# Multi-omics Data Preprocessing and Functional Clustering

Chi Yen Tseng [1], Emilio S. Rivera [1], John R. Tipton [2], Trevor Glaros [1] | [1]Biochemistry and Biotechnology Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA ; [2]Statistical Sciences, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA

Bioscience

## Introduction

### 1. EVALUATE NORMALIZATION METHODS IN MULTI-OMICS DATASETS

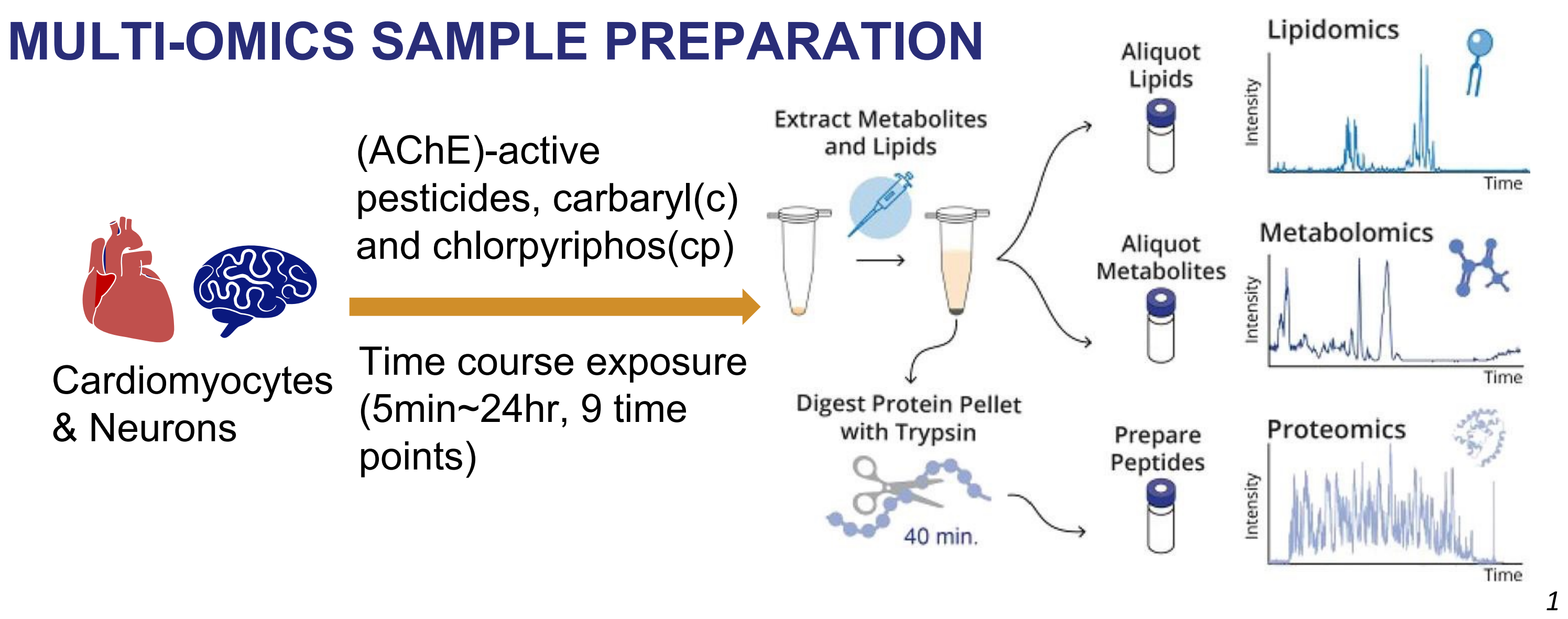
- Examining normalization strategies is critical for multi-omics data preprocessing to reduce systematic error and discover biological differences.
- In this study, multi-omics datasets were acquired from the cardiomyocyte and motor neuron cells in a time-course exposure study to acetylcholinesterase (AChE)-active chemicals. We compared different normalization methods and assessed the effectiveness by observing if a normalized dataset could improve QC feature consistency and treatment-related variance while preserve time-related variance.

### 2. FINITE MIXTURES FOR FUNCTIONAL CLUSTERING

- We use Bayeian hierarchical modeling (BHM) framework for functional clustering. Each time-varying omic feature is assumed to belong to a latent cluster while capturing uncertainty and hierarchical structures.

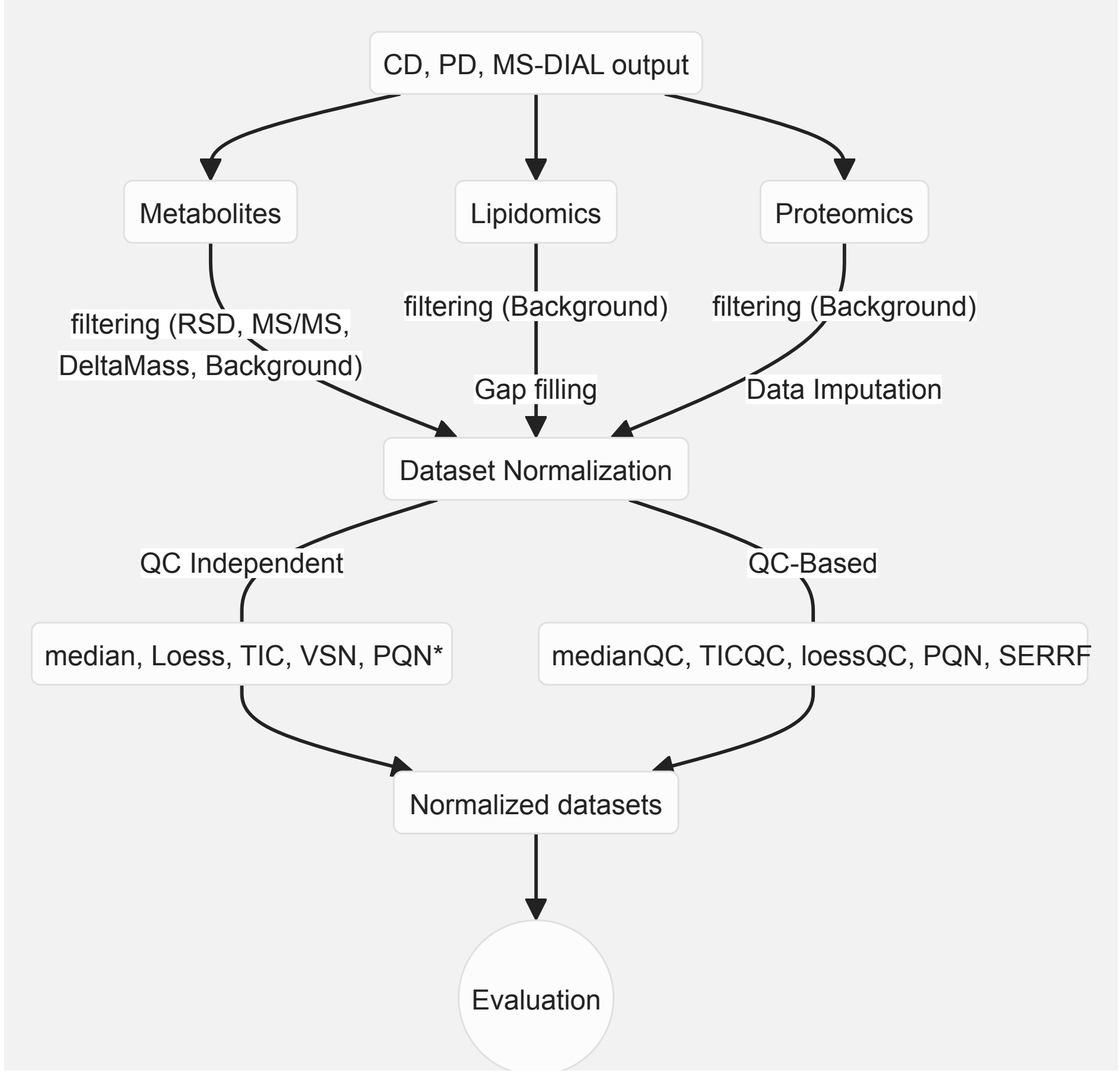
## Material and Methods

### MULTI-OMICS SAMPLE PREPARATION


Cardiomyocytes & Neurons
(AChE)-active pesticides, carbaryl(c) and chlorpyriphos(cp)
Time course exposure (5min~24hr, 9 time points)
Extract Metabolites and Lipids
Digest Protein Pellet with Trypsin
40 min.
Aliquot Lipids
Aliquot Metabolites
Prepare Peptides
Lipidomics
Metabolomics
Proteomics
Intensity
Time

[1]

### OMICS PREPROCESSING WORKFLOW



- 4k to 8k features in each omic.

### NORMALIZATION EVALUATION

- QC feature consistency (RSD < 0.2)
- The change in variance explained by time or treatment after normalization.

### FUNCTIONAL CLUSTERING

- Proteomics:

105 significant features (chlorpyriphos vs control) were selected for functional clustering.

### MODEL STATEMENT

### 1. VARIANCE EXPLAINED BY TIME OR TREATMENT

- PERMANOVA MODEL
- *Main effects of Time, Treatment, and their interaction (Bray-Curtis Distance)*

The adonis2() result includes:

- R²: Proportion of variance explained by each predictor.
- F-value: Ratio of explained to unexplained variance.
- p-value: Statistical significance

### 2. FUNCTIONAL CLUSTERING

- Let $Y\_ir\ (t)$ be the expression value of omic feature i for replicate r at time t transformed to $\log_2$FC relative to time 0

1. Spline Representation:

For each cluster $k$, we have a vector of spline coefficients $\boldsymbol{\beta}_k \in \mathbb{R}^B$ and a B-spline basis where $\mathbf{B}(t)$, which gives the functional mean for that cluster:

$$\mu_k(t) = \mathbf{B}(t)\boldsymbol{\beta}_k$$

2. Cluster Membership:

Each omic feature $Y_{i*}(t)$ is assumed to belong to a cluster indexed by a latent indicator $z_i$ where $z_i \in \{1, \ldots, K\}$.

3. Data Likelihood:

Given the cluster membership $z_i$, the observation $Y_{ir}(t)$ is centered around the cluster-specific mean $\mu_{z_i}(t)$, with shared noise $\sigma$:

$$Y_{ir}(t) \mid z_i, \sigma^2 = k \sim \mathcal{N}(\mu_k(t), \sigma^2)$$

4. Prior Distributions

- Mixture proportions $\pi \sim$ Dirichlet($\boldsymbol{\alpha}$)

-Bayesian smoothing spline prior for each cluster:

$$\beta_{k1} \sim \mathcal{N}(0,2)$$

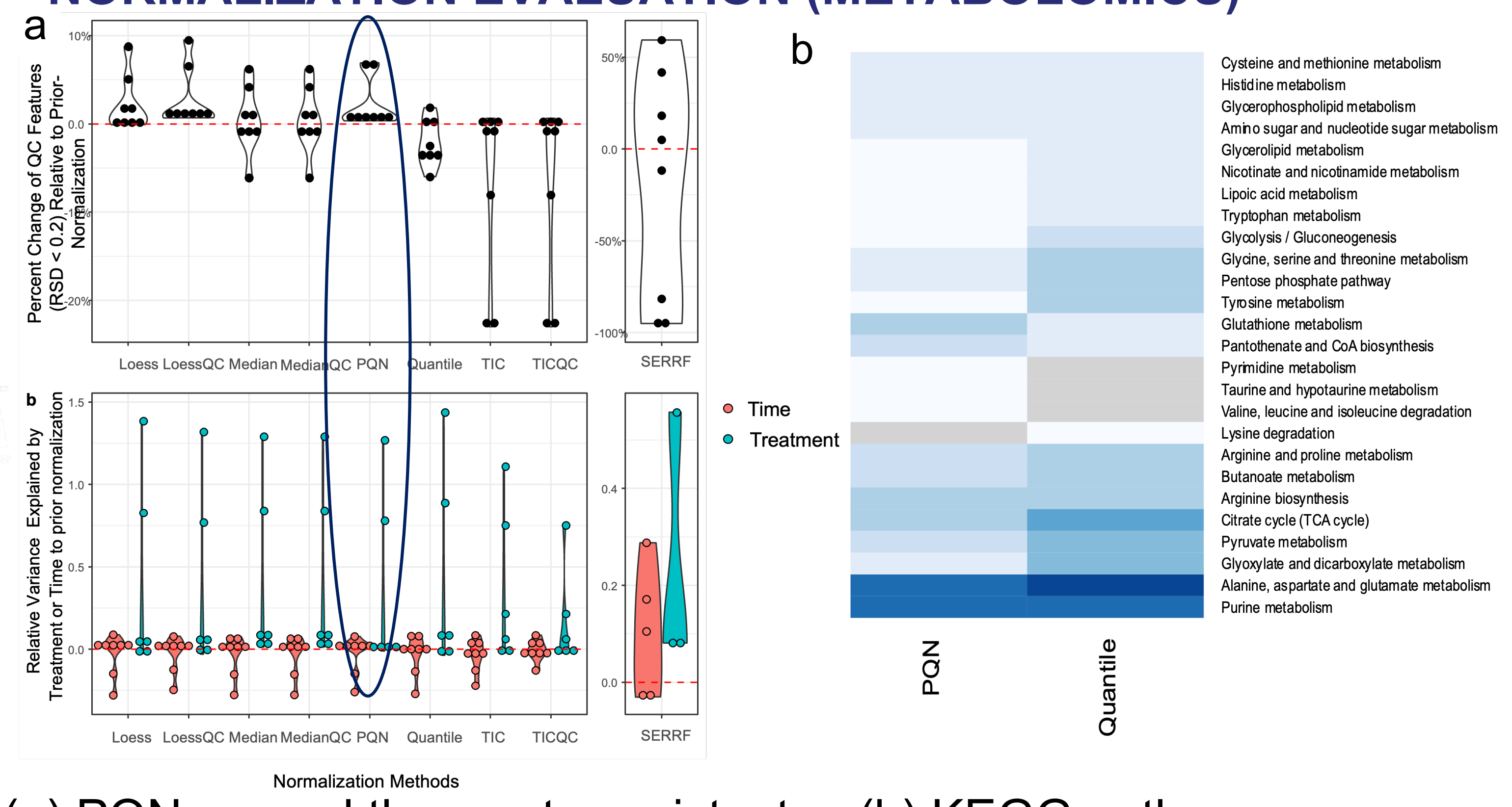
$$\beta_{kj} \sim \mathcal{N}\left(\beta_{kj-1}, \sigma_\beta^2\right), \quad j = 2, \ldots, p$$

-Observation noise parameter:

$$\sigma \sim \mathcal{N}^+(0,1)$$

## Results and Discussion

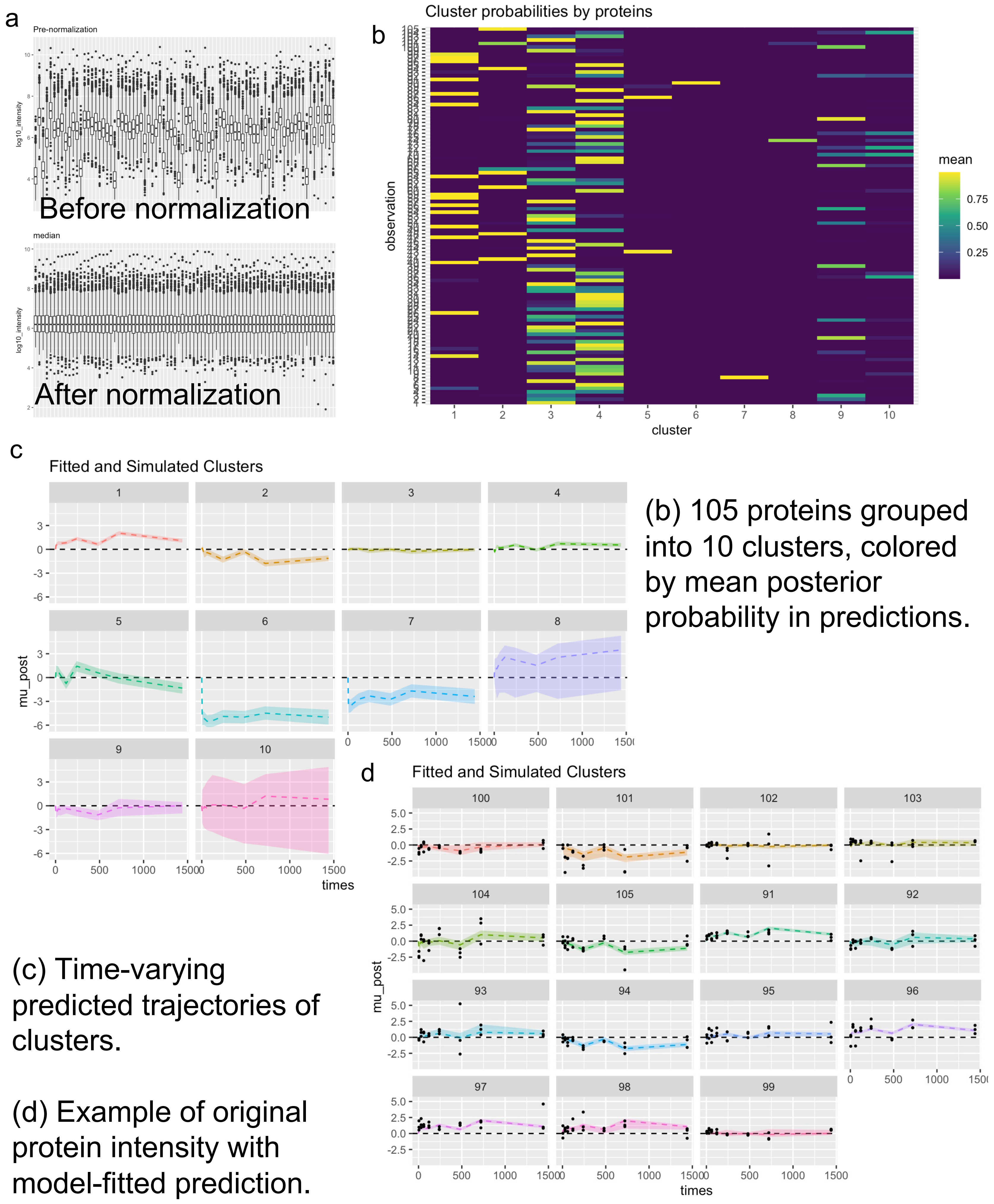
### NORMALIZATION EVALUATION (METABOLOMICS)



(a) PQN caused the most consistent change in QC feature consistency and variance explained by treatment.

(b) KEGG pathways PQN > Quantile

### FUNCTIONAL CLUSTERING (PROTEOMICS)



(b) 105 proteins grouped into 10 clusters, colored by mean posterior probability in predictions.

(c) Time-varying predicted trajectories of clusters.

(d) Example of original protein intensity with model-fitted prediction.

## Conclusion

We identified the most effective normalization methods for multi-omics datasets and demonstrate a clustering strategy that accounts for the uncertainty and hierarchical structures of time-varying omics features.

## Acknowledgements

The Los Alamos National Laboratory Directed R&D Fund, LDRD, grant number 20230084DR, funded this work. The funders had no role in the study design, data collection, analyses, data interpretation, or the decision to publish.

## Reference

1. Muehlbauer, L. K., et al. Anal. Chem. 2023, 95 (2), 659–667.

MS4IO x Los Alamos NATIONAL LABORATORY NNSA National Nuclear Security Administration

Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

LA-UR-25-21610