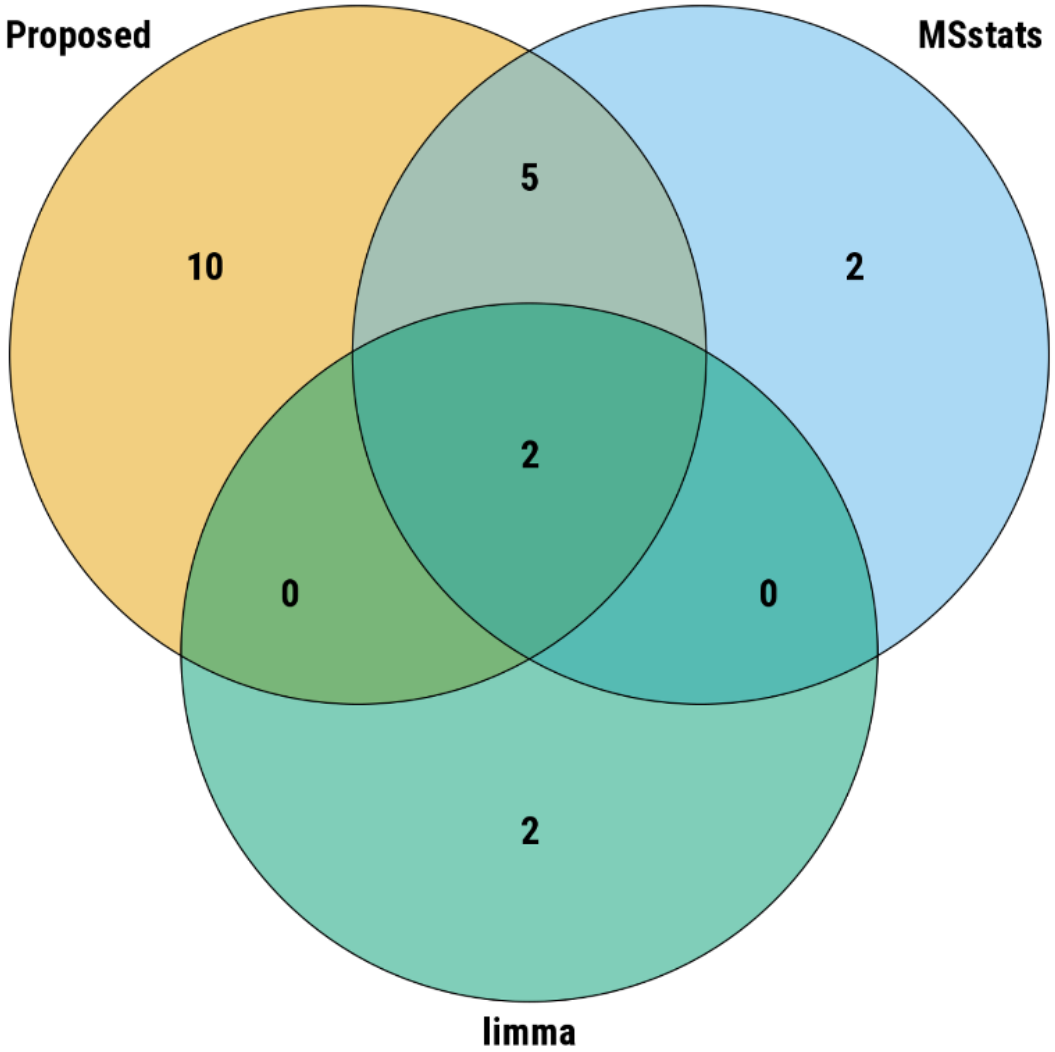
Abdel-Haleem *et al*. Brain, In press. (2025).

Clear area of poorly quantified runs near end of collection



The proposed approach identified more differential proteins compared to existing methods

- Subset subjects into two groups – Low and High Clinical Dementia Rating (CDRSB)
- Test for differentially abundant proteins between CDRSB groups



Conclusions

- Peak quality model automatically detects poorly quantified measurements without relying on double-dipping strategies
- Shown to increase power in highly variable datasets while broadly reducing FDR
- Beta implementation in MSstats and preprint out shortly

Acknowledgements

Northeastern University

OLGA VITEK LAB

Statistical Methods For Studies Of Biomolecular Systems

Northeastern

Olga Vitek
Sarah Szvetecz
Tony Wu
Anshuman Raina

Genentech

Veronica Anania
Mrittika Bhattacharya
Ozge Karayel Eren
Manuel Magana

Mugla Sitki Kocman University

Eralp Dogu

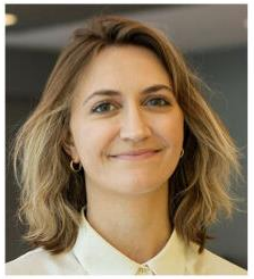
University of Wrocław

Mateusz Staniak

Monday poster
Yinyue Zhu *et al*
MP 419: TIMSImaging: a Python package for trapped-ion mobility spectrometry imaging processing

Monday poster
Ethan Rogers *et al*
MP 417: Statistical principles define an open-source analysis workflow for MSI with complex designs

Monday poster
Sai Lakkimsetty *et al*
MP 422: Teadrop: Unsupervised co-registration of H&E and MSI experiments with neural networks



Wednesday poster
Sarah Szvetecz *et al*
WP 319: Semi-parametric models improve detection of drug-protein interactions in chemoproteomics