Probabilistic Modeling and Functional Clustering in Multi-Omics Data Integration

Chi Yen Tseng [1], John R. Tipton [2], Emilio S. Rivera [1], Tara Harvey [1], Joshua Breidenbach [1], Brett Blackwell [1], Salvator J. Palmisano [1], Grace Thornhill [1], Emilia Solomon [1], Claire Sanders [3], Kes Luchini [1], Ethan M. McBride [1], Jessica A. Salguero [1], Francie E. Rodriguez [1], Phillip Mach [1], Trevor Glaros [1] | [1]Biochemistry and Biotechnology Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA; [2]Statistical Sciences, Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA; [3]Microbial and Biome Sciences Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA

Introduction

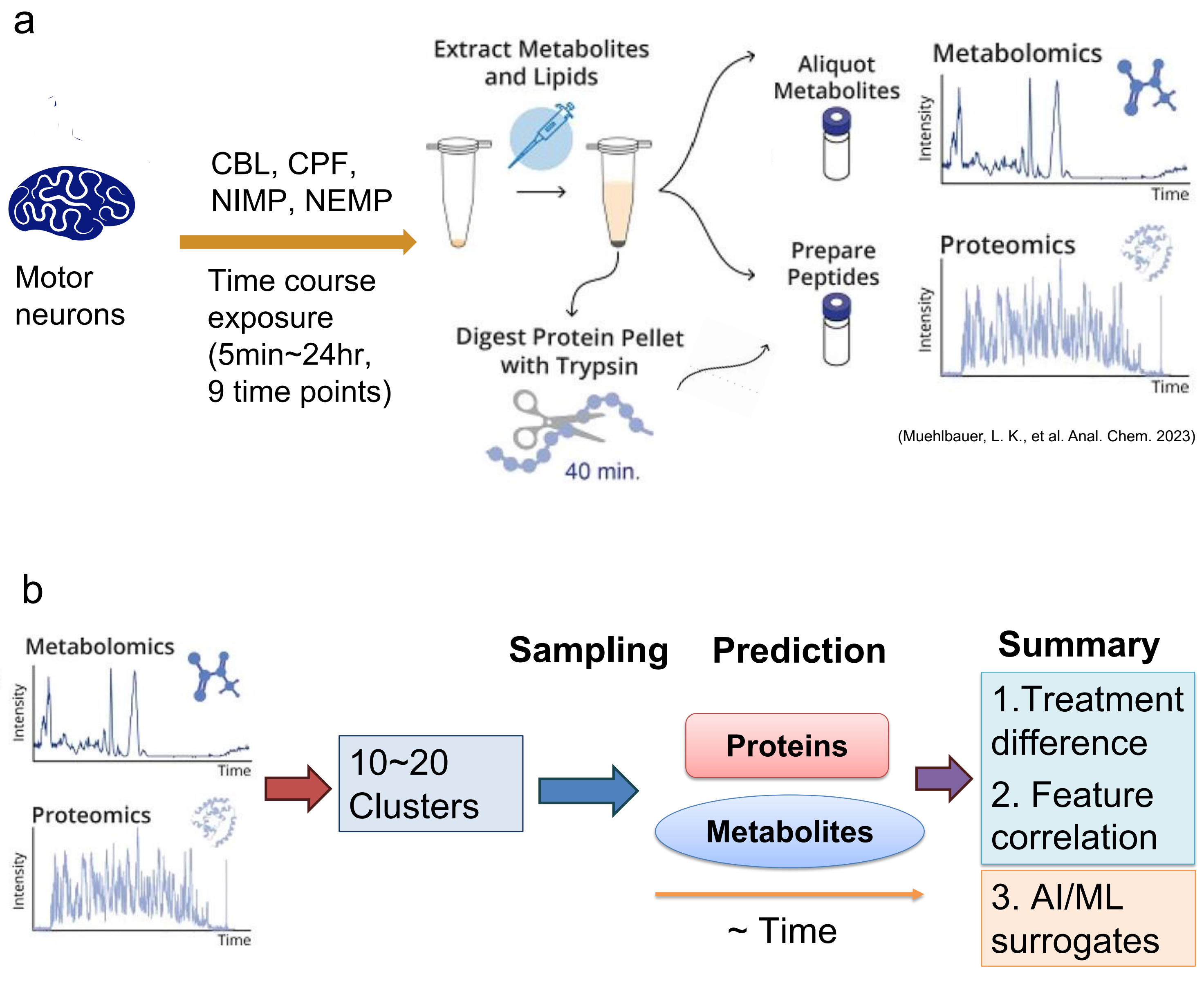
- Longitudinal multiomics data is complex and challenging to model.
- There is a need to model feature-level variation and capture inherent biological variability
- Bayesian hierarchical modeling (BHM) pools information from similar longitudinal trajectories of metabolites and proteins to improve estimation
- Bayesian posterior sampling enables summary statistics for:

  - Similarity of metabolites and proteins across the time course.

  - Longitudinal difference among chemical-exposed omics datasets.

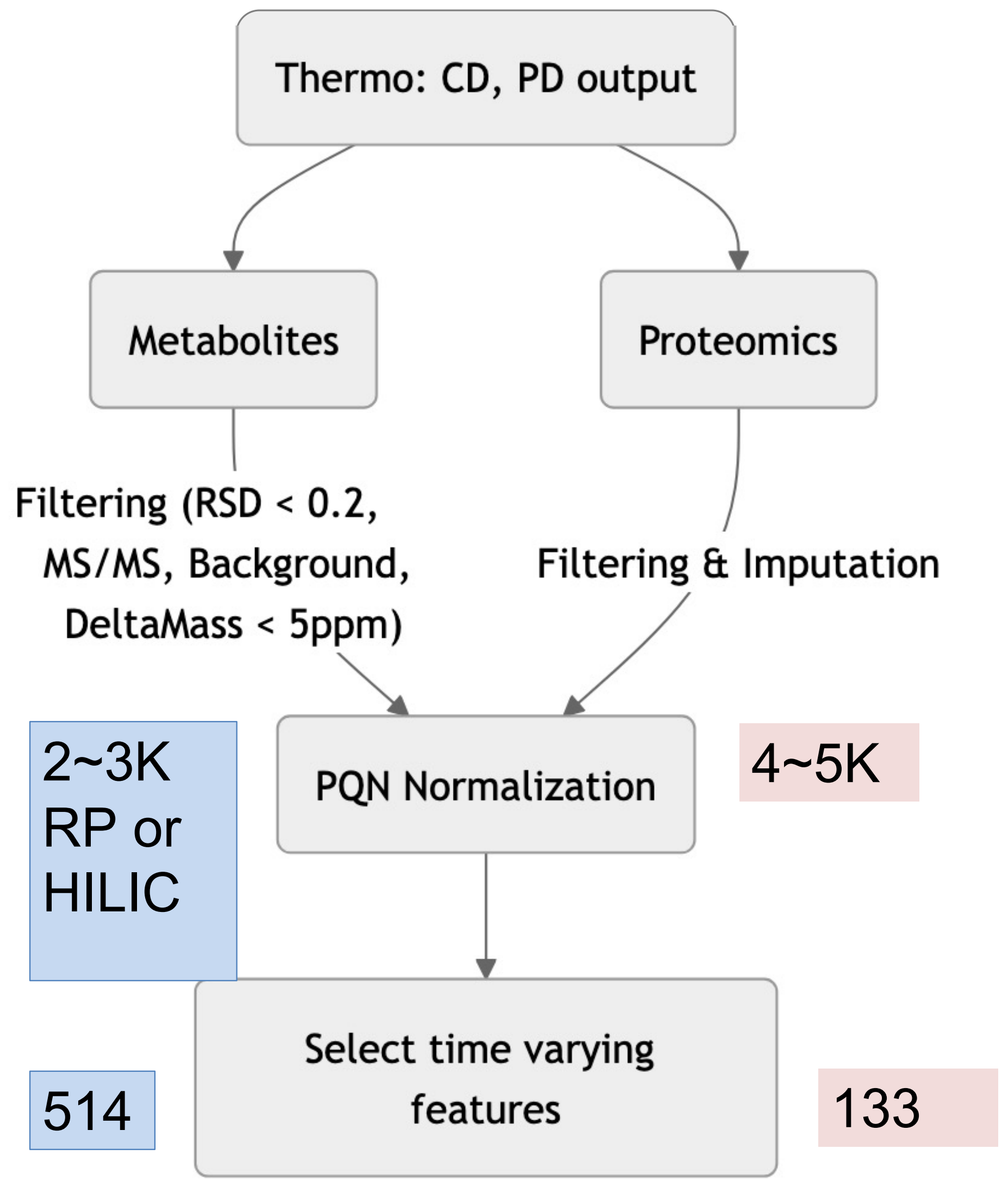
Material and Methods

EXPERIMENTAL SETUP, DATA PROCESSING & MODELING



(Muehlbauer, L. K., et al. Anal. Chem. 2023)

Figure 1. (a) Metabolomics and proteomics were acquired from human motor neurons derived from Induced pluripotent stem cells exposed to either AChE- active pesticides (CBL and CPF), sarin surrogate (NIMP), and VX surrogate (NEMP). (b) Bayesian hierarchical modeling (BHM) framework for functional clustering. Each time-varying protein or metabolite is assumed to belong to a latent cluster while capturing uncertainty and hierarchical structures. Summary statistics from temporal simulations for each treatment could be used to compare treatment effects, correlations between proteins and metabolites, and train surrogate models.


Thermo: CD, PD output

Metabolites

Proteomics

Filtering (RSD < 0.2, MS/MS, Background, DeltaMass < 5ppm)

Filtering & Imputation

2~3K RP or HILIC

PQN Normalization

4~5K

Select time varying features

514

133

Figure 2. Data pre-processing: metabolomic and proteomic datasets were filtered, imputed, and normalized using probabilistic quotient normalization (PQN). The log fold-change (logFC) of each feature (i) relative to media control depends on a smooth, nonlinear function of time*treatment interaction. 514 significant proteins were selected for clustering and 133 significant metabolites were pooled from reverse phase and HILIC, both positive and negative mode.

Result and Discussion

FUNCTIONAL CLUSTERING: e.g., NEMP-treated data

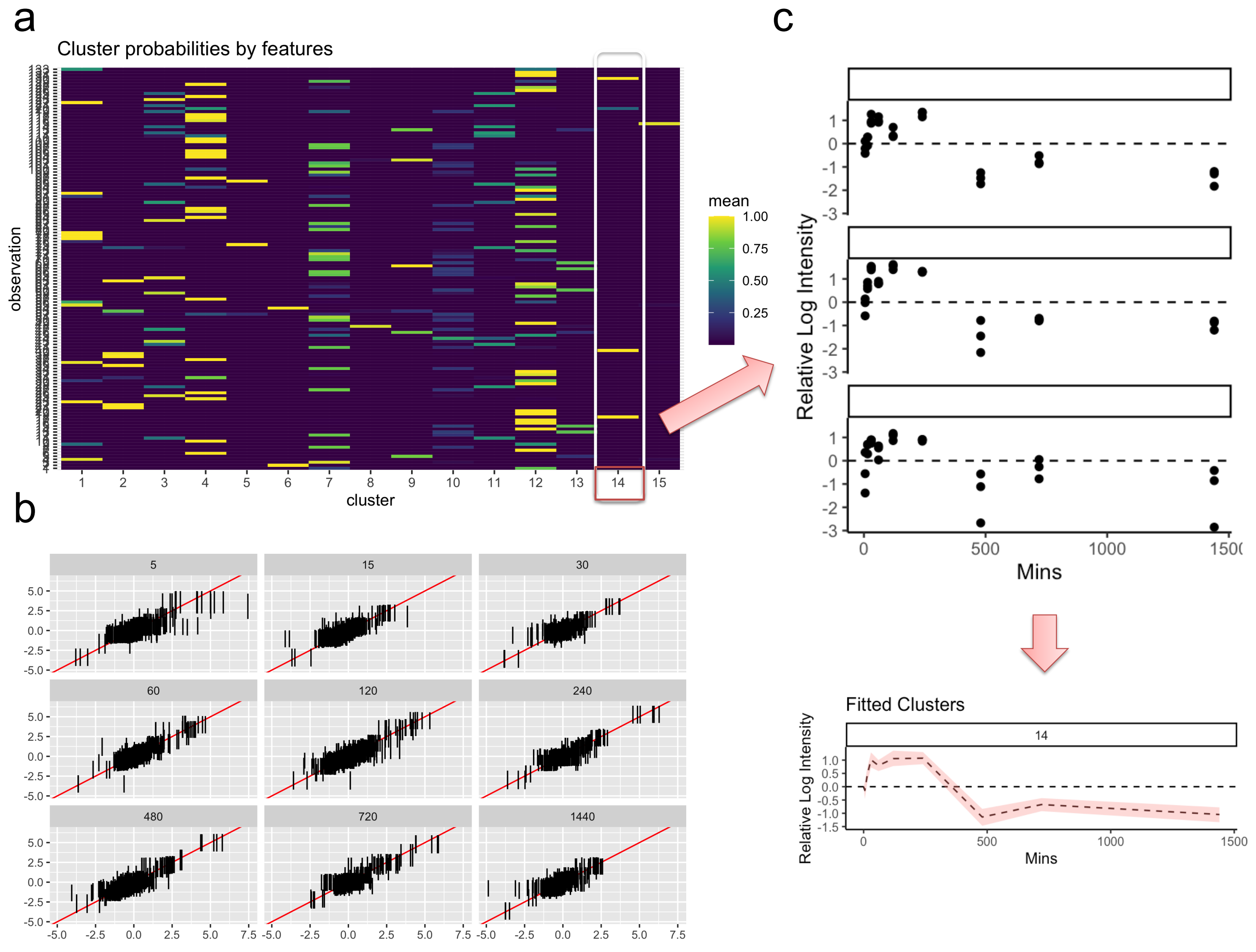
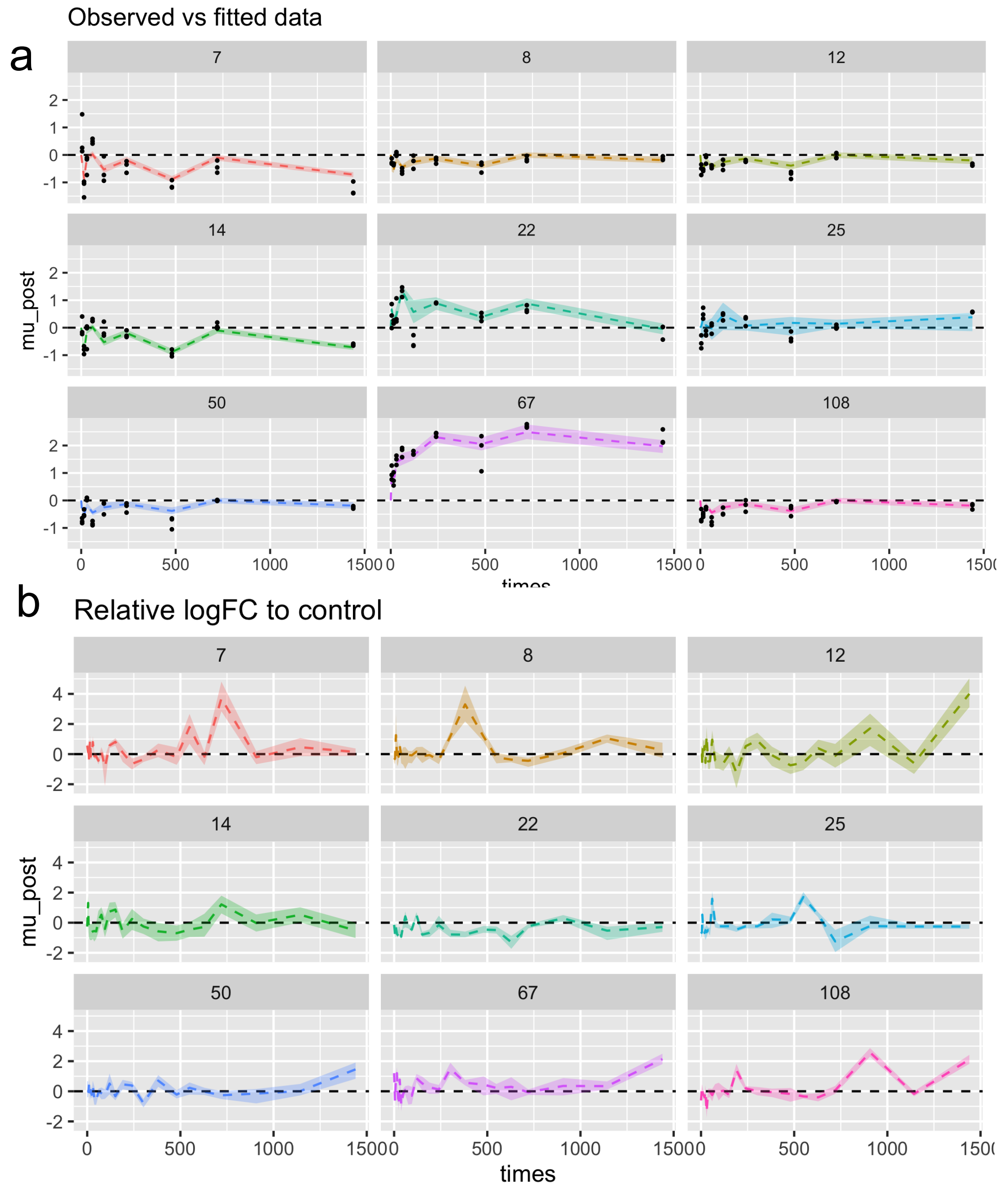


Figure 3. (a) 113 time-varying metabolites were assigned to each cluster. (b) Posterior predictive distribution check showed that simulated data closely matched the observed data at each time point. (c) Three metabolites exhibiting similar trajectory were grouped into cluster 14.



VALIDATION & POST-PROCESSING

- Estimate implied variation.
- Model is robust to extreme observations.
- Identify differentiated metabolites.

Figure 4. (a) Predicted metabolite intensities from the model, overlaid with observed data across the time course.

(b) Implied variations were shown in predicted, relative intensity (logFC) to control of above individual metabolites.

TREATMENT DIFFERENTIATION & SUMMARY STATISTICS

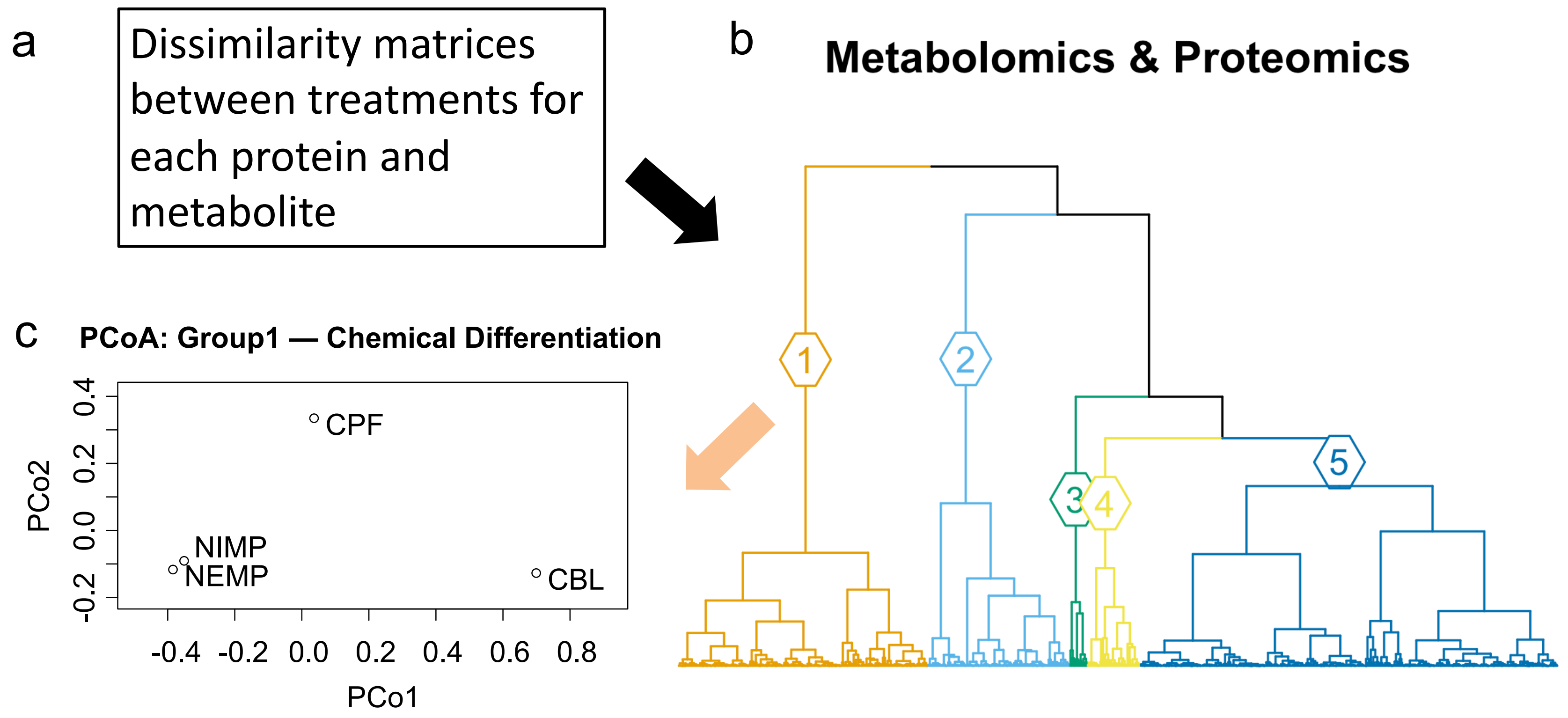


Figure 5. Predicted time-varying omic features are summarized: (a) Dissimilarity matrices in integrated mean squared error (IMSE) across the time points between different treatments. (b) Hierarchical clustering by dissimilarity matrices. (c) Correlated features are assumed to exhibit similar temporal trajectories.

Conclusion

We demonstrate a clustering strategy that accounts for limited observations, uncertainty, and the 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.