Northeastern University
Khoury College of Computer Sciences

Bayesian statistical modeling reveals missing value mechanisms in label-free Mass Spectrometry-based proteomics experiments

Devon Kohler and Olga Vitek

Khoury College of Computer Sciences

Northeastern University, Boston, MA

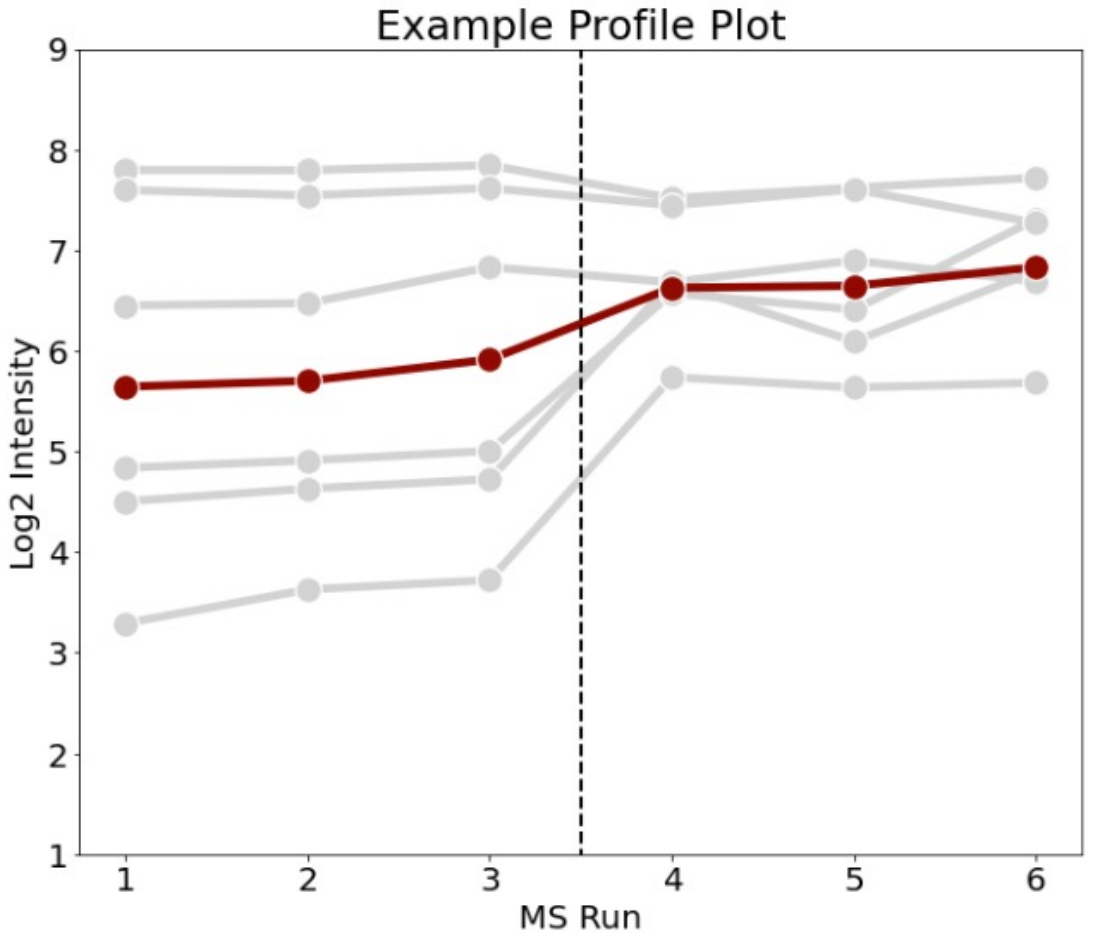
The authors declare no conflicts of interest

Presentation Outline

- Problem statement
- Background
  - MSstats model
  - Missing value imputation
  - Summarization
- Proposed method
- Evaluation
  - Missing mechanisms
  - Imputation and summarization
  - Differential analysis

Accurate relative quantification of proteins in MS-based proteomics is challenging and requires specialized techniques

- Challenges
  - Indirect measurements (peptide fragments)
  - Missing peptides across samples
- Analytical/statistical techniques
  - Normalization
  - Imputation of missing peptide intensities
  - Summarization
- Feature definition
  - precursor ions in DDA
  - transitions in SRM
  - fragments in DIA



Presentation Outline

- Problem statement
- Background
  - MSstats model
  - Missing value imputation
  - Summarization
- Proposed method
- Evaluation
  - Missing mechanisms
  - Imputation and summarization
  - Differential analysis

Experimental design and MSstats model for one protein

Subplot	Whole plot: $Condition_1$				...	$Condition_I$			
	$BioReplicate_1$	$BioReplicate_2$	...	$BioReplicate_J$	...	$BioReplicate_{(I-1)J+1}$	$BioReplicate_{(I-1)J+2}$	...	$BioReplicate_{IJ}$
$Feature_1$	y	y	...	y	...	y	y	...	y
$Feature_2$	**NA**	y	...	y	...	y	**NA**	...	y
...	...	...	...	...	...	...	...	...	...
$Feature_L$	y	**NA**	...	y	...	y	y	...	y

Whole plot | Subplot

$$y_{ijl} = \mu + Condition_i + BioReplicate(Condition)_{j(i)} + Feature_l + BioReplicate * Feature_{ijl}$$

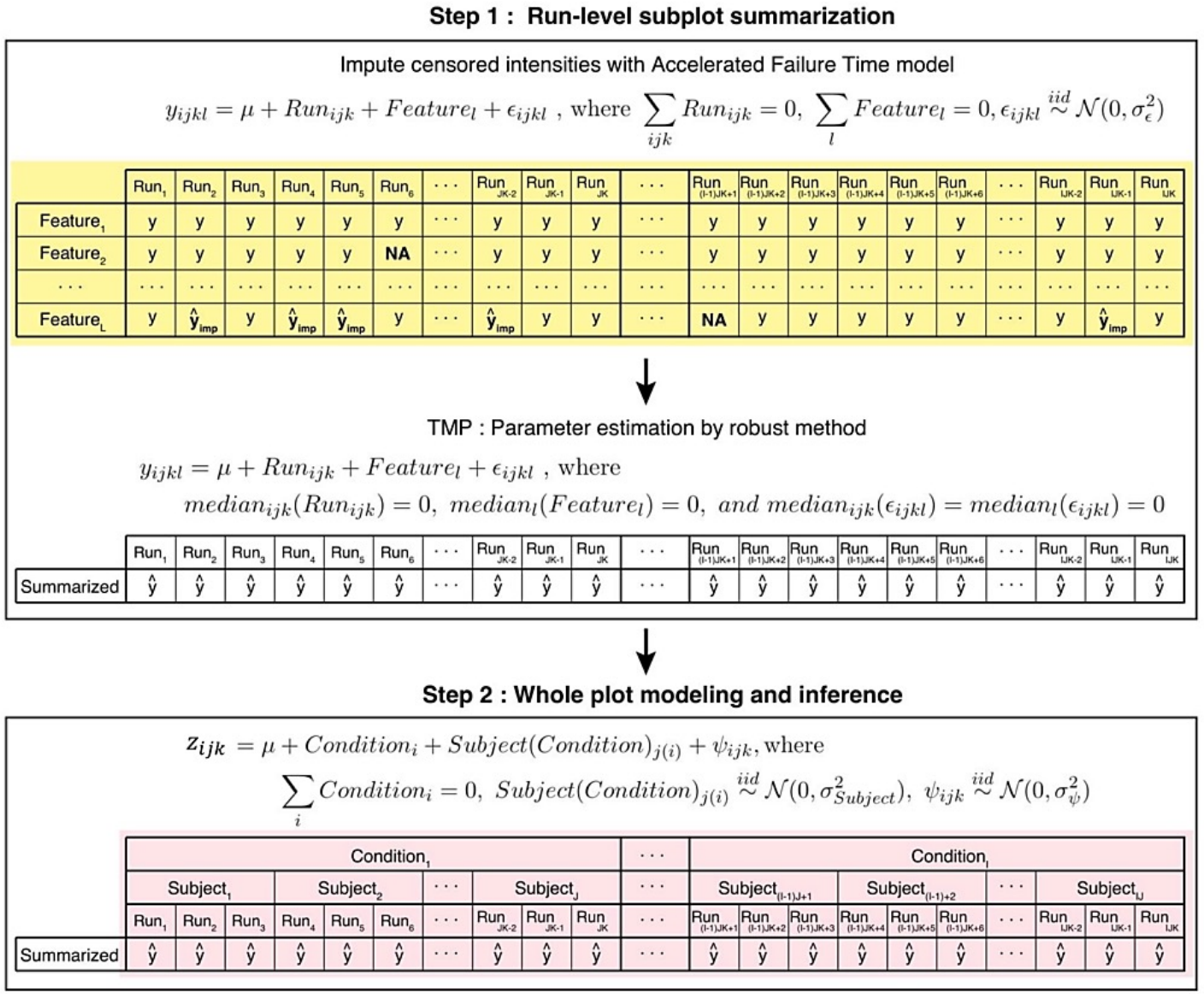
$$where \quad \sum_{i=1}^{I} Condition_i = 0, \sum_{l=1}^{L} Feature_l = 0$$

$$BioReplicate(Condition)_{j(i)} \sim^{iid} N(0, \sigma^2_{BioReplicate})$$

$$Run * Feature_{ijl} = \epsilon_{ijl} \sim^{iid} N(0, \sigma^2_{\epsilon})$$

Kohler et al. Journal of Proteome Research. 2023.

Two-step model in MSstats



Protein-level quantification

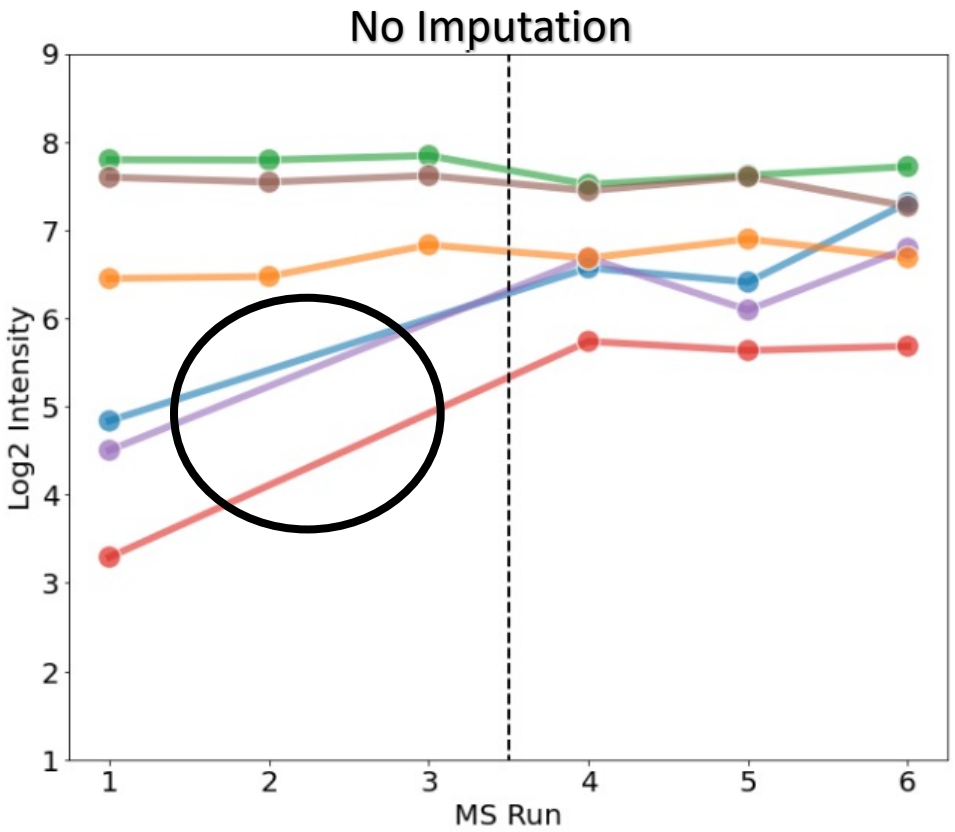
Differential Analysis

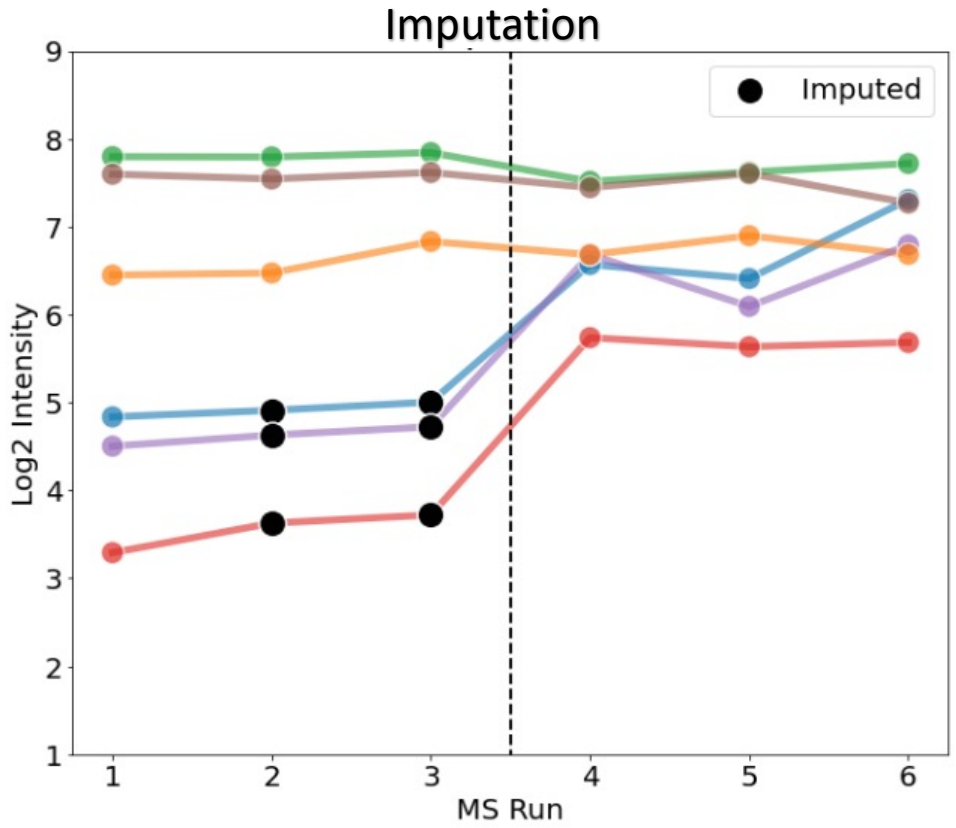
Kohler et al. Journal of Proteome Research. 2023.

Presentation Outline

- Problem statement
- Background
  - MSstats model
  - Missing value imputation
  - Summarization
- Proposed method
- Evaluation
  - Missing mechanisms
  - Imputation and summarization
  - Differential analysis

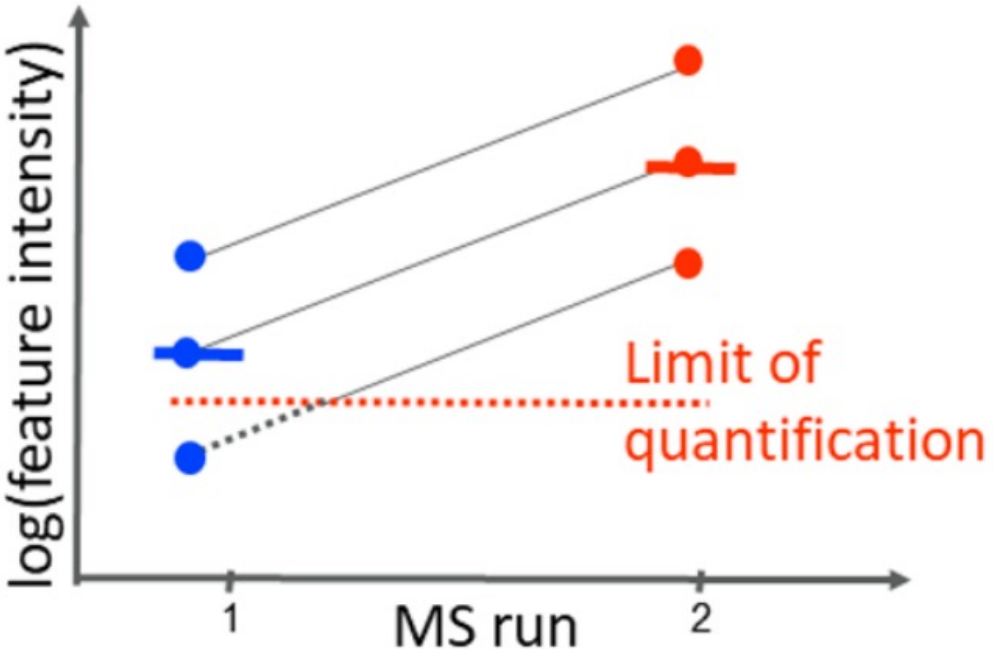
Missing values can have a major impact on protein-level inference





Missing value imputation can create bias, data homogeneity, and are overly confident

- Can create bias if the imputation method is incorrect
  - Missing completely at random (MCAR)
  - Missing not at random (MNAR)
- Causes data to be more homogenous
- Classic imputation returns a point estimate, removing the uncertainty in the estimation process



Kohler et al. Journal of Proteome Research. 2023.

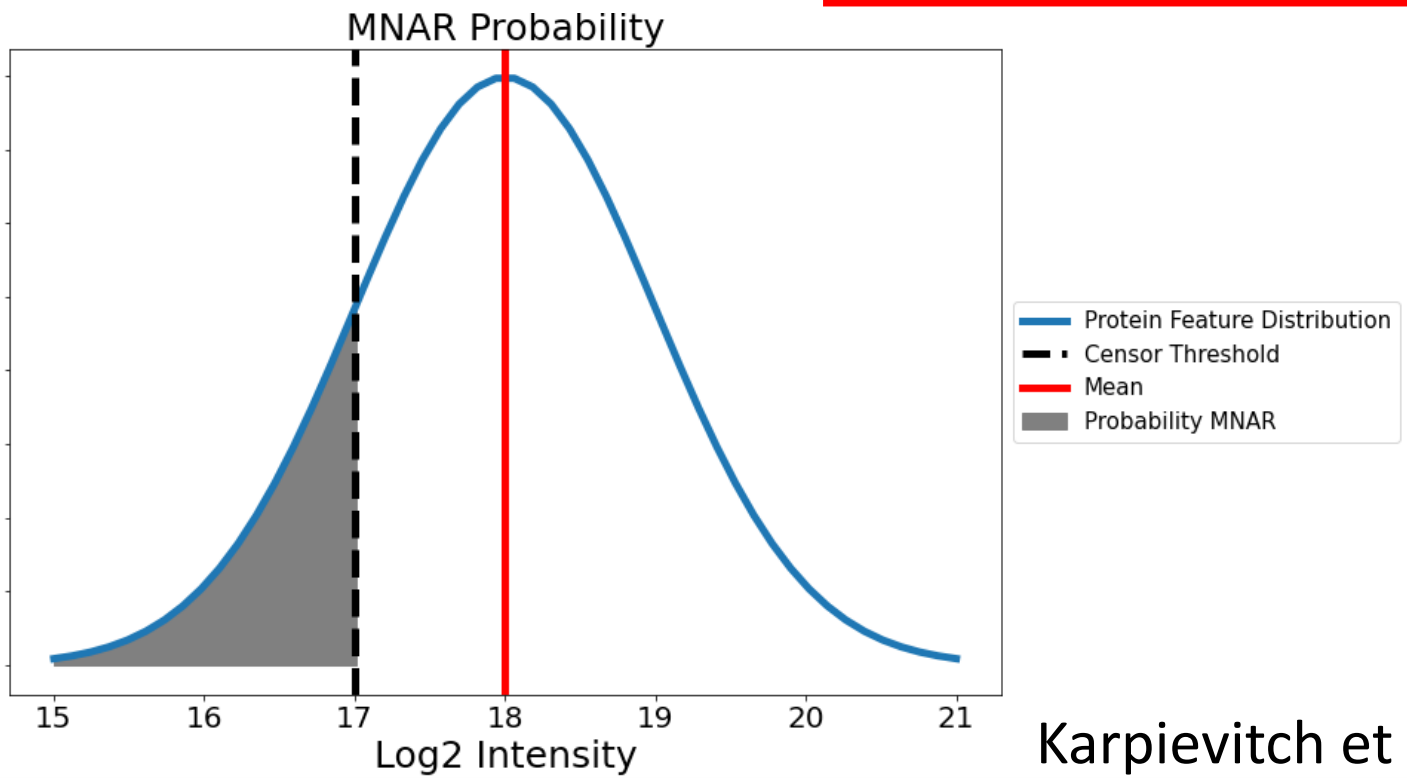
Build missing mechanisms into model

$$P(W_{ijkl}=1)=1-(1-\pi_l)\left(1-\Phi\left((c_{ij}-\mu_{ijk})/\sigma_{ij}\right)\right)$$

$$=\pi_l+(1-\pi_l)\Phi\left((c_{ij}-\mu_{ijk})/\sigma_{ij}\right)$$

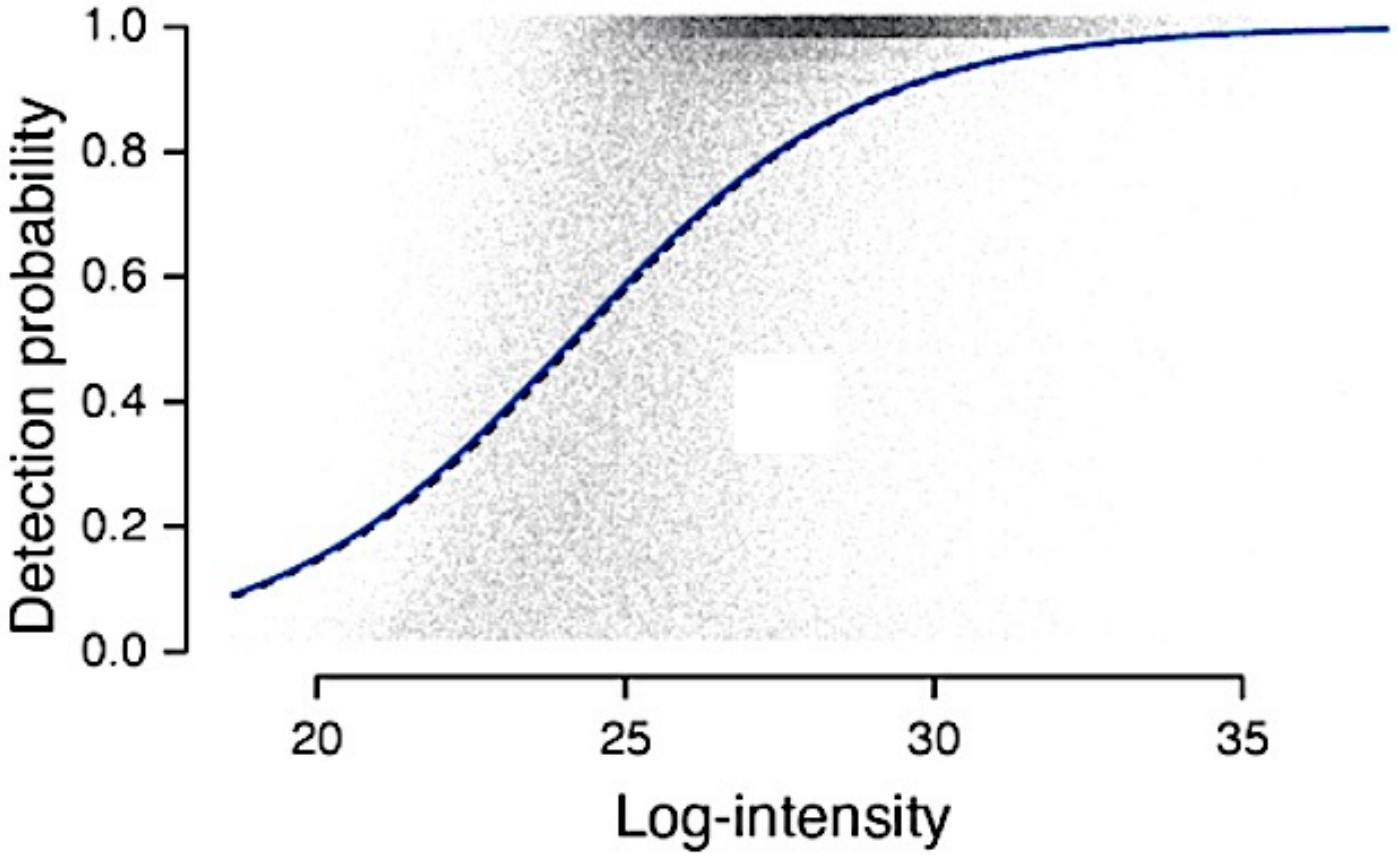
MAR Probability

MNAR Probability



Karpievitch et al., *Bioinformatics.* 2009.

Detection probability curve better represents MNAR data generating process

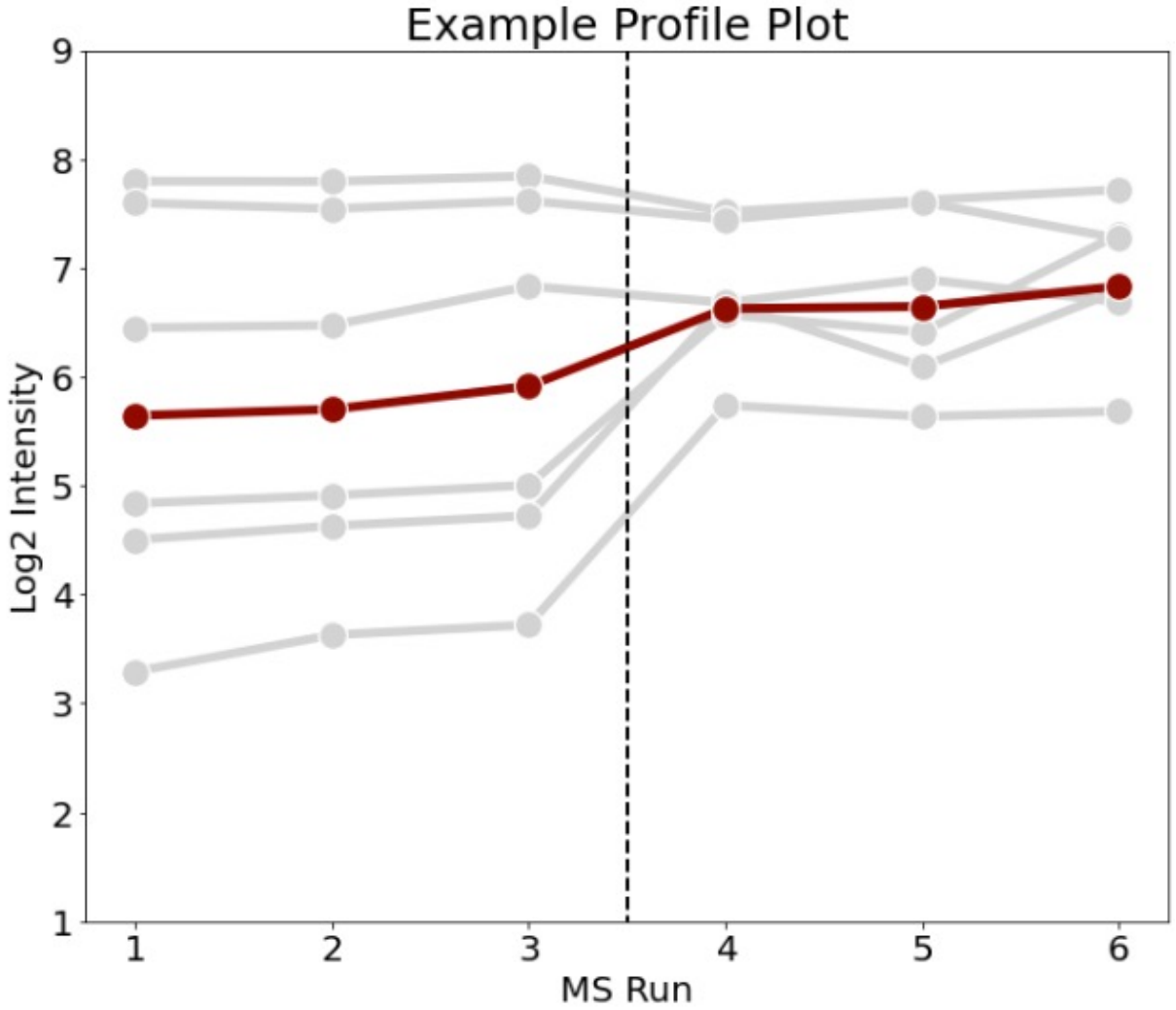


Li and Smyth, *Bioinformatics.* 2023.

Presentation Outline

- Problem statement
- Background
  - MSstats model
  - Missing value imputation
  - Summarization
- Proposed method
- Evaluation
  - Missing mechanisms
  - Imputation and summarization
  - Differential analysis

Protein summarization results in point estimates that eliminate uncertainty in the estimation



Presentation Outline

- Problem statement
- Background
  - MSstats model
  - Missing value imputation
  - Summarization
- Proposed method
- Evaluation
  - Missing mechanisms
  - Imputation and summarization
  - Differential analysis

We propose a Bayesian statistical modeling method that quantifies the uncertainty in the imputation and summarization workflow

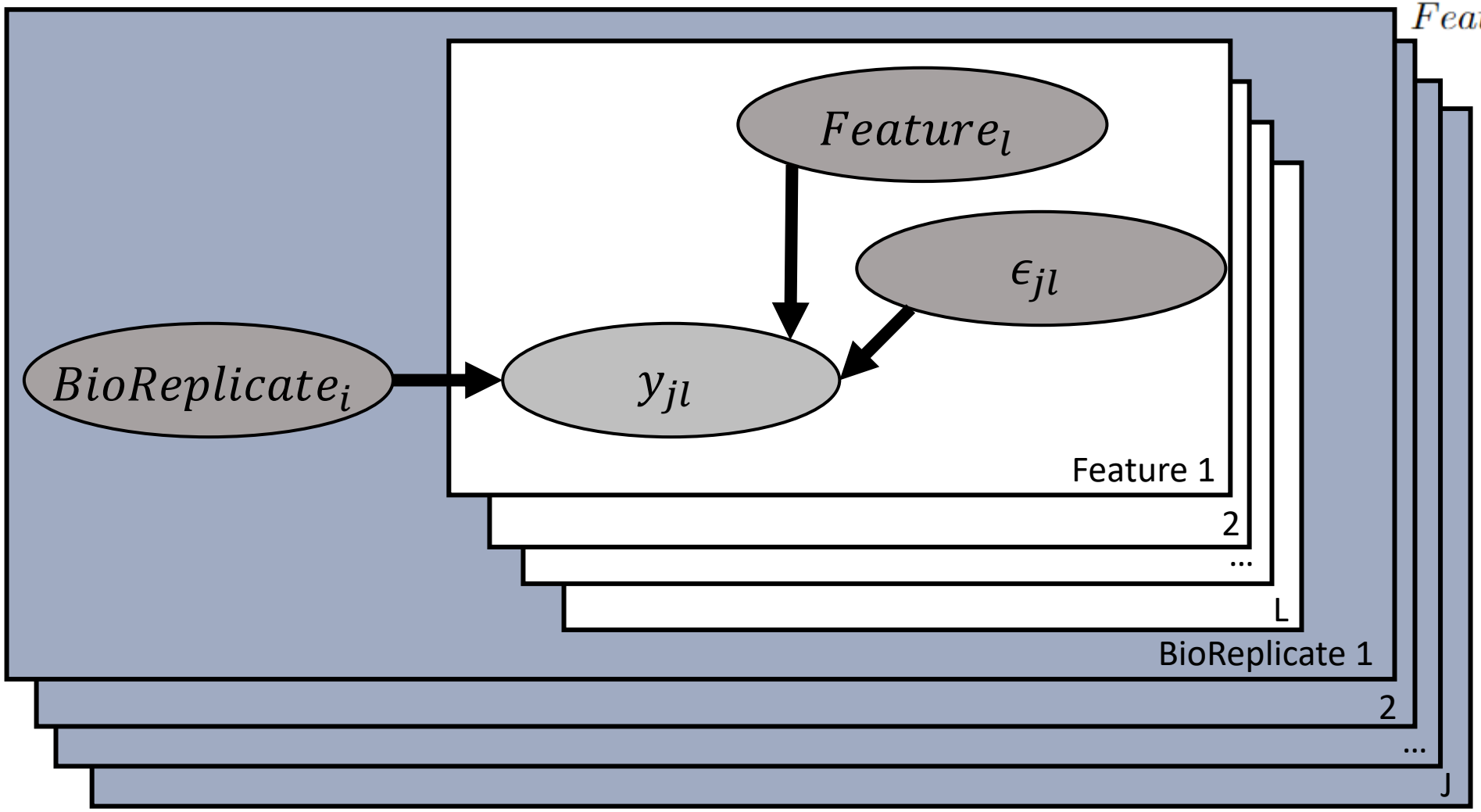
- Counters bias stemming from missing value imputation
- Flexible to a variety of experimental designs and acquisition methods
- Quantifies the uncertainty via posterior estimation

Reformulate subplot model into graphical format

$$y_{jl} = \mu + BioReplicate_j + Feature_l + \epsilon_{jl} \quad where \quad BioReplicate_j \sim^{iid} N(0, \sigma^2_{BioReplicate})$$

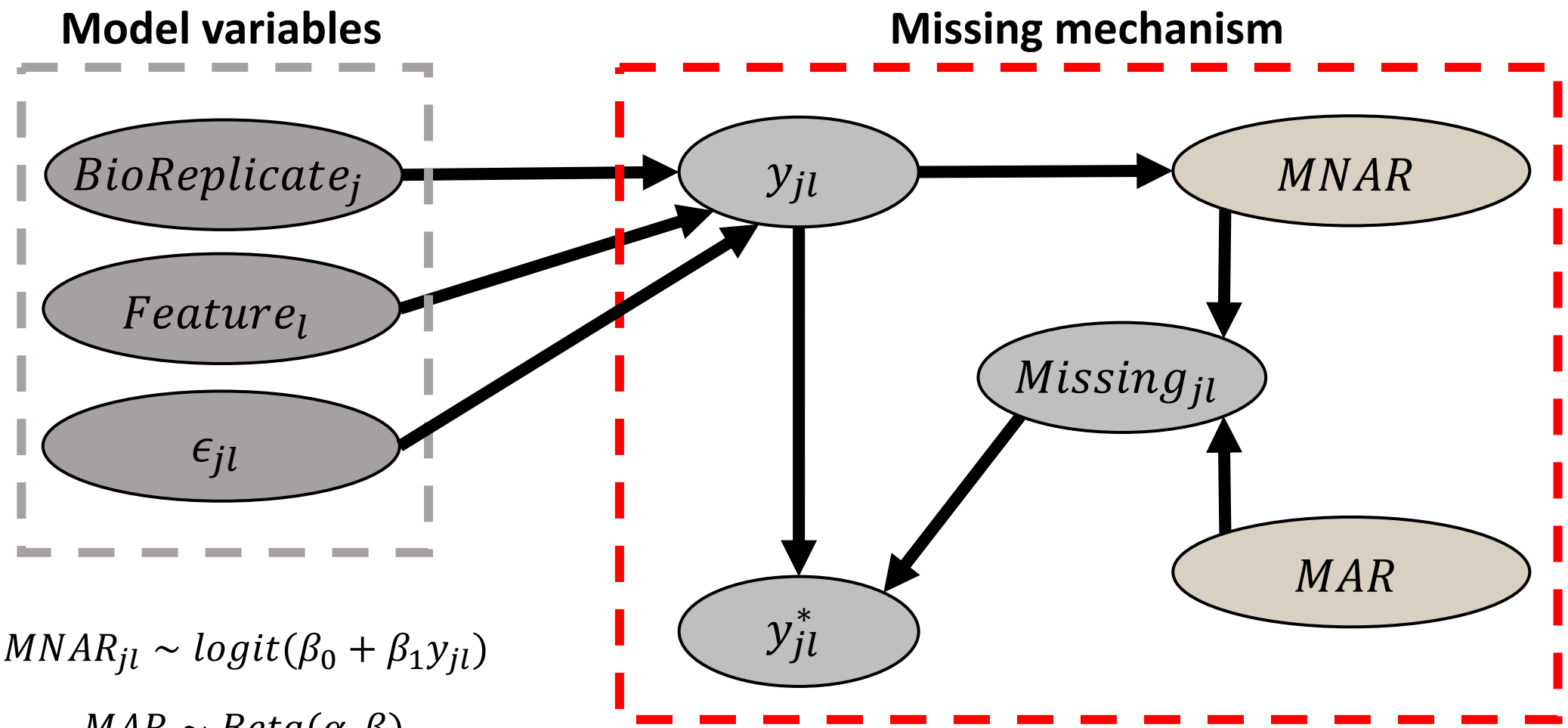
$$Feature_l \sim^{iid} N(0, \sigma^2_{Feature})$$

$$\epsilon_{jl} \sim^{iid} N(0, \sigma^2_{\epsilon})$$



Specify priors over random variables

Graphical encoding of missing mechanism



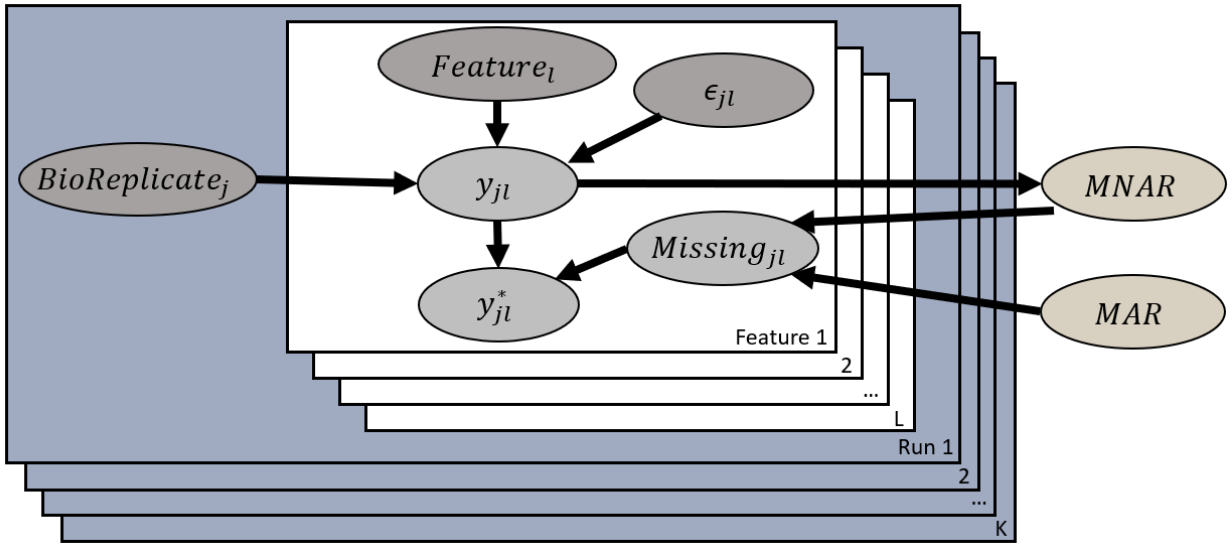
$$MNAR_{jl} \sim logit(\beta_0 + \beta_1 y_{jl})$$

$$MAR \sim Beta(\alpha, \beta)$$

$$Missing_{jl} = B(MAR + (1 - MAR) * MNAR)$$

$$y^*_{jl} = \begin{cases} y_{jl} & Missing_{jl} = 0 \\ NA & Missing_{jl} = 1 \end{cases}$$

Graphical and mathematical model



$$y_{jl} \sim N(\mu + BioReplicate_j + Feature_l, \epsilon_{jl})$$

$$where \;\; BioReplicate_j \sim^{iid} N(0, \sigma^2_{BioReplicate})$$

$$Feature_l \sim^{iid} N(0, \sigma^2_{Feature}), \epsilon_{jl} \sim^{iid} N(0, \sigma^2_{\epsilon})$$

Missing mechanism

$$y_{jl} \sim N(\mu + BioReplicate_j + Feature_l, \epsilon_{jl})$$

$$MNAR_{jl} \sim logit(\beta_0 + \beta_1 y_{jl})$$

$$MAR \sim Beta(\alpha, \beta)$$

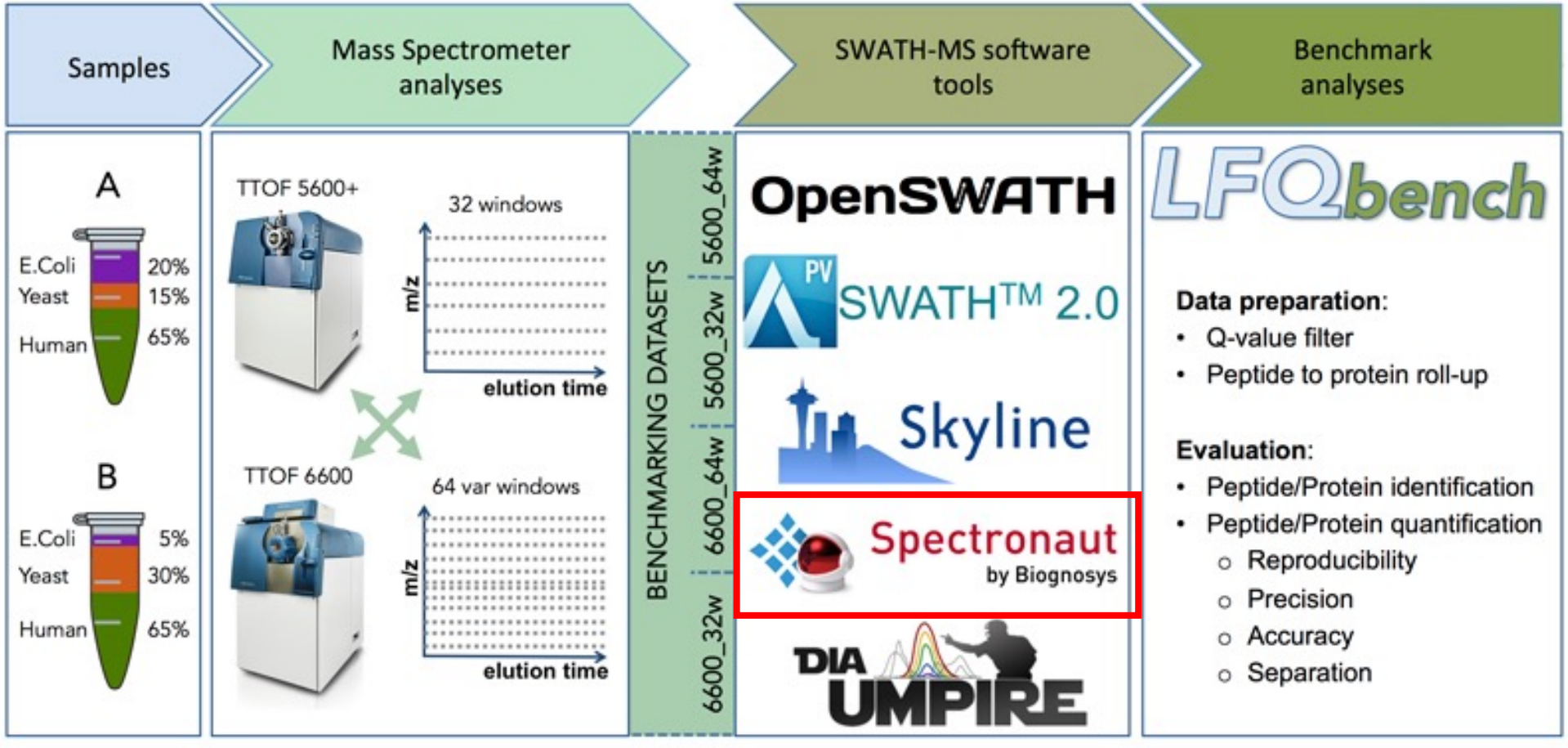
$$Missing_{jl} = \mathrm{B}(MAR + (1 - MAR) * MNAR)$$

$$y^*_{jl} = \begin{cases} y_{jl} & Missing_{jl} = 0 \\ NA & Missing_{jl} = 1 \end{cases}$$

Presentation Outline

- Problem statement
- Background
  - MSstats model
  - Missing value imputation
  - Summarization
- Proposed method
- Evaluation
  - Missing mechanisms
  - Imputation and summarization
  - Differential analysis

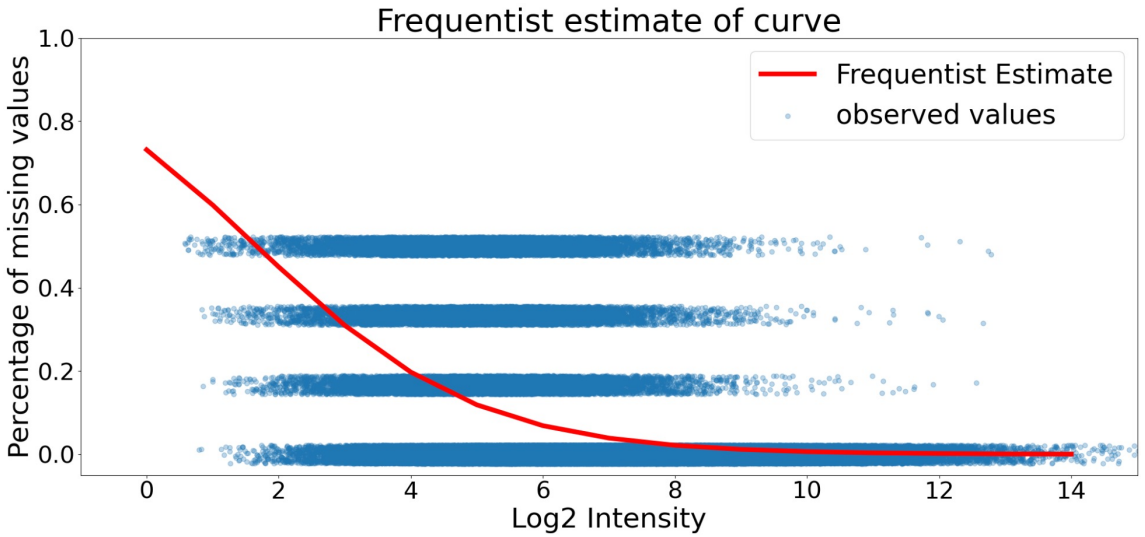
Benchmark dataset

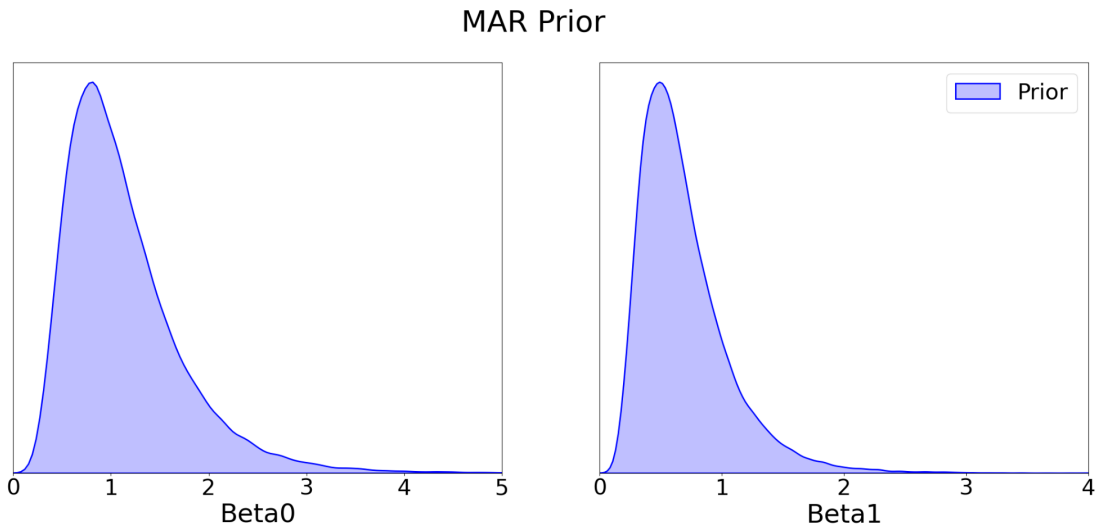


Navarro et al. Nat. Biotech. 2016

Estimate MNAR missing mechanism

Step 1. Learn prior from the data



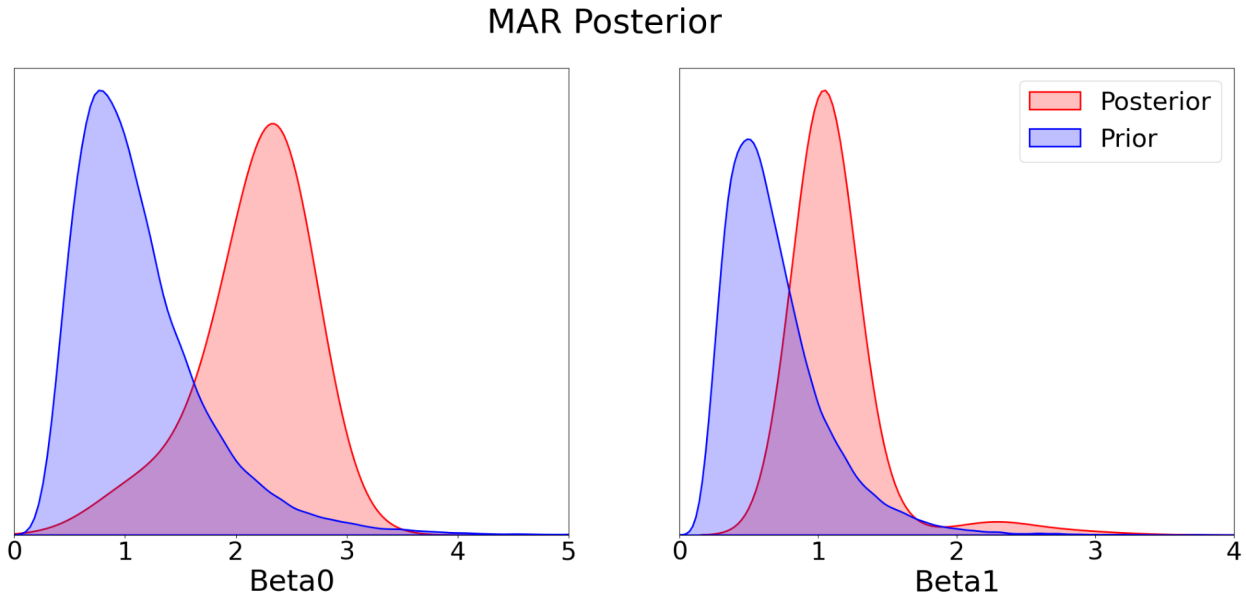


$$MNAR_{jl} \sim logit(\beta_0 + \beta_1 y_{jl})$$

$$\beta_0 \sim LogNormal(\mu_{\beta_0}, \sigma^2_{\beta_0})$$

$$\beta_1 \sim LogNormal(\mu_{\beta_1}, \sigma^2_{\beta_1})$$

Step 2. Estimate posterior

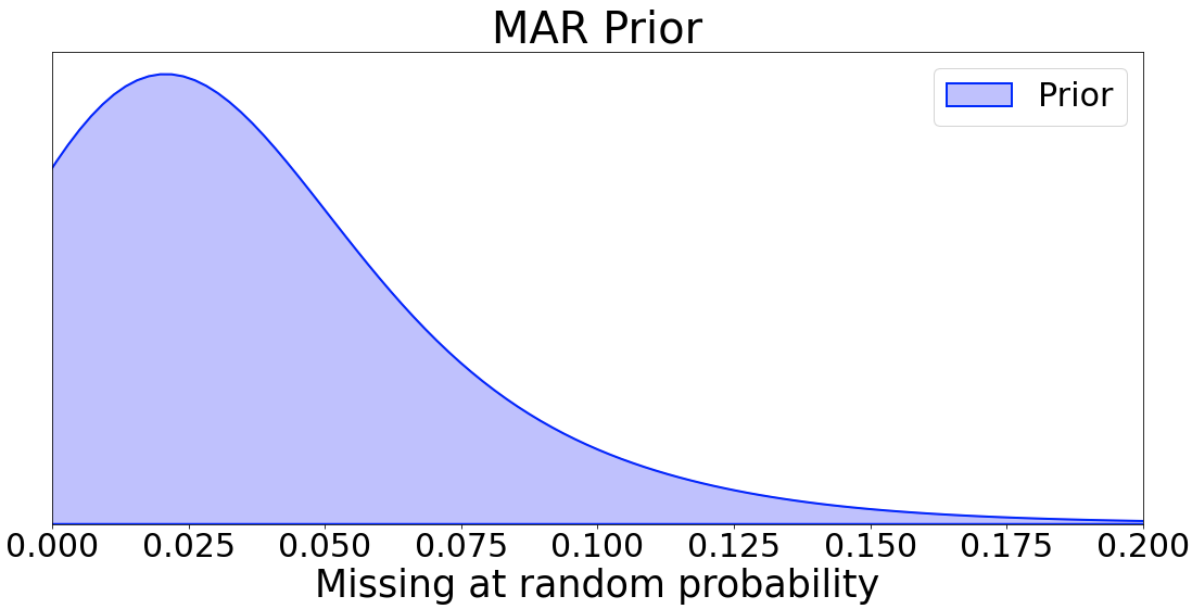


Estimate MAR mechanism

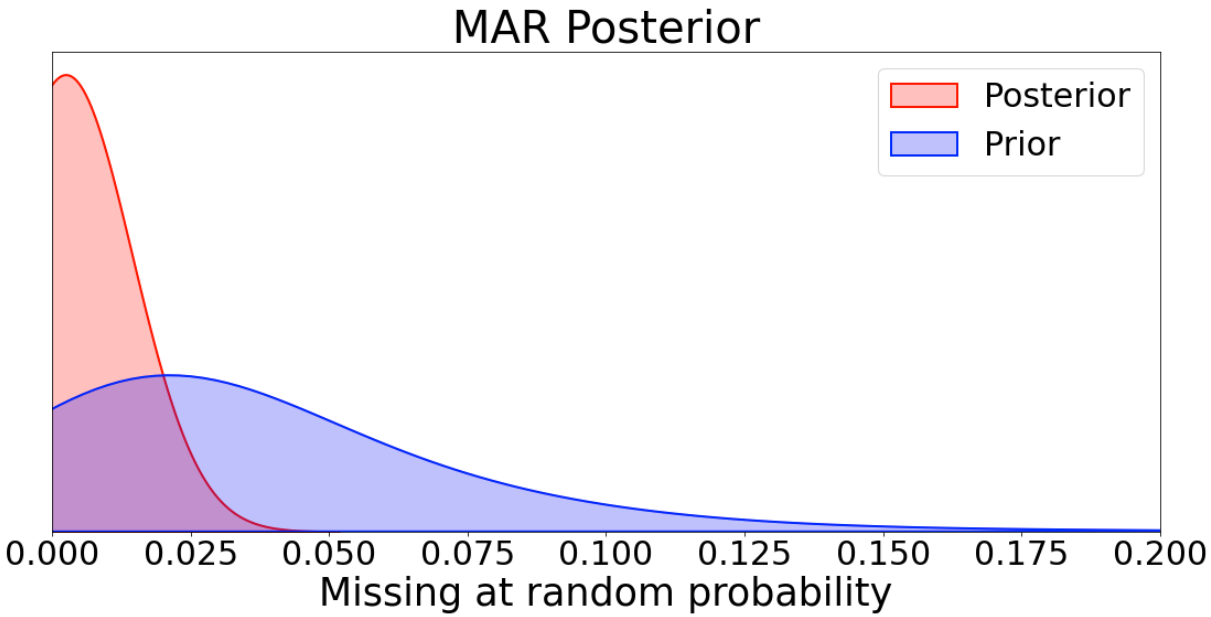
$$MAR \sim Beta(\alpha, \beta)$$

Step 1. Learn prior from the data

Quantile	Avg. log intensity	Avg. missing %
90th	8.48	2.8%
95th	9.15	2.2%

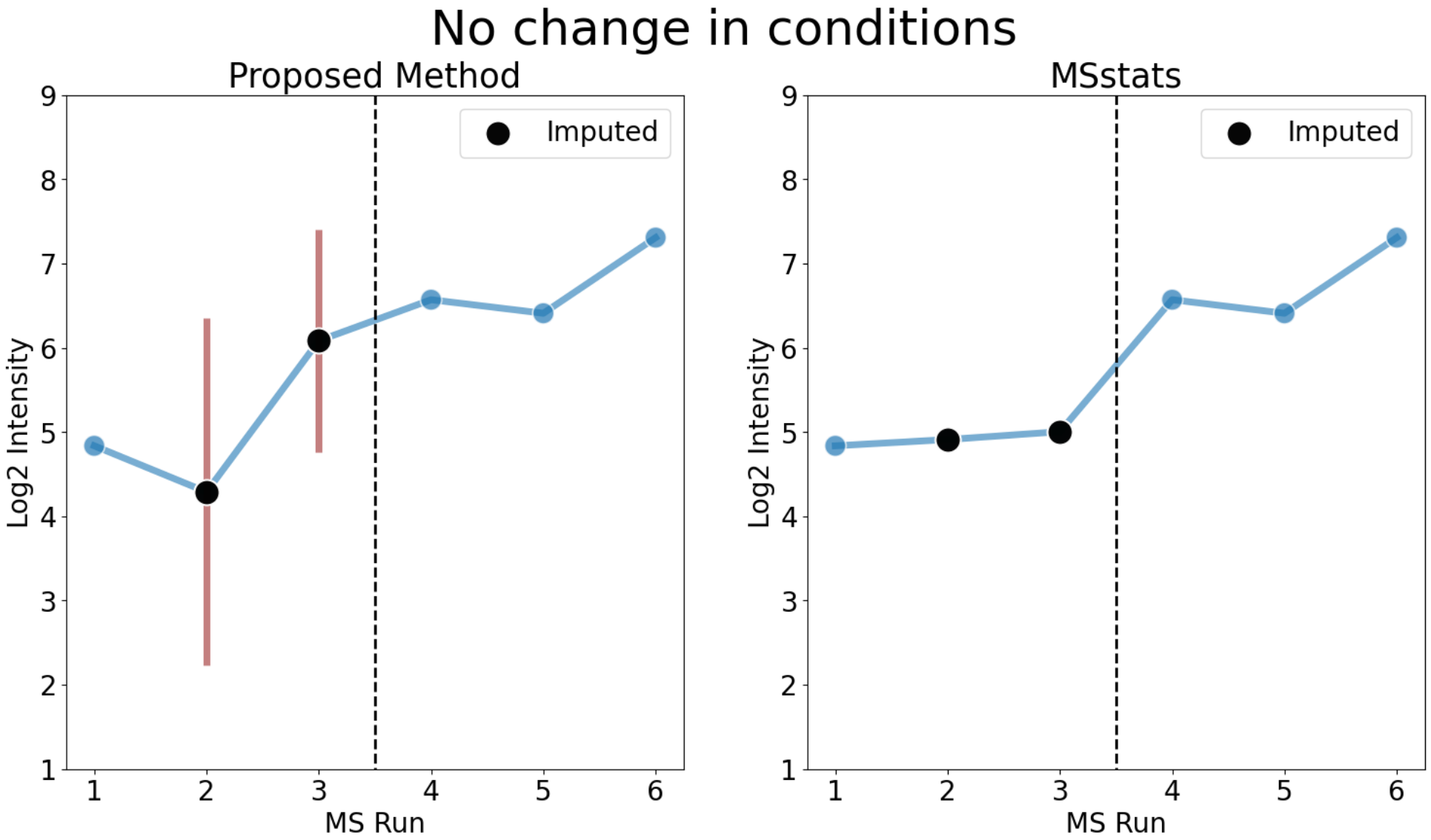


Step 2. Estimate posterior

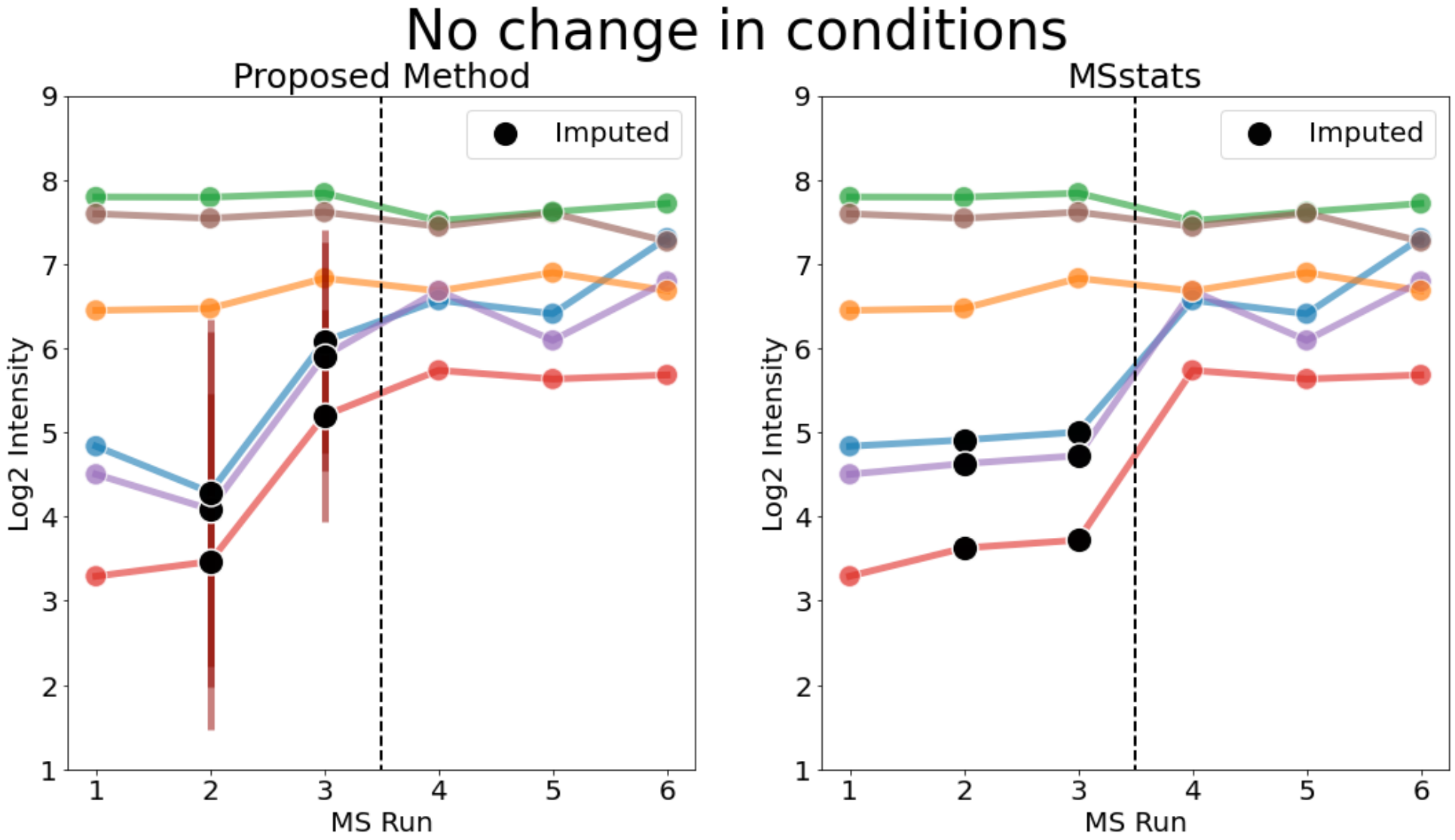


What do these posteriors mean for imputation?

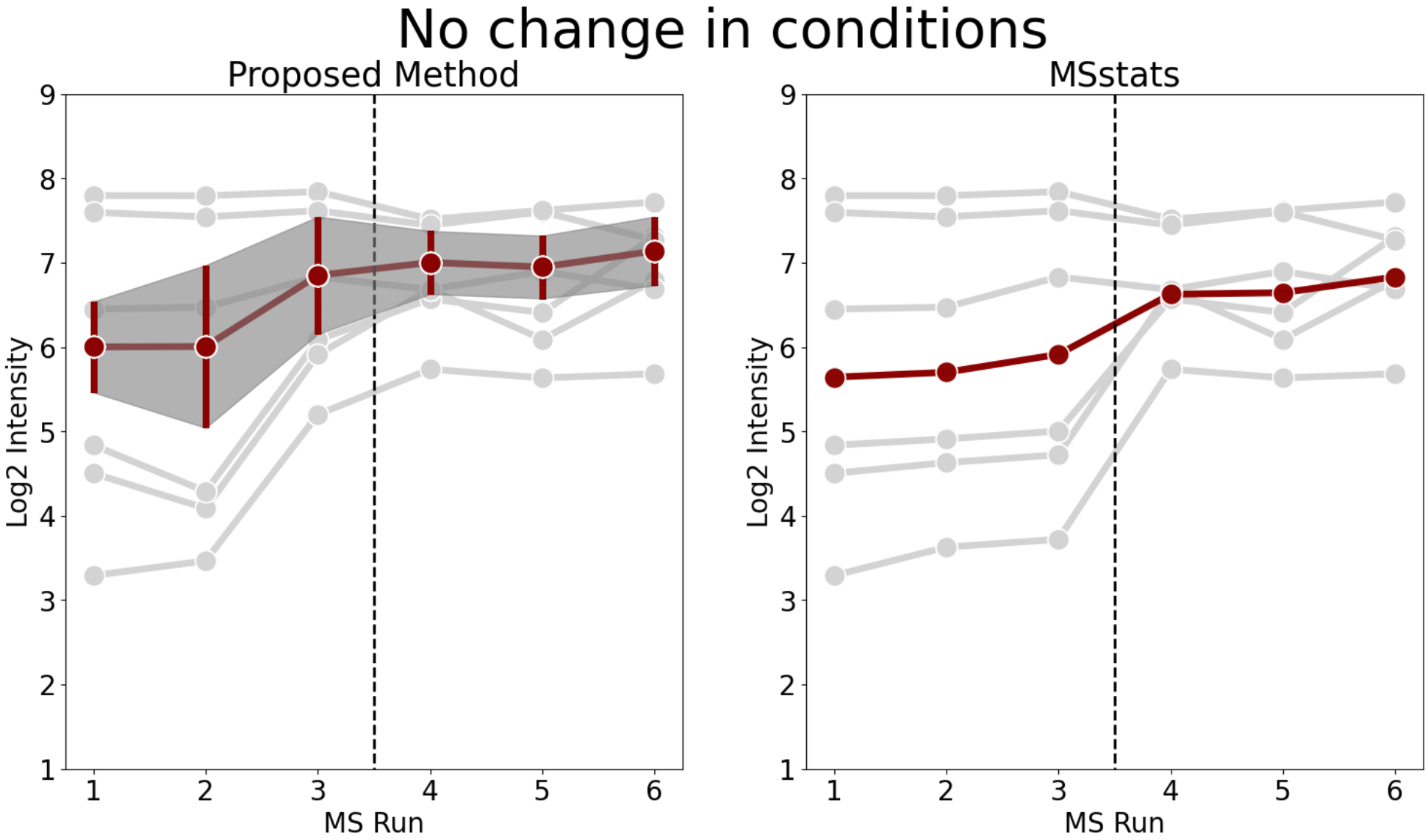
Imputation comparison with MSstats



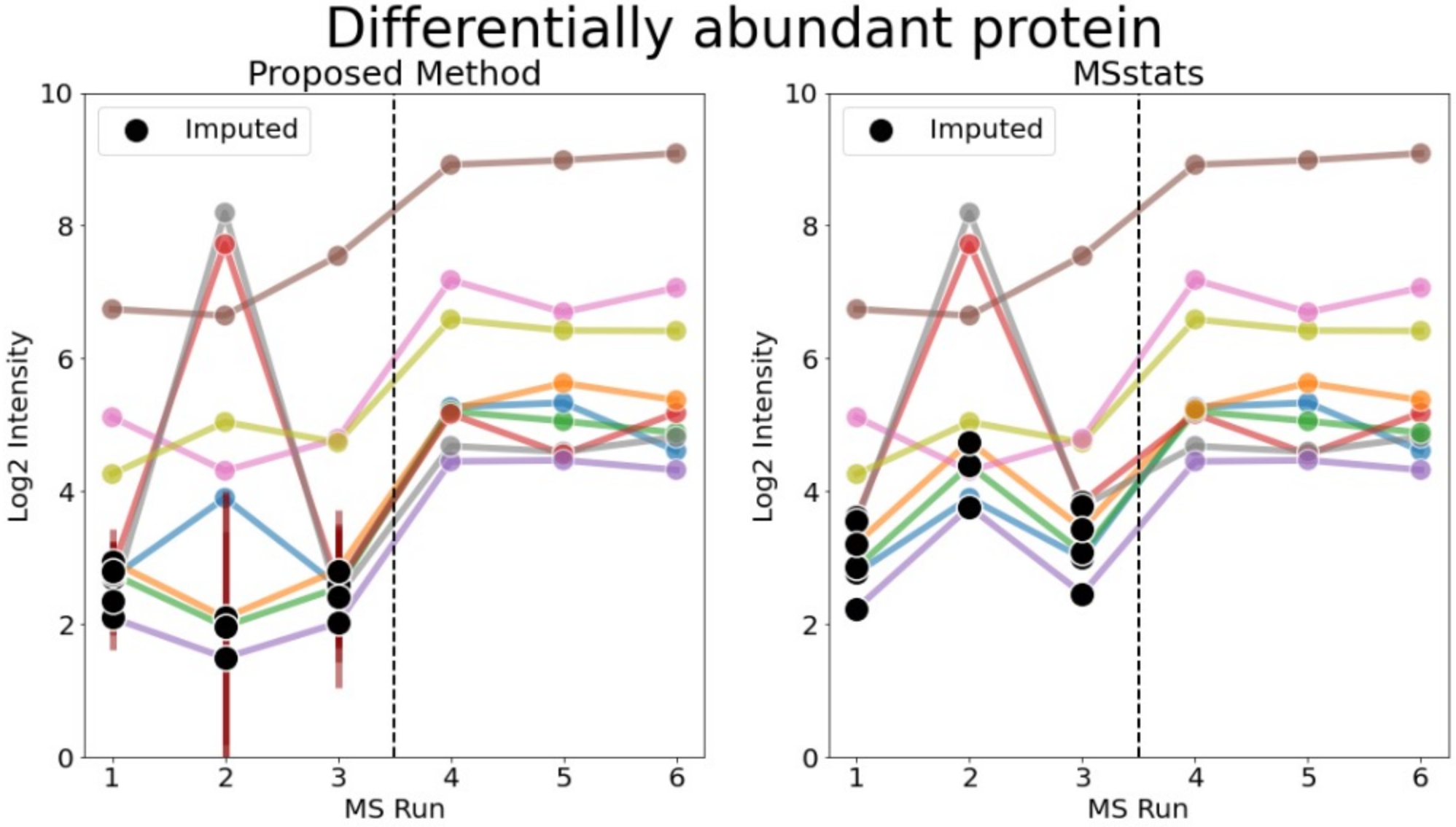
Imputation comparison with MSstats



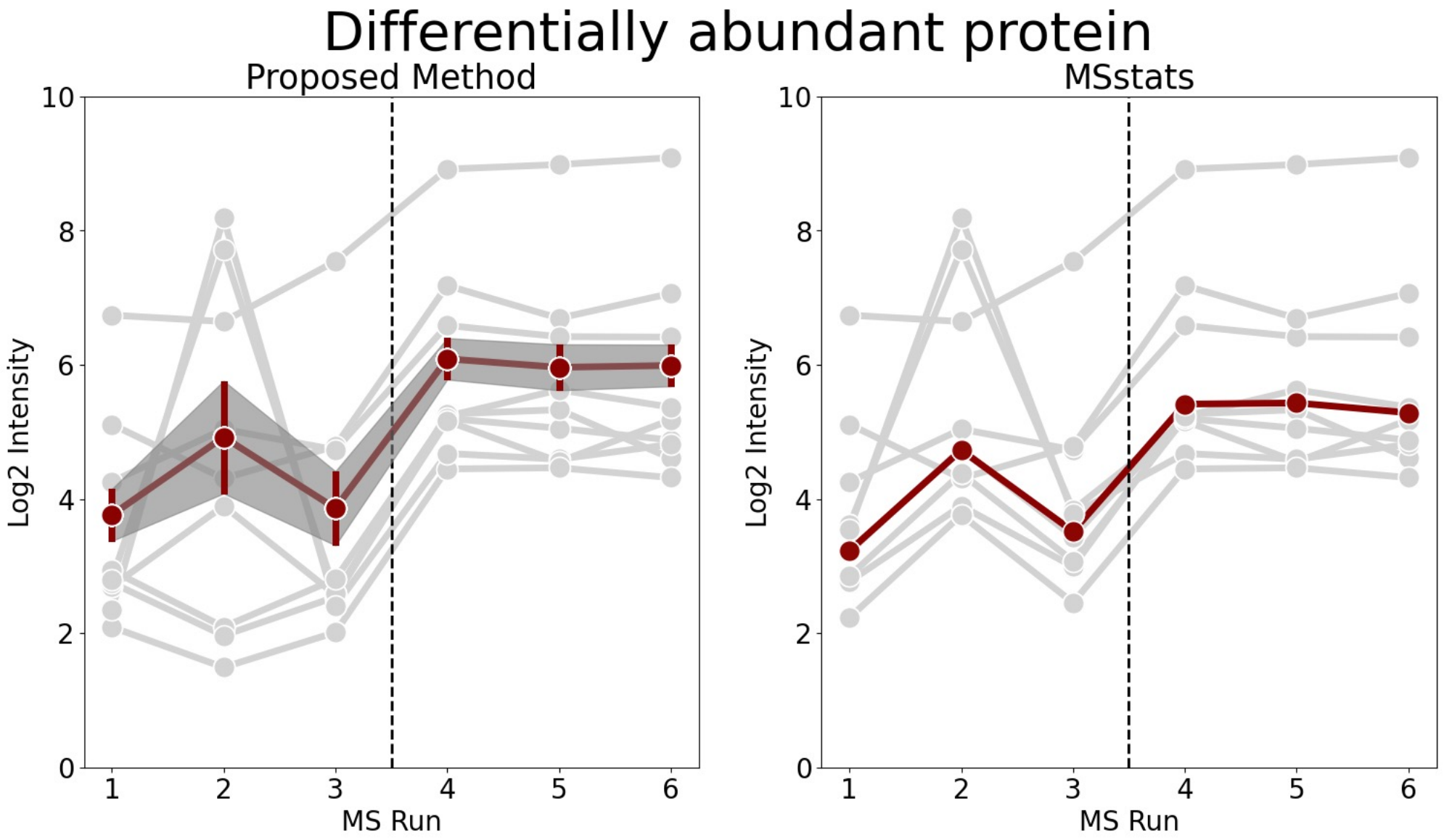
Summarization comparison with MSstats



Imputation comparison with MSstats



Summarization comparison with MSstats



Overview

- We now have uncertainty measurements associated with both the imputed values, as well as the protein-level intensity estimation per run
- What can we do with these posteriors?
  - Use in other analysis applications
  - Incorporate into differential analysis

Differential analysis with uncertain inputs

- Several options to do this
  - Building condition effect directly into the Bayesian model (one step)
  - Weighted linear regression with unequal error variances (two step)

Weighted least squares regression with unequal variances

MLE linear regression equation with unequal variances

$$z_{ij} = \mu + Condition_i + \epsilon_{ij}$$

$$\textit{where} \quad \sum_{i=1}^{I} Condition_i = 0$$

$$\epsilon_{ij} = \begin{bmatrix} \sigma_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{ij}^2 \end{bmatrix}$$

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

$$w_{ij} = \frac{1}{\sigma_{ij}^2}$$

Define weighted loss function

$$\sum_{n=1}^{ij} w_n (z_n - Condition_i)^2$$

Weighted least squares regression with unequal variances

MLE linear regression equation with unequal variances

$$z_{ij} = \mu + Condition_i + \epsilon_{ij}$$

$$\textit{where} \quad \sum_{i=1}^{I} Condition_i = 0$$

$$\epsilon_{ij} = \begin{bmatrix} \sigma_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{ij}^2 \end{bmatrix}$$

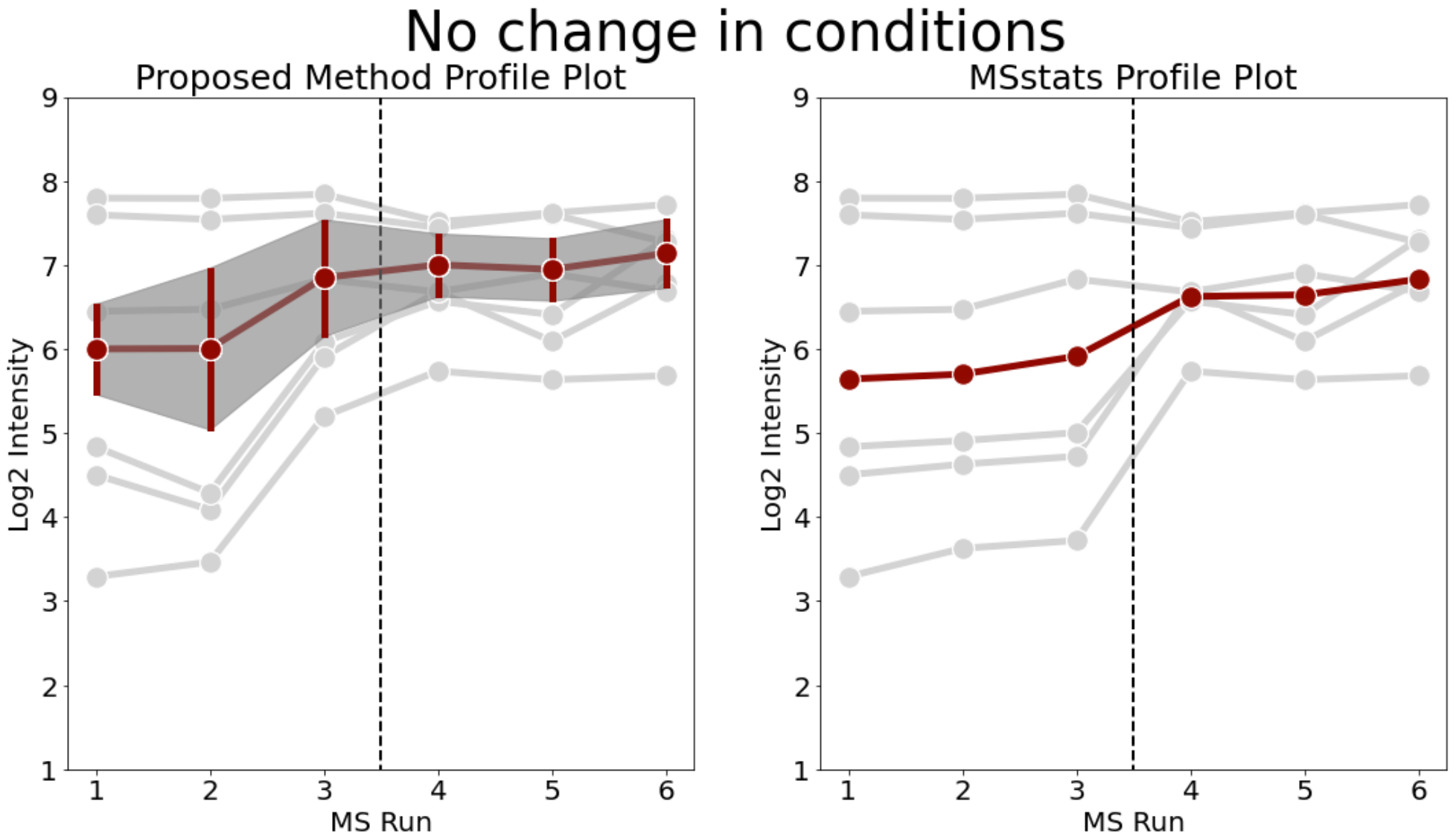
Define weight w_i as the reciprocal of the σ_i^2 in maximum likelihood

$$w_{ij} = \frac{1}{\sigma_{ij}^2}$$

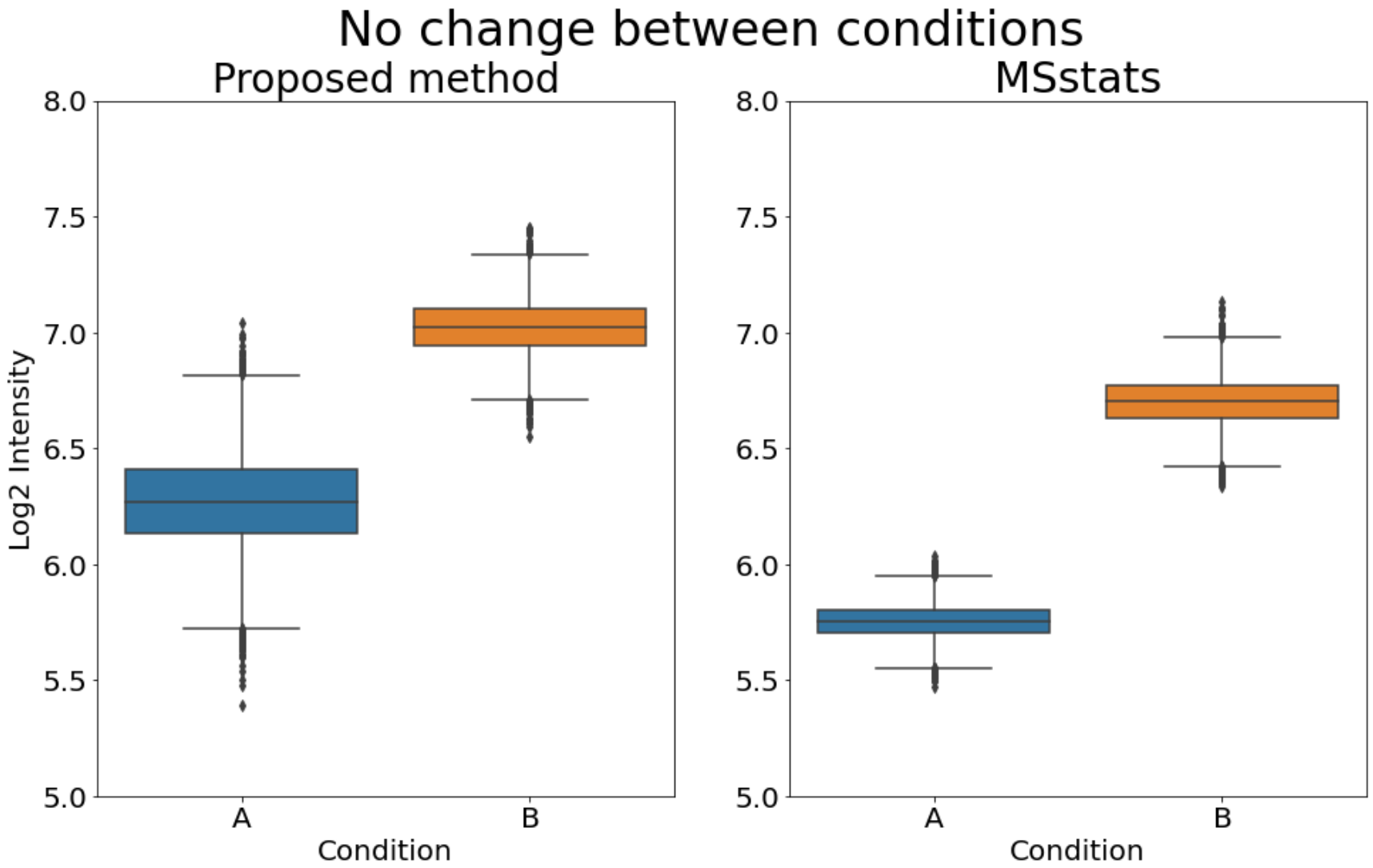
Define weighted loss function

$$\sum_{n=1}^{ij} w_n (z_n - Condition_i)^2$$

Summarization comparison

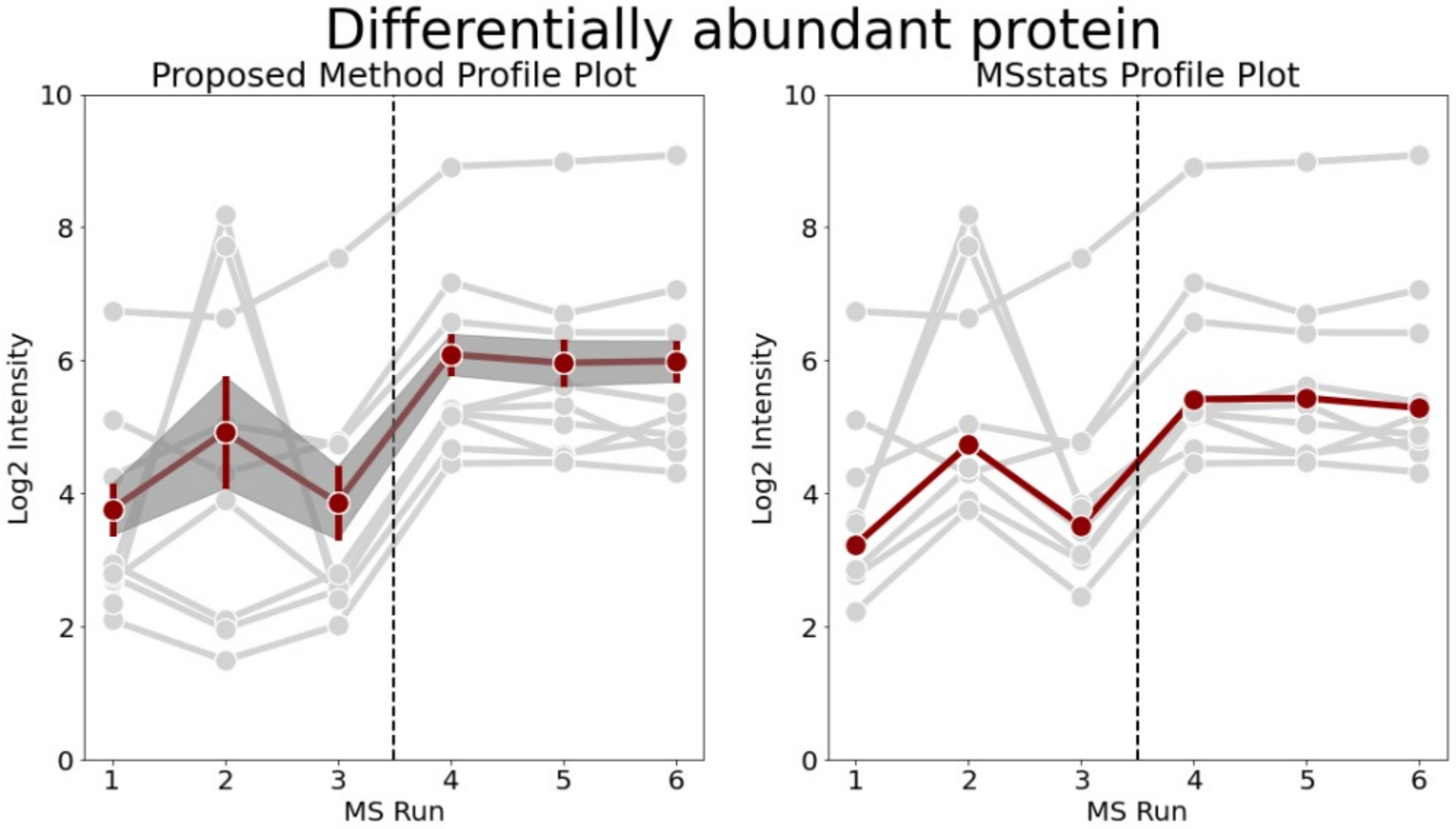


Statistical testing results

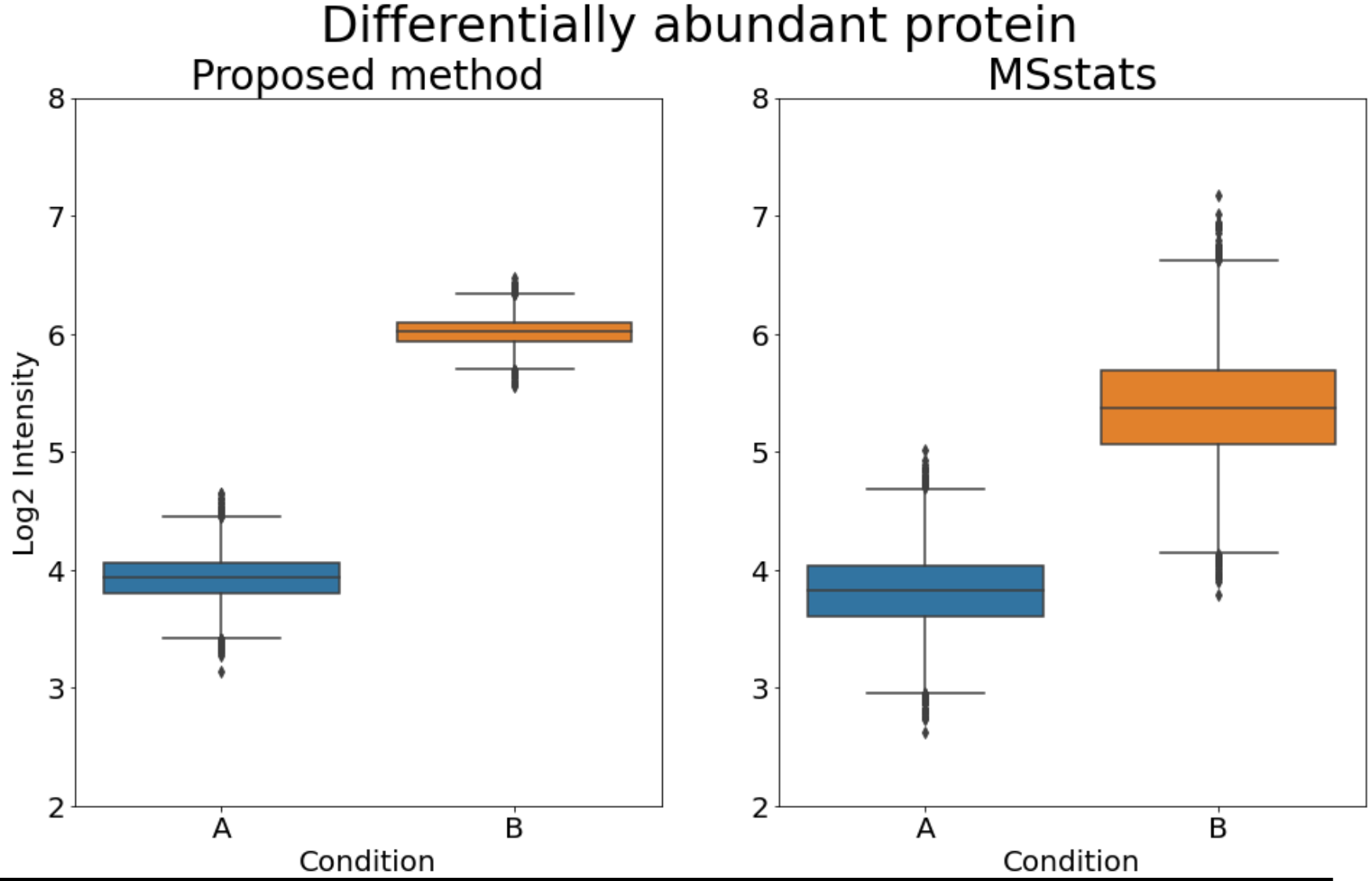


Adjusted p-value	.062	.0023

Summarization comparison



Statistical testing results



Adjusted p-value	.0024	.053

Concluding points

- Incorporate underlying uncertainty across upstream data processing workflow which was previously lost
- Learns the underlying probabilities in the missing value mechanisms
- Corrects differential analysis in cases of high uncertainty
- Flexible and applicable to a variety of complex experimental designs

Acknowledgements

Northeastern University

OLGA VITEK LAB

Statistical Methods For Studies Of Biomolecular Systems

Lab Members

Kylie Bemis
Sara Mohammad Taheri
Ritwik Anand
Vartika Tewari
Sai Srikanth Lakkimsetty
Mateusz Stankiak

Collaborators

Meena Choi

Jeremy Zucker

Karen Sachs

Barnett Institute travel award