KNOWLEDGE DISCOVERY FROM *IN VIVO* METABOLIC DYNAMICS

by

YUE WU

(Under the Direction of ARTHUR S. EDISON & JONATHAN ARNOLD)

ABSTRACT

Living organisms have a biochemical network that responds perturbations, including food intake, environmental changes, and interactions with other organisms. Knowledge of regulation and dynamics in biological networks is crucial for understanding disease mechanisms and optimizing industrial fermentation. Metabolomics techniques can profile the metabolome and the biochemical network. However, most current research focuses on a single time point without capturing its complete dynamics and researchers often miss considerable proportions of the curated network. This is mainly because of difficulties in collecting time-series data and the lack of computational approaches for high-dimensional dynamics. This leaves us with an incomplete understanding of the network and its role in disease.

Here, I present a workflow to extract biological and chemical knowledge from the dynamic *in vivo* metabolome. The workflow is composed of experimental profiling of metabolic dynamics, feature extraction of complex data, and knowledge discovery from the time series. First, continuous *in vivo* metabolism by nuclear magnetic resonance (CIVM-NMR) was built to record the *in vivo* metabolome through time in multiple organisms, particularly *Neurospora crassa*. After perturbations in oxygen levels and

carbon sources, different response profiles were recorded and analyzed. From CIVM-NMR, we often collect multiple perturbation datasets, where each is composed of multiple spectra, and each spectrum at one time point has hundreds of peaks, thus producing a complex data structure. Second, I designed several computational approaches to extract features from complex datasets. I built Ridge Tracking-based Extract (RTExtract) to extract NMR features from the time-series spectra, even in cases of highly overlapped and crossing peaks. To improve accuracy in overlapping regions and promote automation, I also built spectral automatic NMR decomposition (SAND), which automates preprocessing and decomposes NMR spectra. With RTExtract and SAND, a complex NMR dataset can be reduced to a table of peaks at different time points. Third, I extracted biochemical knowledge regarding the *in vivo* metabolome from this high-dimensional time-series dataset by dimensionality reduction and network construction. An end-to-end workflow was built to extract knowledge from *in vivo* perturbed systems and the workflow will be applied to broader time-series studies, particularly in precision medicine.

INDEX WORDS: Metabolomics; Biological network; Time series; Metabolic modeling; NMR; *Neurospora crassa*; Metabolic quantification; *In vivo*; Data integration; Feature extraction

KNOWLEDGE DISCOVERY FROM *IN VIVO* METABOLIC DYNAMICS

by

YUE WU

B.S., Nanjing University, China, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

KNOWLEDGE DISCOVERY FROM *IN VIVO* METABOLIC DYNAMICS

by

YUE WU

Major Professor:	Arthur S. Edison
	Jonathan Arnold
Committee:	Heinz-Bernd Schuttler
	Juan Gutierrez
	Thiab R. Taha
	Suchendra M. Bhandarkar

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2022

DEDICATION

To all the people that give me five minutes when they do not have them and the ones that like to share and help when they do not have to.

ACKNOWLEDGEMENTS

I would first like to first thank my advisors, Arthur S. Edison and Jonathan Arnold. They allowed me considerable freedom when I was trying out new ideas. They were also supportive and offered help when I needed connections or suggestions. During my graduate school, I tried plenty of both good and bad ideas and failed a lot. They were always there supporting me during my trip to the dawn.

I also want to thank all my committee members for their technical support and diverse perspectives. In particular, I want to thank Heinz-Bernd Schuttler for all of his helpful technical discussions on Friday nights when he could have gone home.

People in my two labs have also been active and helpful. Discussions with Michael T. Judge helped me gain better perspectives on metabolomics and experiments, Gonçalo J. Gouveia always provided help with figure making, and Max Colonna was always supportive.

I also want to thank Karen Howard for her help on logistics, Laura Morris for all the computer equipment and resources I requested, and Sandra and April for help navigating departmental and graduate school regulations. Without support from you all, I would have had much less precious time for the deep research work.

I also like to thank all my collaborators. I'm lucky to have worked with all of you and have learned a lot about science and life. I'd also like to thank all people that I have asked for comment and feedback on my papers, abstracts, and applications. Your input has been invaluable.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION TO TIME-SERIES METABOLOMICS AND RELATED COMPUTATIONAL APPROACHES

Metabolome and Metabolic Network

Living organisms contain a dynamic reaction network, the metabolome, that connects different biological functions and fulfills multiple needs. This network links genetic information to phenotypes, responds dynamically to media environments (e.g., carbon sources), and closely indicates the current cellular state. In particular, perturbation in carbon sources and gene mutations has been studied in multiple organisms, and many associations have been found, including those of medical relevance (Chadeau-Hyam et al. 2010; Fuhrer et al. 2017; Judge et al. 2019). The metabolic network is also directly related to multiple diseases, industrial fermentation, and biofuel production (Abedi and Hashemi 2020; Hanahan and Weinberg 2011; Ivanov et al. 2013; McNerney et al. 2015; Show et al. 2015; B. Yu et al. 2019). For instance, the role of metabolism in cancer has been studied for decades (Hanahan and Weinberg 2011; Peng et al. 2018; Warburg 1956), and researchers have found links in central metabolism (DeBerardinis et al. 2007; Mehrmohamadi et al. 2014; Son et al. 2013) and cancer biomarkers (Bifarin et al. 2021; Maughon et al. 2022). Cohort studies have also been performed regarding the metabolome, presenting population diversities, associations with different diseases, and interventions (Bar et al.

2020; Beckonert et al. 2007; Holmes et al. 2008; Liang et al. 2020; B. Yu et al. 2019).

The metabolome is controlled by a complex network, composed of reactions, regulation, and transporters. For example, in *Neurospora crassa*, there are 1238 enzyme reactions, 190 transport reactions, 845 enzymes (among 9960 protein-coding genes), and 1357 metabolites (https://cyc.pnnl.gov) (Dreyfuss et al. 2013). The network is divided into different functional parts, but the different parts are highly interconnected. Through genetic sequencing, gene annotation, and database mapping, the reaction pathway can be curated in MetaCyc and KEGG (Kanehisa et al. 2010; Karp et al. 2002; Yandell and Ence 2012). However, information on regulation and dynamic response to perturbations is largely missing (A. M. Al-Omari et al. 2022). The main reason is the lack of perturbation experiments and difficulty in time-series metabolomics profiling. Computational approaches suitable for large datasets and high-dimensional time series are also needed. Currently, our understanding of the metabolic network is simplified and static, and this prevents us from correctly interpreting and integrating omics data, particularly metabolomics.

We need novel systematic approaches to advance our understanding of the metabolic network and its dynamics. First, we need experimental techniques to measure dynamic responses of the metabolic network to different perturbations. Second, efficient processing and feature extraction methods are needed for the complex dataset composed of hundreds of samples with diverse patterns. Third, new computational methods are needed to extract clusters and

biochemical knowledge from the high-dimensional dataset. The following three sections will cover the background.

Quantifying the Metabolome Through Metabolomics

Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are two popular approaches to studying metabolomics. When hyphenated with chromatography (e.g., high-performance liquid chromatography, HPLC), MS can detect thousands of features with high sensitivity (Bauermeister et al. 2022; Dunn et al. 2011). While NMR has less sensitivity, it provides non-invasive or even *in vivo* measurements that are crucial for studying metabolic dynamics (Edison et al. 2021). The capabilities of absolute quantification and standard annotation procedure through 2D NMR are also crucial advantages. Meanwhile, time-series sampling and quantification are often difficult for MS, though this has been relieved by in-line extraction and direct injection (Link et al. 2015). My thesis focuses on NMR-based analysis of *in vivo* metabolism, including data processing, feature extraction, and modeling.

Metabolomics is often divided into targeted and untargeted approaches. Targeted analysis focuses on specific compounds of interest (Patti et al. 2012). This is suitable for hypothesis testing experiments, where the target of interest is known a priori. Untargeted approaches profile the whole metabolome, analyzing both known compounds and unknown features. Most of the features are often unknown in the analysis, and they need to be computationally mapped to chemicals or associated with biological functions (Edison et al. 2021; Monge et

al. 2019). Hence, untargeted analysis is more complex but provides a global picture of the metabolome.

NMR has been used to profile individual biological samples and dynamic responses. The conventional approach to studying the dynamic response after perturbations is sampling at multiple time points on a fixed schedule, which is labor-intensive and introduces sampling variances. Flow NMR was built to monitor culture systems, where the media was pumped through the NMR probe (Friebel et al. 2019; Gonzalezmendez et al. 1982). Agar-embedding of cells was also used as it preserved the sample homogeneity for NMR and kept cells viable (Koczula et al. 2016). High-resolution magic angle spinning (HRMAS) provides an easier approach to measuring the cultured organism. Through a multiphase approach, metabolites, proteins, cell membranes, and exoskeletons were measured in fresh water shrimps (Mobarhan et al. 2016; Mobarhan et al. 2017a). We also published continuous *in vivo* metabolism by NMR (CIVM-NMR), where we measured dynamic metabolic responses of *Neurospora crassa* under aerobic and anaerobic conditions (Judge et al. 2019). Using less effort than conventional sampling, multiple time-series datasets were collected, each with hundreds of peaks and around 50 time points. Multiple perturbations were tested, including different carbon sources, oxygen availabilities, densities, and mutations (Yue Wu et al. 2022). We also tested isotope-labeling in CIVM-NMR, where different biochemical processes and fluxes of glucose were distinguished (Judge et al. 2019).

Time-series metabolic profiling can be readily applied to cohort samples. In cohort studies, the metabolome is often profiled through biofluids, including serum and urine, and the metabolite levels often respond to events like food intake, exercise, environmental factors, and diseases (Contrepois et al. 2020; Psychogios et al. 2011). Profiling metabolic dynamics can help us understand the relevant biological processes and eventually promote precision medicine through longitudinal studies of a single person (Ahadi et al. 2020; Rose et al. 2019; Sailani et al. 2020).

Computational Processing of NMR Metabolomics

Metabolic samples produce complex spectra that require multiple-step processing (Fig. 1.1). Biological samples have complex compositions, including hundreds to thousands of metabolites and some macromolecules (e.g., proteins) even after proper extraction. Through NMR experiments, we collect time-domain signals (free induction decay, FID), which are the sum of multiple decaying sinusoids. Instead of direct visualization, researchers often Fourier transform (FT) the FID into the frequency-domain data, where each decaying sinusoid is converted to a Lorentzian peak, and the manual inspection is much easier. However, each metabolite often has multiple Lorentzian peaks in the frequency domain and decaying sinusoids in the time domain. Therefore, NMR spectra often contain a considerable number of overlapping peaks after FT (Fig. 1.1). Additionally, macromolecules produce broad peaks, and solvents produce dominant peaks, and both often overlap with metabolite peaks of interest (Fig. 1.1). Experimental spectra also experience phase distortion, which will cause

peaks to deviate from the absorption mode and affect quantification. Correct quantification and annotation of metabolites need processing steps to resolve these aforementioned issues.

Preprocessing methods for NMR spectra have been developed to solve some of the aforementioned problems and are established. The process often includes FT, phase correction, baseline correction, solvent signal removal, and referencing (Jeffrey C. Hoch and Stern 1996).

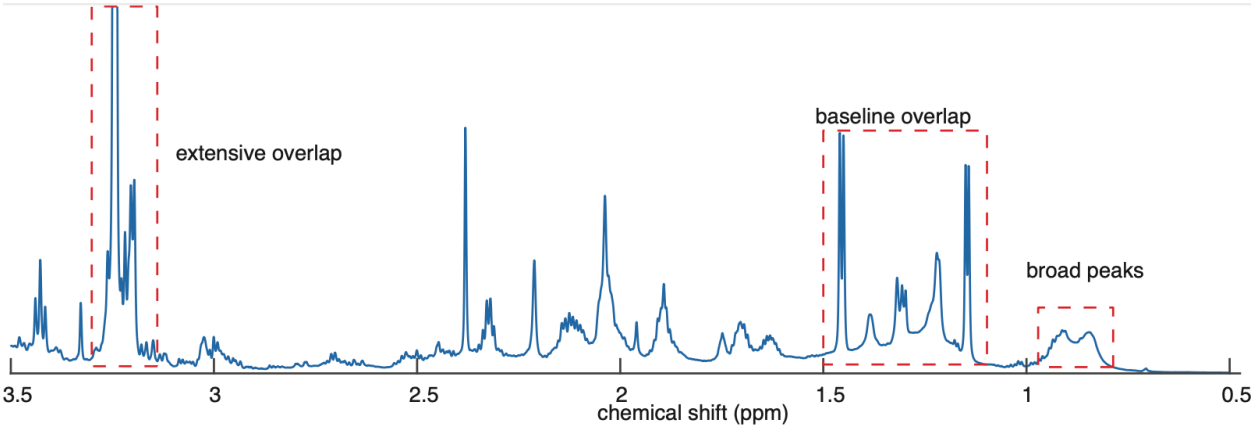


Figure 1.1: Example regions of experimental NMR spectra. An example NMR spectra of *Caenorhabditis elegans* is presented with difficult regions highlighted. The X-axis is chemical shift (ppm).

After preprocessing, the NMR spectra often display absorption mode peaks, a relatively flat baseline, and reduced solvent signals. The same peaks in different samples are close in chemical shift, except for pH-induced peak movement.

Preprocessing can be done automatically, but it often needs spectral-based manual refinement (F. Delaglio et al. 1995a). Overlap between spectral features is the remaining dominant problem to be solved.

It is still a challenge to quantify untargeted metabolomics data automatically and reliably. Among the hundreds of quantifiable peaks, many of them overlap with each other and with broad peaks, thus biasing the quantification from peak intensity or integral. Library-based peak fitting relies on reference libraries and cannot match features that are not in the library (Hao et al. 2012; Hao et al. 2014a; Ravanbakhsh et al. 2015). We implemented a new computational approach, spectral automatic NMR decomposition (SAND), to decompose untargeted metabolomics spectra into different peaks and will present it in detail in Chapter 4. Compared with similar approaches (J. C. Hoch 1989; Krishnamurthy 2013; D. V. Rubtsov and Griffin 2007; Denis V. Rubtsov et al. 2010), SAND improves automation and performs consistently for untargeted metabolomics of different biological samples.

Time-series NMR spectra add an additional time dimension and more complexities in processing. In particular, the same peaks in different samples or time points need to be matched for downstream analysis. Chemical shifts not associated with pH can often be mapped through alignment, while pH-induced peak movement is more complicated. Chemical shifts of many peaks change as a titration curve in response to pH and divalent ions (Takis et al. 2017; Tredwell et al. 2016b). Conventional alignment methods often fail in this difficult issue because contributing factors like pH are often unknown, and peaks cross each

other (Vu and Laukens 2013a). Fortunately, there is a simpler solution for time-series NMR, where pH changes continuously through time and the peaks move continuously like a ridge. We designed Ridge Tracking-based Extract (RTExtract) to quantify peaks through time (More details in Chapter 2). This solution relies on the continuity constraint imposed by time but can be expanded when such constraint is available because of other reasons, including pH titrated samples or spectra reordering through leading peaks. In the latter case, cohort metabolic samples with unknown pH can be reordered by chemical shifts of pH-affected peaks, and many peaks will form similar titration curves that can be tracked and quantified.

Data Integration in Metabolomics

High-dimensional time series can be obtained from CIVM-NMR experiments and RTExtract feature extraction. However, efficient interpretation of such datasets has not yet been achieved. The ultimate goal is to clarify the role of metabolites in biological functions and diseases, and this requires multiple subgoals to be resolved first.

Feature annotation needs to be improved in order to expand the scope of analysis. Many features in metabolomic studies remain unknown, even after 2D experiments and manual investigation (Judge et al. 2019; Monge et al. 2019). The annotation also only covers a small subset of the metabolic pathway (Fig. 1.2). In the CIVM-NMR publication, more than half of the features remains unknowns and they cover less than 10% of the metabolic pathway (Judge et al. 2019). This is because of incomplete libraries and the intensive overlap in NMR

spectra (Fig. 1.1). The annotation process is also labor-intensive and expert-dependent. Therefore, automation in proposing candidate annotations and database retrieval is highly needed. Statistical total correlation spectroscopy (STOCSY) has been commonly used to find peaks chemically associated with a targeted peak (Cloarec et al. 2005). STOCSY can be expanded by automatically clustering metabolic features into chemical associated groups measured by correlation. Extracted 1D NMR features from RTExtract and SAND are ideal input to the clustering approach and can be used to build an automatic workflow for annotation. 2D NMR improves the confidence in annotation and complex mixture analysis by NMR (COLMAR) provides multiple tools for validation (Bingol et al. 2016). Matching currently unknown features to metabolites can improve the biological impact of NMR metabolomics.

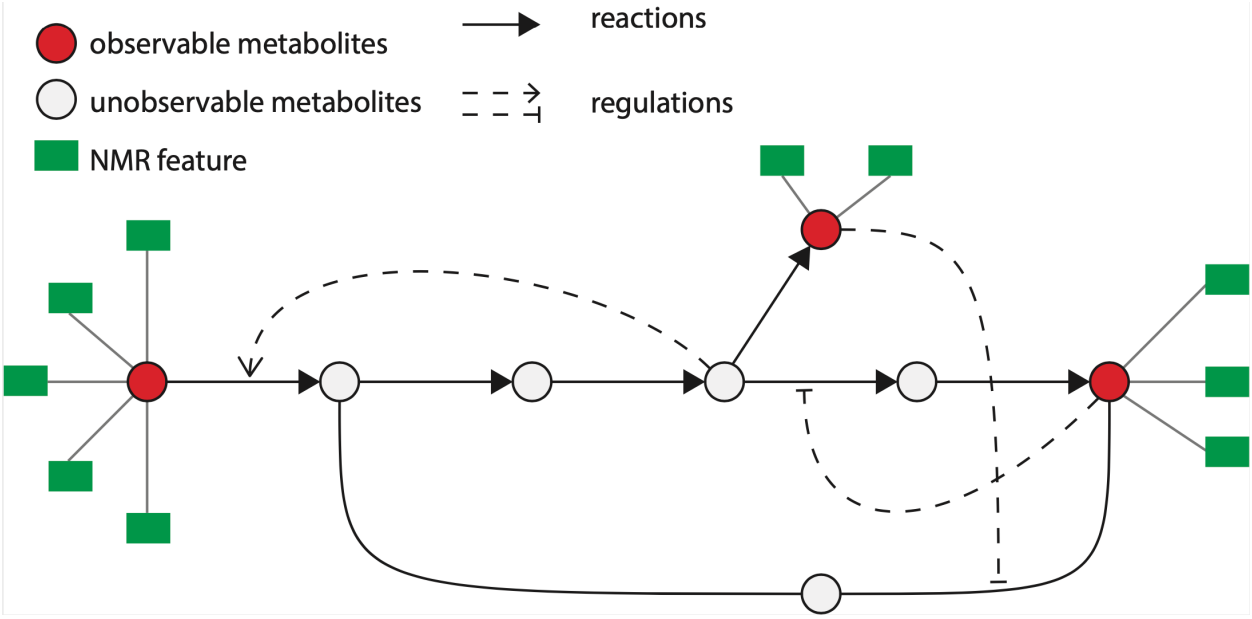


Figure 1.2: A diagram of experimental observation of a metabolic network. In a pathway, a small section of compounds (red circles) can be measured in an experiment while others (empty circles) remain unobservable. Activation and inhibition regulations (dotted lines) are often missing in the database. Each measured compound often has multiple NMR peaks (green rectangle).

Dynamics need to be considered when analyzing metabolic data. The *in vivo* metabolome responds dynamically to different perturbations. Concentrations of metabolites, reaction rates, regulation, and levels of pathway functions change on the scale of hours or even seconds (Judge et al. 2019). However, many computational approaches depend on fixed time points, steady-state assumptions, or a simplified model without considering the kinetic structure (Ashburner et al. 2000; Edwards et al. 2002; Fell and Small 1986; Subramanian et al. 2005). Researchers also built mechanistic models to capture, explain, and compare dynamic behaviors. Direct modeling is a nice choice if the model complexity is constrained, enough data are provided, and partial observation of the network is limited (Battogtokh et al. 2002; Brown and Sethna 2003; Pfister et al. 2019a; Raue et al. 2013). However, metabolic networks are the opposite of those assumptions, as they contain thousands of reactions (equations) and reactants (variables), receive limited perturbation experiments, and are largely experimentally unobservable (Fig. 1.2). These imperfections make direct modeling and optimization computationally intensive and less informative. An

alternative approach is to extract statistical dependencies between NMR features without explicit modeling of the huge network. Partial correlation has shown promising results in finding biochemical connections when features are annotated and combined into compounds (de la Fuente et al. 2004; Krumsiek et al. 2011). Additionally, a recent publication presented CausalKinetiX to extract stable and predictive kinetic associations between time series (Pfister et al. 2019a).

Metabolic features need to be associated with biological functions. The success of this goal relies on the solution of the former two goals, compound annotations and proper dynamic representations, but also has its own difficulties. The species-specific information (e.g., regulation) in the context of metabolic pathways is often missing (Fig. 1.2). This complicates analysis and is one of the reasons that this effort is needed. The model topology built from pathway knowledge tends to be incomplete, and the wrong assumption affects modeling results (Fig. 1.2). Researchers have tried to search for the correct model topologies, but it is still a computationally hard endeavor (Meyer et al. 2014). This further argues against direct modeling and favors empirical studies, where novel biological associations are extracted from the data without strong assumptions on the network. New developments in graph modeling (Nelson et al. 2019), time series association (Pfister et al. 2019a), and dimensionality reduction (G. Montana et al. 2011a; Ramsay and Silverman 2005; Ramsay et al. 2009; Yue Wu et al. 2022) provide approaches to extracting knowledge from high-dimensional time series.

Description of Study

In Chapter 2, I present CIVM-NMR, an efficient experimental approach to recording time-series metabolic dynamics. In Chapters 3 and 5, two different but related methods for data processing and feature extraction are presented. In Chapter 3, I design and implement RTExtract to efficiently extract NMR features in time-series spectra, even in the case of highly overlapped and crossing peaks. In Chapter 5, I provide the next-step solution, SAND, to the spectral overlap problem. SAND applies to a broad range of 1D NMR data and enables further automation in NMR metabolomics. Based on the experimental method and computational workflow, high-dimensional metabolomics time series are collected under different conditions. In Chapter 4, I provide empirical approaches to resolving the complex datasets and extracting biochemical and chemical knowledge.

CHAPTER 2

CONTINUOUS *IN VIVO* METABOLISM BY NMR [1]

[1]: Wu Y, Judge MT, Tayyari F, Hattori A, Glushka J, Ito T, Arnold J, Edison AS. Continuous in vivo Metabolism by NMR. Frontiers in Molecular Biosciences. 2019;6(26).

Reprinted here with permission from the publisher.

Foreword

This chapter is reprinted from Wu Y, Judge MT, Tayyari F, Hattori A, Glushka J, Ito T, Arnold J, Edison AS. Continuous in vivo Metabolism by NMR. Frontiers in Molecular Biosciences. 2019;6(26) and is available at https://www.frontiersin.org/articles/10.3389/fmolb.2019.00026/full. Michael T. Judge and I were equally contributing co-first authors. My contribution to this work consisted of: (1) building MATLAB script to process, quantify, analyze and visualize the data, (2) building tutorials on compound ridge quantification for future users, (3) mapping feature extraction and compound annotation and accessing the quality, (4) interpreting the biological process, and (5) writing and editing the manuscript and addressing reviewer comments. The contribution of Michael T. Judge is as follows: (1) preparing *Neurospora* samples, (2) performing *in vivo* NMR experiments and preprocessing the data, (3) contributing to the MATLAB script and visualizing the results, (4) interpreting the biological process, and (5) writing and editing the manuscript, and addressing reviewer comments. Arthur S. Edison and Jonathan Arnold designed the project goal, edited the manuscript, and responded to the reviewers. Fariba Tayyari contributed to spectral annotation by 2D experiment and analysis. Ayuna Hattori and Takahiro Ito performed *in vivo* carbon-labeled measurements on human cells. John Glushka advised and supported NMR experiments and data preprocessing. The research was supported by National Science Foundation awards (NSF 1713746 and NSF ERC 1648035), Georgia Research Alliance, American Cancer Society

(RSG-17-03201DDC), and National Institute of Health (1S10OD021623-01). The supplementary materials in the chapter are listed in Appendix A.

Abstract

Dense time-series metabolomics data are essential for unraveling the underlying dynamic properties of metabolism. Here we extend high-resolution-magic angle spinning (HR-MAS) to enable continuous in vivo monitoring of metabolism by NMR (CIVM-NMR) and provide analysis tools for these data. First, we reproduced a result in human chronic lymphoid leukemia cells by using isotope-edited CIVM-NMR to rapidly and unambiguously demonstrate unidirectional flux in branched-chain amino acid metabolism. We then collected untargeted CIVM-NMR datasets for *Neurospora crassa*, a classic multicellular model organism, and uncovered dynamics between central carbon metabolism, amino acid metabolism, energy storage molecules, and lipid and cell wall precursors. Virtually no sample preparation was required to yield a dynamic metabolic fingerprint over hours to days at ~4-min temporal resolution with little noise. CIVM-NMR is simple and readily adapted to different types of cells and microorganisms, offering an experimental complement to kinetic models of metabolism for diverse biological systems.

Introduction

Metabolic time-series data are invaluable for the development and validation of high-quality models that accurately describe the dynamics of metabolism (Link et al. 2014; Giovanni Montana et al. 2011b; Sefer et al. 2016). Information about the metabolic state of an organism typically requires extensive

time, resources, and sample material. As such, researchers must choose between variables such as the number of replicates, the number of time points, and the time resolution for time-series. Furthermore, traditional metabolomics experimental designs face the challenges of extraction biases (Sitnikov et al. 2016) and the confounding of biological and analytical variance (Anaraki et al. 2018). While many studies employ sample preparation and extraction approaches effectively, direct or *in vivo* measurements are fundamentally simpler to obtain and interpret. Likewise, while carefully designed (Rhoades et al. 2017) and executed studies with large sample sizes undeniably yield powerful insights into the dynamics of biological systems (Cannon et al. 2018; Krishnaiah et al. 2017; Sengupta et al. 2016), continuous and repeated measurements on the same living sample are invaluable for monitoring and confirming these dynamics.

Small molecules and their fluxes have been measured *in vivo* using NMR (Bastawrous et al. 2018b), and methods have recently been developed that begin to address the need for a continuous time dimension in metabolomics data. For example, long-standing flow NMR techniques allow monitoring of secretion and uptake of extracellular metabolites for organisms grown in liquid culture (Bastawrous et al. 2018b). Link et al. recently achieved high temporal resolution on many metabolites by developing an automated real-time metabolomics platform that samples liquid cultures of single cells and directly injects them onto a time-of-flight mass spectrometer every 15-30s (Link et al. 2015). The group have more recently probed the interactions between biomass synthesis and cell division in *E. coli* using this method (Sekar et al. 2018). Koczula et al. conducted

in vivo measurements changes in media composition with 4-8 min resolution for chronic lymphoid leukemia. Sedimentation and line broadening are major factors that limit standard NMR measurements of complex samples like cells. Koczula et al. were able to mitigate sedimentation by immobilizing the single cells in agarose (Koczula et al. 2016).

Alternatively, HR-MAS enables high-resolution NMR measurements on mixed-phase samples such as tissues (Beckonert et al. 2010), or more recently, living organisms (Augustijn et al. 2016; Bastawrous et al. 2018b; Mobarhan et al. 2016; Righi et al. 2014; Sarou-Kanian et al. 2015) with minimal line broadening. In this study, we extended HR-MAS to real-time continuous *in vivo* measurements of metabolism in cells. Using isotope editing, CIVM-NMR was able to reproduce and more directly observe a surprising branched-chain amino acid (BCAA) flux result reported last year in human myeloid leukemia cells (Hattori et al. 2017). We found that CIVM-NMR was not only easier but faster and more conclusive than traditional approaches for flux measurements in human cell cultures. We then applied CIVM-NMR to the multicellular filamentous fungus, *N. crassa*, in both aerobic and anaerobic environments. We observed highly reproducible dynamics in central carbon and amino acid metabolism with ~4 min resolution over 11 hours. The continuous nature of these measurements facilitated metabolite annotation, and semi-automated peak tracing provided relative quantification of known and unknown compounds. We developed several new MATLAB functions and workflows, freely available through GitHub, for the analysis and visualization of these novel data. As CIVM-NMR can be applied

widely to cells, tissues, and small multicellular organisms, it enables new opportunities in fields such as developmental and chronobiology for monitoring high-resolution metabolic time-series data. Importantly, it will enable more robust and experimentally-based kinetic metabolic models for diverse biological systems.

Materials and Methods

Human leukemia cell culture and preparation for HR-MAS NMR

The human BC-CML cell line K562 was obtained from ATCC, and cell line authentication testing was performed by ATCC-standardized STR analysis to verify their identity. After cell counting and washing with PBS, K562 cells were resuspended and labeled in a custom-made Iscove's modified Dulbecco's Medium (IMDM) without BCAAs supplemented with 10% dialyzed FBS, 100 IU/ml penicillin, 100 µg/ml streptomycin and the following amino and keto acids. For ^{13}C-KIV (keto-isovalerate) tracer experiment, isoleucine, leucine and valine were supplemented at 170 µM. For ^{13}C-valine tracer experiment, isoleucine, leucine and KIV were added at 170 µM. Cell suspension (54 µl) was loaded in a clean 4 mm diameter zirconia HR-MAS rotor (Bruker BioSpin), and then either [(U)-^{13}C]-ketoisovalerate or [(U)-^{13}C]-valine solution in D_2O was added to a final concentration of 170 µM. Rotor was closed with a Kel-F rotor cap (Bruker BioSpin).

Preparation of growth media and slants for *N. crassa*

Ingredients for Vogel's Media (3 % Glucose) (Glucose, 0.334 M; Biotin, 0.614 µM; Arginine, 1.95 mM; Na_3 Citrate, 9.74 mM; KH_2PO_4, 36.7 mM; NH_4NO_3,

25.0 mM; $MgSO_4$, 0.811 mM; $CaCl_2$, 0.680 mM; $ZnSO_4$, 34.8 µM; Fe $(NH_4)_2$ $(SO_4)_2$, 5.10 µM; $CuSO_4$, 2.00 µM; $MnSO_4$, 0.592 µM; H_3BO_3, 1.62 µM; Na_2MoO_4, 0.413 µM) were dissolved in ddH_2O in a large glass bottle, filter-sterilized (0.22µm Steritop Threaded Bottle Top Filter, 500mL, Millipore EMD), stirred using a magnetic stir bar, then aliquoted into clean, sterile 500-mL bottles. Ingredients for Vogel's Media with Agar (same as above, with the addition of 1.5 % agar, w/v, and using 1.5 % Glucose, w/v) were combined in a beaker. Agar was dissolved by heating in a microwave oven. The dissolved mixture was aliquoted to 15-mL or 5-mL glass test tubes, stoppered with cotton, and sterilized by autoclaving.

Vogel's media for NMR and wash solution

2X Vogel's Media (minus glucose), DSS solution, and D_2O were combined to make a concentrate, which was split into two aliquots. To prepare Vogel's Media for NMR (1.5 % Glucose), filter-sterilized D-glucose solution (0.5mg/µL) was added to the smaller aliquot to a final composition of Glucose, 0.167 M; DSS, 1mM; Biotin, 0.614 µM; L-arginine, 1.95 mM; Na_3 Citrate, 9.74 mM; KH_2PO_4, 36.7 mM; NH_4NO_3, 25.0 mM; $MgSO_4$, 0.811 mM; $CaCl_2$, 0.680 mM; $ZnSO_4$, 34.8 µM; Fe $(NH_4)_2$ $(SO_4)_2$, 5.10 µM; $CuSO_4$, 2.00 µM; $MnSO_4$, 0.592 µM; H_3BO_3, 1.62 µM; Na_2MoO_4, 0.413 µM in 95 ddH_2O/5 D_2O (v/v). Wash Solution was prepared by adding ddH2O in place of D-glucose solution to the larger aliquot.

Preparation and storage of *N. crassa* conidial suspension

A frozen bd1858 (A) stock obtained by the Fungal Genetics Stock Center (McCluskey et al. 2010) was used to inoculate two Growth Slants (Vogel's Media Agar, 1.6% Glucose w/v, 3 mL in 15 mL glass test tubes stoppered with sterile cotton plugs). These were incubated for 2 days at 30°C, then placed under a benchtop lamp at 25 °C for 2 days to induce maturation of conidia. Conidia were collected from both tubes sequentially by suspension in 12 mL Vogel's Media (no glucose) and filtration through sterile cotton. Concentration of the resulting conidial suspension was found to be 6.47 x 10^7 cells/mL using a Nexus Cellometer Auto 2000 (Nexelcom Bioscience; Lawrence, MA, USA). The conidial suspension was kept at 4 °C over the course of the experiments (4 weeks).

Growth of *N. crassa* mycelia

Vogel's Media (50 mL, 3 % Glucose w/v) in a 250-mL Erlenmeyer flask, covered in aluminum foil was inoculated under aseptic conditions with Conidial Suspension to a total concentration of 2.7 x 10^4 cells/mL, (21 µL Conidial Suspension). Liquid cultures were grown with orbital shaking (~237 rpm) at room temperature (~25 °C) under constant cool white light (7 µmol L^{-1} s^{-1} m^{-2}) for 32h. At that point mycelia consistently formed a single, cohesive mass. The entire culture was transferred to a 50 mL Conical Tube (Sarstedt; Newton, NC, USA) for transport to the NMR facility (15-30min).

Preparation of *N. crassa* mycelia

Under aseptic conditions, a section of mycelium from the edge of the main mycelial mat was cut off using a sterile tube cap and trimmed to fit the volume of

approximately 126 µL using a pre-marked microcentrifuge tube. Mycelia were handled from this point using clean, sterile tweezers (cleaned with 70 % EtOH on a lint-free single-ply lab tissue (kimwipe) and dried in an aseptic environment). The section of mycelium was then patted dry on autoclaved filter paper (Whatman Filter Paper #3; GE Healthcare, USA) atop a layer of folded kimwipes, and was washed by placing in a sterile microcentrifuge tube containing 1 mL Wash Solution and vortexing briefly (~10 s) until the mycelium had fully absorbed the media. Washing was repeated with fresh Wash Solution for a total of 4 washes. The mycelium was reduced to ~63 µL (0.9 x volume of rotor + plug), measured in a second microcentrifuge tube pre-marked to that volume. The mycelium was pat-dried in a sandwich of sterile filter paper folded into kimwipes, pressing firmly three times (until no liquid spots were visible on the filter paper). The dried mycelium was then weighed in a separate microcentrifuge tube. The dry mycelium was 9.04-10.13 mg in our experiments (µ = 9.62 mg; SD = 0.32 mg). We observed a reduction in mass of ~30 % as conidia, loose filaments, and other debris are removed along with waste products and glucose during wash steps. In our hands, the prep process took between 4 and 13 min., during which time the organism was immersed in a low-glucose environment.

Loading *N. crassa* mycelia into the rotor

The dried, weighed mycelium was then placed in a microcentrifuge tube containing fresh Vogel's Media for NMR (500 µL, 1.5 % Glucose), and vortexed briefly until the mycelium had fully absorbed the media. The mycelium was then transferred to a third, pre-marked microcentrifuge tube (63 µL). By

adding/removing media, the volume was adjusted to the 63 µL volume mark. Sterile tweezers were used to transfer the mycelium to a clean 4 mm diameter zirconia rotor (Bruker BioSpin) cleaned by rinsing with bleach solution, tap water, 70 % ethanol, tap water, and ddH_2O x 4). The mycelium was pushed to the bottom, taking caution not to lose liquid. The remaining liquid in the tube was added to the rotor and one tweezer prong was used to position the mycelium to remove larger air bubbles, although small bubbles occurred with no issues in the NMR. A Teflon sealing plug (Bruker BioSpin) was then inserted to ~2 mm below the edge of the rotor. For the aerobic condition, a Kel-F rotor cap (Bruker BioSpin) modified with a 0.016-inch diameter hole drilled using a lathe was lined on the inside with three layers of Rayon breathable microplate sealing tape (QuickSeal Breathable Film, Thomas Scientific, USA) to prevent spore escape. The cap was fully inserted to push the sealing plug into its final position. The cap was then removed, and the insides of the cap and plug were inspected to ensure that no liquid was lost and that an airspace existed between the plug and the sample. The rotor was then re-capped, the bottom edge marked with a permanent marker, and dropped into the bore of the magnet (cap facing up). In our hands, this process typically takes 15-30 min. For the anaerobic condition, media was added to fill all airspaces and an unmodified cap was used to prevent gas exchange. For the ^{13}C labeling experiment in partially anaerobic conditions, an airspace was left and fresh Vogel's Media for NMR was prepared with 3% (w/v) ^{13}C-labeled Glucose (99% labeled; Cambridge Isotope Laboratories; Tewksbury, MA, USA) was used in place of 1.5% (w/v) glucose.

NMR parameters

For human ML cell experiments, a hsqcetgpsisp.2 gradient Heteronuclear Single Quantum Coherence Spectroscopy (HSQC) experiment run as a 1D experiment was used with the following parameters: Data Points: 7272. Dummy Scans: 4 at the beginning of the run. Number of Scans: 128 /timepoint. O1 offset: 4.699 ppm. O2 offset: 30.000 ppm. Acquisition Time 0.3999600 s. Delay: 1.5 s. fid Resolution: 2.500250 Hz. Receiver Gain: auto (101). Temperature: 298 K = 25 °C. Spinning Speed: 3100 Hz. A standard noesypr1d protocol (Bruker) was used for *N. crassa* non-labeled real-time metabolomics measurements. The following parameters applied to all samples and timepoints: Data Points: 42856. Dummy Scans: 8 at the beginning of the run. Number of Scans: 64 /timepoint. Spectral Width 11904.762 Hz. Acquisition Time 1.7999520 s. Delay: 1.5 s. fid Resolution 0.555570 Hz. Receiver Gain: auto (101). Temperature: 298 K = 25 °C (calibrated using a deuterated methanol standard (Van Geet 1970)). The following parameters were optimized for each sample: O1 offset for water suppression: 2817.24 - 2818.24 Hz. PWL9 water suppression power: 43.87 - 44.42 dB (μ = 44.23 dB, SD = 0.19 dB). P1 pulse width: 12.49 - 13.30 μs (μ = 12.78 μs, SD = 0.29 μs). Spinning Speed: 6000 Hz. Notably, this variation in pulse width between samples manifested as a difference in temporal resolution (i.e. longer pulse widths resulted in time points slightly farther apart). The effect was measurable (on the order of minutes) over hundreds of measurements. The average experiment took 4.23min +/- 0.004min (SD).

For measurement of ^{13}C in the labeled glucose experiment, a modified hmqc Heteronuclear Multiple-Quantum Correlation (HMQC) experiment with additional phase cycling was used. The following parameters were used: Data Points: 3636. Dummy Scans: 8 at the beginning of the run. Number of Scans: 64 /timepoint. O1 offset: 2826.24 Hz. O2 offset: 12070.62 Hz. Acquisition Time 0.2545200 s. Delay: 1.5 s. fid Resolution 3.928964 Hz. Receiver Gain: auto (101). Temperature: 298 K = 25 °C. Spinning speed: 6000 Hz. All Bruker parameter files are available with the raw data at http://www.metabolomicsworkbench.org/.

Automated data acquisition and post-experiment sample preparation

For human ML cells, spectra were collected sequentially using the multizg command in TopSpin (v4.0.1; Bruker).

For *N. crassa* samples, the noesypr1d experiment, optimized for the sample, was imported into IconNMR in TopSpin (v4.0.1; Bruker). The solvent was set to "D2O_H2O+salt". The "iterate" command was used to queue 1024 identical, sequential noesypr1d experiments (each taking ~4.6 min) on a Bruker NEO equipped with a 4-mm CMP probe. Dummy scans were only implemented for the first experiment. Experiments generally ended after ~12 h, though some were allowed to continue as long as 37 h. By spinning *N. crassa* at 6 KHz, spinning sidebands (Maricq and Waugh 1979) were eliminated in the spectral region of 0-10 ppm. At the end of each run, the mycelia were transferred from the rotor to a sterile microcentrifuge tube with clean, sterile tweezers. All liquid from

the rotor was also transferred to the tube. This was either extracted and assessed for growth immediately, or was allowed to sit on the bench for one day.

Survival assessment

Sterile tweezers were used to tear a piece of mycelium from the rotor contents; this was used to inoculate a growth slant. All growth slants were assessed for 24 h or longer post-inoculation for growth. Photographs were taken using a 16MP digital camera on an LG G5 cell phone in Manual Mode.

Extraction

The remaining rotor contents were transferred with a pipette to a microcentrifuge tube containing a mixture of zirconia beads (1 mm, 167 µL or ~375 mg; 0.7 mm, 334 µL or ~1314 mg; 500 µL total) on dry ice. The old tube was rinsed by briefly vortexing with 800 µL MeOH (80 % in ddH_2O), which was added to the beads. This mixture was either processed immediately or frozen on dry ice for up to 3 days. Contents were twice homogenized on dry ice for 180 s @1800 rpm using a MP FastPrep 96 (MP Biomedical; USA) adapted for microcentrifuge tubes, adding dry ice each time. The homogenate was centrifuged at 14k rpm at 4 °C for 5 min (18220 x g; Centrifuge 5417C; Eppendorf, USA). The supernatant was transferred to a separate microcentrifuge tube and kept on dry ice while the pellets were back-extracted with 500 µL MeOH (80 %), homogenized once for 180 s @1800 rpm, and centrifuged an additional 5 min. Supernatants from both extractions were combined, then dried to completion in a CentriVap Concentrator/CentriVap Cold Trap -105 °C system (Labconco, Kanasas City, MO, USA) for 4-6 h. Pellets for two samples were

combined during resuspension in D_2O (DSS, 1/9 mM) for each condition. Two replicates from each condition were pooled and pipetted into 1.5 mm NMR tubes (Norell; Morganton, NC, USA).

Annotation

For each pooled sample representing the anaerobic and the aerobic conditions, noesypr1d, ^{13}C-HSQC, Total Correlation Spectroscopy (TOCSY), and ^{13}C-HSQC-TOCSY spectra were collected on a 600 MHz Bruker magnet equipped with a 5mm cryoprobe and an Avance III HD console at the University of Georgia NMR Facility. 2D data were processed in nmrPipe (System Version 9.4 Rev 2017.340.17.07 64-bit) and submitted to COLMARm (Bingol et al. 2016) for putative compound identification. After manual inspection, metabolites were assigned a confidence level ranging from 1 to 5, with 5 being the highest. The scale is defined (Jacquelyn M Walejko et al. 2018b) as follows: (1) putatively characterized compound classes or annotated compounds, (2) matched to literature and/or 1D reference data such as HMDB (Wishart et al. 2007) and BMRB (Ulrich et al. 2008) (3) matched to HSQC, (4) matched to HSQC and validated by HSQC–TOCSY (COLMARm (Bingol et al. 2016)), and (5) validated by spiking the authentic compound into sample. Identifications from extracted 1d spectra were manually mapped to real-time *in vivo* noesypr1d data. An additional score was assigned to each mapped compound: 0 (unannotated), 1 (annotated only), 2 (qualitatively assessed), or 3 (relatively quantifiable) in the real-time data. This score depended on number of observed peaks, baseline, peak overlap, and sensitivity. Both metabolite confidence levels are reported in Table S1. All raw

and processed data files are available at http://www.metabolomicsworkbench.org/ and matching can be run on COLMARm (Bingol et al. 2016) directly.

Batch processing in nmrPipe for *in vivo* NMR data

Parameters were optimized based on agreement between spectra from several time points for a given sample. A custom bash script ran nmrPipe (Frank Delaglio et al. 1995b) using the optimized parameters on all spectra for a given sample. This script included all necessary nmrPipe commands for file conversions and NMR data processing. In brief, the following were implemented: line broadening, Fast Fourier Transform, 0- and 1st-order phasing, end removal, and baseline correction using automatic polynomial fitting. All raw data, parameter files and code are available at http://www.metabolomicsworkbench.org/.

Additional processing in MATLAB for *in vivo* NMR data

For each sample, custom scripts were written in MATLAB R2017b (The MathWorks, Inc., Natick, Massachusetts, USA), to load the processed spectra, ppm vectors, and measurement start times from .ft and .acqus files. Spectra were then referenced to DSS semi-automatically, stored as a matrix, and saved as a MATLAB workspace in .mat format. Using custom MATLAB scripts, .mat files from individual experiments were combined into a "sampleData" structure. Metadata (e.g. condition, pulse width, time shift between inoculation and start time) were added to each sample by manual entry or by automated retrieval from the Bruker acqus files for each sample. Spectral ends outside of [-0.5,10] ppm

were removed. The spectral region containing the water signal [4.7,5] ppm was replaced by zeros. Measurements for time points >11 h were removed in all experiments for consistency. Each spectrum was normalized to its DSS peak intensity as a formal step to allow for relative quantification. Finally, every three spectra were summed starting from the first timepoint for improved signal-to-noise. The resulting structure was saved as a .mat file (~2 Gb). All data and scripts are available at http://www.metabolomicsworkbench.org/ and at https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA.

Relative quantification of NMR resonances

A combination of a Gaussian smoothing filter with user-defined sigma in the ppm and time dimensions and peak picking script was used to identify peak maxima for a given region of ~0.5-1 ppm in a given sample, allowing some noise to be picked. Agglomerative clustering based on single linkage of Euclidean distances was then used to cluster the picked points in the chemical shift (ppm), time, and intensity space. Weights for each dimension in the clustering, as well as the number of clusters, were manually optimized for each region and sample. Clusters were quality-controlled by interactive visual inspection. If multiple ridge points existed for the same time, the one with highest intensity was retained. Peak positions at temporal gaps were estimated using linear interpolation between the two closest existing ridge points. Ridges on the smoothed data were mapped to the unsmoothed data for each time point by choosing the maximum within a small window around the peak position obtained from the smoothed data. A window size of 10 indices (~2.9 x 10^{-3} ppm) worked for all but a few

ridges, whose optimal mapping windows ranged between 6 and 60 indices (between 1.7 x 10^{-3} and 1.7 x 10^{-2} ppm). All ridges were visually inspected for good tracing, well-defined peaks, and minimal overlap by plotting on real spectra. To combine the trend information from multiple ridges annotated to the same compound, intensities of constituent ridges were scaled such that the ridge means across the time points shared by the highest number of ridges were equal. Lastly, the mean across scaled ridges at each time point was taken, yielding a single composite trajectory for each compound.

Titration of a citrate standard for estimation of *in-vivo* pH changes

A 10 mM solution of citric acid (A104-500; Fisher Scientific, USA) containing 1mM DSS reference standard was prepared, and 600 µL were added to a 5 mm NMR tube (Norell; Morganton, NC, USA). The pH of the solution was adjusted in-tube in ~0.25 pH increments by addition of 0.5-2 µL volumes of dilutions of concentrated NaOH and HCl and four rounds of inversion and vortex mixing. For each pH point, exact pH was measured in-tube using a calibrated accumet AB150 pH meter (Fisher Scientific, USA), then a 1D noesypr1d spectrum was collected (DS = 2; NS = 16) on a 600 MHz Bruker magnet equipped with a 5mm cryoprobe and an Avance III HD console at the University of Georgia NMR Facility. Data were phased and referenced to DSS in TopSpin (v3.5pl7; Bruker). Custom Matlab scripts were used to obtain the most upfield citrate peak position for each pH. A 3rd-order polynomial was fit to the positions (R^2 > 0.99) and used with the ridge belonging to the same peak to estimate the pH of each culture at each timepoint.

Results

Isotopic CIVM-NMR measurements confirm unidirectional KIV-to-valine flux in ML cells

Branched-chain amino transferase-1 (BCAT1) is a reversible enzyme, but in most cells the reaction degrades BCAAs and makes branched-chain keto acid (BCKA)s. However, we recently demonstrated that BCKA transamination by the BCAT1 enzyme builds up the BCAA pool in myeloid leukemia (ML) cells, essentially running in the reverse direction(Hattori et al. 2017). When α-keto-isovalerate (KIV; one of the substrates of BCAT1) was ^{13}C-labeled, valine (the expected product of BCAT1) containing ^{13}C accumulated. Labeled KIV was not observed when ^{13}C-labeled valine was supplied, indicating a non-canonical, unidirectional flux from KIV to valine (Hattori et al. 2017). In that study, metabolic fingerprints were acquired via a traditional, labor- and material-intensive sampling scheme involving months of sample preparation and several dozen samples. One reason for the large number of samples in this or similar studies is the biological and technical variation due to sample preparation steps; these factors make it more challenging to compare time-series data without large numbers of replicates. We sought to replicate the result of the original Hattori et al. study using real-time *in vivo* metabolomics.

First, we cultured myeloid leukemia cells as previously described (Hattori et al. 2017), then pelleted and resuspended them in IMDM media without KIV or valine. Working quickly, we loaded the cells into an HR-MAS rotor and added either ^{13}C-labeled KIV or valine to make a total volume of ~60 µL, capped the

sample, and inserted the rotor into the magnet. We recorded 1D HSQC spectra every 4.2 minutes while spinning at 3500 Hz at the magic angle (54.7°) (Beckonert et al. 2010). A hole in the rotor cap allowed for gas exchange (Mobarhan et al. 2016).

By monitoring the intensity of the methyl peaks of both KIV and valine, we observed that ^{13}C-labeled KIV decreased in intensity and fell close to the limit of detection within about 60 min (Fig. 2.1A). The ^{13}C-labeled valine peak grew with an inversely proportional trajectory, providing real-time, *in vivo* evidence of KIV-to-valine conversion. As the reaction rate depended on the concentration of the cells in the rotor, cell density was adjusted to accommodate measurement of the rapid reaction and provide greater detail about reaction kinetics. As reported previously, labeled KIV was not observed when ^{13}C-labeled valine was supplied (Fig. 2.1B). Thus, the original results from Hattori et al. that took months of sample prep and data collection were reproduced with real-time resolution in one afternoon. We did not need to adapt the culture media (Link et al. 2015) or embed the cells (Koczula et al. 2016) to get these results. Additionally, the combined rate of uptake and conversion of valine could be measured with precision, where measurements at only a few time points were taken previously.

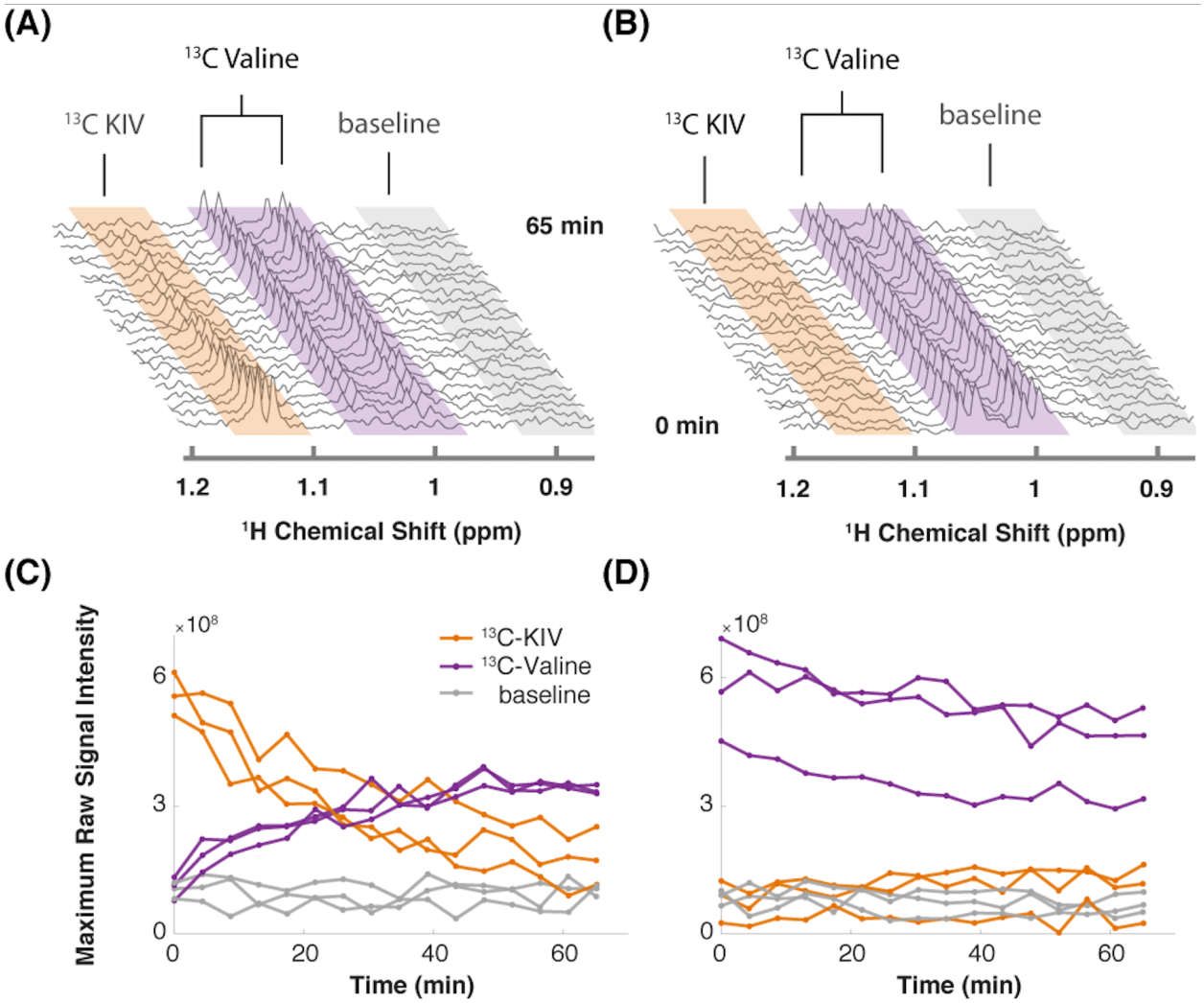


Figure 2.1. Targeted isotopic CIVM-NMR measurement of metabolic flux in human myeloid leukemia cells. (A) ^{13}C-labeled keto-isovalerate (KIV) was converted to valine. **(B)** ^{13}C-labeled valine was not converted to KIV, confirming unidirectional flux in ML cells.

Untargeted CIVM-NMR measurements of *N. crassa* growth

Given the utility of CIVM-NMR for the targeted monitoring of known reactions in mammalian cells, we applied it to the continuous measurement of the metabolic dynamics of the filamentous fungus *N. crassa* over 11 h in both aerobic and anaerobic environments. *N. crassa* is an obligate aerobe but will live under low oxygen conditions (C. L. Slayman 1965; C. L. Slayman and Slayman 1968; CL Slayman et al. 1973). We grew *N. crassa* tissue in a nutrient-rich liquid medium (Fig. 2.2A). After 30 h, a piece of tissue with a volume of ~50 µL was taken from the main mycelial mass, rinsed, and put into a 4-mm HR-MAS rotor with fresh media. The rotor was sealed with a cap with a hole filtered with rayon culture tape punches ("aerobic") (Mobarhan et al. 2016) or no hole ("anaerobic"), placed in the HR-MAS probe, and spun at 6000 Hz at the magic angle for the duration of each experiment (Fig. 2.2B). Each individual scan of a standard 1dnoesypr experiment took ~3.97 s. Scans were recorded and summed continuously, and free induction decays (fids) were written to a file once every 64 scans, establishing our shortest temporal resolution at 4.23 min (Fig. 2.2C). After data acquisition, properly phased and Fourier-transformed frequency-domain data were added together to increase the signal-to-noise ratio (S/N) at the expense of temporal resolution (Fig. 2.2D). The organism was assessed for survival after each experiment (ranging between 11 h to 4 days). In every case (n = 9), mycelia did not sediment, were intact, and grew significant hyphae within hours of being placed on standard nutrient agar (Supplementary Figure 2.1). Thus, *N. crassa* survived the CIVM-NMR experiments and could be used in

downstream experiments or processing steps (Fig. 2.2E). Custom shell scripts allowed for batch processing of NMR data (Fig. 2.2D) using NMRPipe (F. Delaglio et al. 1995a). Normalizing to the stable 1 mM DSS reference resonance (0.0 ppm) allowed for relative comparison of peak intensities across time points and samples. To improve S/N, sequential spectra were time-averaged, resulting in 12.7-min temporal resolution for all downstream analyses (Fig. 2.2D).

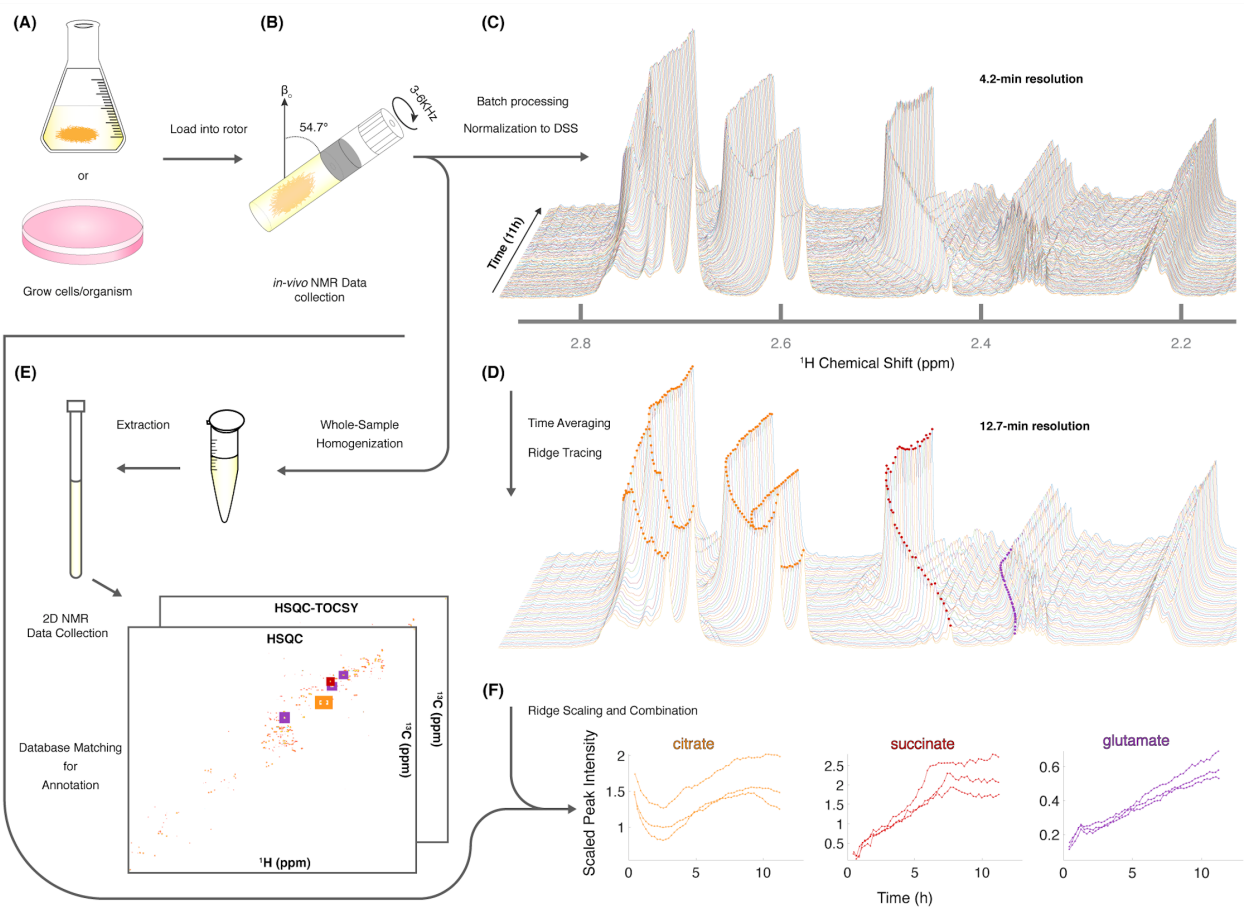


Figure 2.2. Sample preparation and analysis for CIVM-NMR experiments.

(A) Samples were first grown to a suitable volume or density in standard media

and **(B)** transferred to the HR-MAS rotor (*N. crassa* is shown). Gas composition (e.g. air availability) was altered using a filtered hole or no hole in the cap, and the rotor was spun at the magic angle. NMR data were collected continuously every 4 minutes over the course of hours, then **(C)** processed and normalized to the DSS reference peak (0 ppm) to yield full-resolution data. **(D)** Every three spectra were time-averaged (summed) for improved S/N, and peak intensities were traced across time using ridge tracing to yield relative quantification of metabolites. **(E)** Following HR-MAS, the rotor contents were homogenized, methanol-extracted, and used for 2D NMR analysis for peak annotation by database matching. **(F)** For annotated metabolites with >1 peak (e.g. citrate), the quantified and annotated trajectories (ridges) for each peak were scaled and combined into a single representative trajectory. Trajectories for each annotated compound in 3 aerobic experiments are plotted to compare time series between biological replicates.

To assist with annotation and compound identification, the organism and media were removed at the end of each run, bead-homogenized, and extracted in MeOH (80%) (Fig. 2.2E). Combined supernatants for representative samples were analyzed using ^{13}C-HSQC, HSQC-TOCSY, and noesypr1d NMR experiments, and the data were matched to an NMR metabolomics database using COLMARm (Bingol et al. 2016). Resulting putative identifications were manually assigned confidence scores as described previously (Jacquelyn M

Walejko et al. 2018b). We mapped 34 metabolites with high confidence scores onto the real-time *in vivo* spectra of *N. crassa* (representative annotations, Fig. 2.2F), including multiple amino acids and metabolites involved in the TCA cycle, glycolysis, and fermentation (Fig. 2.3A-C; Supplementary Table 2.1). Several metabolites overlapped with those found in a previous NMR study in *N. crassa (J. D. Kim et al. 2011)*. We created MATLAB functions for visualization of time series data for samples individually (Fig. 2.2C-D) or as interactive mirror images (Fig. 2.3). We found that the latter approach facilitated comparison between samples, revealing several differences in metabolism between the aerobic and anaerobic conditions (Fig. 2.3) that were reproduced in replicate samples (Supplementary Figure 2.2).

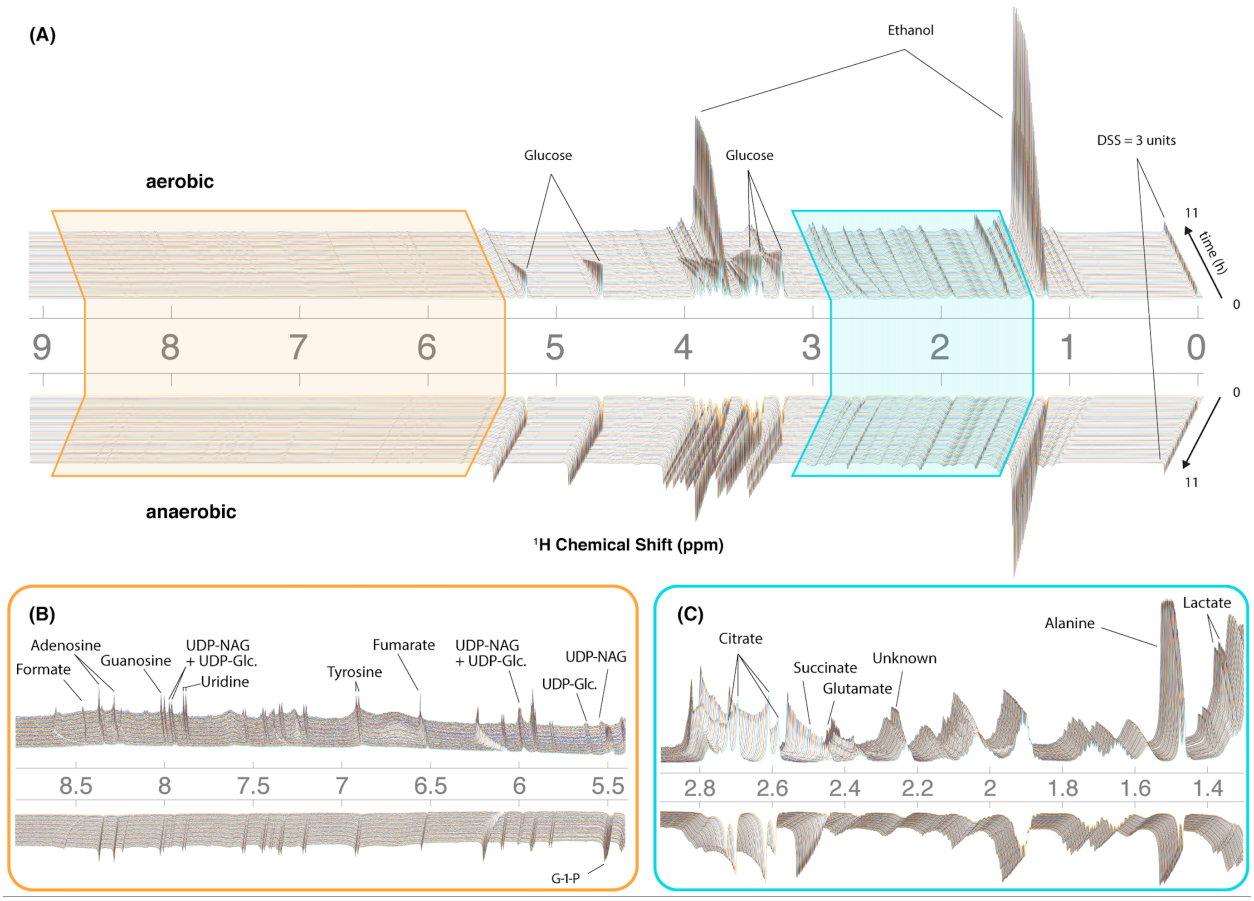


Figure 2.3. CIVM-NMR measurements of N. crassa metabolism under aerobic and anaerobic conditions. ^{1}H NMR data for one aerobic replicate (top) and one anaerobic replicate (bottom) plotted interactively as a 'mirror plot' for direct comparison between conditions by peak height and position at a given time. To improve the S/N, data were analyzed at 12.7 min resolution. Annotations are shown for select peaks of interest for **(A)** the entire spectrum, and expansions of **(B)** the aromatic region and **(C)** the aliphatic region. Several peaks change position and intensity over the course of the experiments. Abbreviations: UDP-NAG, UDP-N-Acetyl Glucosamine; UDP-Glc, UDP-Glucose; G-1-P, Glucose-1-Phosphate.

Relative quantification of metabolites by ridge tracing

The 34 compounds that were mapped to *in-vivo* data were assigned a second confidence score for quantifiability. For 21 highly scoring metabolites (Supplementary Table 2.1), we obtained relative quantification (Supplementary Figure 2.3) by tracing peaks across time with a ridge-tracing algorithm (Fig. 2.2D and 2.4A). With our current algorithm that is limited to peaks with low overlap, we traced over 170 peaks across all of our spectra, including ~150 that are currently un-annotated. We combined the information from ridges of sufficient quality when assigned to the same compound (Fig. 2.4B), leveraging the information about compound concentration from multiple measurements. Replicates of dense, continuously repeated measurements on the same sample offer benefits (Sefer et al. 2016) that would be eliminated by taking time-wise averages or standard errors. We are developing more comprehensive and robust statistical treatments of these unique data within a modeling framework to address this need.

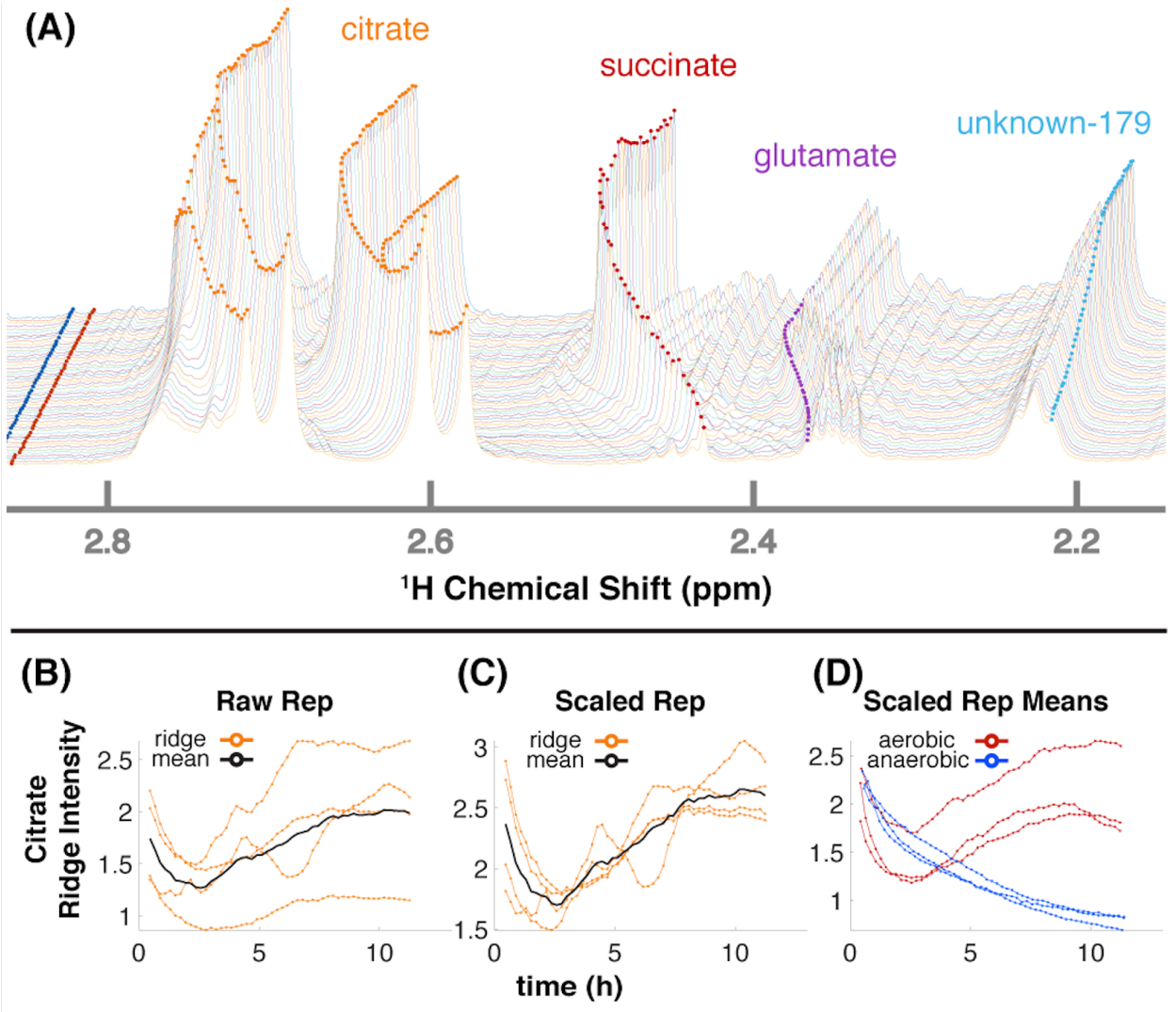


Figure 2.4. Ridge tracing produces concentration dynamics of metabolites. **(A)** Multiple traced ridges for a single aerobic replicate. Peak maxima at each time point were located using a peak-picking algorithm that includes an adjustable Gaussian filter. Maxima were connected to form ridges along the time dimension using a single linkage hierarchical agglomerative clustering based on Euclidean distances between the points in chemical shift, time, and intensity space. Metabolites typically have several characteristic NMR peaks, e.g. the 4 orange ridges in citrate **(A)**. A simple time-wise average represented by the black

line in **(B)** only gives the average intensity over time but loses valuable information on actual dynamic trends. To more accurately extract trends for a particular metabolite, we first integrate each peak in that metabolite over time to obtain its mean value. Then, each peak trajectory is scaled by ratio of the highest mean to its own mean, yielding the 4 orange lines in **(C)**. The mean of these trajectories is shown in black in **(C)** and represents the relative concentration over time for that metabolite in that replicate. The 3 aerobic (red) and 3 anaerobic (blue) replicates for citrate are shown in (**D)**.

Glucose-dependent changes in pH

NMR chemical shifts are sensitive to pH and metal ion content (Tredwell et al. 2016a; Ye et al. 2018), typically requiring peak alignment algorithms that are prone to creating artifacts. The positions of peaks clearly changed across time in our data (Fig. 2.3B-C, 2.4A), particularly in the aerobic samples. Because these changes were monitored continuously, peak identity across time was unambiguous, eliminating the need for alignment and facilitating annotation and quantification even as changes in peak position affected overlap with other peaks. Changes in peak position for organic acids in our samples were compared with reported titration curves (Koczula et al. 2016; Tredwell et al. 2016a; Ye et al. 2018), in-house titrations for citrate (Supplementary Figure 2.4), and Bruker AssureNMR software (Bruker Biospin, USA; Supplementary Table 2.2) to estimate pH of the sample at each timepoint. Our data indicate that the pH

of the aerobic cultures began at 6.2-6.4, then dropped to 5.2-5.4 coincidentally with glucose consumption. Furthermore, this acidification reversed after glucose depletion at 6-7h, and pH increased to 5.5-5.7 by the end of our experiments. In the anaerobic samples, the pH decreased from 6.2-6.3 to 5.7-5.9. Although we did not perform high-resolution titrations for glutamate, succinate, and fumarate, their reported shifts were consistent with the trends for citrate (Supplementary Table 2.2).

Maintenance of characteristic differences in pH is well-accepted between organelles, the cytoplasm, and the extracellular milieu (Bencina 2013; Casey et al. 2010; Magnuson and Lasure 2004). Filamentous fungi including *N. crassa* (Vrabl et al. 2012) secrete large amounts of organic acids such as citrate, fumarate, and succinate, to acidify their extracellular environment (Dorsam et al. 2017; Kubicek et al. 2010; Magnuson and Lasure 2004), and the two latter acids are taken up by carbon-limited *N. crassa*, with maximal uptake occurring around pH 5.5 .

Activation of central carbon metabolism in aerobic conditions

Under aerobic conditions, glucose and trehalose were consumed within the first 6h while ethanol and lactate were produced. Glucose fell below our limit of detection by ~6h, after which it was depleted as the concentrations of ethanol and lactate plateaued (Fig. 2.3A-B, 2.5). The inverse trends between sugars and fermentation products indicate that most carbon was shunted to fermentation. In fact, we observed that most ^{13}C in labeled glucose was channeled to ethanol and succinate (Supplementary Figure 2.5), consistent with known *N. crassa*

biochemistry (Colvin et al. 1973; Greenfield et al. 1988). Low levels of trehalose, another energy metabolite in the aerobic samples are consistent with vegetative growth, while increasing levels in the anaerobic samples may reflect a developmental shift to production of conidia in response to stress (Hill and Sussman 1964).

Four TCA cycle metabolites were detected in our experiments (Fig. 2.3B-C). Fumarate and succinate increased in the aerobic condition, and both accumulated slightly faster around 6h following glucose depletion and remained abundant (Fig. 2.5). This dynamic could be a result of glyoxylate cycle (Voet and Voet 2011) or mixed acid fermentation (Dreyfuss et al. 2013) activation, and would not be observable without a continuous, densely sampled time series. Standard replicate averaging with extracted samples at different times would average out much of this detail. Next, one of the clearest signs of low oxygen levels in the anaerobic sample was the slight reduction in succinate compared to the drastic reduction in fumarate. Succinate levels in the aerobic condition are comparable to those in the anaerobic condition, while fumarate accumulates much more in the aerobic condition (Fig. 2.5). This is explained by the fact that conversion from succinate to fumarate depends on oxygen reduction in the electron transport chain (Dreyfuss et al. 2013; Kanehisa et al. 2016). Finally, citrate was abundant in the aerobic condition and followed a complex trend, while malate was observed in endpoint extracts (Supplementary Figure 2.6A). Taken together, these trajectories constitute strong evidence for active glycolysis and fermentation in the presence of glucose and oxygen, while alternative pathways

such as the glyoxylate cycle were active after glucose depletion. Similar trends with lower rates were observed in the anaerobic samples, except for differences in citrate and glucose-1-phosphate (G-1-P) (Fig. 2.5).

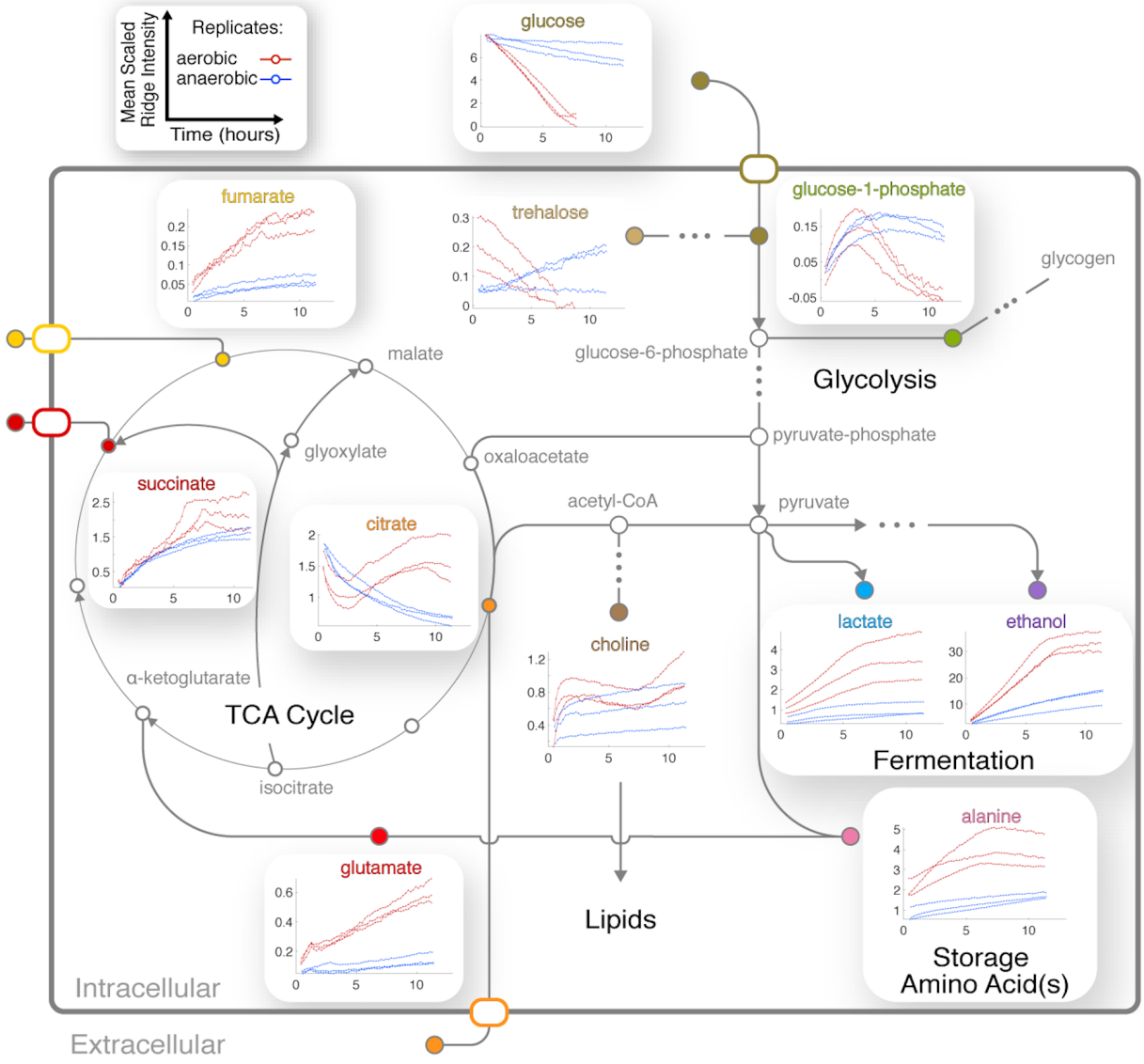


Figure 2.5. Integration of central metabolic pathways. Arrows correspond to one or more reactions, and nodes correspond to metabolites (Dreyfuss et al.

2013; Kanehisa et al. 2016). Nodes are filled for observed metabolites. Plots show the means of scaled peak/ridge intensities for a given compound in a given replicate over traceable times, where red and blue trajectories represent aerobic and anaerobic conditions, respectively. Arrows indicate typical reaction directions. The glyoxylate cycle is shown as a shunt through glyoxylate embedded in the TCA cycle.

Interplay between amino acid, central carbon, and nitrogen metabolism

The dynamics of glutamate were different between the two conditions (Fig. 2.5). Glutamate accumulates while synthesis of glutamine is repressed in *N. crassa* in nitrogen-sufficient conditions (Kanamori et al. 1982). We could not annotate glutamine with confidence because of overlap (Supplementary Table 2.1B). However, resonances consistent with glutamine increased after ~3h (Supplementary Figure 2.6B), indicating potential nitrogen insufficiency in the aerobic culture. Arginine levels correspond to those of glutamate in the aerobic condition (Supplementary Figure 2.3). Glutamate is produced from arginine degradation (Voet and Voet 2011); for instance, arginine has been reported as an abundant amino acid in extracted samples of actively growing *N. crassa* cultures (Kanamori et al. 1982; J. D. Kim et al. 2011) and is thought to be catabolized to glutamate during conidiation (J. D. Kim et al. 2011).

Trends for alanine (Fig. 2.5) and an unknown in the aliphatic region (Fig. 2.4A) were very similar to that of ethanol and lactate (Fig. 2.5), indicating that

their metabolic fluxes are closely dependent on intermediates or energy produced by glycolysis and fermentation. This hypothesis is supported by the fact that alanine is synthesized from glutamate and pyruvate by alanine transaminase (Kanamori et al. 1982; Radford 2004). Glutamate levels increased and were unaffected by glucose, but alanine first accumulated and then decreased upon glucose depletion (Fig. 2.5). We conclude that alanine synthesis was limited by a lack of pyruvate caused by glucose depletion. Glutamate levels are maintained during starvation (Voet and Voet 2011), and Kanamori et al. (1982) suggested that alanine serves as a storage for pyruvate and nitrogen *via* glutamate in favorable conditions (Kanamori et al. 1982). Therefore, the observed decrease in alanine suggests that it was utilized for pyruvate and glutamate when glucose concentrations were low in the aerobic condition (Fig. 2.5). Our CIVM-NMR data therefore supports glutamate as a hub between central carbon and amino acid pathways and confirms the maintenance of glutamate stores even under starvation.

Complex trends reveal dynamics between energy storage and cell wall synthesis pathways

CIVM-NMR data revealed significant changes that preceded glucose depletion at ~6 h for compounds such as citrate, choline, adenosine, and valine, which all had similar trends in the aerobic condition (Fig. 2.5). Citrate decreased at the start of all experiments. Under aerobic conditions it began to accumulate again around 2.5 h and surpassed initial levels, while in anaerobic conditions it decreased at an exponential rate to a very low amount (Fig. 2.5). In the

anaerobic samples, citrate was utilized while succinate accumulated to levels similar to those in the aerobic samples. However, fumarate levels remained low. Lack of oxygen could explain low rates of conversion from succinate to fumarate, while glyoxylate cycle activity can occur in anaerobic conditions (Rude et al. 2002; Wayne and Lin 1982) and yields succinate and malate without fumarate.

N. crassa does not survive on citrate as a sole carbon source (Wolfinbarger and Kay 1973), and to our knowledge extracellular citrate utilization has not been reported for *N. crassa*. However, citrate levels were observed well below the initial amount present in the media alone (9.74 mM) in both conditions, strongly indicating that external citrate was consumed in both experiments. Isotopic labeling experiments will more directly test this hypothesis.

Pyruvate and acetyl-CoA both serve as crossroads between major energy metabolites and lipids. Although we did not observe pyruvate and acetyl-CoA directly, most accumulating metabolites in pathways emanating from pyruvate exhibited strikingly similar trends (Fig. 2.5), suggesting flux through pyruvate. Curiously, citrate and choline did not follow this pattern, indicating activity from pathways that consume and replenish their pools. However, the rates of change of these metabolites were clearly opposed in both aerobic and anaerobic samples. This opposition suggests that flux from acetyl-CoA was being channeled differentially between citrate and choline synthesis and demonstrates a major carbon and energy exchange between central metabolism and lipid precursors (Markham et al. 1993).

The accumulation of citrate and choline after glucose depletion in the aerobic samples was puzzling. Prior work has indicated that under low oxygen or glucose depletion *N. crassa* cells become vacuolated (Clifford L Slayman et al. 1994; C Slayman and Potapova 2006). The synthesis of membranes for the vacuoles and their membranes under anaerobic conditions would explain the rise in choline. A concordant decrease in G-1-P at ~3h may indicate a shift of carbon flux to glycolysis from glycogen, caused by sensing of extracellular glucose levels (Wang et al. 2017) or limitations of glycogen capacity. Glucose conversion to G-6-P (Glucose 6-phosphate) is the first step of glycolysis (Voet and Voet 2011), which was clearly active in the first stages of our aerobic condition (Fig. 2.5). High levels of G-6-P drives its conversion by phosphoglucomutase to G-1-P (Voet and Voet 2011), which is converted by UDP-glucose pyrophosphorylase and UTP hydrolysis to the direct glycogen precursor UDP-glucose (Voet and Voet 2011). The latter is the rate-limiting step in glycogen synthesis, which is an endergonic process. If G-6-P levels were high and flux were shunted to glycogen, high levels of G-1-P would be expected. In fact, G-1-P levels increased in the aerobic samples until around 3 h then decreased, while UDP-Glucose was also observed. Therefore, we conclude that glycogen synthesis occurred in the first half of the aerobic experiments, but slowed or reversed after 3 h. This conclusion is further supported by the fact that glycogen synthesis occurs during high rates of growth in *N. crassa*, and wanes during slow growth (Brody and Tatum 1967; de Paula et al. 2002; Virgilio et al. 2017).

On the other hand, high levels of G-1-P are unlikely to be produced by the relatively low levels of glycolysis observed in the anaerobic samples and can indicate glycogen degradation. G-1-P accumulated to comparable levels in both conditions, but it remained in the anaerobic samples. Glycogen degradation is exergonic and releases G-1-P and glucose directly in an approximate 9:1 ratio (Voet and Voet 2011). Furthermore, UDP-Glucose was not observed in the anaerobic samples. Thus, high levels of G-1-P in the anaerobic conditions may indicate glycogen degradation.

The primary chitin cell wall building block UDP-N-acetylglucosamine (UDP-GlcNAc) (Milewski et al. 2006) increased in only the aerobic cultures (Fig. 2.3, Supplementary Figure 2.6C), although overlap and low intensity prevented quantification. UDP-GlcNAc is synthesized via the unidirectional Leloir pathway (Milewski et al. 2006), and the only known uses for UDP-GlcNAc in *N. crassa* are chitin/cell wall biosynthesis and UDP-GalNAc production (Edson and Brody 1976; Milewski et al. 2006). Filamentous fungi such as *N. crassa* produce chitinases (Patil et al. 2000) and could utilize these for autolysis under stress conditions. However, if an increase in UDP-GlcNAc indicated cell wall degradation (i.e. due to stress or autolysis), those resonances would be expected to increase in the anaerobic condition; however, they were barely detected (Fig. 2.3, Supplementary Figure 2.6C). Curiously, a recent study suggested that *N. crassa* utilizes alternative chitin catabolism pathways that would not result in increased GlcNAc-derived UDP-GlcNAc (Gaderer et al. 2017). Considering the

above dynamics, we conclude that resources were allocated between energy storage and cell wall synthesis pathways in glucose-rich conditions.

Discussion

We have demonstrated the use of CIVM-NMR to monitor metabolic dynamics in cells and whole microorganisms. An uninterrupted, high-resolution time series of NMR data allows observation of rapid but reproducible metabolic events. In contrast, using traditional studies of different replicates for each time point, the biological and technical variation often obscure details of dynamics. The lack of extraction removes a major source of technical variation found in typical MS and NMR metabolomics workflows and can facilitate inter-study comparisons.

NMR has relatively low sensitivity, but it is a quantitative and reproducible technique, and conventional NMR cryoprobes allow routine proton detection of compounds at concentrations as low as about 5 μM ^{1}H. HR-MAS probes are generally less sensitive. However, the temporal dimension of CIVM-NMR data allows for more confident assignment of peaks with surprisingly low signal-to-noise ratios. By taking advantage of this unique property of CIVM-NMR data, we detected peaks as low as ~24-62 μM ^{1}H (Supplementary Figure 2.7). The sensitivity of CIVM-NMR is therefore particularly well-suited for observation of the major sources, sinks, and bottlenecks of metabolism in an organism or cells (e.g. for metabolic engineering). For instance, absolute quantification of 103 metabolites in *E. coli* by LC-MS/MS revealed intracellular concentrations ranging from 0.13 μM to 96 mM. Of these, 61 were found in concentrations of 100 μM or

higher (Bennett et al. 2009), placing them well within the detection limits of CIVM-NMR.

Only 20-70 µL of sample is needed with no sample preparation to yield an entire time series for various metabolites, and the sample can be used in downstream *in vivo* or chemical analyses following NMR data collection. These factors make CIVM-NMR ideal for scarce samples that would not otherwise be possible to study by time-series metabolomics (Sefer et al. 2016). With an internal rotor radius of 1.4 mm spinning at 6000 Hz, our samples experienced up to 200,000 x g of acceleration. As sedimentation was not observed, it is possible that a low relative density of *N. crassa* mycelia compared to the media may have resulted in a lower effective radius of rotation. While some samples, including the leukemia cells in Fig. 2.1, are less stable at high spinning rates, microorganisms such as *E. coli* and *S. cerevisiae* can grow under different amounts of hypergravity, even with cellular and organellar sedimentation (Deguchi et al. 2011). Furthermore, methods have been developed to obtain HR-MAS data with slow spinning (Mobarhan et al. 2017b), which could allow monitoring in ~1500 x g or less. The lack of perfusion and a limited sample volume are both factors that need to be considered with regard to nutrient depletion and waste accumulation. Lastly, identification of spectral features and deconvolution of overlap are still challenging in any NMR or LC-MS metabolomics study. However, temporal continuity clearly provides information (e.g. as seen in Fig. 2.3 and Supplementary Figure 2.7) that will be helpful in addressing these problems.

Full utilization of the time series data from CIVM-NMR will require a modeling-based method, and our data underscore the need for accurate and experimentally-based kinetic models of metabolism. With the potential for <4-minute resolution by using fewer scans before saving fids at the cost of signal-to-noise ratio, CIVM-NMR provides a unique opportunity for probing flux changes as well as allosteric regulation (Link et al. 2013) with kinetic models (Link et al. 2014; Link et al. 2015) for abundant metabolites. Each replicate can be seen as a single, complete model with different initial conditions, which is significantly better than a time series of averages. Previous real-time methods (Koczula et al. 2016; Link et al. 2015) have equal or greater temporal resolution at the expense of disadvantages such as being destructive (Link et al. 2014), limitation to cell suspensions (Koczula et al. 2016; Link et al. 2014), primarily measuring the media(Koczula et al. 2016; Sengupta et al. 2016), or having combined biological and technical variance. CIVM-NMR minimizes noise by eliminating sampling and extraction variance. Batch effects for each replicate are eliminated since all experimental and NMR parameters are consistent across timepoints. Analytical drift is eliminated because the detector never contacts the samples, and the sample is not perturbed by measurement. These factors in turn facilitate optimization of modeling parameters (Ghasemi et al. 2011).

CIVM-NMR can also allow continuous monitoring of the metabolic state before, during, and after a range of environmental and genetic perturbations. For instance, we observed a shift from a glucose-rich environment to starvation at ~6h. Sequential utilization of alternative carbon sources such as quinic acid

(Tang et al. 2011) is a natural extension of this work. In studies involving targeted pathways in mammalian cells, CIVM-NMR can facilitate the testing of mutants, gene knockdowns, or small molecule substrates, inhibitors, or activators. We are exploring the use of alternative gas mixtures for spinning the rotor, allowing adjustment of O_2:CO_2:N_2 ratios during experiments. This presents another environmental shift and facilitates monitoring of cells (e.g. mammalian) which require controlled gas compositions. Real-time temperature control could also be used to probe temperature shifts or assess the effects of a temperature-sensitive mutation on the metabolism of an organism. Perturbations such as these will be critical to exploring and refining dynamics in kinetic models with an empirical basis.

Acknowledgments

We thank the A. Simpson lab for discussions and advice on drilling holes in rotor caps, M. Case, J. Griffith, C. Slayman, and Z. Lewis for discussions on *Neurospora* stress and metabolism, and J. Walejko and G. Gouveia for discussions on method development. L. Morris provided guidance in bash scripting for automation of NMRPipe commands. B. Schuttler and L. Mao provided stimulating discussions on *N. crassa* modeling.

CHAPTER 3

RTEXTRACT: TIME-SERIES NMR SPECTRA QUANTIFICATION BASED ON 3D SURFACE RIDGE TRACKING [2]

[2]: Wu Y, Judge MT, Arnold J, Bhandarkar SM, Edison AS. RTExtract: time-series NMR spectra quantification based on 3D surface ridge tracking. Bioinformatics. 2020;36(20):5068-75.

Foreword

This chapter is reprinted from Wu Y, Judge MT, Arnold J, Bhandarkar SM, Edison AS. RTExtract: time-series NMR spectra quantification based on 3D surface ridge tracking. Bioinformatics. 2020;36(20):5068-75 and is available at https://academic.oup.com/bioinformatics/article/36/20/5068/5870445. My contribution to this work as the first author consisted of: (1) building RTExtract algorithm and program, (2) validating and testing RTExtract on simulated and experimental datasets, (3) producing tutorials for future users, (4) producing visualizations, and (5) writing and editing the manuscript and addressing reviewer comments. Arthur S. Edison, Jonathan Arnold, Suchendra M Bhandarkar, and Michael T Judge edited the manuscript and responded to the reviewers. Michael T Judge also contributed to visualization. Suchendra M Bhandarkar and Jonathan Arnold provided technical support on differential geometry. The research was supported by National Science Foundation (NSF 1713746) and Georgia Research Alliance. The supplementary materials in the chapter are listed in Appendix B.

Abstract

Motivation

Time-series NMR has advanced our knowledge about metabolic dynamics. Before analyzing compounds through modeling or statistical methods, chemical features need to be tracked and quantified. However, because of peak overlap and peak shifting, the available protocols are time consuming at best or even impossible for some regions in NMR spectra.

Results

We introduce RTExtract (Ridge Tracking based Extract), a computer vision-based algorithm, to quantify time-series NMR spectra. The NMR spectra of multiple time points were formulated as a 3D surface. Candidate points were first filtered using local curvature and optima, then connected into ridges by a greedy algorithm. Interactive steps were implemented to refine results. Among 173 simulated ridges, 115 can be tracked (RMSD < 0.001). For reproducing previous results, RTExtract took less than two hours instead of ~48 hours, and two instead of seven parameters need tuning. Multiple regions with overlapping and changing chemical shifts are accurately tracked.

Availability

Source code is freely available within Metabolomics toolbox GitHub repository (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA/tree/master/metabolomics_toolbox/code/ridge_tracking) and is implemented in MATLAB and R.

Introduction

Experimental approaches have been developed in time-series metabolic measurements by both NMR (nuclear magnetic resonance) and MS (mass spectrometry) (Bastawrous et al. 2018a; Judge et al. 2019; Koczula et al. 2016; Link et al. 2015; G. Montana et al. 2011a; Tabatabaei Anaraki et al. 2018). These experimental methods provide opportunities to understand metabolic dynamics, including metabolic changes under variation in carbon sources or oxygen levels

(Judge et al. 2019; Link et al. 2014; Link et al. 2015). Among existing approaches, CIVM-NMR (continuous *in vivo* monitoring of metabolism by NMR) provided high time-resolution, *in vivo* measurements of metabolites in *Neurospora crassa* under aerobic and anaerobic conditions (Judge et al. 2019). These measurements covered a large proportion of pathways in central metabolism, and interesting dynamics in compound concentration were observed.

NMR provides a highly reproducible way to identify and quantify compounds. In an NMR spectrum, different metabolites are represented by different peaks (features), and peak height (intensity) is proportional to compound concentration, when peak shape doesn't change. Peak resonance frequency is sensitive to the local electronic structure and some environmental variables. Resonance frequency is reported as chemical shift, δ, which is derived by dividing the frequency in Hz by the spectrometer frequency in MHz and thus has units of parts per million (ppm). The dependence on local electronic structure allows for reliable compound identification. Additionally, some metabolites are sensitive to changes in the local chemical environment (e.g. pH or metal ion concentration) and systematically change their chemical shift, providing a useful way to measure these environmental factors (Takis et al. 2017; Tredwell et al. 2016b).

Metabolism yields changes in metabolite concentration and local pH, resulting in NMR peak intensity and chemical shift changes. These changes provide important information about metabolic dynamics but also complicate

feature extraction. Moreover, peaks in NMR spectra often overlap, which affects both compound annotation and quantification. The combination of systematic chemical shift, overlap, and amplitude changes makes peak tracking and quantification a difficult problem. A practical, stable computational approach is needed to track and quantify peaks over time, regardless of overlap, amplitude and chemical shift changes.

Traditional alignment-based methods are popular for processing NMR spectra from different samples and aligning shifting peaks. However, these methods often introduce artifacts and are unreliable for the regions where peaks cross (Csenki et al. 2007; Vu and Laukens 2013b). In CIVM-NMR data, the pattern of peak shifting is less noisy and more continuous than in discrete extracted samples by traditional methods. These properties provide new information for quantifying crossing peaks as discussed below.

Multiple methods have been implemented to track peaks in time-series NMR spectra. The TSATool can track a peak by peak picking and a predefined function describing shifting trajectory (Koczula et al. 2016; Ludwig and Gunther 2011). It was used to track NMR peaks in leukemia cells and interesting hypoxia metabolic response was observed. This method, though capable of tracking individual peaks, does not provide a general solution for quantifying multiple peaks efficiently. In our initial CIVM-NMR study, a better framework for multiple peak tracking was introduced. Peak tracking was achieved by a smoothing filter to reduce noise (filtering step) and hierarchical clustering (connection step) to connect candidate peaks (Judge et al. 2019). While this method tracked peaks

with chemical shift variation, substantial manual effort was still needed in parameter tuning to accommodate different spectral regions. For instance, the proper scaling factor for the extent of chemical shift variation and the number of expected clusters were crucial parameters but difficult to optimize. About 48 hours of work were needed for the original CIVM-NMR quantification, which can be a significant bottleneck and cost (Judge et al. 2019). Additionally, none of the aforementioned methods can deal with crossing or severely overlapped peaks.

Computer vision methods have been adapted to solve other spectroscopy (Klukowski et al. 2015; Klukowski et al. 2018) and biological object tracking problems (Steger 1998; Tinevez et al. 2017). It can also be implemented here to promote efficiency. The steps of both filtering candidate points and their subsequent connection can be improved by treating NMR peak extraction as a ridge tracking problem. Time-series NMR data can be viewed as a 2D matrix (or a 3D surface if we treat matrix elements as height) with each row being a spectrum at one time point and each column being the intensity of a particular resonance frequency across time. As the same peaks change continuously through time, they can be conceptualized as surface ridges, for which efficient detection algorithms exist (Suk and Bhandarkar 1992). Surface segmentation techniques have been implemented in computer vision to classify 3D surface points based on their local curvature into qualitative surface types: *inter alia*, ridge, peak, and valley (Supplementary Fig. 3.1) (P. J. Besl and Jain 1986; P. J. Besl and Jain 1988; Suk and Bhandarkar 1992).

In RTExtract (Ridge Tracking based Extract), to filter candidate peak points, we combined ridge classification with other information such as local maxima. This combination of multiple filters provided cleaner results with fewer false positives, and tuning parameters were fewer and more intuitive. Candidate points were then connected by a 2-step greedy method, which is composed of simple local optimal connections without global evaluations; this is possible because of the better filter on candidate points. Additionally, manual refinement steps were introduced to expand flexibility in tracking and increase tracking accuracy.

In this paper, we present our new method (RTExtract) to extract and quantify time-series NMR spectra. We simulated time-series NMR data specifically presenting the challenges that limited previous methods. We also conducted a direct comparison of our previous method and RTExtract on experimental datasets (Judge et al. 2019) and found that RTExtract was faster and easier than our previous approach. Previous tracking results were reproduced in less than two hours instead of ~48 hours. Additionally, we were able to track complex spectral regions, such as those with high amounts of overlap, that were impossible with previous published methods. RTExtract therefore significantly expands the utility of the rich data collected in CIVM-NMR and accelerates its analysis. Furthermore, we show that it can be applied to other time series NMR methods such as pH titration analysis (Brockerman et al. 2019; Edison et al. 1999; Joshi et al. 1997; Liebeke et al. 2013; Zachariah et al. 2001).

Methods

Ridge point classification

Local curvature was used to classify ridge points and functions as one of the filters for candidate points. Including a ridge point filter with local optima filters reduced noise levels in selecting candidate points and increased accuracy in ridge tracking. The following section describes the ridge point filter.

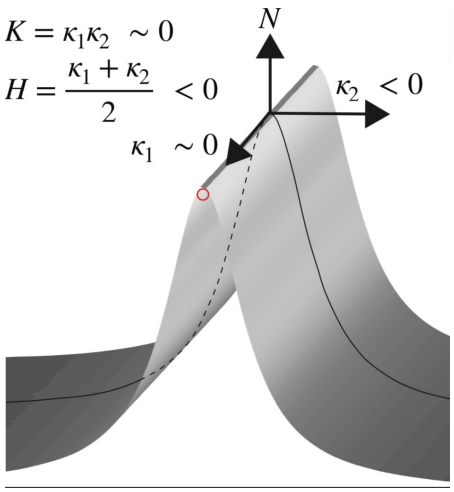


Figure 3.1: Illustration of the concept of H and K curvature. N is the normal vector and the curve is one of the intersecting curves. The two principal curvatures, κ_1 and κ_2, correspond to the two vectors. Gaussian curvature (K) and mean curvature (H) are calculated by κ_1 and κ_2. For the ridge surface shown here, $\kappa_2 < 0$ and $\kappa_1 \approx 0$ which results in $H < 0$ and $K \approx 0$. Computation of H and K curvature is in Methods 2.1. Other surface types are illustrated in Supplementary Fig. 3.1.

For each point on the 3D surface (Fig. 3.1), a normal vector ($\boldsymbol{N}$) can be defined. All planes that contain $\boldsymbol{N}$ (the normal planes) will intersect the 3D surface along a curve, and for each such curve at the point of interest, the curvature can be computed. The maximum and minimum curvature values, denoted by κ_1 and κ_2 respectively, correspond to two mutually orthogonal orientations of the normal planes and are referred to as the principle curvatures of the 3D surface at that point. From κ_1 and κ_2, the Gaussian curvature (K) and mean curvature (H) are defined (Equation [3.3.1] & [3.3.2] and Fig. 3.1) (P. J. Besl and Jain 1986; P. J. Besl and Jain 1988; Suk and Bhandarkar 1992). The curvatures H and K can be used to classify 3D surface points locally into qualitative types, including, *inter alia*, peak, ridge, and valley. Specifically, when $K \approx 0$ and $H < 0$, the surface is classified as a ridge, and the central point of the surface is the candidate point (Supplementary Fig. 3.1).

$$K = \kappa_1 \kappa_2 \qquad [3.1]$$

$$H = \frac{\kappa_1 + \kappa_2}{2} \qquad [3.2]$$

As an alternative to Equations [3.3.1] and [3.3.2], the values of H and K can also be derived through the fundamental form matrices G and B, which provide a practical computation process (Stoker 1969). Let $z = f(x, y)$ be the surface and $X = (x, y, f(x, y))$ be a point on it. The first fundamental form G of the surface and the second fundamental form B of the surface can be computed from partial derivatives (e.g. $X_x = \frac{\partial X}{\partial x}$) and the unit surface normal vector ($\boldsymbol{n}$) (Equations [3.3.3], [3.3.4], [3.3.5]) (Suk and Bhandarkar 1992).

$$G = \begin{bmatrix} X_x \cdot X_x & X_x \cdot X_y \\ X_y \cdot X_x & X_y \cdot X_y \end{bmatrix} \quad [3.3]$$

$$B = \begin{bmatrix} \boldsymbol{n} \cdot X_{xx} & \boldsymbol{n} \cdot X_{xy} \\ \boldsymbol{n} \cdot X_{xy} & \boldsymbol{n} \cdot X_{yy} \end{bmatrix} \quad [3.4]$$

$$\boldsymbol{n} = \frac{X_x \times X_y}{\| X_x \times X_y \|} \quad [3.5]$$

We use a discrete biorthogonal second-order Chebychev polynomial with the interaction term ignored to approximate the local 3D surface within a 7 by 7 window (P. J. Besl and Jain 1986; Haralick et al. 1983). Using biorthogonal polynomials instead of a more general fitting process increased computational speed. As the surface of interest was large (e.g. ~ 50 spectra $\times$ 35000 points for each spectrum in our experimental data set) (Judge et al. 2019), this approximation was necessary to incorporate real-time analysis input within the workflow (wait time < 5 seconds). From the biorthogonal polynomial approximation, the first and second order derivatives, the fundamental forms, and the curvatures (H and K) were computed in order (Suk and Bhandarkar 1992).

Multiple surface types were generated from a second-order polynomial with the interaction term in a 101×101 window (Equation [3.3.6]. Supplementary Fig. 3.1). The parameters A, B, and C in Equation [3.3.6] were varied to produce different surface types, including saddle ridge, minimal surface, saddle valley, ridge, flat surface, valley, peak, and pit. The curvatures H and K were computed for the central point in the window to check with the expected values. The expected H and K curvatures were derived based on Equation [3.3.6] and computed with Equations [3.3.7] and [3.3.8].

$$Z = AX^2 + BXY + CY^2 \quad [3.6]$$

$$H = A + C \quad [3.7]$$

$$K = 4AC - B^2 \quad [3.8]$$

Feature Quantification by Ridge Tracking

The entire workflow of RTExtract is presented in Fig. 3.2. The steps include filtering candidate points, connecting candidate points into initial ridges, ridge refinement, and manual ridge selection.

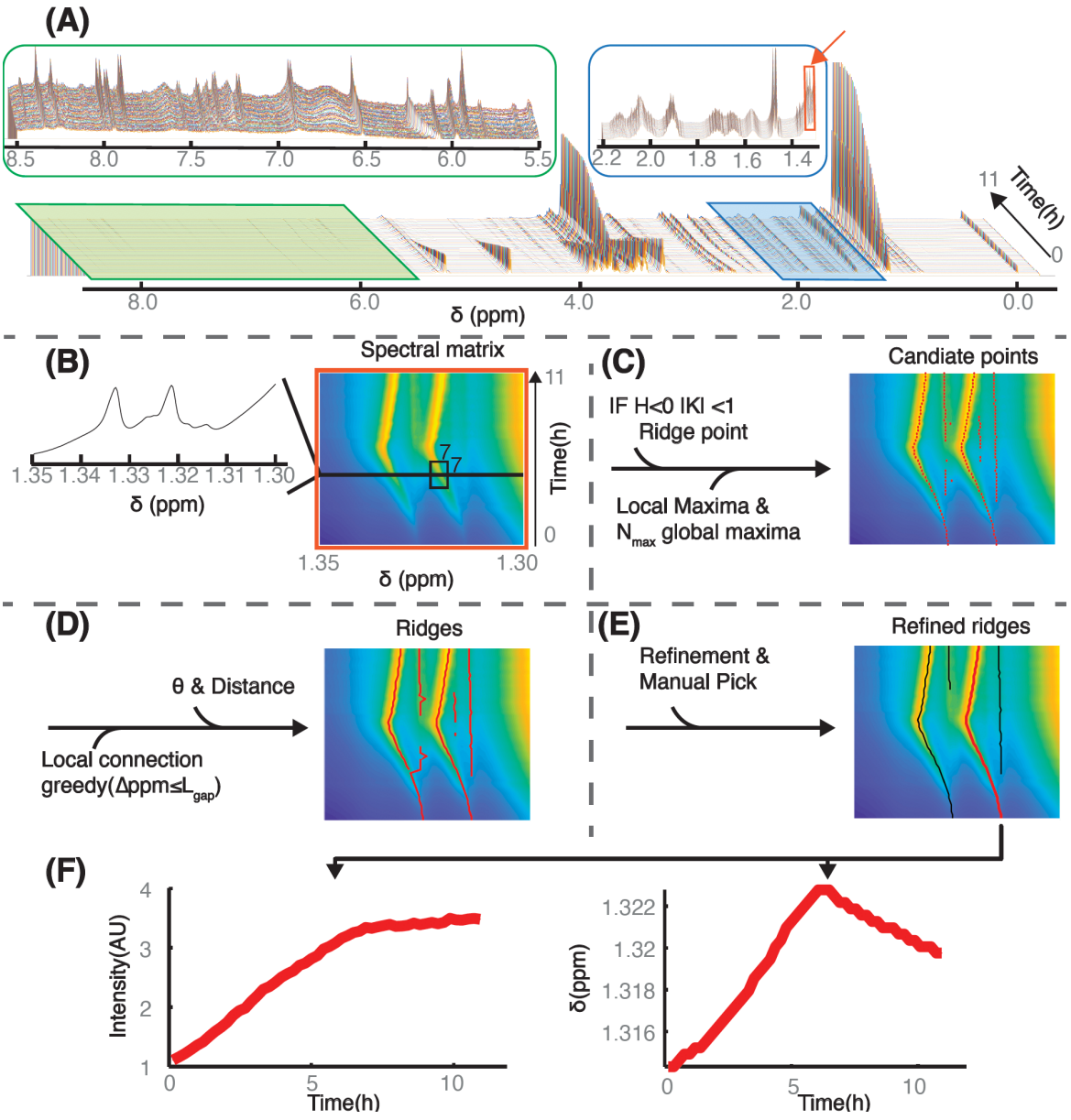


Figure 3.2: Illustration of the RTExtract algorithm. The algorithm is presented based on an example time-series NMR data set, which was measured under aerobic conditions (A) (Judge et al. 2019). The stack spectral plot (A) shows changes of whole NMR spectra (X-axis chemical shift δ) through time (Y-axis), and each time point is distinguished by a different line color as in all stack spectral plot in this paper. Two regions (ppm [5.5, 8.5] and [1.3, 2.2]) are expanded to show spectral complexity. (B) A column slice (ppm [1.30, 1.35], indicated as orange box and arrow in (A) is chosen as show case for RTExtract (B-F). H and K curvatures were computed based on the 7 by 7 window (B) around each point, which was used to define ridge points (*H<0* and $|K| < 1$). (C) To form candidate points (red points), these ridge points are intersected with local maxima and combined with the first N_{max} global maximal points for each spectrum. (D) These points were connected in two steps by a greedy method to form ridges (red line). The first step was based on chemical shift distance (L_{gap}) between points, and the second step was based on distance and angles. (E) Ridges were refined and manually selected for feature quantification. For the red ridge in (E), intensity and chemical shift (δ) were plotted against time (F). The intensity is measured in arbitrary units (AUs), time is measured in hours, and chemical shift (δ) is measured in ppm. Details on the RTExtract algorithm and tuning parameters can be found in Method 2.1 and 2.2.

The tested experimental data sets contain ~50 spectra acquired at different time points (~11 hours), and each of them is comprised of ~35000 points in chemical shift resolution. The original time-series data sets were collected with finer time resolution, and the averaged (denoised) data sets were used to evaluate RTExtract, identical to our original study (Judge et al. 2019). In NMR spectra, even a small region can exhibit high complexity, and peaks of interest also differ considerably from each other in intensity (Fig. 3.2A, green and blue boxes). A region of interest (ROI) (ppm [1.3, 1.35] (Orange box in Fig. 3.2A) was selected as an example to illustrate the computational pipeline (Fig. 3.2B-E). The ROI is presented as a surface, in which different intensities can be visualized as different colors like a topographic map (Fig. 3.2B). Each row of the surface matrix is a single spectrum acquired at one time point. To filter candidate points, information from curvature, local maxima, and a controlled number (N_{max}) of global maxima were combined (Fig. 3.2C). Local maxima are defined for each spectrum and several local maxima (including true peaks and noise) exist in a realistic NMR spectrum. Points on the surface were classified as ridge points (Set S_R) if they satisfied the curvature criteria in Equation [3.3.9] in the 7 by 7 window, for which no changes in the thresholds (1 and 0) were needed to accommodate different spectral regions. Besides ridge points, candidate points (global maxima) were also supplied through N_{max} (the number of highest local maxima to add for each spectrum), which define the set $S_G(N_{max})$. For each selected ROI, N_{max} more local maximal points were added as candidate points for each spectrum from high to low in intensity. These candidate points were then

intersected with local maxima (set S_L) to filter out points which did not correspond to true peaks. The combination of the three criteria ($(S_R \cup S_G(N_{max})) \cap S_L$) helped identify most ridges and improved accuracy.

$$|K| < 1$$

$$H < 0 \qquad [3.9]$$

A two-step greedy connection procedure was implemented to connect candidate points into ridges for quantifying individual peaks through time (Fig. 3.2D). This procedure assumes that chemical shift variation of peaks at nearby time points is local and continuous, which is typically the case in time-series measurements. First, points adjacent in time and with the closest chemical shift distance within L_{gap} (largest step size in chemical shift dimension) were connected into segments. Second, these segments were connected into ridges to cover the entire time-range for the peak. The segment connection was based on the shortest distance and a user-adjustable threshold on angle ($\leq 60°$ default) between them. The angle threshold ensured smooth shift pattern in ridges. The order of the segment connection was ranked from high to low based on their average intensity.

In the ridge tracking process, only the parameters $N_{\max}$ and L_{gap} required tuning. The remaining parameters in the program required no modification for the simulated and experimental data sets we tested. Choices for $N_{\max}$ and L_{gap} values were also intuitive. In the majority of cases, we recommend the same small L_{gap} for most regions. The parameter L_{gap} can be increased when there is peak shifting and can be decreased when there are peaks that are close to each

other. For N_{max}, we recommend using N_{max}=1 plus the number of ridges expected but not yet tracked. The values used in the script ($N_{max} = 1$, $L_{gap} = 10$) can be used as an initial guess for other data sets (Supplementary File 3.1). More details in parameter tuning for RTExtract and the previous method (Judge et al. 2019) are described in S3.3.

Refinement of the tracking results

While most peaks can be tracked without refinement, in some cases, the tracking is imperfect, which can be solved in the refinement step (Fig. 3.2E). Chemical expertise adds value to this step, especially in regions with a low signal-to-noise ratio (SNR). Besides removing short ridges (default minimum ridge length is 5 time points), the refinement steps also include retracking for small regions, manual ridge selection, and removal of imperfect ridge ends (Supplementary File 3.1).

When multiple peaks overlap and change frequency, the greedy connection method (Methods 2.2) had difficulty deciding which direction to continue through peak crossing. This was ameliorated by local retracking. In retracking, we imposed a more stringent constraint that the peaks tend to maintain their original directions when they cross. For each time point, a small search window (length 5) for connecting next candidate points was centered at the linear extrapolation of previous (last 5 time points) chemical shift values. That is, ridges are assumed to be locally linear, which is a reasonable constraint locally. In the global tracking process, however, there are indeed rapid changes in chemical shift, so in this case the stringent constraint is not imposed.

Combining different procedures for global and local tracking increases flexibility when necessary.

The automatic ridge tracking procedure often generates false positives (Supplementary file 3.1 Fig. 3.2-7) and these false ridges can be easily distinguished by the analyst. Hence, the interactive step (manual ridge selection) boosted performance by allowing analysts to select peaks with high confidence according to their knowledge. Moreover, compound quantification can also be improved by selecting peaks with good peak shapes, minimal overlap, and high SNR. The user can also record the annotated compound name and indicate whether the tracked peaks should be used for quantification in later steps.

When the peak intensity decreases to near 0, the ridge tracking sometimes extends to noisy regions with no peak for a few time points. These imperfect ridge ends can also be removed by the ridge end removal interactive step.

The feature quantification workflow provides an interface to walk the user through ridge tracking, retracking for overlapping regions, manual ridge picking, and manual end removal. The user can decide to apply certain steps depending on their needs. All information related to the tracking process, both explicit (parameter choices) and implicit (manual tuning records), is logged in the data structure for documentation and reproducibility. The manual refinement process is easy to use and no laborious peak picking is needed. Subsequently, peak intensity and chemical shift can be plotted through time (Fig. 3.2F). More details can be found in the MATLAB tutorial (Supplementary File 3.1).

Feature mapping and quantification

Tracked ridge features need to be matched to compounds and quantified. In the simulated spectra, after ridge tracking, the time-series data for each peak were automatically matched to compounds by differences in chemical shift. The difference (D) was evaluated using Equation [3.3.10], and only the pairs with the smallest difference for all ridges and all compound peaks were selected as the final matches. In Equation [3.3.10], Ω is a set of time points overlapping between the simulated compound peaks and the tracked ridge peaks, and L is the size of the set Ω. The variables v_{ppm}^{compd} (simulated compound peak) and v_{ppm}^{ridge} (tracked ridge peak) are the corresponding chemical shift vectors within Ω. D is the sum of the squared differences between extracted and simulated chemical shifts within the overlapping time range and is normalized by the range size (L). Compounds were quantified by peak intensities normalized by the intensity of DSS peak (3-(Trimethylsilyl)-1-propanesulfonic acid sodium salt, a chemical shift reference and intensity internal standard). We also computed RMSD (Root Mean Square Deviation) between simulated and extracted chemical shift ($\delta_{sim,i}$ and $\delta_{ex,i}$, N is the length of the ridge) for each ridge (Equation [3.3.11]). The annotation and quantification of the experimental data sets follows our published methods (Judge et al. 2019). Intensity is chosen for quantification for simplicity and peak shape doesn't change for both experimental and simulated spectra.

$$D = \frac{1}{L}\sum_{i}^{\Omega}\left(v_{ppm,i}^{compd} - v_{ppm,i}^{ridge}\right)^2 \qquad [3.10]$$

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\delta_{sim,i} - \delta_{ex,i}\right)^2} \quad [3.11]$$

Data and software

Programs were written in MATLAB and R. They are shared through the Edison lab metabolomics toolbox GitHub repository (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA/tree/master/metabolomics_toolbox/code/ridge_tracking). The example experimental data can be found in Metabolomics Workbench (https://www.metabolomicsworkbench.org PR000738) and other used data can be found in Supplementary Data. We also provide a tutorial on the workflow (Supplement File 3.1).

The programs were extensively run and tested in MATLAB 2018b and R (RStudio Version 1.1.456 and R Version 3.5.1) on a macOS (Mojave 10.14.5) system.

Results

Comparison of simulated and experimental time-series NMR spectra

By their very nature, experimental datasets cannot be used to unambiguously validate the results of an algorithm such as RTExtract, especially in regions with overlap and noise. Therefore, we employed simulated datasets for method evaluation. We first briefly evaluate the complexity of both dataset and how this affects ridge tracking. Complexity in time-series NMR spectra was evaluated in SNR, peak intensity, and chemical shift variation (Supplementary Table 3.1 and method S3.2). Besides the SNR value in the main simulation,

multiple SNR levels were tested, and the workflow can still track ridges accurately in lower SNR levels (Supplementary Fig. 3.2A-D). The peak in was tracked automatically for $SNR \geq 99.24$ (Supplementary Fig. 3.2A-C) and needed some manual tuning for $SNR = 19.97$ (Supplementary Fig. 3.2D). In practice, most peaks in the experimental spectra possessed good enough SNR for tracking if they were visible in a stack plot. As an example, the valine multiplet centered at ppm 2.267 had low SNR, and 6 peaks of the multiplet could still be tracked (Supplementary Fig 3.2E). The ridge tracking method had robust performance under a large range of different SNRs.

For the complexity metrics in variation of intensity (C_{scale} and $C_{dynamics}$) and chemical shift (C_{shift}), the values are similar between experimental and simulated data sets (Supplementary Table 3.1 and S3.2). We note that using our metrics, the aerobic sample is more complex than the anaerobic sample, in agreement with our original qualitative conclusion (Judge et al. 2019).

Performance evaluation of ridge tracking on the simulated data sets

RTExtract was first tested on the simulated data sets (Supplementary Fig. 3.3 and method S3.1). Time-series NMR spectra were simulated with known concentrations and chemical shifts (simulated value), which were used to evaluate the ridge tracking result (extracted value). Extractions were evaluated by RMSD in chemical shift and most peaks were tracked accurately (low RMSD) in simulated data sets by both RTExtract and the previous method (Supplementary Table 3.2 and Supplementary Fig. 3.4).

We simulated a mixture of 15 metabolites, which yielded a possible 173 ridges with all the multiplets (S3.1). Of these 173 potential ridges, 58 were essentially overlapped in the final simulations, allowing for 115 distinct ridges for analysis. Of these, 61 had some overlap and 54 were without overlap. We plotted extracted concentrations ($[C]_{ex}$) against simulated concentrations ($[C]_{sim}$) and observed that peaks without overlap were all near the diagonal (Fig. 3.3A). Small differences in intensity from the line broadening function created small deviations from the slope of 1. Nearly half of the ridges were not overlapped and could be accurately quantified.

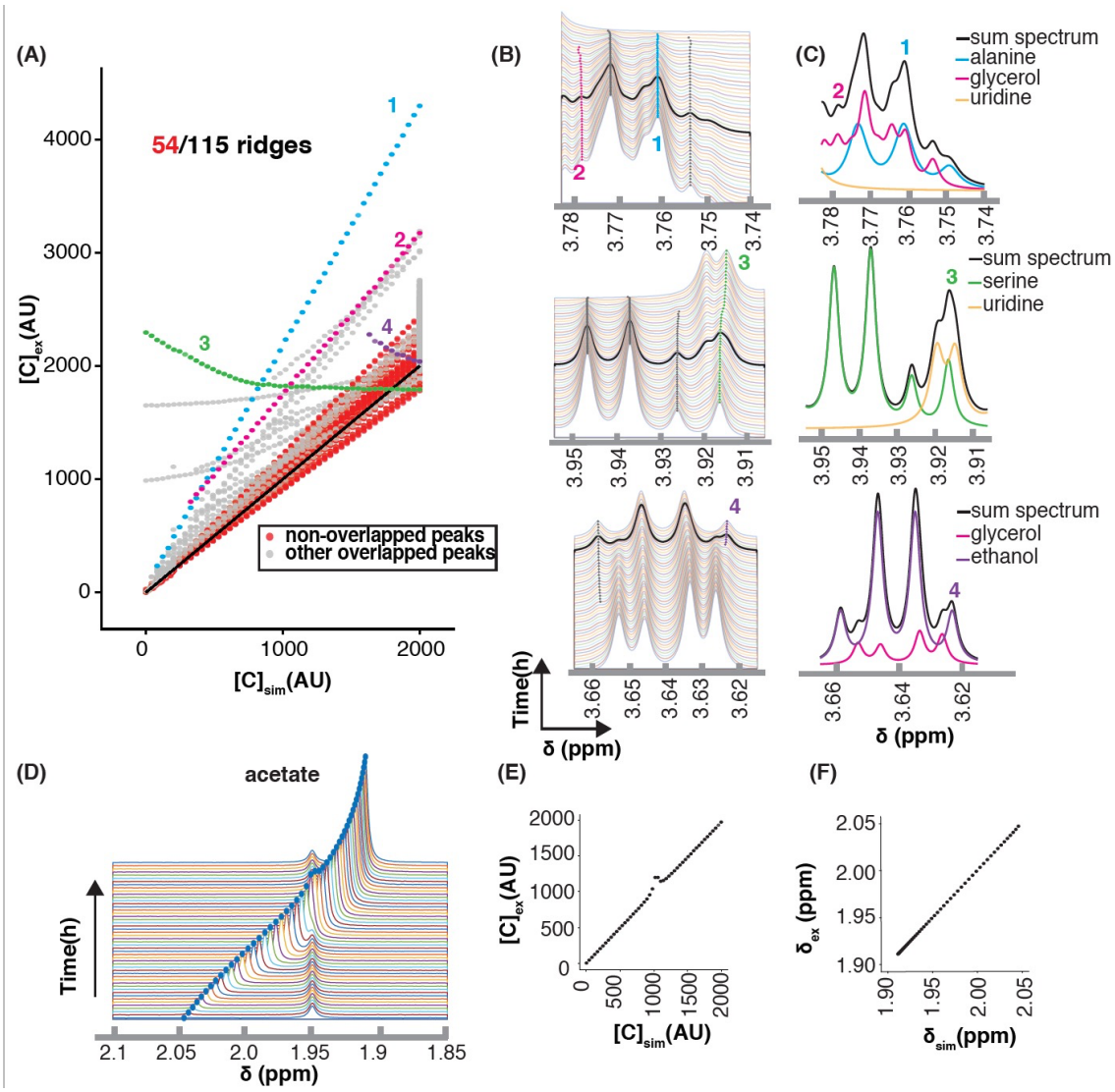


Figure 3.3: Evaluation of the RTExtract algorithm on simulated data sets.

(A) Compound quantification for all ridge points were plotted. For each ridge peak, extracted compound concentrations normalized to DSS (Y-axis, $[C]_{ex}$) were plotted against simulated compound concentration (X-axis, $[C]_{sim}$). The black diagonal represents perfect quantification and its slope is 1. Four overlapped peaks mapped to four compounds (1: alanine (blue); 2: glycerol (pink); 3: serine (green); 4: ethanol (purple)) were selected as examples (A-C). All other ridge overlapped peaks are in grey. (B) In the stack spectra, one

spectrum was highlight in black, which was then decomposed into compound peaks involved in the overlap (C). In (C), the black line is the full simulated spectrum with all the compound peaks. (D) the acetate peak was simulated with chemical shift variation under different pH conditions and overlapped with another peak. (E) the extracted concentrations ($[C]_{ex}$) were plotted against simulated concentrations ($[C]_{sim}$). (F) extracted chemical shifts (δ_{ex}) were plotted against simulated chemical shift (δ_{sim}). Performance of concentration and chemical shift estimation with more compounds can be found in Supplementary Fig. 3.5 and Supplementary Fig. 3.6. Compound concentration was simulated with arbitrary unit (AU) and chemical shift δ was evaluated in ppm.

For peaks with overlap, quantification was affected. For example, the alanine H^a peak (peak 1 in Fig 3.3A-C), $[C]_{ex}$ was overestimated with a linear curve. Alanine peaks are overlapped with glycerol peaks, leading to inaccurate quantification for both metabolites (peak 1 and 2 in Fig. 3.3A-C). Overlaps with glycerol also caused quantification of the ethanol peak to change in the opposite direction (peak 4 in Fig 3.3A-C). In this case, the intensity variation of the small side peak from ethanol is dominated by the variation in glycerol peak.

Besides intensity estimation, overlap can affect chemical shift (e.g. peak 3 of serine in Fig. 3.3A-C). In this overlapping region, the uridine concentration increased through time, and serine concentration decreased. This resulted in the green line 3 in Fig. 3.3A-B, which had a superposition of increasing and

decreasing intensities. This kind of continuous shift between two features not only caused incorrect quantification but also an incorrect chemical shift variation. Even though neither uridine nor serine had clear peak shifts for the pH range considered ([4.0, 6.0]), the overlapped peak shifted smoothly. Through the changes of relative intensities of the two peaks, the chemical shift of the overlapped peak changed.

When peaks are separated enough to be distinguishable, ridge tracking is often accurate in chemical shift estimation (Fig. 3.3D). For the acetate peak, the concentration ($[\mathrm{C}]_{\mathrm{ex}}$) was over-estimated in the overlapped region (Fig. 3.3E), but the chemical shift (δ_{ex}) was estimated accurately (Fig. 3.3F). Both peaks could be tracked for the entire range through the overlapped region, and the relative intensity between the two overlapping peaks did not affect tracking capability. Comparison of extracted and simulated values in concentration and chemical shift for more compounds can be found in Supplementary Fig. 3.5 and 3.6.

Performance evaluation of ridge tracking on the experimental data sets

The ridge tracking method was next tested on experimental data sets and compared with the approach used in the initial CIVM publication (Judge et al. 2019). We first assessed the agreement on quantification of regions without much overlap (Fig. 3.4 and Supplementary Fig. 3.7). For quantification under aerobic and anaerobic conditions, the residuals between the two methods were close to zero for most compounds (19/22, Supplementary Fig. 3.7), and RTExtract reproduced the earlier results with much less time and manual input. The few differences were attributed to how negative values were dealt with

between the two methods in computing Mean Scale Ridge Intensity (Fig. 3.4 and Supplementary Fig. 3.7). In RTExtract, time-series data with negative intensities are shifted to positive first, which was not done in the original publication (Judge et al. 2019).

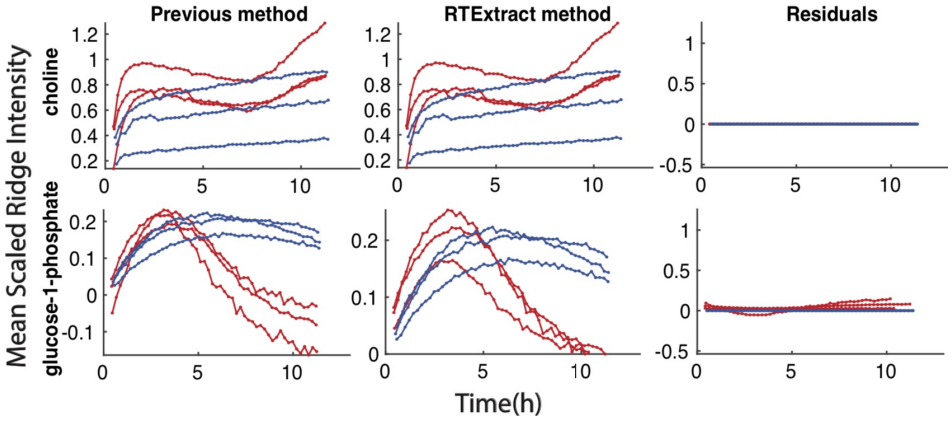


Figure 3.4: Reproduction of compound quantification results from our previous method. In each plot, the X-axes indicates time, and the Y-axes indicates Mean Scaled Ridge Intensity. Red curves are from aerobic data sets, and blue curves are from anaerobic data sets. Details in computing Mean Scaled Ridge Intensity can be found in the previous publication (Judge et al. 2019)). Each row indicates quantification for one compound, including the previous, the RTExtract methods and residuals (differences between the quantifications by the two methods). Comparison for more compounds can be found in Supplementary Fig. 3.7.

RTExtract also worked for more complex regions that were difficult for the previous method (Fig. 3.5 and Supplementary Fig. 3.8) (Judge et al. 2019). The complex regions contained different degrees of overlap and/or peak shifting. When two peaks are closely overlapped with each other, the original method often produced tracking results that “jumped” between the two (e.g. Fig. 3.5A and Supplementary Fig. 3.8). In RTExtract, these peaks were tracked with no jumps, resulting in fewer errors in chemical shift and intensity estimation. Parameter tuning, particularly in complex regions, was difficult in the original publication but is much easier in RTExtract. The glutamate region (ppm [2.3, 2.44], Fig. 3.5B) was another difficult case, in which the six peaks from glutamate shifted with pH and overlap with an unknown peak (yellow). By the previous method, only a small side peak in the multiplet was tracked for glutamate, so the quantification had a low SNR (Judge et al. 2019). By RTExtract, all the six glutamate peaks in the multiplet could be tracked with the retracking approach.

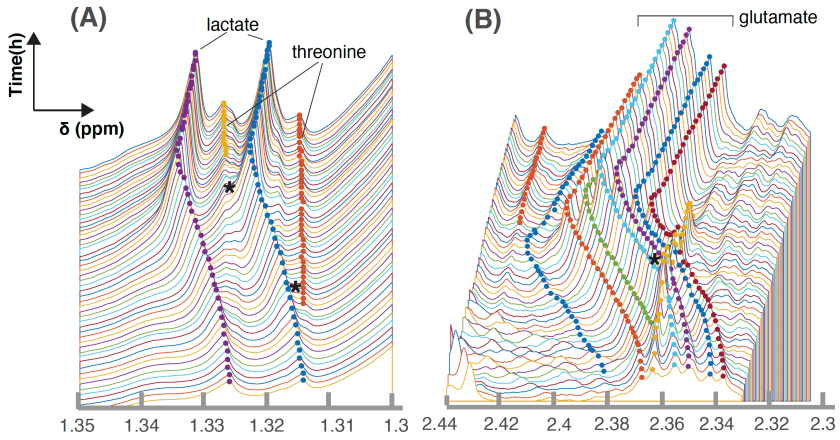


Figure 3.5: Evaluation of RTExtract on complex overlapping regions on the experimental data sets. Two ROIs (A-B) were selected as examples. Peaks in these ROIs can be precisely tracked, and the parts that are problematic in the

previous method are indicated with stars. The middle six peaks in B are annotated to glutamate. Different point colors indicate different tracked peaks. Performance of the algorithm for less complex regions is in Fig. 3.4 and Supplementary Fig. 3.7. More example results from RTExtract and comparison with the previous method are in Supplementary Fig. 3.8. Tracking for B is given as an example in Supplementary File 3.1.

Discussion

RTExtract, a computer vision-based approach is introduced in this paper to quantify time-series NMR spectra. RTExtract takes less time and exhibits better performance on complex regions than our original, less automated, approach (Judge et al. 2019). It provides a more practical way to process time-series NMR spectra and analyze *in vivo* metabolic dynamics of an organism.

RTExtract is an improvement from multiple perspectives. First, we reduced the number of tuning parameters from seven to two, which reduces the interactive time and is more intuitive to optimize. Second, the refinement steps allow fine-tuning of the ridge tracking process and easily remove imperfect regions. Instead of exploring a huge parameter space, the user can fix the imperfect regions through simple manual steps. With these two improvements, the published results can be replicated within 2 hours by RTExtract instead of ~48 hours by the original method (Judge et al. 2019). Finally, RTExtract is also capable of dealing with more complex regions, especially with peak overlap and

peak shifting (Fig. 3.5 and Supplementary Fig. 3.8). It is now possible to track most peaks in these difficult regions, without merging multiple peaks into one. Subsequently, more tracked features can be used for downstream modeling and statistical analyses. We also see that when two peaks are highly overlapped and their concentrations change in opposite way, the overlapped peak might seem to change in chemical shift (Fig. 3.3B peak 3). This could be mistaken for chemical shift change due to pH variation and seems to also occur on the glucose peaks (ppm region [5.2, 5.26]) under the aerobic condition in experimental data sets (Judge et al. 2019).

We still offer the option of manual interaction in the workflow, which helps produce accurate results but still requires expertise, time, and manual effort. Future versions of the workflow will incorporate statistical filters accompanied by higher degrees of automation. A clustering-based method can be implemented to remove artifact ridges, which are characterized by random changes in intensity and chemical shift. Implementing this step might fully remove the manual procedure and make the full process much faster.

The RTExtract can also combine with spectral deconvolution for overlapping feature quantification. From RTExtract, chemical shift and intensity of individual overlapping peaks can be obtained and subsequently fed into the deconvolution methods. Based on the information of intensity and chemical shift, a Bayesian-based deconvolution approach can compute the underlying peak intensity (Hao et al. 2014a; Krishnamurthy 2013).

In principle, as long as peaks are changing in a continuous manner, they can be tracked by RTExtract. The experimental data tested in this paper is from the CIVM method (Judge et al. 2019), and provides dense, continuous, time-series measurements. Other time-series NMR methods, such as flow NMR and *in vitro* sampling from NMR can also provide proper candidate measurements (Foley et al. 2014). Possible applications go beyond time-series measurements as long as the continuity constraint is met between neighboring spectra. For example, we used RTExtract to track peaks in a citrate pH titration experiment, fit the Henderson-Hasselbalch equation for the chemical shift changes under different pH, and estimate pK_a (Supplementary Fig. 3.9) (Edison et al. 1999; Szakacs et al. 2004; Tredwell et al. 2016b; Zachariah et al. 2001). Likely, similar peaks could be tracked in pH or ligand-binding titrations of proteins (Brockerman et al. 2019; Joshi et al. 1997). A preprocessing by chemical shift sorting can even make independent samples of urine data accessible to RTExtract (Liebeke et al. 2013).

Conclusion

RTExtract is introduced in this paper to quantify dense time-series NMR spectra by ridge tracking. It is faster, easier to use, and can deal with more complex regions than previously published methods. The extraction is accurate even in complex overlapping regions. As the ridge tracking method relies on the continuity of peaks at neighboring spectra, it can be further applied to other suitable data types.

Acknowledgements

We thank Heinz-Bernd Schuttler, Leidong Mao, and Juan B. Gutierrez for helpful discussions on the computational method, Peter Kner for discussion on image processing, and John N. Glushka for discussion on NMR spectral processing. Goncalo Gouveia, Amanda Shaver, Mario Uchimiya, and Nicole Holderman have provided helpful suggestions on figures. Laura Morris has provided technical support on computers. Rahil Taujale, Mario Uchimiya, Maxwell Colonna and Sicong Zhang have provided helpful feedback on using the RTExtract tutorial.

CHAPTER 4

UNCOVERING *IN VIVO* BIOCHEMICAL PATTERNS FROM TIME-SERIES METABOLIC DYNAMICS [3]

[3]: Wu Y, Judge MT, Edison AS, Arnold J. Uncovering in vivo biochemical patterns from time-series metabolic dynamics. PloS one. 2022;17(5):e0268394.

Foreword

This chapter is reprinted from Wu Y, Judge MT, Edison AS, Arnold J. Uncovering in vivo biochemical patterns from time-series metabolic dynamics. PloS one. 2022;17(5):e0268394. and is available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0268394. I and Michael T. Judge were equally contributing co-first authors. My contribution to this work consisted of: (1) designing the process of analyzing high-dimensional metabolomics time series, (2) building the statistical methods for dimensionality reduction and information extraction, (3) producing visualizations, (4) interpreting the biological process, and (5) writing and editing the manuscript and addressing reviewer comments. The contribution of Michael T. Judge was as follows: (1) performing *in vivo* NMR experiments and preprocessing the data, (2) visualizing the results, (3) interpreting the biological process with experimental insights, and (4) writing and editing the manuscript, and addressing reviewer comments. Arthur S. Edison and Jonathan Arnold designed the project goal, edited the manuscript, and responded to the reviewers. The research was supported by National Science Foundation awards (MCB-2041546 and NSF 1713746 and NSF ERC 1648035). The supplementary materials in the chapter are listed in Appendix C.

Abstract

System biology relies on holistic biomolecule measurements, and untangling biochemical networks requires time-series metabolomics profiling. With current metabolomic approaches, time-series measurements can be taken

for hundreds of metabolic features, which decode underlying metabolic regulation. Such a metabolomic dataset is untargeted with most features unannotated and inaccessible to statistical analysis and computational modeling. The high dimensionality of the metabolic space also causes mechanistic modeling to be rather cumbersome computationally. We implemented a faster exploratory workflow to visualize and extract chemical and biochemical dependencies. Time-series metabolic features (about 300 for each dataset) were extracted by Ridge Tracking-based Extract (RTExtract) on measurements from continuous *in vivo* monitoring of metabolism by NMR (CIVM-NMR) in *Neurospora crassa* under different conditions. The metabolic profiles were then smoothed and projected into lower dimensions, enabling a comparison of metabolic trends in the cultures. Next, we expanded incomplete metabolite annotation using a correlation network. Lastly, we uncovered meaningful metabolic clusters by estimating dependencies between smoothed metabolic profiles. We thus sidestepped the processes of time-consuming mechanistic modeling, difficult global optimization, and labor-intensive annotation. Multiple clusters guided insights into central energy metabolism and membrane synthesis. Dense connections with glucose 1-phosphate indicated its central position in metabolism in *N. crassa*. Our approach was benchmarked on simulated random network dynamics and provides a novel exploratory approach to analyzing high-dimensional metabolic dynamics.

Author Summary

Metabolic networks are composed of metabolites and reactions, with or without regulation, as nodes and edges. These networks behave dynamically and adjust rapidly to the changing needs of the organism. Understanding these dynamic changes in metabolite concentrations and reactions is crucial for research in basic biology, disease development and progression, and bioprocess control in industrial applications (e.g., fermentation). Conventional approaches to analyzing metabolic dynamics have focused on direct simulation and optimization of differential equation systems. They have been computationally intensive and restricted by comprehensive, labor-intensive annotation of the metabolome. Here we provide an exploratory approach that is faster and expands incomplete metabolic annotation. From experimental time-series metabolic measurements of *Neurospora crassa*, we extracted dominant metabolic trends, expanded annotation by a correlation network, and identified metabolic clusters consistent across different conditions. From these extracted patterns (trends and network clusters), we uncovered processes related to central energy metabolism and membrane synthesis. Our approach provides an alternative to interpreting metabolic dynamics in the context of incomplete metabolite annotation and quantification, thus allowing full utilization of untargeted time-series metabolomics data.

Introduction

Living organisms rely on a complex metabolic network, composed of thousands of metabolites and reactions (Caspi et al. 2020). Recent developments in experimental approaches (Judge et al. 2019; Koczula et al. 2016; Link et al. 2015) and feature extraction (Y. Wu et al. 2020) have enabled direct observation of complex metabolic dynamics, the ultimate phenotypic response linking genes to metabolism (Beadle and Tatum 1941; DeRisi et al. 1997). Here, we describe novel computational tools to extract biological knowledge from such high-dimensional metabolic time-series datasets in a data-driven approach (Ideker et al. 2001).

Time-series metabolomic datasets are rich and complex, but current statistical methods are inadequate. Metabolic measurements often have hundreds to thousands of features, making it a high-dimensional problem. Sampling as a function of time by novel tools, such as continuous *in vivo* monitoring of metabolism by NMR (CIVM-NMR), further complicates the problem by introducing dynamics for each feature (Judge et al. 2019). Many features also have variable patterns in peak locations and overlap. Fortunately, the latent dimensionality of the metabolome is constrained, as changes in metabolites tend to be smooth over time, and dependencies exist between metabolic features. Novel statistic methods utilizing the time-series dependencies are needed and they will inform chemical identification and biological discoveries, which are important goals of metabolomics.

One conventional approach is to model the metabolome explicitly. A metabolic network is a dynamic system with metabolites (nodes) dependent on each other through reactions and regulation (edges). Direct modeling of such a network often involves flux balance approaches (FBA) (Edwards et al. 2002) or, when dynamics are needed, simulation of ordinary differential equations (ODEs) (Battogtokh et al. 2002; Raue et al. 2013; Y. Yu et al. 2007). Such direct approaches face challenges in computation and data accessibility. Including a realistic topology and making appropriate parameter estimations are computationally expensive (A. M. Al-Omari et al. 2022; Battogtokh et al. 2002; Brown and Sethna 2003; Meyer et al. 2014; Raue et al. 2013). Compromises in topological structure (e.g., regulation) can result from a lack of pathway knowledge or enough data and significantly reduce a model's utility. Specifically, many compounds cannot be observed or confidently annotated and quantified, and useful ODE solutions might not be discovered due to the additional uncertainty (Judge et al. 2019; S. Li et al. 2013; Link et al. 2015). Typical time-series metabolomic measurements leave most nodes in metabolic networks unobserved because of limitations in detection and annotation (Judge et al. 2019).

Our alternative approach enables interactive exploration of metabolic dynamics and extracts biological and chemical information from the time series. Based on the framework of functional data analysis (FDA), we smoothed the time-series metabolic features and projected them into lower dimensions to visualize the dominant metabolic trends (G. Montana et al. 2011a; Ramsay and

Silverman 2005; Ramsay et al. 2009). We also enabled a quick comparison of experiments under different perturbations and revealed metabolic adaptations. Furthermore, we identified networks and clusters based on correlation and dynamical associations between time series separately (Cloarec et al. 2005; Hackett et al. 2020; Klimovskaia et al. 2016; Newman and Girvan 2004; Pfister et al. 2019a). The correlation network expanded chemical annotation and the empirical dynamic network revealed *in vivo* biochemical functions (Caspi et al. 2020). Our workflow thus prioritizes NMR features for further biological investigation.

Results

Workflow to analyze time-series NMR spectra

In this study, we built a workflow to enable knowledge discovery from the new type of data, time-series NMR. CIVM-NMR measures *in vivo* metabolism through time, producing rich and complex spectral data. Such datasets can reveal the underlying dynamic metabolic process, profile metabolic adaptations to perturbations, facilitate annotations and uncover biochemical regulation. These are accomplished through dimensionality reduction and network construction.

The data were collected continuously on the model filamentous fungus, *N. crassa*, which lived in the NMR probe for about 12 hours (Fig. 4.1A) (Judge et al. 2019). We recorded NMR spectra of the living organism as it metabolized, which led to a complex spectral surface with coordinates in parts per million (ppm) for the NMR axis and hours for the time axis (Fig. 4.2A). In this study, we worked with experiments of different carbon sources and labeling: six experiments

feeding on glucose (three of an aerobic condition and three of an anaerobic condition) (Judge et al. 2019) and two experiments feeding on ^{13}C uniformly labeled pyruvate in an aerobic condition.

Each time-series dataset was processed in a global and untargeted way through RTExtract (Fig. 4.1B) (Y. Wu et al. 2020). About 300 metabolic features were extracted for each dataset, consisting of around 10^5 data points per experiment. RTExtract efficiently quantified time-series NMR spectra even for peaks with overlap and pH-induced chemical shift changes (Judge et al. 2019; Y. Wu et al. 2020).

We first projected the metabolic features into a lower dimension to visualize differences in dynamics (Fig 4.1C). Compounds were grouped, showing different biochemical processes (Fig. 4.2 and Supplementary Fig. 4.2). The same compound was also compared under different conditions to show responses to perturbation (Fig 4.3 and Supplementary Fig. 4.3). We then constructed networks based on time series and clustered them to search for chemical and biological associations in the living sample (Fig. 4.1D-E). The correlation network produced 34 clusters and seven validated compound annotations without requiring extraction or 2D experiments (Figs 4.1D, 4.4 and Supplementary Fig. 4.4). This can also include those compounds that were consumed and below detection level at the extraction time point. The functional network presented *in vivo* metabolic processes, including central energy metabolism and phospholipid metabolism (Figs 4.1E, 4.5, 4.6, and Supplementary Fig. 4.5).

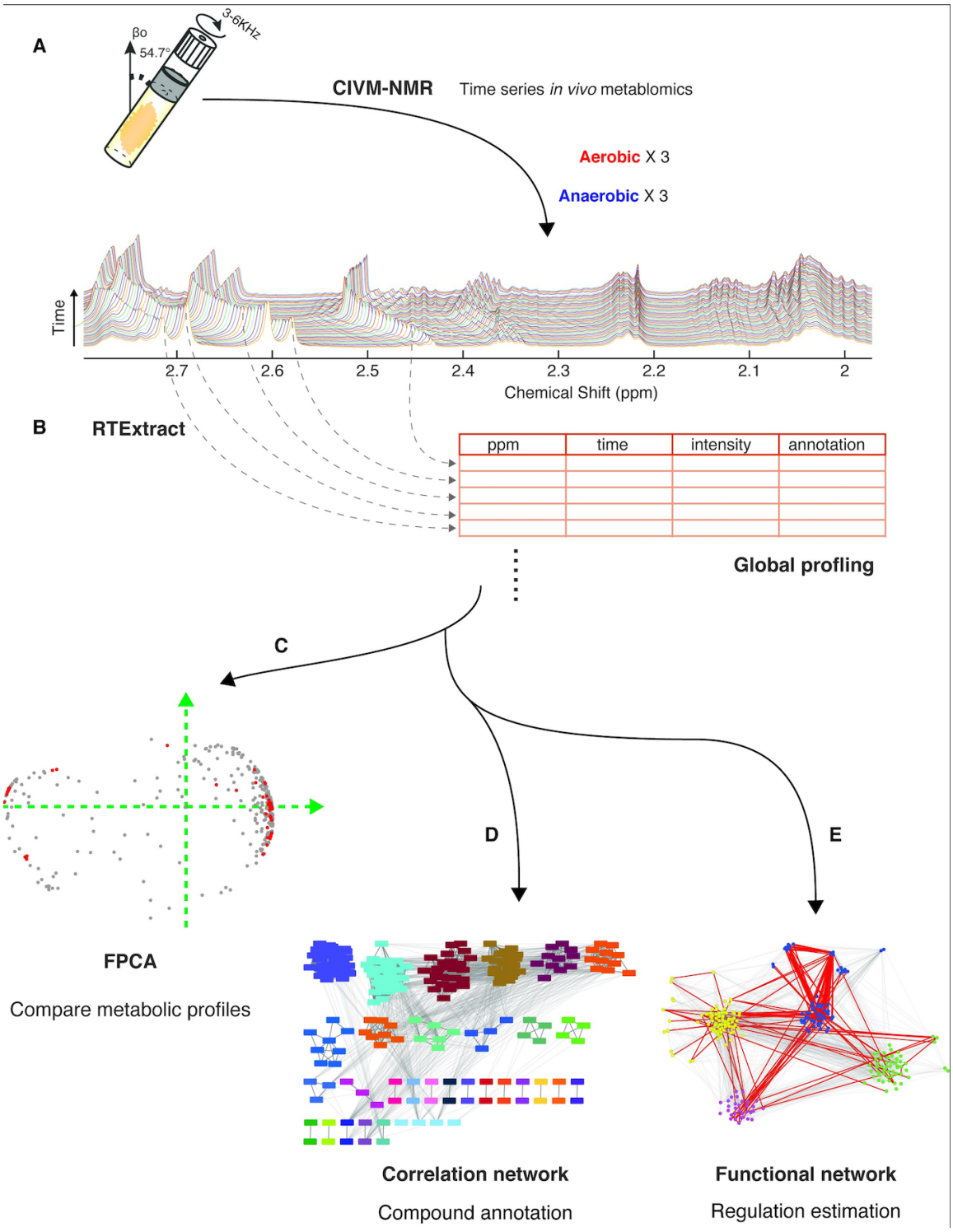


Figure 4.1: The analytic workflow for time-series NMR data. A: time-series *in vivo* measurements were collected through CIVM-NMR under different conditions, including oxygen accessibility (aerobic vs anaerobic) and different

carbon sources. B: RTExtract efficiently extracted peak information and enabled a global untargeted metabolic profiling (~300 time-series features/sample). The data were analyzed by multiple approaches (C-E). C: Dimensionality reduction using FPCA compared metabolic trends under different conditions. D: A correlation network provided annotations. 34 clusters were found by correlating the time series. E: Functional groups and regulation were estimated through dependencies between time-series features. Four clusters were found through CausalKinetiX and community clustering of the *in vivo* time series. Detailed information can be found for each approach: C (Figs 4.2 and 4.3; Supplementary Figs 4.2 and 4.3), D (Fig. 4.4 and Supplementary Fig. 4.4) and E (Fig. 4.5 and Supplementary Fig. 4.5).

Dimensionality reduction of metabolic dynamics

Time-series NMR spectra have high dimensionality, and the patterns are diverse for different metabolites (Fig. 4.2B). Projection into lower dimensions through the functional principal component analysis (FPCA) visualizes different groups and trends between metabolites and between conditions.

We first compared metabolic features in aerobic (Fig 2B) and anaerobic conditions (Supplementary Fig. 4.2A). In both conditions, the first two principal components (PCs) explained the major variance (PC1 about 80%) (Fig 4.2B; Supplementary Figs 4.1 and 4.2B). The PC1 eigenfunction indicates an increasing trend added upon the mean curve of all metabolic features (black

curves in Fig. 4.2B and Supplementary Fig. 4.2C; Details in Methods). Metabolites that were consumed appear as negative values along PC1, while those that were produced have positive values along the same axis. PC1 captures the dominant metabolic trend of consuming carbohydrates (e.g., glucose and trehalose) and producing amino acids (e.g., tyrosine and alanine) and fermentation products (e.g., ethanol). The PC2 eigenfunction is orthogonal to PC1 and has a decreasing and then increasing pattern (Fig. 4.2B), which captures more complex dynamics including glucose 1-phosphate (G1P) and choline. PC2 accounts for more variance in the aerobic condition (13%) than in the anaerobic condition (8%) (Fig. 4.2B and Supplementary Fig. 4.2C), in agreement with the inspection of the NMR spectra (Judge et al. 2019). Most metabolites have multiple NMR peaks, and these are clustered together in the score plot, demonstrating the stability of FPCA separation (Fig. 4.2B and Supplementary Fig. 4.2A). For example, glucose nodes are clustered and similar to the trends of trehalose (Fig. 4.2B).

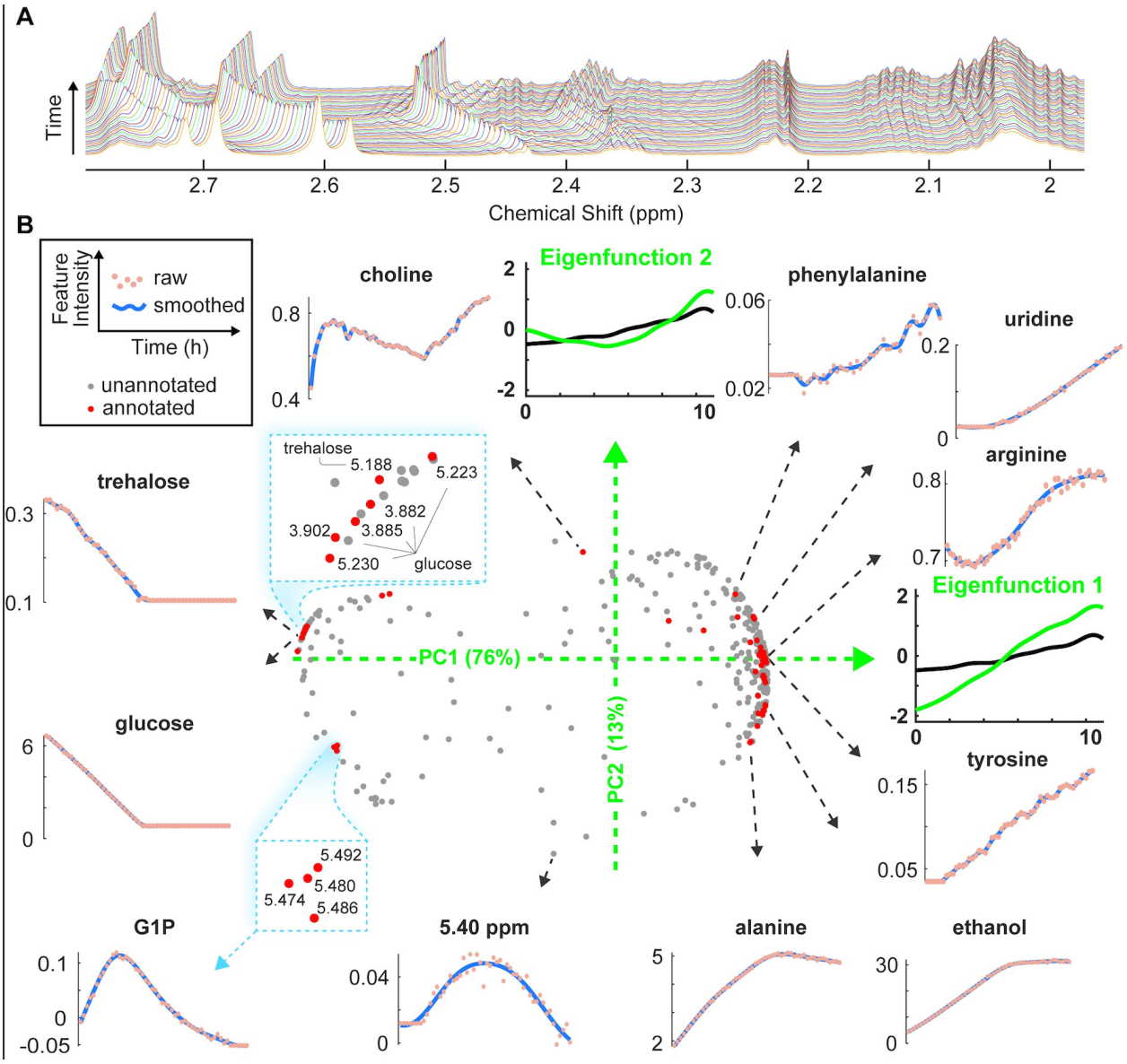


Figure 4.2: Dimensionality reduction through FPCA captures dominant variations in metabolic dynamics within individual samples. A: An example CIVM-NMR spectral region of *N. crassa* under the aerobic condition. The stack plot shows changes in a region of the NMR spectra through time. Each time point is distinguished by a different line color. B: Time series of *N. crassa* NMR features collected under aerobic conditions were smoothed and visualized in two dimensions. In the middle panel, each NMR feature trajectory (unannotated,

grey; annotated, red) is summarized as a combination of the two dominant eigenfunctions (green axes) which define the axes of the FPCA scores plot. The X (Y) axis represents scores for PC 1 (2), and the variance percentage of each PC is given in parentheses. Eigenfunctions in FPCA are analogous to loading vectors in multivariate PCA. While loading vectors are often presented as effects of each feature, eigenfunctions are presented as the smoothed effect added to the non-constant mean curve. The green curves represent adding a fraction (square root of eigenvalue) of the corresponding eigenfunction to the mean curve of all NMR features (black curve). Selected time-series features are presented around the scores plot (raw NMR peak intensities, points; FDA-smoothed intensity profiles, blue lines). In the scores plot, two inset figures (blue boxes) show details of glucose and G1P clusters with chemical shifts of the NMR features. All nodes in the G1P cluster belong to G1P. Each NMR feature was mean-centered and scaled by standard deviation before the FPCA analysis. Time is in hours. Percentages of explained variance of different PCs are presented in Supplementary Fig. 4.1. Results for one anaerobic dataset can be found in Supplementary Fig. 4.2. Details on preprocessing, smoothing and FPCA can be found in Methods.

Time series in the four FPCA quadrants represent different trends in metabolism (Fig. 4.2B). Curves with positive PC1 and negative PC2 increased and plateaued, and they correspond to compounds related to fermentation and

storage (e.g., alanine and ethanol). Curves with positive PC1 and PC2 continued to increase over time (e.g., phenylalanine and uridine), and several amino acids are in this group. Curves with negative PC1 and positive PC2 decreased and plateaued (e.g., glucose and trehalose). These carbohydrates were consumed until they fell below our detection limits of about 50 μM in CIVM-NMR ^{1}H spectra (Judge et al. 2019). Other curves with relatively high magnitude in PC2 values had more nonlinear patterns (e.g., choline and the unknown feature at 5.4 ppm), indicating more complex metabolic processes. In addition to the general patterns represented by each quadrant, the scores themselves also followed a continuous trend as the variance percentages of the PCs change between points (Fig. 4.2B). For example, the arc of scores from alanine to phenylalanine mirrored the changes in corresponding curves as PC2 gradually increased from negative to positive.

Compare metabolic experiments through FPCA

The FPCA projection can be expanded to the comparison of metabolic profiles under different experimental conditions. We can visualize metabolic adaptations by comparing the dynamics of the same compound with different media or mutants. To illustrate the flexibility of this approach, we compared five different CIVM-NMR datasets, all grown aerobically (Fig. 4.3 and Supplementary Fig. 4.3). Three used natural abundance glucose as the carbon source and two used uniformly ^{13}C-labeled pyruvate as the carbon source. The glucose replicates were all grown at a high density of about 10 mg per rotor volume

(Judge et al. 2019). The pyruvate replicates were grown at two different densities: one at 10 and the other at 6 mg per rotor volume.

In an isotopic labeling study, CIVM-NMR allows to simultaneously monitor both labeled and unlabeled metabolites in one experiment (Judge et al. 2019). We accomplished this by interleaving a standard 1D ^{1}H experiment that detects all ^{1}H atoms in the sample and a 1D ^{13}C-HSQC that selects only ^{1}H atoms with directly bonded ^{13}C (Details in Methods). This approach enables us to detect different pools of the same metabolite that originate from different metabolic pathways, as the organism is unlabeled at the start of the isotopic labeling experiment.

We compared patterns of the same compounds under different conditions in FPCA and showed corresponding extracted ridges for the same compound (Fig. 4.3 and Supplementary Fig. 4.3). The pyruvate consumption pattern is similar to that of glucose (Supplementary Fig. 4.3). The ethanol features from high-density cultures all fell on the positive PC 1 axis, regardless of carbon source and labeling. In contrast, the unlabeled ethanol from the low-density pyruvate culture is negative on the PC 1 axis, clearly showing a density dependence of ethanol production (Fig. 4.3). The ethanol produced in the high-density glucose cultures is tightly clustered for all the replicates.

The ethanol in the ^{13}C-labeled pyruvate cultures has interesting dynamics. In both densities, the unlabeled ethanol (green labels, Fig. 4.3) is on the positive PC 2 axis, and it was first consumed and then produced after 3 or 6 hours for the high and low densities, respectively (Fig. 4.3B). In contrast, the ^{13}C-labeled

ethanol patterns (brown labels, Fig. 4.3) are identical for the two culture densities: they increased and then plateaued at about 4 hours. A similar pattern exists in glucose culture, where ethanol plateaued when the major carbon source, glucose, was exhausted.

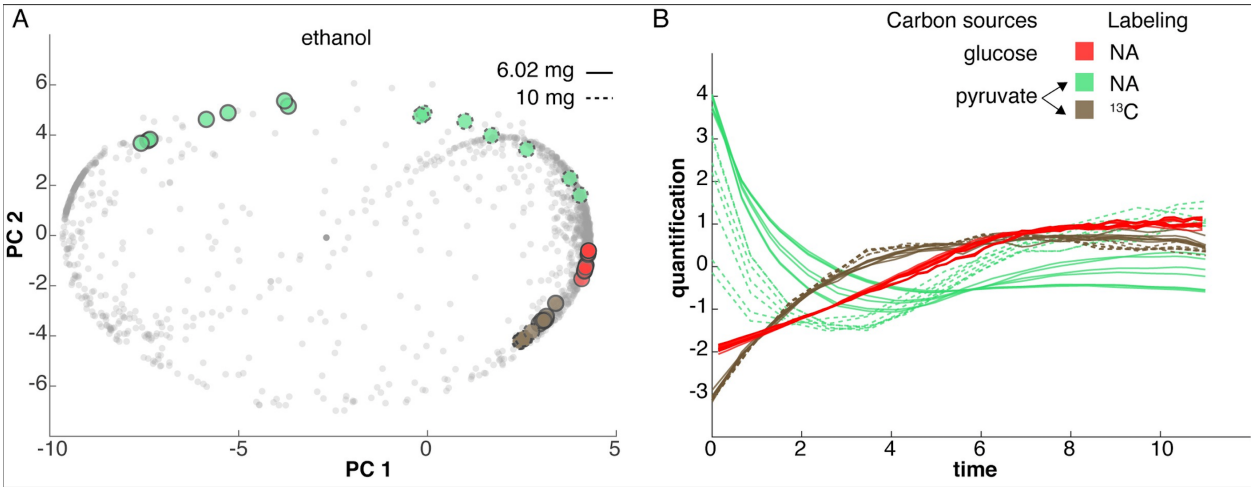


Figure 4.3: Comparing ethanol profiles under different growth conditions through FPCA. Two different carbon sources were compared: natural abundance glucose (3 experiments) and uniformly ^{13}C-labeled pyruvate (2 experiments). The glucose experiments were all done at a high density (10 mg/63 mL), and the pyruvate experiments were done at low (6mg /63 mL, solid lines) and high (10 mg/63 mL, dashed lines) densities. The ethanol features from the glucose experiments are shown in red. The ^{13}C-labeled ethanol produced in the ^{13}C-pyruvate experiments is shown in brown. The unlabeled ethanol produced in the ^{13}C-pyruvate experiments is shown in green. A: FPCA score plot indicates the overall patterns of ethanol in each of these experiments. Each point

represents one ridge in one sample. The small, grey points correspond to all the other ridges detected in these experiments. The X (Y) axis represents scores for PC 1 (2). B: Time trajectory (hours) of the highlighted features from A. Each curve was centered and scaled for A and B. Similar plots for other compounds can be found in Supplementary Fig. 4.3.

Expand metabolite annotation through a correlation network

Among approximately 300 features in each CIVM-NMR dataset, about 60 features (20 compounds) were annotated at the expense of time-consuming experiments and expert labor in our original study (Judge et al. 2019). We note that due to the nature of the CIVM-NMR experiment and RTExtract algorithm (Y. Wu et al. 2020), our extracted features are individual components of J-coupled multiplets, which should have perfect linear correlations in ideal cases. With the extracted ridges, additional information can be directly obtained through clustering a correlation network (CN) of ridge intensities without any 2D experiments. Similar to statistical correlation spectroscopy (STOCSY) (Alves et al. 2009; Cloarec et al. 2005), we constructed a CN using the Spearman correlation of extracted CIVM-NMR ridges in the 6 glucose feeding experiments, found 30 well-separated clusters in the CN, searched each cluster in 1D database and validated candidates for annotation (Fig. 4.4 and Supplementary Fig. 4.4; Details in Method).

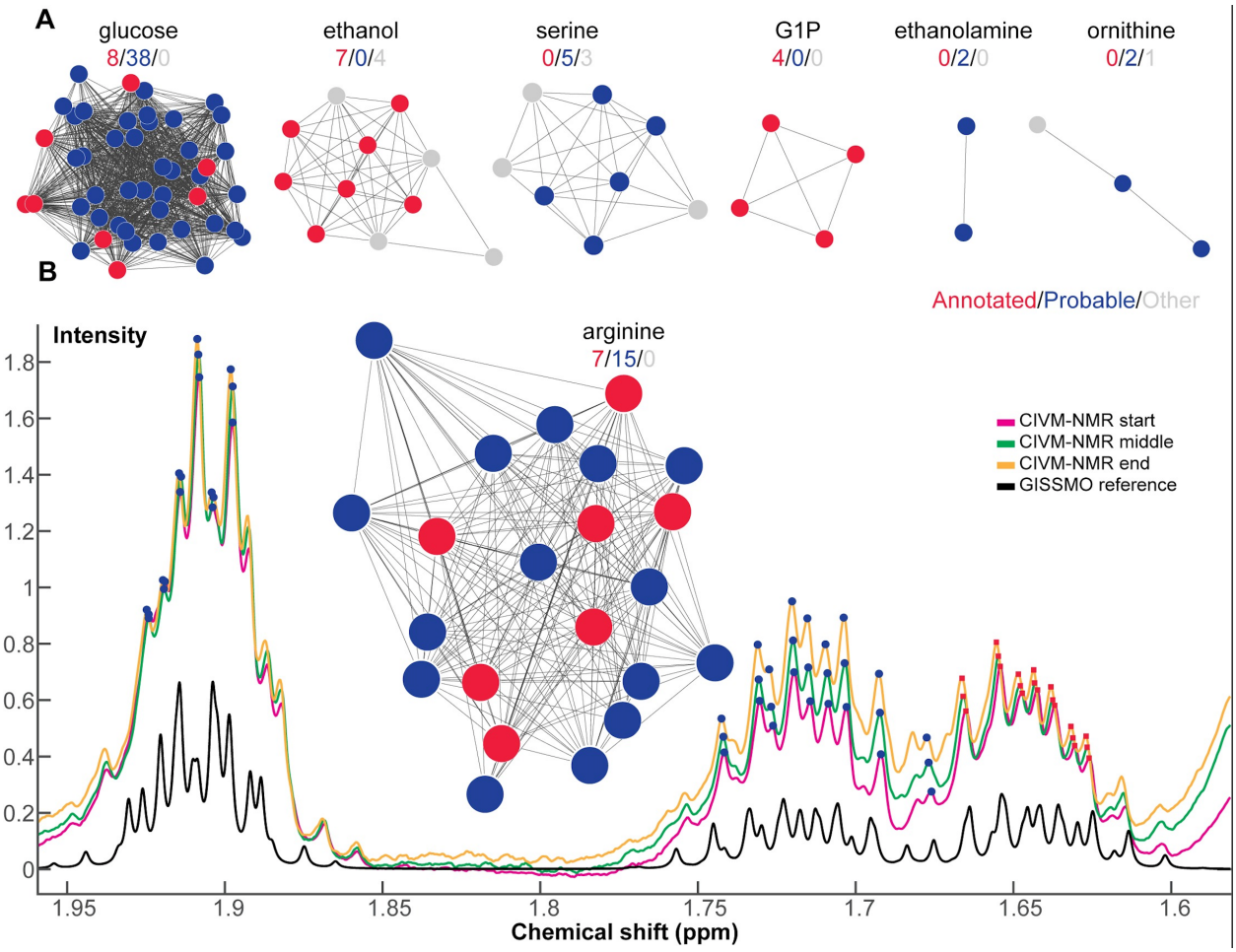


Figure 4.4: Clusters in correlation networks contribute to annotation. A correlation network was built using Spearman correlation between time-series features in the glucose feeding experiments, and clusters were discovered (More details in Methods). Clusters in the correlation network agreed with previous annotations. A: Selected clusters are co-visualized with annotations from our prior publication (Judge et al. 2019). Some compounds can be directly annotated by correlation network clustering. Red nodes represent NMR features with prior annotations (Judge et al. 2019); blue nodes represent NMR features that were consistent with reference peaks from the assigned compound but could not be confidently annotated due to spectral overlap; gray nodes represent other

unannotated NMR features. The numbers above each cluster indicate the number of nodes for each type. B: The cluster for arginine is co-visualized with experimental and reference spectra. The X (Y) axis represents chemical shift (intensity). The pink-, green-, and yellow-colored spectra are from three representative time points in the CIVM-NMR dataset. The black spectrum is the same region of the GISSMO (Dashti et al. 2017) reference spectrum for arginine. Red squares (blue circles) represent features with confident (probable) annotation to arginine and correspond to the color in the network cluster. The entire clustered CN is displayed in Supplementary Fig. 4.4.

Metabolite identification and database matching are always challenging in metabolomics (Edison et al. 2021; Monge et al. 2019). Our lab (J. M. Walejko et al. 2018a) and others (Sumner et al. 2014) proposed different confidence scales for annotation. The traditional annotation approach (J. M. Walejko et al. 2018a) was based on metabolomic sample extraction, 2D NMR experiments (heteronuclear single quantum coherence spectroscopy, HSQC; Total Correlation Spectroscopy, TOCSY) and COLMARm (Bingol et al. 2016), which was used in the original CIVM-NMR publication (Judge et al. 2019). Such a process is standard and powerful but has drawbacks when the goal is to annotate and quantify ridges in time-series NMR. First, besides the additional work of extraction and 2D NMR, the sampling time point(s) for extraction depends on the dynamics of the feature of interest, and multiple time points

might be needed. The compound of interest needs to exist for COLMARm annotation, and metabolites that decrease over time in a CIVM-NMR run may not be detectable in a sample extracted at the end of the run. Second, sample extraction will differentially enhance or diminish some compounds and introduce chemical shift changes depending on the extraction solvent. Changes in relative peak intensity and chemical shift complicate the next step, and mapping from COLMARm to CIVM-NMR spectra is not perfect. CN clustering of CIVM-NMR ridges covers compounds through the whole time range, needs no further experiments or mapping, and has no extraction biases. CN clusters provide a practical way to augment annotation and can be used with 2D NMR at specific time points. We also defined a new confidence scale for annotations for CIVM-NMR datasets (Details in Method) (Judge et al. 2019; J. M. Walejko et al. 2018a).

Among the 30 clusters, we presented seven examples here (Fig. 4.4 and Supplementary Fig. 4.4). Glucose, ethanol, G1P and arginine were annotated through both mapping from COLMARm (Fig. 4.3 in CIVM-NMR publication) (Judge et al. 2019) and CN clusters. Glucose, ethanol, and arginine were at level 4 (See Methods), and G1P was raised to level 5 with spike-in validation (Supplementary Fig. 4.9). CN clusters provided more peaks for quantification (blue nodes in Fig. 4.4A). Those peaks were not assigned in our previous publication because of overlaps with other compounds in the extracted sample. However, in CIVM-NMR data, those peaks were highly correlated and clustered with high confident peaks (red nodes). A detailed example of such an assignment is presented for arginine with experimental NMR spectra and the database

standard (Fig. 4.4B). 2D NMR of the extracted sample indicates lysine and leaves arginine only quantifiable for a subset of peaks at around 1.65 ppm (red points in Fig. 4.4B) (Judge et al. 2019). However, the CN cluster of other peaks (blue points) with confidently annotated peaks (red points) indicates the low concentration of lysine that allows for reliable quantification of arginine (Fig. 4.4B and Supplementary Fig. 4.4).

In addition to expanding the peak assignment for annotated compounds, three CN clusters with blue points but no red points (Fig. 4.4A and Supplementary Fig. 4.4) revealed new compound annotations (e.g., serine, ethanolamine and ornithine) (Judge et al. 2019). These compounds appeared in 2D experiments on methanol-extracted samples but could not be mapped and quantified in the CIVM-NMR study because of overlap, noise, and poor peak shapes (Judge et al. 2019). Particularly, relative compound concentration and spectral overlap pattern in an extracted sample were different from those in CIVM-NMR spectra, and mapping the two is not trivial. Nonetheless, CN revealed highly correlated clusters of CIVM-NMR peaks for those compounds (level 4).

Derivatives of metabolic profiles provide additional insights into metabolic dynamics

Besides chemical associations, biological associations can also be uncovered from the high-dimensional dataset. The first derivative of the ridge intensities with respect to time yields rates of change in metabolite concentrations and provides further information on metabolic states (Fig. 4.5A).

Rates (Fig. 4.5A, the second row) can be visualized and co-analyzed with corresponding intensity curves (Fig. 4.5A, the first row). Derivative-based analysis in the following two sections will focus on the glucose feeding experiments. Glucose was consumed at a relatively constant rate and fell below the limit of detection at around six hours (Judge et al. 2019). Uridine accumulated throughout the experiment but had two distinct intervals between which the rate changed at about six hours when glucose was exhausted. In the first six hours, the rate of uridine accumulation increased; after six hours, that rate was constant. The intensity and first derivative of G1P were complex: it was initially produced (positive rate); the intensity reached a maximum just after two hours when it started to decrease; consumption reached a maximum rate at around four hours. Choline accumulated rapidly in the first hour, and its rate remained relatively low afterward. However, the first derivative estimation of choline is relatively noisy, and a higher smoothness penalty (See Methods) might improve the estimation after two hours. We expanded the analysis of associations between ridge intensities and rates through model searching and network construction in the next section.

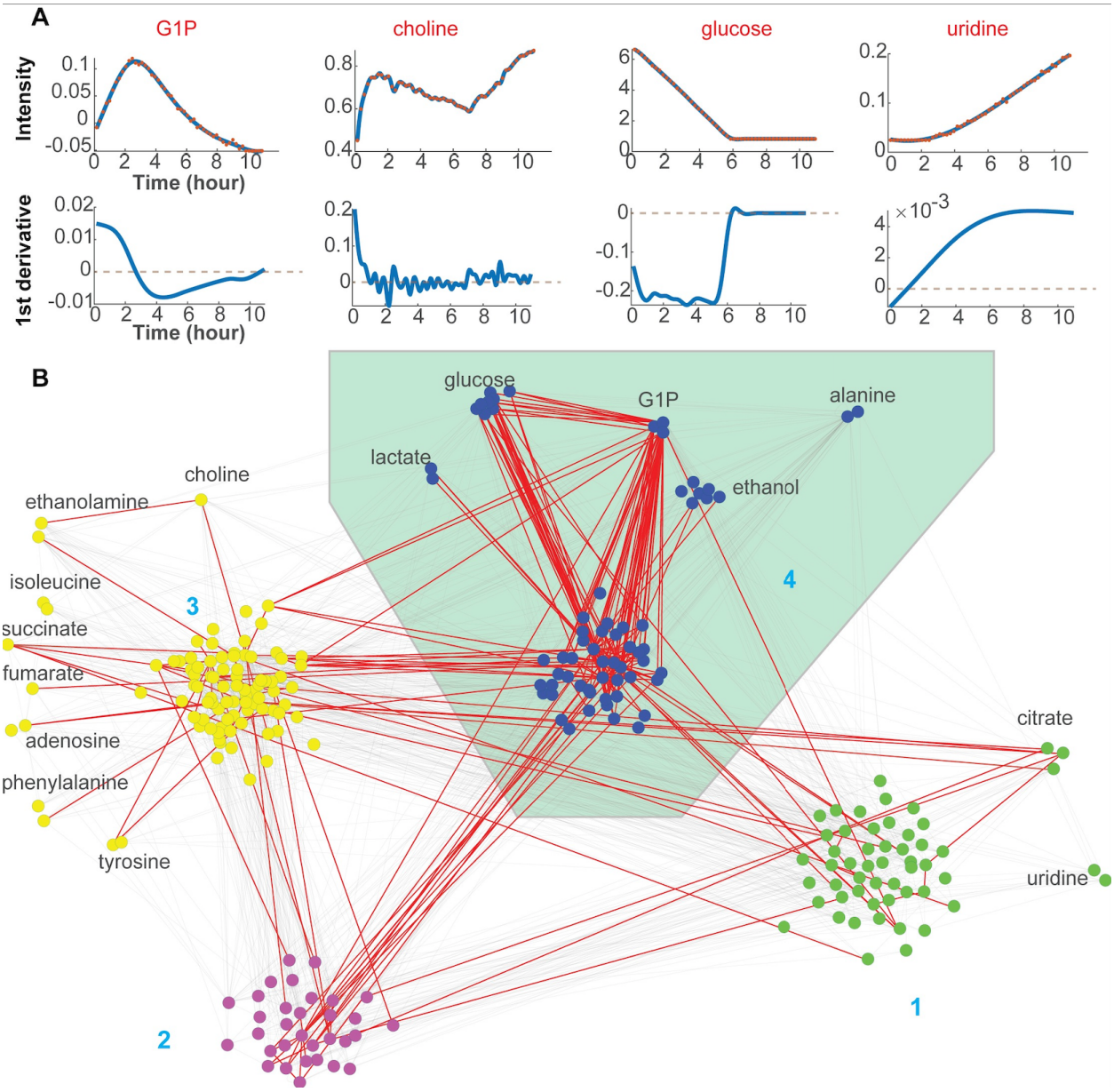


Figure 4.5: Analyzing metabolic dynamics from derivatives in the glucose feeding experiments. A: A few examples corresponding to points in Fig. 4.2B are visualized. Intensities (row one) and corresponding 1st derivatives (row two) are presented. The X (Y) axis represents time (value). The blue curve represents a smoothed curve from FDA, and the red points represent raw measurements. Dotted grey lines indicate zero derivatives. B: An empirical metabolic network

and clusters were inferred from time series. Time-series features were summarized into four clusters through CausalKinetiX and community clustering (More details in Methods). Each node represents one time-series feature, and each edge represents one inferred link from CausalKinetiX. Red edges are supported through bootstrapping at the cutoff of 40%. Clusters are distinguished by color. Cluster 4 is overrepresented by metabolites in central energy metabolism and highlighted with a green background. Some nodes are attached with annotations and were manually moved outside for better visualization. Bootstrapping details can be found in Supplementary Fig. 4.5 Fig and Methods.

Define metabolic associations through network clusters

In theory, one could individually analyze all 300 NMR curves and search for biological relationships. Not only would that be inefficient, but it would miss critical patterns that are inherent to the interconnected metabolic pathways, in which the rate of change of one compound often depends on concentrations of several other compounds. Through integrating different perturbations and network construction, these *in vivo* metabolic connections and regulation can be discovered. We used CausalKinetiX and community clustering (Newman and Girvan 2004; Pfister et al. 2019a) to search for the model. We built a stable and predictive network from CIVM-NMR datasets collected under aerobic and anaerobic conditions and with glucose as the carbon source (Fig 5B). We assumed that the network topology is stable and that different conditions will lead

to changes in the edge strengths through reaction or regulation (Pfister et al. 2019b; Pfister et al. 2019a).

The estimated functional network (FN) edges indicate associations between rates and concentrations, which can be biochemical reactions or regulation. The performance of edge estimation depends on the number of perturbation conditions (Supplementary File 4.1; Supplementary Figs 4.6-4.8), and the estimation can be noisy with fewer conditions. Hence, we relied on the clustering of undirected networks to improve signal detection and to find clusters representing different biochemical processes. Four clusters were recovered through community clustering, among which Cluster 4 is well-supported by bootstrapping (Fig. 4.5B and Supplementary Fig. 4.5; More details in Methods) and includes several critical metabolites in central energy metabolism in *N. crassa* (Fig. 4.6A). Glucose, G1P and ethanol were consistently grouped into one cluster in bootstrapping (Supplementary Fig. 4.5). Metabolic features with inverse trends, primary carbon source (e.g., glucose) and product (e.g., ethanol, lactate and alanine), were clustered together (Figs 4.2B, 4.5B and Supplementary Fig. 4.5), delineating the primary flux under the experimental conditions. Cluster 3 is relevant to amino acid and phospholipid metabolism, and more time-series data with perturbations, such as mutations, are needed for cleaner separation. In the FN, most peaks annotated to the same compound are consistently in the same cluster (Fig. 4.5B). Confidence in our estimations of network edges and clusters was assessed using a benchmark dataset (Supplementary File 4.1; Supplementary Figs 4.6-4.8).

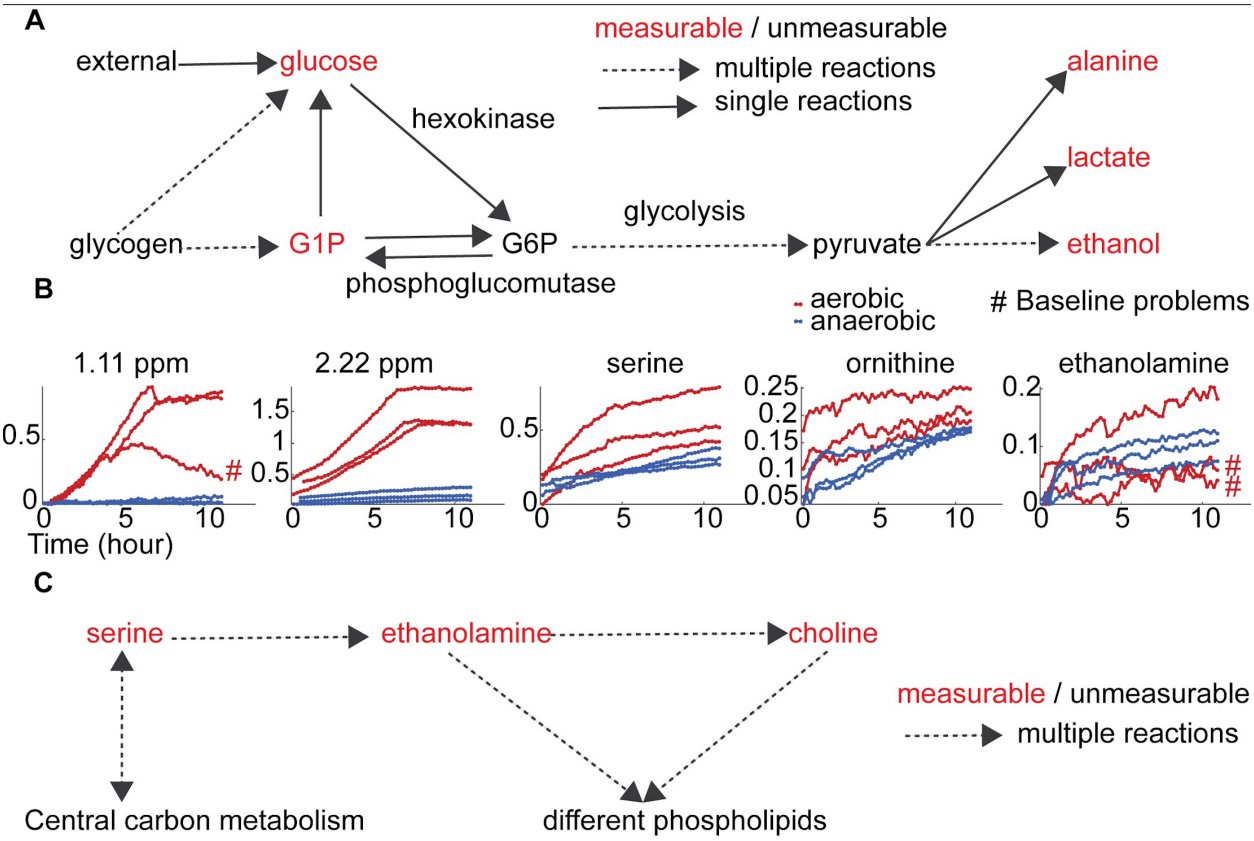


Figure 4.6: Biochemical connections for central energy metabolism and membrane synthesis in the glucose feeding experiments. A: Hypothesized pathway through the central energy metabolism: glycolysis, glycogen degradation and fermentation. Measured (unmeasured) compounds are visualized in red (black). Steps with multiple (single) reactions are represented by dotted (solid) lines. B: Time-series quantification for selected features: high degree unannotated features in Fig. 4.5B and compounds related to phospholipid synthesis and amino acid metabolism. The X (Y) axis represents time (means of scaled ridge intensities). Red (blue) represents aerobic (anaerobic) conditions.

The '#' symbol indicates that the quantified peaks are highly affected by regional baseline changes. C: The hypothesized pathway for phospholipids synthesis.

Discussion

Our exploratory workflow to uncover metabolic associations consists of dimensionality reduction and network analysis by both correlation and CausalKinetiX. FPCA enables metabolic profile comparison, the CN expands incomplete annotation, and the FN discovers metabolic connections, which is important in most practical cases, where pathway knowledge is incomplete. Our workflow prioritizes metabolic clusters for further isotopic labeling and perturbation experiments, as part of the discovery cycle of system biology (Ideker et al. 2001).

Uncover *in vivo* biochemical functions in central energy metabolism

FN Cluster 4 was found to be associated with central energy metabolism, containing glucose, lactate, G1P, ethanol and alanine (Fig. 4.5B and Supplementary Fig. 4.5). In the glucose feeding experiments, glucose was consumed, and fermentation products accumulated. This is largely represented by PC1 in FPCA (Fig 4.2B and Supplementary Fig. 4.2) and agrees with our previous publication (Judge et al. 2019). Additionally, several biochemical connections were found as follows.

G1P can indicate the functional level of glycolysis and central energy metabolism. We saw dense connections between G1P and most other NMR

features in the reconstructed FN (Fig. 4.5B; Supplementary Table 4.1). G1P can be produced by glycogen degradation and converted to glucose or glucose 6-phosphate (G6P) (Fig. 4.6A). G6P is an intermediate metabolite in glycolysis, which is central in many metabolic branches. Except for glucose, glycolysis metabolites (e.g., G6P and pyruvate) were below our detection limit, preventing direct observation of glycolysis. The reaction between G1P and G6P is catalyzed by phosphoglucomutase and is reversible (Fig. 4.6A) (Dreyfuss et al. 2013; Judge et al. 2019; Park et al. 2019), so G1P can be an alternative indicator of glycolysis fluxes, explaining its high associations with many NMR features in the FN. Additionally, the two transition points (indicated by changes in rate) of G1P at about two and four hours did not coincide with glucose exhaustion at six hours (Fig. 4.5A), indicating alternative regulation mechanisms. Specifically, the initial accumulation of G1P can be caused by rate limitation in glycolysis during a transition from starving to high rates of glucose consumption (Judge et al. 2019; Park et al. 2019). Based on the FN analysis, our working hypothesis is that the metabolic associations with G1P depend on phosphoglucomutase (gene id: *NCU10058*) and its reversible connection to glycolysis. In the future, we will test the effects of its mutants (strain id: FGSC #18976) on central energy metabolism.

We also found that alanine might serve as an alternative carbon source after glucose exhaustion. Alanine is in the increase-and-plateau quadrant in the FDA scores plot but has a slight decreasing pattern after glucose exhaustion at about six hours (Fig. 4.2B). In FN, it is in Cluster 4 and has dense links with glucose (Fig. 4.5B; Supplementary Table 4.1). Alanine is directly connected to

pyruvate through a transamination reaction and can be used for carbon and nitrogen storage when they are abundant (Judge et al. 2019; Kanamori et al. 1982). The observation that alanine consumption started when glucose was exhausted, suggests the hypothesis that alanine was an alternative carbon source in our experiments.

Diverse patterns of ethanol under various conditions (Fig. 4.3) were uncovered through metabolic profile comparison by FPCA. We were able to compare different carbon sources (glucose and pyruvate), different starting densities (6 or 10 mg per 63 mL) and isotopic labeling. In the ^{13}C-pyruvate data, we were able to monitor two distinct pools of ethanol, one that was isotopically labeled with ^{13}C and the other that was at natural abundance carbon. These have very different patterns, and the distribution on the FPCA scores plot allows us to understand their relationship. For example, the unlabeled ethanol data from the ^{13}C-pyruvate experiments was initially consumed and then released after 2-6 hours, depending on density. But the ^{13}C-labeled ethanol in the same ^{13}C-pyruvate experiments increased and then plateaued at about 4 hours. Similar comparisons could be made with studies of other genotypes, carbon sources, or isotopic labeling.

Interesting unannotated features were selected for further experiments. Many high-degree nodes in FN Cluster 4 (Fig. 4.5B) were not annotated (Judge et al. 2019). Many of them belong to glucose or are considerably affected by glucose concentration because of overlap (Fig. 4.4A). Among other unannotated nodes, features at 2.22 ppm and 1.11 ppm have consistent FN connections and

different dynamics between aerobic and anaerobic conditions (Fig 4.6B and 4.5B; Supplementary Table 4.1). Both nodes are functionally connected to glucose and G1P, and 2.22 ppm is also connected to fermentation products, including ethanol and lactate. These connections are supported by bootstrapping. Both features increased and plateaued in aerobic conditions and have a much lower level in anaerobic conditions. The peak at 2.22 ppm also exists in both unlabeled and ^{13}C-labeled forms in the pyruvate dataset. We suspect that the 2.22 ppm feature corresponds to an N-acetyl functional group and are working on its full annotation.

Fluxes towards membrane synthesis and amino acid metabolism

Phospholipid synthesis was also active in our experiments. Three compounds related to this process were annotated: serine, ethanolamine and choline (Figs 4.6B, 4.4A and 4.2B). The former two were newly annotated here by searching CN clusters using the COLMAR 1D databases (Robinette et al. 2008). Serine is associated with glucose in the FN (Fig. 4.5B; Supplementary Table 4.1). Its synthesis starts from 3-phospho-D-glycerate, a glycolysis intermediate, and its degradation produces pyruvate. Serine can flow to phospholipid synthesis, ethanolamine (and phosphatidylethanolamine) and then to choline (and phosphatidylcholine) (Fig. 4.6C) (Dreyfuss et al. 2013; Horowitz et al. 1945; Radford 2004). Genes *chol-5*, *chol-8*, *chol-11* and *chol-12* were previously found related to the first step, and *chol-1* was for the second step (Galagan et al. 2003). An association between ethanolamine and choline was also found in the FN. Both choline and ethanolamine are important precursors for

phospholipid and membrane synthesis. Different dynamics of the two precursors might indicate different fluxes to corresponding membrane phospholipids. Changes in phospholipid ratio during development have been observed previously (Beck and Greenawalt 1977), and the formation of vacuoles in stressed (e.g., starving) *N. crassa* cells is well-known (C Slayman and Potapova 2006).

Ornithine was also newly annotated and quantified in CIVM-NMR spectra through searching CN clusters in the COLMAR database (Robinette et al. 2008). It was connected to valine and tyrosine in the FN, though its quantification is relatively noisy. Connection to glutamate was expected (Radford 2004) but not found, and this might result from glutamate's intense connections to many other amino acids and dependencies on them. The FN connection to valine and tyrosine might indicate balance and regulation among different amino acids.

Technical improvements on network-based dynamic analysis

Even though our prior analysis did not require extensive pathway mapping, including known pathway knowledge can yield further biological insights. Specifically, the empirically estimated FN clusters can be compared or merged with conventional metabolic pathways for interpretation. Starting from compounds of interest, possible paths can also be searched in conventional pathways and compared with those in the empirical FN (S. M. Kim et al. 2017, 2020). New developments in graph neural networks can also be applied to merge information from the FN networks and pathways (Nelson et al. 2019).

Including more perturbation experiments can considerably improve the FN method (S1 File). Edges, directions and clusters can be more confidently estimated when more diverse perturbations are available. Although CIVM-NMR and RTExtract simplify measuring time-series metabolic features (Judge et al. 2019; Y. Wu et al. 2020), collecting many perturbations is still expensive and time-consuming. Metabolic dynamics highly depend on both the media condition and prior culturing process (Judge et al. 2019), so a refined experimental procedure is needed to maintain consistency between perturbation experiments. Effective exploration of interesting perturbations is also crucial, and the results in this paper provide valuable next targets.

CN clustering provides a simple initial annotation of the CIVM-NMR dataset, complementing the traditional approach involving extraction and 2D experiments. Currently, we have searched CN clusters in the COLMAR database to expand annotation (Robinette et al. 2008). However, many features remained unannotated because of the considerable overlap in NMR spectra and limited coverage in the database. We can improve this by combining CN connections with a probabilistic graph model. Biochemical associations in FNs and metabolic pathways also help reduce the searching space of possible compounds.

Integration of our different analyses (FPCA, CN and FN) and perspectives can facilitate the utility of our tools by researchers with broad backgrounds. We are currently working on an efficient web interface to integrate these analyses (Franz et al. 2016). In the new interface, with a few clicks, biologists will be able to select one cluster in the FN or the CN and highlight nodes in the other

network, visualize network clusters in the FPCA scores plot or vice versa, and search for interactions between a group of network nodes in metabolic pathways (S. M. Kim et al. 2017, 2020; Paull et al. 2013). Based on biochemical pathways, information from other omics experiments, such as transcriptomics (DeRisi et al. 1997) and proteomics (Ideker et al. 2001) can also be co-analyzed with metabolomics (A. M. Al-Omari et al. 2022).

We created a framework to summarize the dominant trends of both annotated and unannotated features and enabled comparisons between perturbation experiments. We then leveraged the unique properties of these data to facilitate and expand annotation. Finally, we integrated this feature set into a network of functional dependencies stable across conditions to yield biological insights. This work helps bridge the gap between untargeted time-series metabolomic data and biochemical integration.

Methods

Neurospora culturing, preparation, and data collection

Briefly, Neurospora bd 1858 mycelia were grown for around 30 hours in 3% Glucose Vogel's minimal media in 50 mL shaking flask cultures under constant light (Judge et al. 2019). An hour before the start of the CIVM-NMR experiment, mycelia were poured with media into a 50 mL conical tube and transported to the NMR facility. Working quickly, a small piece of mycelium was pulled from the mycelial mass and washed four times in 1.5 mL conical tubes containing 1 mL minimal media without carbon sources. The mycelium was pat-dried, weighed and adjusted to around 11 mg, then resuspended in 500 µL NMR

media (1.5% glucose). The total volume of mycelia and media was adjusted to around 63 µL, and contents were transferred to a 63 µL (4 mm) zirconia HR-MAS rotor, which was capped with a breathable cap. All experiments were collected on a 600 MHz Bruker NEO equipped with a 4-mm CMP-MAS probe running TopSpin (v4.0.1; Bruker). Details in sample growth, preparation, data collection, and preprocessing are described in (Judge et al. 2019). For 13C-pyruvate experiments, the same procedure was carried out with the following adjustments:

(1). No citrate was used in the Vogel's media.

(2). Mycelia were grown in 1.5% Sucrose instead of 3% Glucose.

(3). The total mass of the rotor contents was carefully adjusted to 50 mg ≈ 50 µL (mycelia and media) by removing/adding media.

(4). Immediately before recording, pyruvate addition was carried out by swapping 10 µL rotor liquid for 10 µL concentrated 13C-pyruvate (220 mM; Aldrich 490717-500MG, Lot # FMBBC3492) in the NMR media for a final concentration of 37 mM in the rotor. Thus, most pyruvate metabolism was observed as early as possible. An unavoidable delay due to rotor spin-up (~3-10 min) still occurred.

(5). The spinning speed was reduced from 6000 Hz in the glucose experiments to 3500 Hz for pyruvate samples (Judge et al. 2019).

(6). Data were recorded and averaged every eight scans (~35 s) for noesypr1d and every 32 scans (~30 s) for hsqcetgsisp2.2 experiments. These were collected in an interleaved manner as described previously and smoothed by moving average (Judge et al. 2019).

NMR feature extraction

We followed the original method in RTExtract (Y. Wu et al. 2020) to extract NMR time-series features. The experimental data were collected with 52 time points, under aerobic and anaerobic conditions, and each condition contains three experiments. Comprehensive regions of interest (ROI) were selected and tracked in each dataset. Some NMR features existed and were extracted in only part of the time range. The empty values in chemical shift and intensity were filled with those of neighboring existing time points. This is used instead of intensity zero filling because there is often a local baseline shift in the spectra.

NMR features were matched across different experiments. Annotated features were matched directly, and unannotated features were matched based on chemical shift differences. In the latter circumstance, distances were calculated between features in different samples and the closest pairs (maximum threshold 0.01 ppm) were matched. Each NMR feature in one sample is represented by its mean chemical shift through time.

For ^{13}C-labeling experiments, we interleaved standard ^{1}H 1D (noesypr1d) with ^{13}C-HSQC 1D (hsqcetgpsisp2.2). The ^{1}H 1D data provides information on all metabolites, and the ^{13}C-labeled species have addition peaks from the large ^{1}H to ^{13}C coupling constants. The ^{13}C-HSQC 1D data selects for only ^{1}H atoms that are directly bound to ^{13}C, which significantly simplifies the data but also eliminates the species that are not isotopically labeled. Our RTExtract algorithm works equally well for either the ^{1}H 1D or ^{13}C-HSQC 1D datasets. Direct comparison of them needs normalization based on the large ^{13}C-labeled pyruvate

signals. Natural abundance pyruvate only has a single peak from the methyl group at about 2.37 ppm in the ^{1}H 1D spectrum. This single peak is split into 4 peaks in ^{13}C-pyruvate because of the large 1 bond and smaller 2 or 3 bond ^{13}C-^{1}H J couplings. We used ^{13}C decoupling during the ^{13}C-HSQC 1D acquisition so that the dataset only has a single peak at about 2.37 ppm. To calculate the normalization factor, we set the intensity of the single peak in the ^{13}C-HSQC 1D spectrum equal to the sum of the intensities of the 4 pyruvate peaks in the ^{1}H 1D spectrum at each time point. The normalization factor was calculated for pyruvate peaks higher than 10% quantile and then averaged. This same normalization factor was then applied to all other compounds in the datasets. Afterward, all ridges were scaled by the maximum of the unlabeled compound and then scaled within the experiment and labeling. This last step reduced variance from multiplets and simplified visualization.

Smoothing and dimensionality reduction

Time-series features were then analyzed with FDA, which is a collection of methods for analyzing curves or functions, including time series. The analysis often starts from smoothing and then can go into multiple directions, including derivative analysis, dimensionality reduction and regression (G. Montana et al. 2011a; Ramsay and Silverman 2005; Ramsay et al. 2009). FPCA and derivative-based regression (Pfister et al. 2019a) were used in our workflow.

Time-series curves were first fitted with B-splines with smoothness penalties (Equation 4.(4.1 as the objective function) (Ramsay and Silverman 2005; Ramsay et al. 2009). The first part of *F* is the sum of squared distance

between the original data value y_j and the smoothed function $x(t_j)$. The second part is the smoothness penalty, where D is the derivative operator. The 1st derivatives represent rates of metabolic activities and were forced to be smoothed, making the smoothness penalty based on the 3rd derivatives (Equation 4.(4.1) (Ramsay and Silverman 2005). The penalty parameter λ was searched in a log scale on 33 values from 10^{-4} to 10^{4} and chosen by minimizing generalized cross validation (Craven and Wahba 1979).

$$F = \sum_j [y_j - x(t_j)]^2 + \lambda \int [D^3 x(t)]^2 dt \qquad (4.1)$$

Dimensionality reduction was done with FPCA (Ramsay and Silverman 2005; Ramsay et al. 2009) based on smoothed curves. Just as PCA provides lower-dimensional representations of the original dataset, FPCA provides a similar smoothed representation of time series. Like a PC vector in PCA, each found FPCA eigenfunction (harmonic, $\xi(t)$) is orthonormal ($\int \xi_i(t)\xi_j(t)dt = 0$ where $i \neq j$ and $\int \xi_i^2(t)dt = 1$) and explained the maximal variances iteratively. Each time-series feature is mean-centered, scaled by standard deviation and smoothed before FPCA. The origin is represented by the mean function (curve) with the mean spline coefficients from all curves. Algorithm details of FPCA can be found here (Ramsay and Silverman 2005; Ramsay et al. 2009).

Correlation network estimation and visualization

A correlation network was constructed from experimental NMR measurements to help compound annotation. Spearman correlations were calculated between time-series features concatenated from different experiments, and the largest 10% were included in the network. The network was

clustered by the Markov Clustering Algorithm (MCL) based on correlation value and granularity parameter 5 (Enright et al. 2002). Chemical shifts of each network cluster were searched through the GISSMO database (Dashti et al. 2017) and COLMAR 1D Query (Robinette et al. 2008) for possible matches. The GISSMO searching process is automatic, with Cytoscape controlled through R by RCy3 (Gustavsen et al. 2019; Ono et al. 2015), GISSMO search through API (http://gissmo.nmrfam.wisc.edu/peak_search) and spectral visualization in MATLAB.

New annotation confidence level for CIVM-NMR experiments

We defined new confidence levels to accommodate the CIVM-NMR experiment and CN clustering (Judge et al. 2019; J. M. Walejko et al. 2018a). This new definition includes the requirement of mapping back to CIVM-NMR spectra and the annotation power of CN clusters. The levels are defined from 1 to 5 with increasing confidence. (1) There is a similarity of 1D ^{1}H spectra (between a standard reference and CIVM-NMR spectra at any timepoint). (2) There is HSQC match using COLMAR (Bingol et al. 2015) from an extracted sample and the compound can be mapped back to CIVM-NMR spectra. (3) Compound annotation can be found for a CN cluster in CIVM-NMR spectra. (4) There is double matching from two sources as well as matching in CIVM-NMR spectra. It can be HSQC match and TOCSY/HSQC-TOCSY validation from extracted data using COLMARm (Bingol et al. 2016). It can also be HSQC match and CIVM-NMR CN cluster validation. (5) The compound is spiked into the CIVM-NMR sample and validated.

Functional network estimation and clustering

In a metabolic network, nodes depend on each other (e.g., reactions or regulation), and densely connected subnetworks represent specific functions. We estimated associations (edges) between metabolic features (nodes) by CausalKinetiX (Pfister et al. 2019a) and searched for functional groups by community clustering (Newman and Girvan 2004). CausalKinetiX can estimate predictive and stable edges in a metabolic network in a time-efficient manner (Pfister et al. 2019a). Unlike conventional methods for extracting time-series dependencies (Granger 1969), CausalKinetiX learns directly from non-stationary time series in which relationships between rates and concentrations hold up under different perturbations (Pfister et al. 2019b). Features were extracted by RTExtract as in the previous section (Y. Wu et al. 2020).

CausalKinetiX (Pfister et al. 2019a) estimates edges, the dependencies between the changing rate of one compound and the concentration of some compound(s) (Equation 4.(4.2, adapted from Equation 1 in (Pfister et al. 2019a)). $X^i(t)$ is the time-series features in experiment *i*. *T* is the target variable feature, and *S* is a subset of features. The dependency *f* holds under all experiments *i* (Equation 4.(4.2). Predicative and stable models (Pfister et al. 2019a) were searched for derivative estimation of each target feature (X_T). Covariate features (X_S) were ranked by importance, and connections with p-values less than 0.05 were selected as estimated edges. Before fitting, target variables were smoothed with a penalty on the 3rd derivatives (Pfister et al. 2019a). To reduce computational complexity, the expected number of terms in each model was set

to at most two, and pre-screening was used (Pfister et al. 2019a). Interaction terms were also included. The model was fitted in the derivative mode. Other parameters were default values (Pfister et al. 2019a). Constructed networks were undirected and clustered by community-based clustering based on topology (Newman and Girvan 2004) using clusterMaker in Cytoscape (Morris et al. 2011; Shannon et al. 2003).

$$\frac{dX_T^i(t)}{dt} = f\left(X_S^i(t)\right) \tag{4.2}$$

Clustering stability was evaluated through bootstrapping. The dataset was resampled 100 times, and, in each iteration, the complete time series for each feature were sampled with replacement among the three experiments of fixed conditions. The network was then constructed and clustered as above based on each new bootstrapped dataset. Frequencies that two features share the same cluster were calculated as bootstrapping support.

Benchmarking dataset simulation

We simulated random networks and corresponding dynamics as a benchmark. Networks and clusters were estimated based on the simulated time series and compared with the ground truth. We generated different random networks with 100 nodes and three clusters (sizes: 40, 20 and 20). Edges were randomly generated for node pairs with higher probability within clusters (0.15) and lower probability among all nodes (0.015).

Temporal dynamics were simulated by ODEs, which were generated based on the random networks with nodes (edges) representing metabolites (reactions or regulation). The reaction was formulated by a sum of regulated

mass actions (Equation set 4.(4.3). *X(t)* are time-series features, where *T* indicates the target variable features, and *r* indicates the regulatory features. Multiple reactants ($1 \leq N_h \leq 2$) are involved in one reaction, and each feature is connected by multiple reactions (*N*). Kinetic parameter *k* is regulated by X_r with the exponent $\alpha \in \{-1, 1\}$. *s* is the stoichiometric factor. Reaction direction was randomly generated, including reversible reactions. Reactions with multiple reactants ($N_h > 1$) were simulated by randomly combining single reactions. The ODEs were simulated under different initial conditions (*X(0)*) as different experimental perturbations. The simulation time grids were set from 0 to 5 with the step size 0.2. Kinetic parameters and initial conditions were uniformly sampled (*U*(0,1)).

$$\frac{dX_T(t)}{dt} = \sum_{j}^{N} [s_j \, {k_j}^* \prod_{h}^{N_h} X_h(t)] \tag{4.3}$$

$$k^* = k \prod_{r} X_r^{\alpha_r}(t)$$

Some procedures were introduced to ensure similarities between simulation and experimental measurement. Duplicated signals belonging to the same compounds were simulated to resemble multiple peaks in NMR. For each times series in the ODE simulation, a random number of signals ($N_S \in \{1,2,3,4,5\}$) were added with each multiplication factor uniformly sampled (*U*(0.3,3)). For each fixed ODE set and initial condition, three replicates were simulated with Gaussian noise (Equation set 4.(4.4) (Pfister et al. 2019a). *Y(t)* is the observation, *X(t)* is the simulation with no noise, and $\sigma(X)$ is the standard deviation function. The partial observation was also simulated to resemble the

incomplete coverage of experimental measurements. 50% of the features were randomly selected to be observable.

$$Y(t) = X(t) + \sigma^* \tag{4.4}$$

$$\sigma^* \sim N(0, a)$$

$$a = 0.02 \cdot \sigma(X) + 10^{-7}$$

Performance evaluation on the benchmark dataset

Edge estimation was evaluated by recall and precision (Equations 4.(4.5 and 4.(4.6). N_{tp} is the number of true positives, N_{pp} is the number of predicted positives, and N_{cp} is the number of observable conditional positives. For compound features with multiple signals, the instances were counted for each compound.

$$Precision = \frac{N_{tp}}{N_{pp}} \tag{4.5}$$

$$Recall = \frac{N_{tp}}{N_{cp}} \tag{4.6}$$

The capability of recovering underlying clusters was evaluated by the match ratio (Equation 4.(4.7). Among all estimated clusters, the one with the most overlapped nodes with the real cluster was selected. $N_{overlap}$ is the number of overlapped nodes, and $N_{cluster}$ is the size of the estimated cluster. For compound features with multiple signals, the instances were counted in terms of each signal.

$$R = \frac{N_{overlap}}{N_{cluster}} \tag{4.7}$$

Code and availability

The program was implemented in MATLAB, R and Python. Codes are freely available through GitHub (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA/tree/master/metabolomics_toolbox/code/net_ana). The experimental data can be found in Metabolomics Workbench (https://www.metabolomicsworkbench.org PR000738). The local programs were implemented under R 3.5.1, MATLAB_R2018b and Cytoscape 3.8.0 on macOS 10.15.7. The extensive simulation was implemented on HPC: R/3.5.0-foss-2019b on Sapelo2 at Georgia Advance Computing Resource Center (GACRC).

Acknowledgements

We thank Abby Moore, Chen Hsieh and Robert Powers for useful discussions. We appreciate computational support from GACRC, especially Shan-Ho Tsai and Zhuofei Hou. Jeremy Zucker and Bill Cannon at Pacific Northwest National Laboratory provided a metabolic database, Hesam Dashti and Hamid Eghbalnia provided assistance with the GISSMO NMR library, and Niklas Pfister provided technical details on CausalKinetiX.

CHAPTER 5

QUANTIFYING METABOLIC SYSTEMS THROUGH AUTOMATIC NMR SPECTRAL DECOMPOSITION [4]

[4]: Yue Wu et al. To be submitted to Nature Methods

Abstract

New developments in untargeted nuclear magnetic resonance (NMR) metabolomics enable the profiling of hundreds to thousands of biological samples and provide opportunities to quantify metabolism in cohort studies. Metabolomics data closely represent the dynamic phenotypes in the sample and so are highly variant. We lack a consistent and automatic processing workflow, particularly in feature quantifications, as a result of NMR spectral overlap. We built spectral automatic NMR decomposition (SAND) to decompose and quantify the NMR spectra of different sample types. The quantification capability of SAND was validated with both simulated and experimental benchmark datasets. Based on the decomposed peaks, annotations were also inferred automatically through correlation networks and clustering. SAND was tested on complex experimental datasets collected from *C. elegans*. SAND decomposed and uncovered annotations successfully even though the *C. elegans* spectra were filled with overlap, broad peaks, and pH-induced peak movement. To further enable automation in NMR metabolomics, SAND was combined with Network of Advanced NMR (NAN)-and NMRbox.

Introduction

Metabolomics techniques have advanced recently, producing untargeted profiling on different biological systems (Bifarin et al. 2021; Choe et al. 2012; Edison et al. 2021; Gouveia et al. 2021; Judge et al. 2019; Link et al. 2015; Maughon et al. 2022; Shaver et al. 2021). Nuclear magnetic resonance (NMR) spectroscopy has enabled noninvasive or even *in vivo* metabolic measurement of

samples (Bastawrous et al. 2018b; Judge et al. 2019). Such untargeted profiling provides a detailed description of the metabolic landscape, metabolic dynamics (Battogtokh et al. 2002), gene-environment interactions (Fuhrer et al. 2017), and possible biomarkers (Bifarin et al. 2021).

Peak Overlap has been a serious problem in automatic NMR spectral analysis. It not only affects peak quantification but also obstructs efficient annotation. For regions with low signal to noise ratios (SNR), quantifying small peaks is highly affected by baseline levels or by overlapping with broad peaks.

Spectral overlap can be relieved but not solved by current experimental approaches. For example, conducting high-performance liquid chromatography (HPLC) on a sample prior to NMR analysis can reduce overlap (Whiley et al. 2019), but this process greatly increases labor cost, complicates quantification, and still leaves many spectra unresolved. Methods like Diffusion Ordered Spectroscopy (DOSY), J-RESolved Spectroscopy (JRES), or other 2D NMR can reduce the overlap effect by spreading the peaks in another dimension (Aue et al. 1976; C. S. Johnson 1999). However, 1D NMR is the most popular approach for large-scale metabolic profiling and produces considerable number of spectra in cohort study. An automatic pipeline is needed for large-scale spectral processing and quantification. The Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence reduces broad peaks but affects quantifications of other peaks and leaves other overlap unresolved (Carr and Purcell 1954; Meiboom and Gill 1958).

Computational spectral decomposition can enable automatic NMR profiling, and existing 1D NMR datasets provide considerable resources for

validation (Haug et al. 2020). Different methods have been proposed, for example spectral quantification by peak intensity or integral, although both are highly affected by overlap. Combined with peak picking and binning, peak intensity and integration can automatically quantify peaks rapidly (Sousa et al. 2013). Reference library-based quantification can decompose routine spectra or compounds but is restricted to available libraries. Bayesian automated metabolite analyzer for NMR (BATMAN) relies on precise template libraries, and its targeted metabolite quantification leaves unknown features unresolved (Astle et al. 2012; Hao et al. 2012; Hao et al. 2014b; Zheng et al. 2011). Bayesil has restriction on sample types and NMR procedures and are limited to a defined list of compounds (Ravanbakhsh et al. 2015). Chenomx provides a commercial solution to library-based fitting, but it needs time-consuming manual interactions (Hao et al. 2014b) and have operator dependent divergences in quantification (Tredwell et al. 2011).

There are also more suitable methods for untargeted metabolomics analysis. These methods often model the NMR data directly in frequency or time domains and estimate peak information through different optimization approaches (de Beer and van Ormondt 1992; Sekihara and Ohyama 1990). Particularly, time-domain based signal modeling is advantageous and direct and current computational resources can support the intensive optimization (Bretthorst 1990b, 1990c, 1990d). In the time domain, nonuniform sampled data can be directly analyzed and baseline problem are often easier to resolve. Based on a Bayesian framework, the complete reduction to amplitude frequency table

(CRAFT) method has been used to decompose 1D and 2D NMR spectra by fitting time-domain data, but manual tuning is still needed for good performance (Bretthorst 1990b, 1990c, 1990d; Krishnamurthy 2013). Similar ideas have been tested previously by researchers including Denis and Julian, but extensive validation and a consistent workflow are not available (Bretthorst 1990a; Matviychuk et al. 2017; D. V. Rubtsov and Griffin 2007; Denis V. Rubtsov et al. 2010). Singular value decomposition (SVD) based methods can also decompose the FID, but the tuning parameter, reduced order, seems hard to get in real-world datasets (Barkhuijsen et al. 1985; Barkhuijsen et al. 1987; Djermoune et al. 2014; J. C. Hoch 1989). In the meantime, global spectral deconvolution (GSD) provides a commercial solution by fitting frequency-domain data, but it tends to overfit spectra by many flexible nonideal peaks. Therefore, an automatic method to decompose untargeted NMR spectra is still needed.

Based on the current development in spectral decomposition, we designed the spectral automatic NMR decomposition (SAND) method and tested it on simulated and experimental *C. elegans* datasets. Spectra were automatically binned and decomposed by optimizing the peak model in the time domain. SAND reliably quantified overlapped peaks and is easily adapted to different sample types with little manual input needed. Annotation can also be automatically generated through correlating and clustering decomposed peaks. Data transfer, processing, and optimization were still computationally intensive, but we relieved this by connecting SAND with Network of Advanced NMR (NAN) and NMRbox (Maciejewski et al. 2017), where data infrastructure and

considerable computational sources are available. This connection enables newly collected metabolic samples to be presented to the researchers in two forms, spectra and peak tables.

Results

The computational workflow for NMR spectral decomposition

SAND automatically converts NMR signals into a feature table for different model systems (Fig. 5.1). SAND was applied and tested in metabolomics datasets in this paper, but it is general and can be expanded to other applications, like protein NMR. NMR data are collected as time-domain signals (free induction decay, FID), and FIDs are often Fourier transformed (FT) into frequency-domain data (Fig. 5.1B). This converts the complex decaying sinusoidal signal in the time domain into individual peaks in the frequency domain and simplifies visualization for chemists. Data in the time and frequency domains are equivalent representations. In metabolomics studies, there are often hundreds of spectra, and each spectrum contains hundreds to thousands of peaks from tens to hundreds of compounds. The spectra are often highly overlapped, not flat in the baseline, and distorted in phase even after processing.

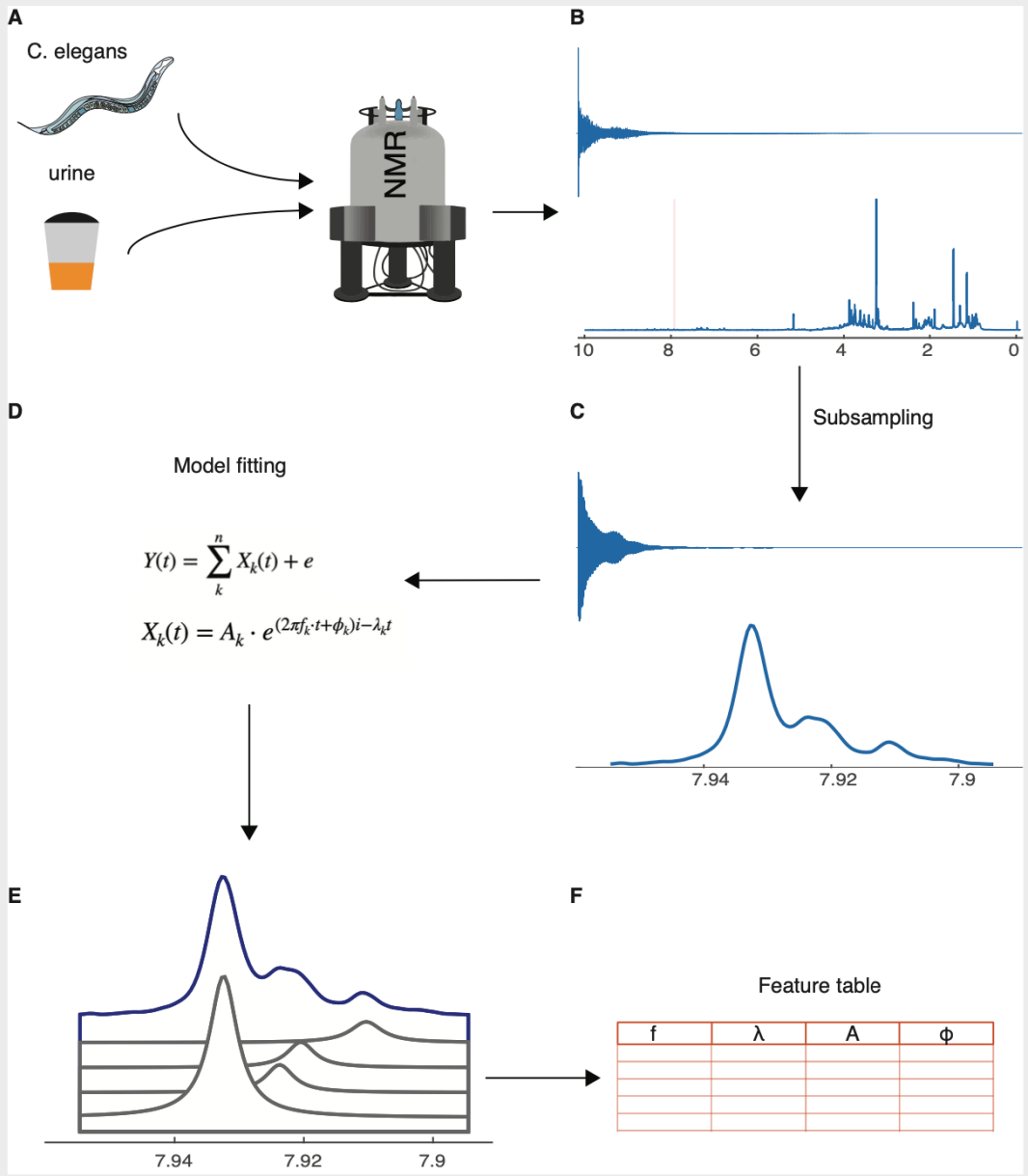


Figure 5.1: The SAND workflow. A: Different biological samples can be measured using NMR, producing spectra with different patterns. B: The time-domain signal is directly collected from NMR and can be Fourier transformed into the frequency domain, where individual peaks are easily visualized. Time and frequency domains are the equivalent representations of the same data. C: The decomposition workflow first subsamples in a frequency range (the pink region in

B) and produces less complex time and frequency-domain data. D: Each peak is then modeled as a complex decaying sinusoidal signal and the spectrum is modeled as a sum of such signals. Peak parameters are searched through optimization. E: The example region (blue line) is decomposed into four overlapping peaks shown as separated spectra (grey). F: A spectrum can be converted into a simpler feature (peak) table with estimated model parameters representing frequency, peak width, amplitude, and phase.

The complex time-domain signal was filtered into frequency ranges and then decomposed (Fig. 5.1C-D). Each peak signal was modeled as one decaying sinusoidal function, which has four optimizable parameters: frequency, peak width, amplitude, and phase. The sum of multiple peak signals was used to fit the FID in a frequency region (Fig. 5.1D). Modeling the whole spectra directly was difficult because of the number of parameters. We instead subsampled the spectrum into frequency regions (details in Methods) and optimized the spectrum in each region (Fig. 5.1C). Such separation also enables the algorithm to be highly parallelizable. In the end, a mixture spectrum was reduced to a feature table and a list of individual peaks (Fig. 5.1E-F).

Evaluation of SAND with benchmark datasets

Multiple datasets were simulated to test SAND performance under different conditions: baseline simulation, simulation with broad peaks, and simulation with phase distortion (Fig. 5.2-5.4). The simulation covered a range of

1.5 ppm, and peak intensities were randomly generated (Fig. 5.2A). Broad peaks were added onto narrow metabolite peaks to resemble experimental spectra (Fig. 5.3). In some simulations, the phase parameter was modified for all peaks to introduce zero-order phase problems.

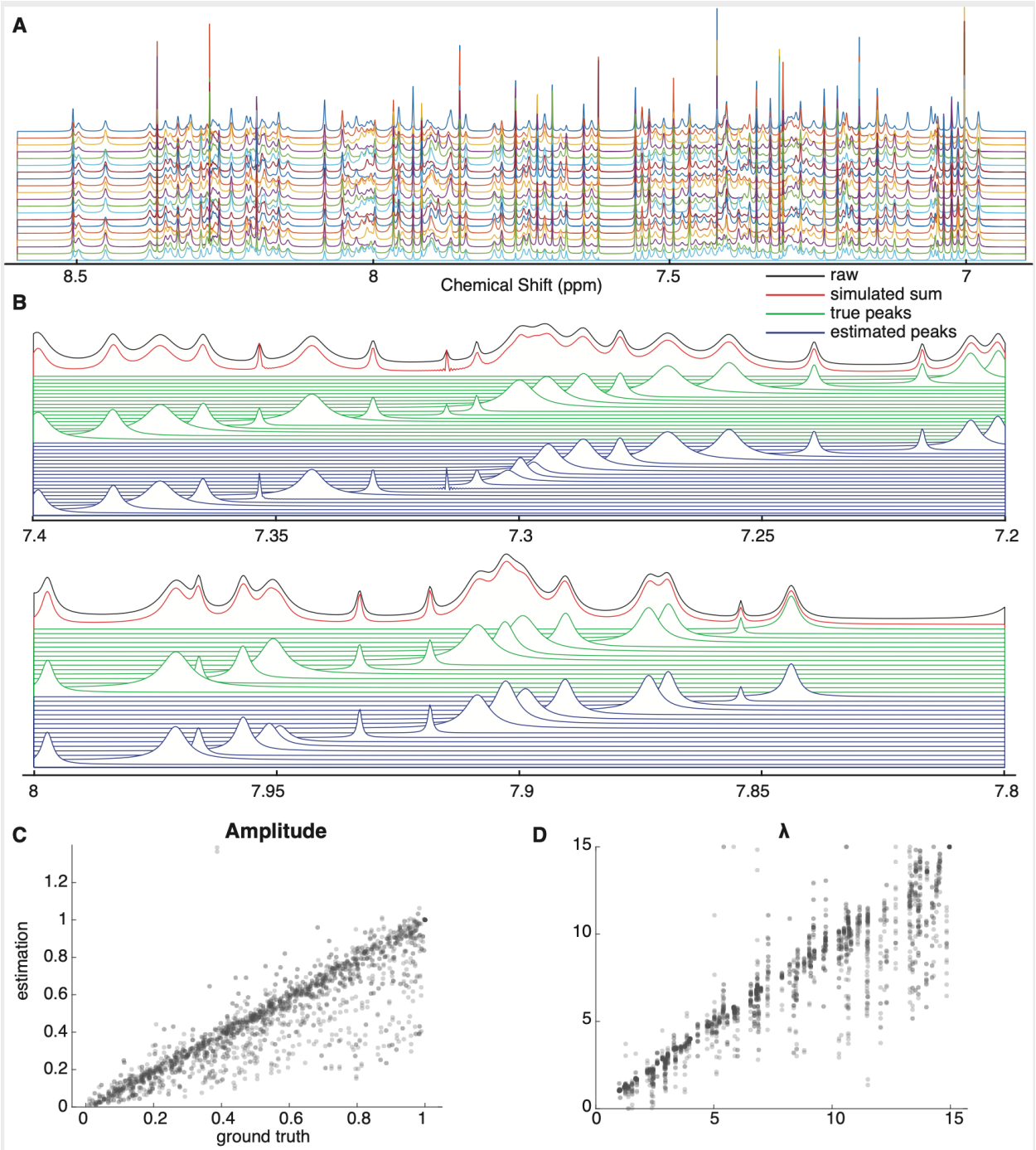


Figure 5.2: Decomposition performance on benchmark spectra. SAND is evaluated with both computational simulated and experimental benchmark spectra. In both cases, the ground truth is known and compared with the SAND estimations. A: Stack plot of 20 simulated spectra with both narrow and broad peaks. Peaks were highly overlapped with randomly generated intensity. The x-axis is the chemical shift and different spectra are distinguished by colors. An example of broad peak overlap can be found in Figure 5.3. B: Two example regions of the decomposition. The black lines are simulated raw spectra; the red lines are the sum of decomposed peaks; the green lines are simulated ground truth peaks; the blue lines are decomposition estimated peaks. C: The scatter plot of ground truth (x-axis) and decomposition estimation (y-axis) for amplitude in the dataset of A. D: The scatter plot of ground truth (x-axis) and decomposition estimation (y-axis) for peak width in the dataset of A. More details can be found in Methods.

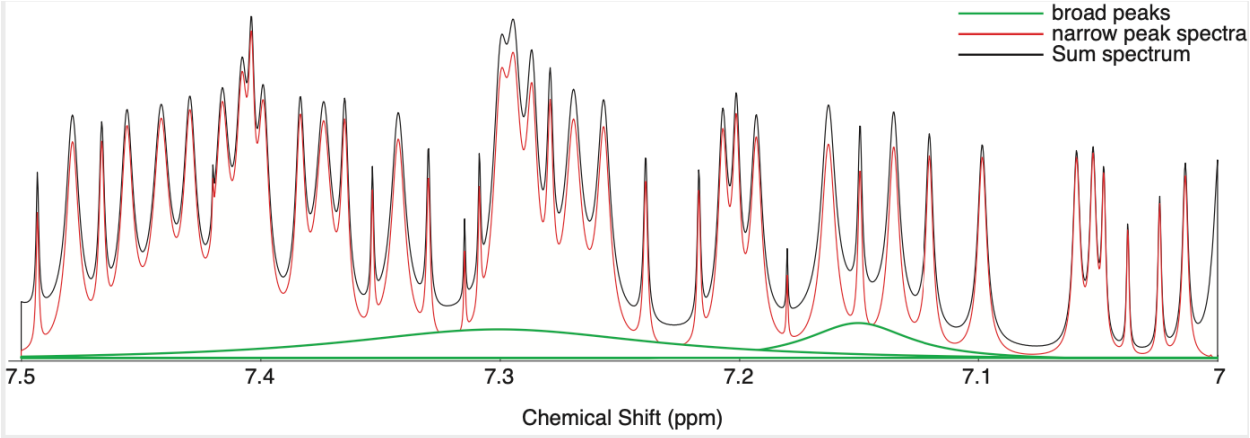


Figure 5.3: Visualization of overlapping between broad and narrow peaks. The x-axis is chemical shift (ppm). The green lines are added broad peaks; the red line is the sum spectra with only narrow peak overlap and no broad peaks; the black line is the spectra with both narrow and broad peaks.

SAND uncovered most peaks accurately (Fig. 5.2B). The estimation (blue peaks) mostly agreed with the ground truth (green peaks), even for highly overlapped regions. Sometimes, SAND had wrong estimation of the number of peaks for highly overlapped regions (Fig. 5.2B), which are also difficult for manual inspection. Globally, the estimation for amplitude and peak width was accurate (Fig. 5.2C-D), proving SAND's quantification capability. The quantification was also good for spectra with slight phase distortion (Fig. 5.4), which is common after automatic phasing. Additionally, we tested cases with high noise and relatively large overlapping peaks (Fig. 5.5). We simulated two fixed broad peaks and one narrow moving peak. All three peaks were successfully decomposed in all conditions, and the small narrow peak was quantified accurately (Fig. 5.5B), even though its overlap with the dominant broad peak biased the quantification.

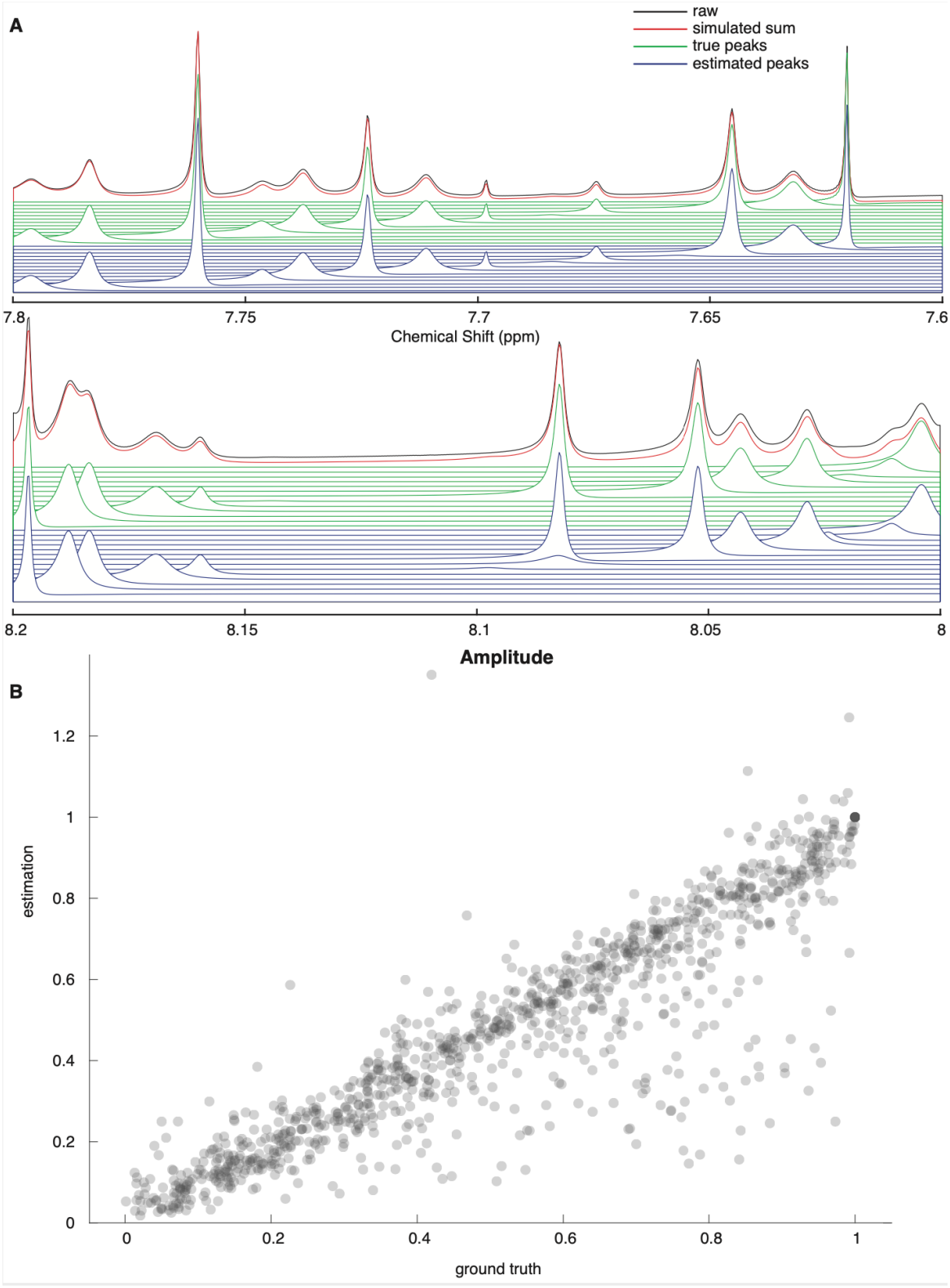


Figure 5.4: Decomposition performance on simulated spectra with phase distortion. A: Two example regions of the decomposition in computationally

simulated spectra with phase distortion. The x-axis is chemical shift (ppm). The black lines are simulated raw spectra; the red lines are sums of decomposed peaks; the green lines are simulated ground truth peaks; the blue lines are decomposition estimated peaks. B: The scatter plot of ground truth (x-axis) and decomposition estimation (y-axis) for amplitude in a phase distorted dataset.

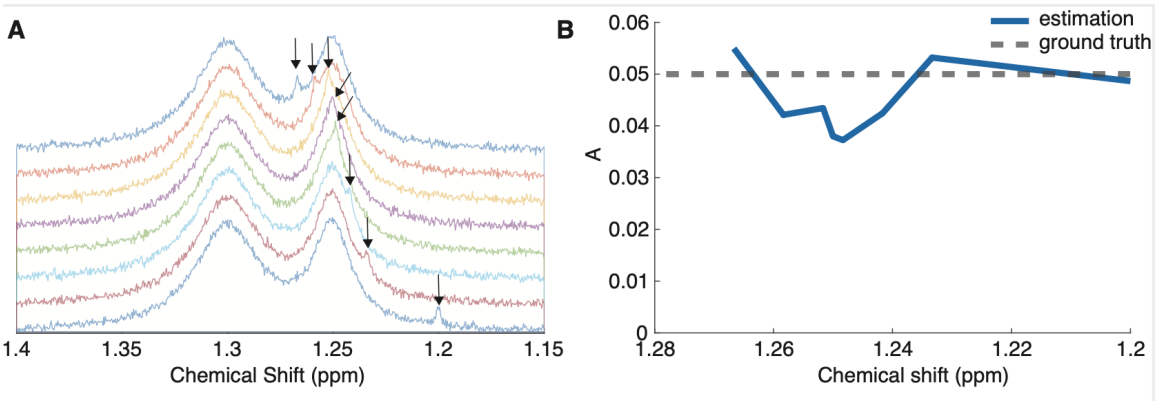


Figure 5.5: A simulation example for overlapping and moving peaks. Two broad peaks and one narrow moving peak are simulated. A: Visualization of the overlapping spectra. The x-axis is chemical shift (ppm), and the different colors shows different simulations. The location of the narrow peak is indicated by arrows. B: Divergence of amplitude estimation of the narrow peak. The x-axis is chemical shift (ppm), and the y-axis is amplitude. The blue line is the estimation, and the grey dotted line is the ground truth.

We also validated SAND on experimental datasets where relative concentrations were known. Ibuprofen and prednisone were mixed with 11 different relative concentrations in DMSO-d_6, and each concentration pair was repeated three times (Krishnamurthy 2013). The relative concentrations were accurately estimated by SAND (Fig. 5.6).

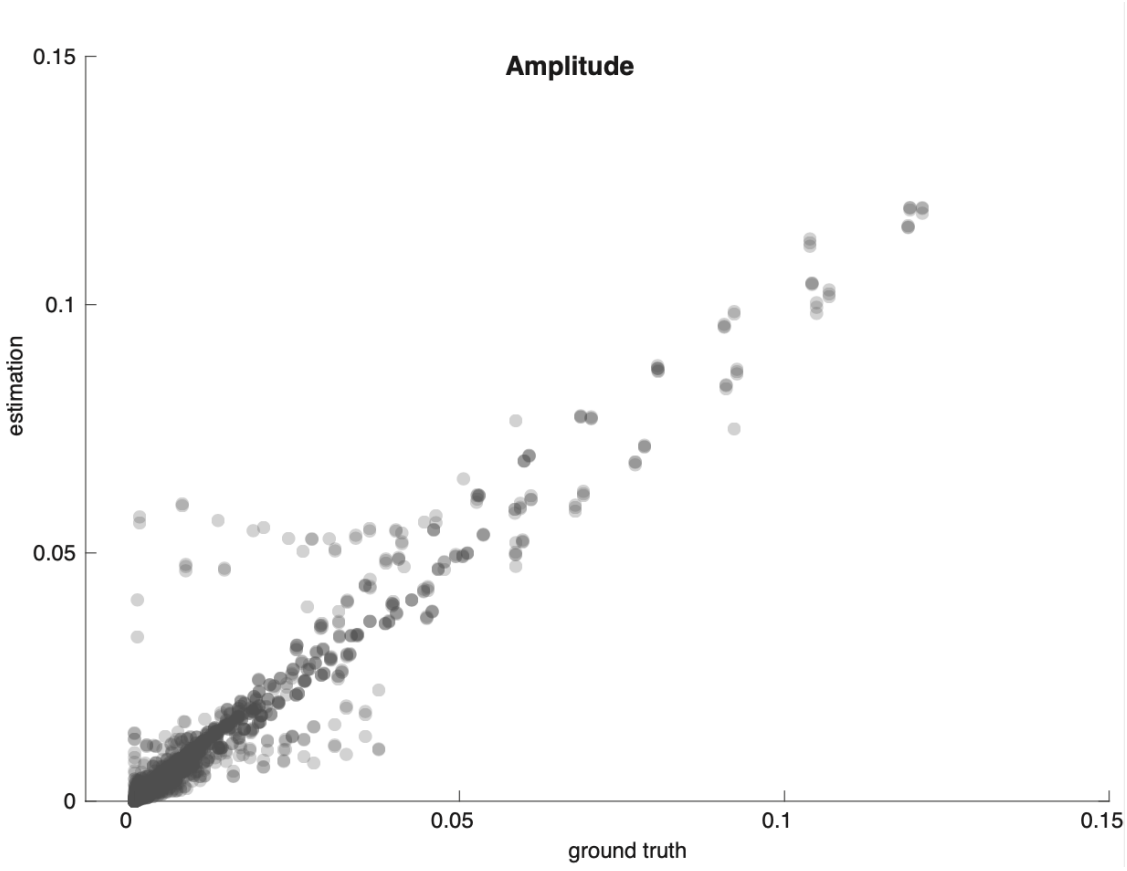


Figure 5.6: The scatter plot for amplitude estimation in the Ibuprofen-prednisone mixture dataset. The x-axis (y-axis) is ground truth (decomposition estimation) in an experimental benchmark dataset with known compound concentrations.

SAND was then compared with other untargeted automatic quantification approaches, intensity and integral. The performance was tested with multiple criteria on the datasets mentioned previously (Table 5.1). Mean squared error (MSE) and relative MSE measures the distance between estimation and ground truth. Slope (k) checks the scaling factor and linearity (details in Methods). In the simulation datasets with around 100 peaks, SAND was consistently better than other methods in all the criteria (Table 5.1). In the ibuprofen-prednisone mixture, three methods had similar and near-perfect performances, probably because of the simplicity of the two-component system.

Table 5.1: Performances of different quantification methods. The benchmark datasets include three computational simulated datasets and one experimental dataset with known relative concentrations. Two of the computationally simulßated datasets have either broad peaks or phase distortions to increase complexity. Three automatic methods, SAND, peak intensity, and peak integral were compared. The performance is evaluated on MSE, relative MSE, correlation, and k. The best performance was highlighted in bold font. Details can be found in Methods.

Methods	Dataset	relative MSE	MSE	correlation	k
SAND	baseline simulation	**1.10E-01**	**2.05E-02**	**0.884**	**0.876**
intensity	baseline simulation	7.02E-01	2.21E-01	0.379	0.756
integral	baseline simulation	3.37E-01	8.50E-02	0.542	0.758
SAND	+ broad peaks	**1.13E-01**	**2.01E-02**	**0.884**	**0.889**
intensity	+ broad peaks	7.03E-01	2.22E-01	0.378	0.756
integral	+ broad peaks	3.43E-01	8.60E-02	0.536	0.762
SAND	+ phase distortion	**1.21E-01**	**1.82E-02**	**0.895**	**0.895**
intensity	+ phase distortion	6.64E-01	2.21E-01	0.421	0.818
integral	+ phase distortion	4.99E-01	9.56E-02	0.606	0.616
SAND	experimental mixture	3.81E-01	6.28E-06	0.972	**1.007**
intensity	experimental mixture	**2.96E-01**	4.07E-06	0.990	1.054
integral	experimental mixture	3.60E-01	**3.29E-06**	**0.990**	1.032

Automatic annotation inference

Besides compound quantification, the decomposed features also inform underlying chemical associations. Statistical total correlation spectroscopy (STOCSY) has been used to find peaks correlated with a targeted peak and annotate the unknown compounds (Cloarec et al. 2005). This utilizes the perfect linearity between different peaks of the same compound under multiple concentrations. Similar to STOCSY, a correlation network can also be built for annotation, where each pair of peaks are connected based on its correlation and network clusters can indicate possible annotations. While overlap can introduce noise and false positives, decomposition cleans the signal by removing the overlap effects.

Annotation was first tested on simulated mixtures with 18 groups of peaks, and in each group, peaks were simulated with fixed relative intensity, like different peaks in a compound. After decomposition, features were matched between samples, correlated, and clustered (Details in Methods). Most peaks in the simulated group (solid lines) can be automatically estimated and clustered (dotted lines) (Fig. 5.7A). For most simulated groups, more than 60% of the ground truth peaks were estimated by the best cluster (Fig. 5.8). In the ibuprofen-prednisone mixtures, two main clusters were uncovered (Fig. 5.7B), and they represent many of the peaks of the compounds. The solid curves represent the spectra with one of the two compounds but still contain some other impurity peaks, such as DMSO and water peaks.

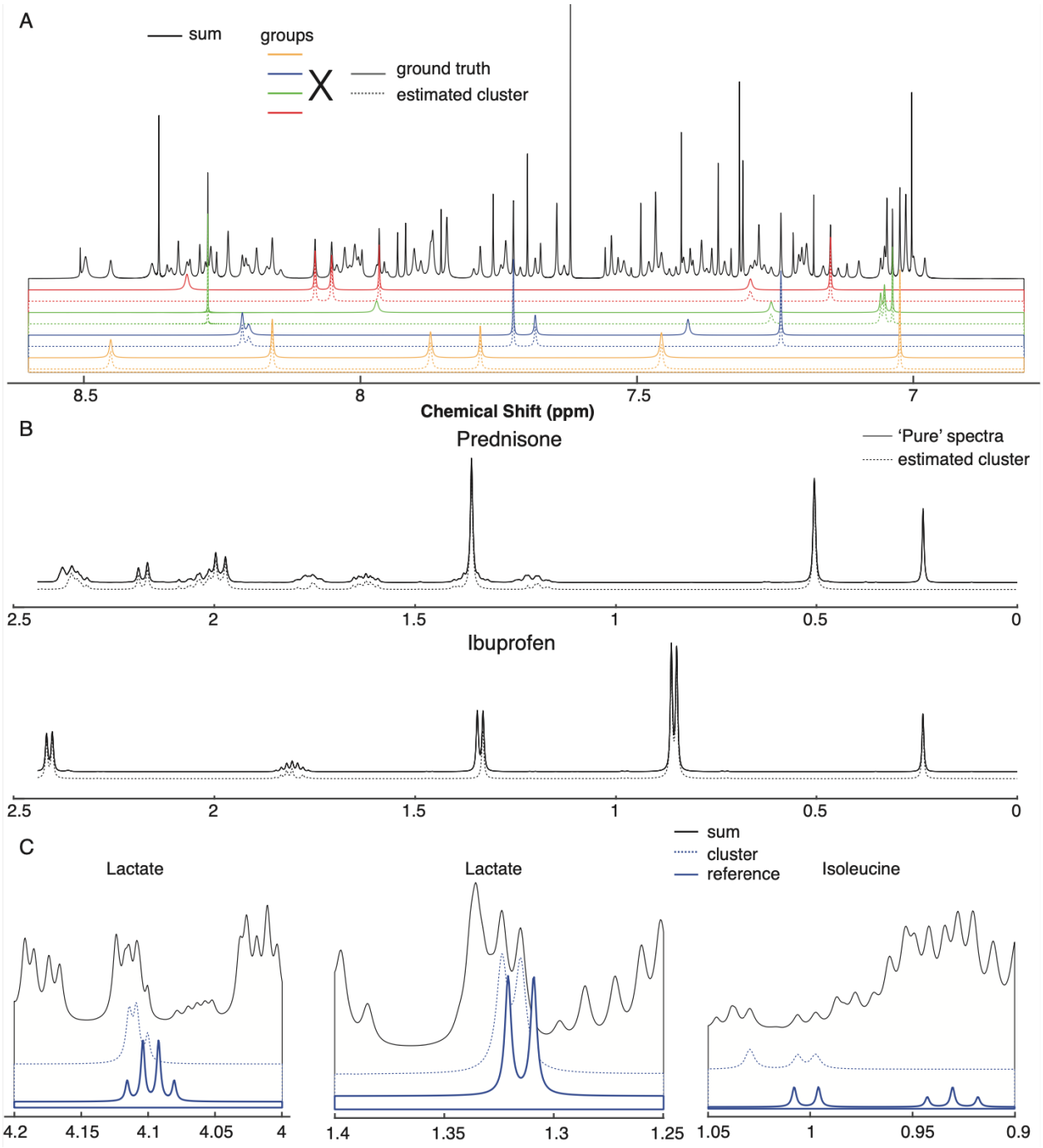


Figure 5.7: Automatic annotation from decomposed spectra. Each spectrum is decomposed into a peak list, and matched peaks among samples were correlated and clustered. Clusters provide peak annotation and are compared with ground truth. A: Recovered clusters in simulated spectra. The spectra were

simulated similarly as in Figure 5.2A but with groups of peaks changing together as different NMR peaks in a compound. The black line is the sum spectra of all peaks. Different color indicates different simulated groups, where solid lines are ground truth peaks and dotted lines are spectra of decomposed peaks from the best-matched cluster. The x-axis is chemical shift (ppm). B: Recovered clusters in an experimental benchmark dataset. Two major compounds, prednisone and ibuprofen, were mixed with different relative concentrations. Recovered clusters were presented in an example region. Solid lines are spectra without the other compound though it is not yet pure solutions. Dotted lines are spectra of clustered decomposed peaks. C: Recovered cluster in a *C. elegans* dataset. Black lines are the raw spectra; blue dotted lines are spectra of clustered decomposed peaks; blue solid lines are the GISSMO reference spectra. There are differences in line broadening between the reference and experimental spectra. Details can be found in Methods.

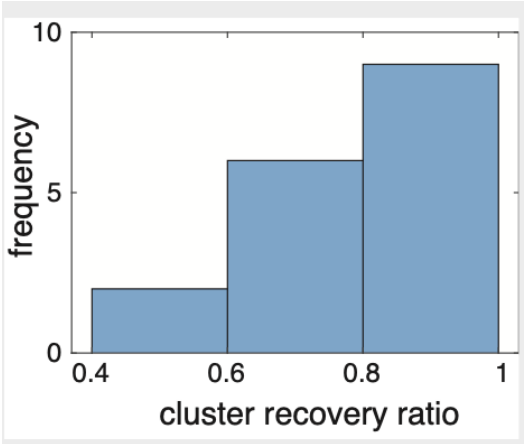


Figure 5.8: performance of recovering clusters in the simulated dataset. The correlation clusters of decomposed peaks were compared with ground truth peak groups. The ratios of recovery were calculated for each group and based on the best-matched clusters. The result is presented as a histogram with the x-axis as the recovery ratio and the y-axis as frequency. Details can be found in Methods.

Decomposition of *C. elegans* NMR samples

Forty NMR spectra were collected on the PD1074 strain. The spectra were highly variable between samples and had many broad peaks and overlap. Each spectrum was decomposed and matched between samples (Fig. 5.9). For such biological samples, no ground truth is available, and it is hard to distinguish close overlap and single peaks. We validated the results partially by annotations, as mentioned in the previous section. A cluster of lactate peaks was found with the multiplets at around 1.3 and 4.1 ppm correlated and clustered (Fig. 5.7C). The relative peak intensity also resembled the GISSMO reference spectra (Dashti et al. 2018), though there were clear differences in chemical shifts as a result of pH and in peak widths as a result of preprocessing. An isoleucine cluster was also uncovered (Fig. 5.7C).

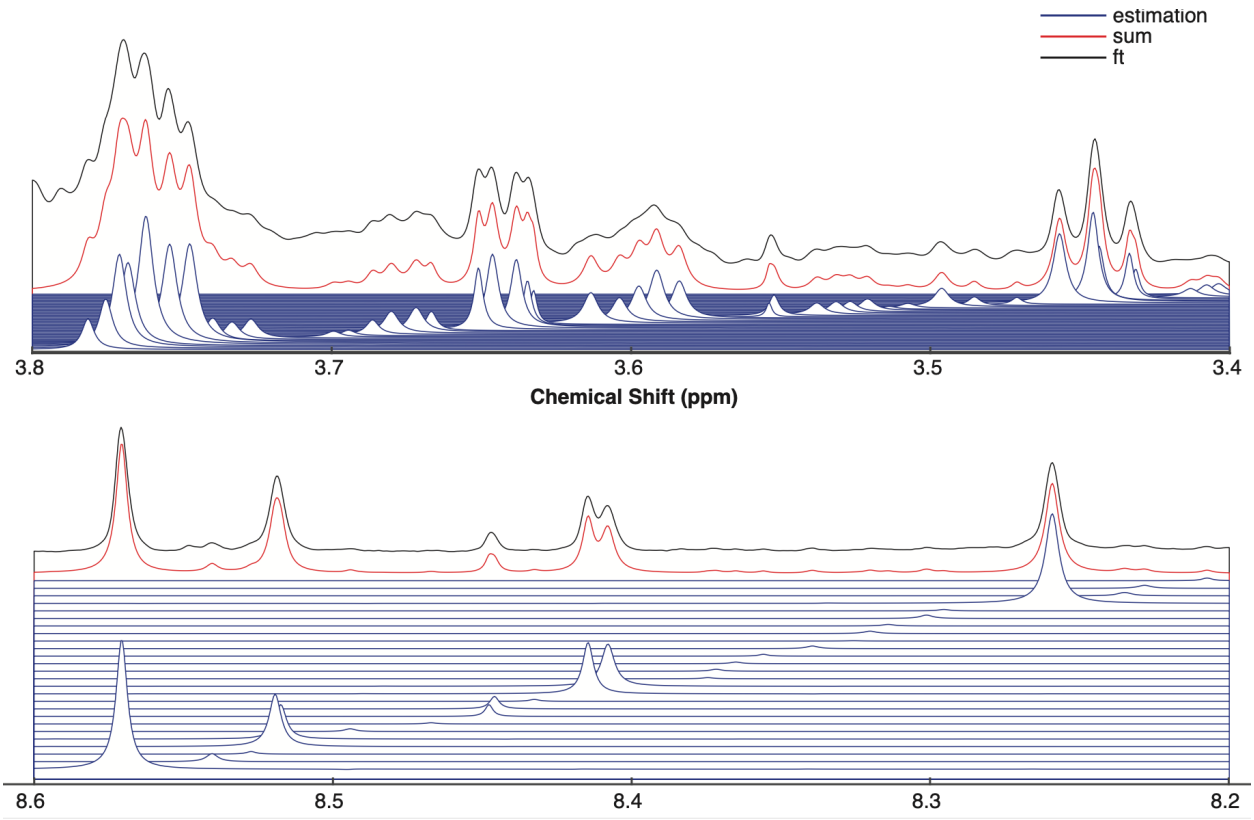


Figure 5.9: Decomposition performance on NMR spectra of PD1074 samples. Two example regions of the decomposition are visualized. The black lines are raw spectra; the red lines are the sum of decomposed peaks; the blue lines are decomposition estimated peaks. The x-axis is the chemical shift.

Discussion

We built SAND to decompose NMR spectra into feature tables. SAND reduces the dataset dimensionality from around 10,000s to 100s, which relieves the curse of dimensionality in the downstream statistical analysis. Such an automatic process also frees researchers from the tedious work of manual inspections. We now discuss a few directions to improve and expand SAND.

Improving peak alignment between samples

Alignment between biological samples has proven to be a difficult problem, especially when the sample pH varies (Tredwell et al. 2016b). Peaks move differently under pH changes, and they sometimes cross each other, which cannot be solved by traditional alignment methods (Vu and Laukens 2013a; Y. Wu et al. 2020). This introduces a crucial peak mapping problem. After SAND, the same decomposed peaks might have different chemical shifts in different samples and need to be mapped correctly. Slight chemical shift variation can be solved by mapping the closest peaks between samples, and pH-induced chemical shift variation can be relieved by sample reordering (unpublished work by Sicong Zhang, University of Georgia). In reordering, pH is assumed to be the major hidden variable and contributes to the titration pattern of different peaks. Leading peaks are manually selected and reordered by chemical shift, after which many peaks can be clarified. The *C. elegans* sample was processed by reordering and many peaks were matched between samples. However, reordering still left many unresolved peaks and was manually intensive, leaving peak matching for NMR profiling samples an unsolved technical problem. Fortunately, a simpler problem, tracking peaks in time-series NMR spectra, was already solved by RTExtract (Y. Wu et al. 2020). We utilized the continuity of chemical shift changes in the time series and tracked peaks through time. In an updated version of SAND, we plan to combine RTExtract and decomposition to quantify overlapped spectral ridges.

The decomposed peak table can also be used for alignment. Currently, NMR alignment methods focus on moving peaks in whole spectra (Vu and Laukens 2013a), but often researchers are only interested in peaks, which is a small subset of the spectral data. We will design a method to align peaks based on the decomposed peak table, which can be computationally faster than conventional approaches because of the reduced dimension of the problem. Estimated peak width might also help match the same peak in different samples.

Performance optimization for SAND

SAND is currently implemented mainly in MATLAB and takes a few hours to finish in a parallel HPC environment for one sample. Although this is an automatic process and computational resources are relatively cheap, an improvement in speed can simplify usage and enable new applications. Signal estimation might also be improved.

SAND optimization can be improved by implementing a more adaptive procedure. Currently, the four parameters of a peak were treated similarly in the optimization process, where Markov chain Monte Carlo (MCMC) randomly selected parameters for each step. However, each peak is a relatively independent identity. Therefore, a better implementation will involve randomly selecting peaks in each step and then proposing changes to relative orthogonal parameters. The more targeted optimization steps can improve both speed and accuracy. We will implement and benchmark the new algorithm in the future version of SAND.

Decomposition as a step of NMR processing

The goal of SAND is to enable objective reproducible automatic quantification of large-scale NMR metabolic studies, like in other omics (Dobin et al. 2013; B. Li and Dewey 2011; Muzny et al. 2012). Besides speed improvement and consistency for different biological systems, such a goal also needs a good interface to existing software to improve research reproducibility.

A C++ rebuild of SAND can improve both speed and usability. MATLAB provides an efficient way to test and visualize the algorithm but lacks adaption to other software and HPC systems. SAND will be rebuilt in C++ with the interface to NMRPipe (F. Delaglio et al. 1995a) and adapted to the resources of NMRbox (Maciejewski et al. 2017) and NAN. NAN also provides a framework to collect, store, transfer, and analyze NMR spectra of large numbers of samples. The metabolomics NMR spectra can be tagged and decomposed automatically in this workflow, and the peak table will be returned to users. Besides NMR profiling samples, SAND can also be combined with RTExtract to quantify time-series NMR features (Y. Wu et al. 2020).

Methods

NMR spectral preprocessing

Each NMR spectrum was preprocessed, binned, and then decomposed. The preprocessing was implemented in NMRPipe and MATLAB, and included zero filling (ZF), FT, phasing, line broadening, referencing, scaling, and format conversion (F. Delaglio et al. 1995a). This is automatic for different samples of one dataset but might require specific changes for different datasets, particularly

because of the instruments used. Binning and decomposition will be covered later.

The FID in NMR spectroscopy is composed of a sum of decaying sinusoidal signals, one for each peak (Equation 5.1). Each signal k is represented by a sequence of complex numbers through time and is defined by four parameters: amplitude (A_k), frequency (f_k), peak width (λ_k), and phase (ϕ_k). The FT of such a signal will produce a Lorentzian peak in NMR frequency-domain data. Frequency and time domains are equivalent representations of the same data.

$$X_k(t) = A_k e^{(2\pi f_k t + \phi_k)i - \lambda_k t} \quad [5.1]$$

Metabolic NMR spectra often contain hundreds to thousands of peaks. Binning separates the spectrum into smaller frequency regions to simplify optimization and parallelization. We automatically generated bins in the frequency domain with slackness 0.99 and bucket size 0.002 (Sousa et al. 2013). The bin boundary is based on local minimum and constrained by the slackness and bucket size parameters. Within each bin, there were one or more peaks, and the neighboring three bins were then combined to include the local pattern of overlap and baseline. The binning was done for all samples together or individually for each sample depending on the variances between samples. Spectra from each bin were then post-processed. The spectrum was shifted to move its minimum to zero to effectively remove the baseline effect. A raised-cosine window function was then applied to each bin, with one in the range of bin and a roll-off shoulder of 0.02 ppm, where the window function smoothly decayed

from one to zero, and zero everywhere else. This window function removes drastic changes in the bin boundary and reduces artifacts in the reverse Fourier transformation (IFT). IFT converts frequency-domain data back to time-domain. In NMRPipe, IFT is preceded by Hilbert transformation (HT) and followed by inverse zero filling (IZF) (F. Delaglio et al. 1995a). Through binning, the raw FID was subsampled into multiple FIDs corresponding to each bin.

NMR spectral decomposition: overview

The FID of each bin was then decomposed independently in parallel. For each FID, the process involved an iteration of initial value estimation, hybrid optimization, and stop criteria checking. Each iteration added one more peak until the stop criteria were met.

Before starting the iteration, the FID was subsampled. Around 6000 points in the beginning were sampled to be used so that noisy end signals were ignored. This setting can be modified according to the case, but it does not have much effect on the results. The new FID was then randomly sampled into three sets: “training” for parameters optimization (70%), “validation” for checking the stop criteria (20%), and “testing” for final evaluation (10%). This separation was intended to infer the reasonable number of peaks in the stop criteria checking step rather than strict cross-validation in forecasting, because time-series samples can be autocorrelated.

NMR spectral decomposition: initial parameter estimation

In each iteration, a new peak was added if the stop criteria were not met. The initial values for the newly added peak were estimated from the residual

signals (Equation 5.2) by sweeping through a frequency grid. The measured FID $Y_j(t)$ of bin j was compared with the sum of $n-1$ simulated signals ($k \in \{1..n-1\}$) to produce the current residual. The signal with maximum amplitude was retrieved from the residual as the initial guess. For such a guess, the frequency was swept with 1000 values and the default value was used for ϕ and λ. The frequency was found by the maximum product of the two complex vectors.

$$r_n(t) = Y_j(t) - \sum_{k=1}^{n-1} X_k(t) \quad [5.2]$$

NMR spectral decomposition: optimization

A hybrid optimization was used to estimate peak parameters by minimizing the objective function (Equation 5.3), where θ was the vector of parameters of all peaks. We used Metropolis sampling for the global parameter search (Battogtokh et al. 2002; Landau and Binder 2014) and interior-point method for the local parameter refinement (Byrd et al. 2000). The interior-point method is the default option in MATLAB for constrained nonlinear multivariable function optimization. Metropolis sampling is one of MCMC methods and constructs a Markov chain to capture the unknown model and generate results consistent with the observation. The stochastic process can explore the complex parameter space with many local optima and arrive at global optima. Great performance has been shown in statistic physics (Landau and Binder 2014) and Bayesian inference (Liu 2001; Salvatier et al. 2016). It has also been applied to biological problems where there are large number of parameters, but the measurement is sparse and noisy (Battogtokh et al. 2002). Introduction of MCMC can be found here (Battogtokh et al. 2002; Landau and Binder 2014; Liu 2001).

$$Argmin_{\theta} \sum_t |Y_j(t) - \sum_{k=1}^{n} X_k(t)|^2 \quad [5.3]$$

The optimization process contained an outer loop (iteration) of MCMC optimization with local refinement in the inner loop (step). One iteration (MCMC sweep) contained $4N$ steps, which is the number of optimizable parameters (four times the number of peaks). In each MCMC step, a random parameter was selected and randomly perturbed. The perturbation depended on the predefined parameter range and relative step size and was bounded by 10% in each step. Afterward, the local optimization occurred every 100 MCMC iterations to refine the optimization solution. The local optimization was not executed for every iteration to save time. The proposed perturbation was accepted or rejected based on the probability (Equation 5.4) where the temperature was defaulted at 80. For each FID, the temperature was adjusted with estimated standard deviation based on the FID end with no signals (Equation 5.5). There were $4N$ steps in one iteration and 2000 iterations (by default) in the optimization of one binned FID. Multiple optimizations (default 3) by different random seeds were started and the best estimation was selected.

$$p = e^{-\frac{\Delta H}{T}} \quad [5.4]$$

$$T_{adj} = T\sigma^2 \quad [5.5]$$

NMR spectral decomposition: stop criteria

The optimization was stopped when any of the criteria were met to control the model complexity. The first criterion regards the improvement from the new signal. It is true if the new model does not improve the objective function by a threshold (default 0.001). The second criterion regards the number of peaks in

the optimized region and the maximum was, by default, 7. This mostly aimed to save computational resources for difficult regions as the computational complexity increased with the maximal number of signals by approximately the power of three. The third criterion regards the initial guess amplitude of the new peak. It is true if the new peak is less than, by default, 10% of the minimum of all existing peaks in the bin.

Benchmark dataset simulation

The simulation was implemented in the time domain with preset parameters (Equation 5.1). We also make the simulation resemble experimental spectra. One urine spectrum was decomposed into a peak table. The peak table was refined by selecting an around 1.5 ppm region, removing distorted peak regions, cleaning peak clusters, randomly generating lambda, and adding the reference peak at 0 ppm. The peak clusters with dense overlap were cleaned by removing close peaks by threshold 0.03 ppm. Around 100 peaks remained after the filtering. Lambda was uniformly sampled from 1 to 15, and the lambda of the reference peak was defined as 4. Concentrations were uniformly sampled, and the spectra were simulated with noise. As in experimental datasets, the first time point of the FID was divided by two and 76 zero time points were added in the beginning. In the simulation with broad peaks, three broad peaks with lambdas greater than or equal to 100 were added. In the simulation with phase distortion, the phase of all peaks was set to be 0.2. In the simulation with groups, 18 groups (compounds) were predefined and within each group, peaks maintained a fixed

relative ratio like those peaks of one compound. Simulated NMR spectra then went through preprocessing and binning steps.

Preprocess experimental NMR spectra

The *C. elegans* dataset was collected with a Bruker NEO 800MHz spectrometer equipped with a 1.7mm TCI cryoprobe and noesypr1d pulse sequence. 0 and 1st order phases were manually corrected. The spectra were preprocessed in NMRPipe and then binned for each sample (F. Delaglio et al. 1995a). The separated binning was necessary as different NMR spectra varied by a large amount.

The prednisone-ibuprofen mixtures were prepared in DMSO-d6 and the NMR data were collected with a Varian Inova 500 spectrometer (Krishnamurthy 2013). The spectra were processed in NMRPipe, including additional formatting, baseline correction, and chemical shift recalibration. The spectra were then binned together because they were simple two-component mixtures.

Evaluation of different quantification approaches

SAND was evaluated on simulated and experimental benchmark datasets and compared with intensity and integral. In each bin before combination, the minimum was shifted to zero, and then maximum intensity was determined, and integral was calculated by Trapezoidal numerical integration. Peaks in the simulated data were normalized to a constant peak at 0 ppm. Relative concentrations were compared for the prednisone-ibuprofen mixtures. MSE (Equation 5.6) was calculated for each matched peak e (total number N_e), for one sample k, and between estimation x and ground truth y. Relative MSE (RMSE)

had each deviation normalized by mean (Equation 5.7). Pearson correlation was used for correlation. Slope k was calculated by fitting a linear regression model between estimation and ground truth with no intercept. Evaluations were calculated within each sample, from which the means and standard errors were calculated.

$$MSE_k = \frac{1}{N_e}\sum_e^{N_e}\left(x_{k,e} - y_{k,e}\right)^2 \quad [5.6]$$

$$RMSE_k = \frac{1}{N_e}\sum_e^{N_e}\left((x_{k,e} - y_{k,e})/\frac{x_{k,e}+y_{k,e}}{2}\right)^2 \quad [5.7]$$

Correlation network and clustering

To build the correlation network, we matched peaks by chemical shift distance between samples and calculated the Spearman correlation. The edges were selected with 0.9 and 0.5 as the correlation threshold in the simulated dataset and in the prednisone-ibuprofen mixtures. For these two datasets, the groups of peaks were clear and separated without clustering.

Analyzing the *C. elegans* samples involves more complexity in peak matching and clustering. Peaks with pH-induced movement were matched by reordering. All samples were reordered by the chemical shifts of a leading peak, and peaks with titration patterns after reordering were matched. Other peaks were matched by chemical shift distance iteratively. Feature pairs with correlations higher than 90% quantile were connected and the Markov Cluster Algorithm (MCL) was used to cluster the correlation network with granularity 4 (Enright et al. 2002; van Dongen and Abreu-Goodger 2012).

Code and data availability

The SAND program has been implemented in MATLAB, R, and C shell. Codes are freely available through GitHub (https://github.com/edisonomics/SAND). It was tested on a local iMac Pro computer, sapelo2 at Georgia Advance Computing Resource Center (GACRC), and NMRbox servers (Maciejewski et al. 2017).

Acknowledgements

We thank Jeff Hoch, Mark W. Maciejewski and Chad Rienstra for useful discussions. We appreciate computational support from GACRC, especially Shan-Ho Tsai and Zhuofei Hou. This work was supported by the NSF 1946970: Mid-scale RI-2 Consortium: Network for Advanced NMR. This study made use of NMRbox: National Center for Biomolecular NMR Data Processing and Analysis, a Biomedical Technology Research Resource (BTRR), which is supported by NIH grant P41GM111135 (NIGMS). We appreciated Gerard Weatherby for supporting the computing in NMRbox.

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

My thesis has focused on analyzing metabolic network and its dynamics through analytical and computational approaches, composed of data collection, feature extraction, statistical analysis, and knowledge extraction. We used CIVM-NMR to collect time-series *in vivo* metabolic profiles, RTExtract and SAND to extract and quantify metabolic features from the complex data, and FPCA and empirical networks to reduce the dimensionality and extract biological knowledge (Judge et al. 2019; Y. Wu et al. 2020; Yue Wu et al. 2022). Such a process gives untargeted and empirical evaluation of the central metabolism and its dynamic response (Link et al. 2015; Yue Wu et al. 2022). Expansion to more samples, more perturbation conditions, and different biological systems is the next step, where the method will also be validated and refined. In particular, profiling metabolic dynamics and associations in precision and cohort study is promising.

Metabolic Dynamics and Precision Medicine

Multiomics approaches, from genomics to proteomics and metabolomics, have been used to measure biological states in population human and associated to different disease conditions (Ahadi et al. 2020; Liang et al. 2020; Peng et al. 2018; Rose et al. 2019; Sailani et al. 2020). Contrary to well-controlled experimental conditions, cohort candidates are perturbed by more events, including food and medicine intake, exercise, environmental conditions,

and infection (Ahadi et al. 2020; Jiang et al. 2018; Rose et al. 2019; Sailani et al. 2020). The internal biological network responds to those events through multiple levels of regulation with metabolomics among the most dynamic one and closest to phenotypes (Edison et al. 2021; Guijas et al. 2018; C. H. Johnson et al. 2016). Hence, metabolomics is a rich place to search for disease associated biomarkers and improve our understanding of the network interaction in disease development. Here, I focus on the technical difficulties and opportunities.

Metabolic profiling in cohort study lacks its coverage in time and in the network. On the one hand, some routine measurements (e.g., blood glucose, heart rates) can be collected continuously through wearable devices (e.g., continuous glucose monitoring system, Fitbit, and Apple Watch), and dynamic profiles are available in response to food intake or exercise (Alavi et al. 2022; Hall et al. 2018; Rose et al. 2019). On the other hand, cohort metabolome can be profiled through metabolomics (e.g., LC-MS) on urine, serum, or plasma at a less frequent schedule because of the difficulty in sampling and expense (Ahadi et al. 2020; Rose et al. 2019). Hence, for many metabolites, even those in central carbon metabolism, the dynamic profile is lacking in human subjects. Meanwhile, compared with other relatively stable omics, the metabolome can change within hours or even seconds. Dense metabolomics sampling tends to be expensive and burdensome for the subject. One possibility is to reduce sample size and simplify the collection process for metabolomics. New developments in magnetic resonance spectroscopy (MRS) also enable observation of multiple metabolites in living systems (Hwang and Choi 2015; Manganas et al. 2007).

Computational approaches can also be developed to relieve the coverage problem. Associations can be drawn between a few metabolites with dense time samples and many more metabolites which are profiled of a single time point. Then, machine learning models can be built to infer dynamic profiles of many compounds (Goldberger et al. 2000; Mercer et al. 2021; Rubanova et al. 2019). Different metabolic states can also be clustered separately for time series and fixed time point profiling. The clusters separate different *in vivo* states in lower dimensions of both cohort individuals and time segmentations, informing disease subtypes and transitions (Mercer et al. 2021; Moon et al. 2020; Rose et al. 2019). Integration with other omics data through Bayesian network or flux analysis can also infer metabolic states (Hackett et al. 2016; Heirendt et al. 2019; Khodayari and Maranas 2016; Macklin et al. 2020; Vaske et al. 2010). Modeling the dynamic biological network with multiple omics, time points, and samples is often computationally intensive. Connection with well-designed cloud computing infrastructure, like AWS, NAN and HTCondor, enables efficient data transfer, privacy protection, and computational resources for the highly parallel computing (Maciejewski et al. 2017; Thain et al. 2005).

BIOGRAPHICAL SKETCH

Yue Wu grew up in a small seaside town in the north part of China. He had early interests in math, physics, and computer science. As an undergraduate at Nanjing University, he chose to major in biotechnology and started to focus on computational work. He started with statistical methods in phylogenetics and evolution. In his senior year, he was exposed to sequencing and bioinformatics and started to collaborate with experimental researchers. Later, he was also exposed to chromatin structure, epigenomics, cancer multiomics, and disease modeling. During the exploration stage, he found that even though considerable amounts of data had been collected from all omics, there lacked an understanding of the biological network. In particular, the metabolic network lacks its coverage in time and compounds. In 2017, he joined the lab of Arthur S. Edison and Jonathan Arnold and started to work on time-series metabolomics. He created multiple computational tools that improve automation in metabolomics and extract knowledge from the time-series data. In the future, he will apply his knowledge in a broader field, including modeling multiomics in precision medicine.

REFERENCES

Abedi, E. and Hashemi, S. M. B. (2020), 'Lactic acid production - producing microorganisms and substrates sources-state of art', *Heliyon,* 6 (10).

Ackerman, J. J., et al. (1996), 'The NMR chemical shift pH measurement revisited: analysis of error and modeling of a pH dependent reference', *Magn Reson Med,* 36 (5), 674-83.

Ahadi, S., et al. (2020), 'Personal aging markers and ageotypes revealed by deep longitudinal profiling', *Nature Medicine,* 26 (1), 83-+.

Al-Omari, A., et al. (2018), 'Discovering Regulators in Post-Transcriptional Control of the Biological Clock of Neurospora crassa Using Variable Topology Ensemble Methods on GPUs', *Ieee Access,* 6, 54582-94.

Al-Omari, A. M., et al. (2022), 'Ensemble Methods for Identifying RNA Operons and Regulons in the Clock Network of Neurospora Crassa', *IEEE Access,* 10, 32510-24.

Alavi, A., et al. (2022), 'Real-time alerting system for COVID-19 and other stress events using wearable data', *Nature Medicine,* 28 (1), 175-+.

Alves, A. C., et al. (2009), 'Analytic Properties of Statistical Total Correlation Spectroscopy Based Information Recovery in H-1 NMR Metabolic Data Sets', *Analytical Chemistry,* 81 (6), 2075-84.

Anaraki, Maryam Tabatabaei, Simpson, Myrna J., and Simpson, André J. (2018), 'Reducing impacts of organism variability in metabolomics via time

trajectory in vivo NMR', *Magnetic Resonance in Chemistry,* 56 (11), 1117-23.

Ashburner, M., et al. (2000), 'Gene Ontology: tool for the unification of biology', *Nature Genetics,* 25 (1), 25-29.

Astle, W., et al. (2012), 'A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures', *Journal of the American Statistical Association,* 107 (500), 1259-71.

Aue, W. P., Karhan, J., and Ernst, R. R. (1976), 'Homonuclear Broad-Band Decoupling and 2-Dimensional J-Resolved Nmr-Spectroscopy', *Journal of Chemical Physics,* 64 (10), 4226-27.

Augustijn, D., et al. (2016), 'Metabolic Profiling of Intact Arabidopsis thaliana Leaves during Circadian Cycle Using 1H High Resolution Magic Angle Spinning NMR', *PLoS One,* 11 (9), e0163258.

Bar, N., et al. (2020), 'A reference map of potential determinants for the human serum metabolome', *Nature,* 588 (7836), 135-40.

Barkhuijsen, H., Debeer, R., and Vanormondt, D. (1987), 'Improved Algorithm for Noniterative Time-Domain Model-Fitting to Exponentially Damped Magnetic-Resonance Signals', *Journal of Magnetic Resonance,* 73 (3), 553-57.

Barkhuijsen, H., et al. (1985), 'Retrieval of Frequencies, Amplitudes, Damping Factors, and Phases from Time-Domain Signals Using a Linear Least-Squares Procedure', *Journal of Magnetic Resonance,* 61 (3), 465-81.

Bastawrous, M., et al. (2018a), 'In-Vivo NMR Spectroscopy: A Powerful and Complimentary Tool for Understanding Environmental Toxicity', *Metabolites,* 8 (2).

Bastawrous, M., et al. (2018b), 'In-Vivo NMR Spectroscopy: A Powerful and Complimentary Tool for Understanding Environmental Toxicity', *Metabolites,* 8 (2).

Bates, R. G. and Pinching, G. D. (1949), 'Resolution of the Dissociation Constants of Citric Acid at 0-Degrees to 50-Degrees, and Determination of Certain Related Thermodynamic Functions', *Journal of the American Chemical Society,* 71 (4), 1274-83.

Battogtokh, D., et al. (2002), 'An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa', *Proc Natl Acad Sci U S A,* 99 (26), 16904-9.

Bauermeister, A., et al. (2022), 'Mass spectrometry-based metabolomics in microbiome investigations', *Nature Reviews Microbiology,* 20 (3), 143-60.

Beadle, G. W. and Tatum, E. L. (1941), 'Genetic Control of Biochemical Reactions in Neurospora', *Proc Natl Acad Sci U S A,* 27 (11), 499-506.

Beck, D. P. and Greenawalt, J. W. (1977), 'Composition and synthesis of cellular lipids in Neurospora crassa during cellular differentiation', *J Bacteriol,* 131 (1), 188-93.

Beckonert, O., et al. (2007), 'Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts', *Nature Protocols,* 2 (11), 2692-703.

Beckonert, O., et al. (2010), 'High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues', *Nat Protoc,* 5 (6), 1019-32.

Bencina, M. (2013), 'Illumination of the spatial order of intracellular pH by genetically encoded pH-sensitive sensors', *Sensors (Basel),* 13 (12), 16736-58.

Bennett, Bryson D, et al. (2009), 'Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli', *Nature chemical biology,* 5 (8), 593.

Besl, P. J. and Jain, R. C. (1986), 'Invariant Surface Characteristics for 3d Object Recognition in Range Images', *Computer Vision Graphics and Image Processing,* 33 (1), 33-80.

Besl, P. J. and Jain, R. C. (1988), 'Segmentation through Variable-Order Surface Fitting', *Ieee Transactions on Pattern Analysis and Machine Intelligence,* 10 (2), 167-92.

Bifarin, O. O., et al. (2021), 'Machine Learning-Enabled Renal Cell Carcinoma Status Prediction Using Multiplatform Urine-Based Metabolomics', *Journal of Proteome Research,* 20 (7), 3629-41.

Bingol, K., et al. (2016), 'Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server', *Analytical Chemistry,* 88 (24), 12411-18.

Bingol, K., et al. (2015), 'Unified and Isomer-Specific NMR Metabolomics Database for the Accurate Analysis of C-13-H-1 HSQC Spectra', *Acs Chemical Biology,* 10 (2), 452-59.

Bretthorst, G. L. (1990a), 'An Introduction to Parameter-Estimation Using Bayesian Probability-Theory', *Maximum Entropy and Bayesian Methods,* 39, 53-79.

Bretthorst, G. L. (1990b), 'Bayesian-Analysis .1. Parameter-Estimation Using Quadrature Nmr Models', *Journal of Magnetic Resonance,* 88 (3), 533-51.

Bretthorst, G. L. (1990c), 'Bayesian-Analysis .2. Signal-Detection and Model Selection', *Journal of Magnetic Resonance,* 88 (3), 552-70.

Bretthorst, G. L. (1990d), 'Bayesian-Analysis .3. Applications to Nmr Signal-Detection, Model Selection, and Parameter-Estimation', *Journal of Magnetic Resonance,* 88 (3), 571-95.

Brockerman, J. A., et al. (2019), 'The pKa values of the catalytic residues in the retaining glycoside hydrolase T26H mutant of T4 lysozyme', *Protein Sci,* 28 (3), 620-32.

Brody, S. and Tatum, E. L. (1967), 'Phosphoglucomutase mutants and morphological changes in neurospora crassa', *Proc Natl Acad Sci U S A,* 58 (3), 923-30.

Brown, K. S. and Sethna, J. P. (2003), 'Statistical mechanical approaches to models with many poorly known parameters', *Physical Review E,* 68 (2).

Byrd, R. H., Gilbert, J. C., and Nocedal, J. (2000), 'A trust region method based on interior point techniques for nonlinear programming', *Mathematical Programming,* 89 (1), 149-85.

Cannon, William, et al. (2018), 'Prediction of Metabolite Concentrations, Rate Constants and Post-Translational Regulation Using Maximum Entropy-Based Simulations with Application to Central Metabolism of Neurospora crassa', *Processes,* 6 (6), 63.

Carr, H. Y. and Purcell, E. M. (1954), 'Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments', *Physical Review,* 94 (3), 630-38.

Casey, J. R., Grinstein, S., and Orlowski, J. (2010), 'Sensors and regulators of intracellular pH', *Nature Reviews Molecular Cell Biology,* 11 (1), 50-61.

Caspi, R., et al. (2020), 'The MetaCyc database of metabolic pathways and enzymes - a 2019 update', *Nucleic Acids Res,* 48 (D1), D445-D53.

Chadeau-Hyam, M., et al. (2010), 'Metabolic Profiling and the Metabolome-Wide Association Study: Significance Level For Biomarker Identification', *Journal of Proteome Research,* 9 (9), 4620-27.

Choe, A., et al. (2012), 'Sex-specific mating pheromones in the nematode Panagrellus redivivus', *Proceedings of the National Academy of Sciences of the United States of America,* 109 (51), 20949-54.

Cloarec, O., et al. (2005), 'Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets', *Anal Chem,* 77 (5), 1282-9.

Colvin, H. J., Sauer, B. L., and Munkres, K. D. (1973), 'Glucose utilization and ethanolic fermentation by wild type and extrachromosomal mutants of Neurospora crassa', *J Bacteriol,* 116 (3), 1322-8.

Contrepois, K., et al. (2020), 'Molecular Choreography of Acute Exercise', *Cell,* 181 (5), 1112-+.

Craven, P. and Wahba, G. (1979), 'Smoothing Noisy Data with Spline Functions - Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation', *Numerische Mathematik,* 31 (4), 377-403.

Csenki, L., et al. (2007), 'Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional 1H NMR data', *Anal Bioanal Chem,* 389 (3), 875-85.

Dashti, H., et al. (2017), 'Spin System Modeling of Nuclear Magnetic Resonance Spectra for Applications in Metabolomics and Small Molecule Screening', *Anal Chem,* 89 (22), 12201-08.

Dashti, H., et al. (2018), 'Applications of Parametrized NMR Spin Systems of Small Molecules', *Analytical Chemistry,* 90 (18), 10646-49.

de Beer, R. and van Ormondt, D. (1992), 'Analysis of NMR Data Using Time Domain Fitting Procedures', in M. Rudin (ed.), *In-Vivo Magnetic Resonance Spectroscopy I: Probeheads and Radiofrequency Pulses Spectrum Analysis* (Berlin, Heidelberg: Springer Berlin Heidelberg), 201-48.

de la Fuente, A., et al. (2004), 'Discovery of meaningful associations in genomic data using partial correlation coefficients', *Bioinformatics,* 20 (18), 3565-74.

de Paula, R., et al. (2002), 'Molecular and biochemical characterization of the Neurospora crassa glycogen synthase encoded by the gsn cDNA', *Mol Genet Genomics,* 267 (2), 241-53.

DeBerardinis, R. J., et al. (2007), 'Beyond aerobic glycolysis: Transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis', *Proceedings of the National Academy of Sciences of the United States of America,* 104 (49), 19345-50.

Deguchi, Shigeru, et al. (2011), 'Microbial growth at hyperaccelerations up to 403,627× g', *Proceedings of the National Academy of Sciences,* 108 (19), 7997-8002.

Delaglio, F., et al. (1995a), 'Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes', *Journal of Biomolecular Nmr,* 6 (3), 277-93.

Delaglio, Frank, et al. (1995b), 'NMRPipe: A multidimensional spectral processing system based on UNIX pipes', *Journal of Biomolecular NMR,* 6 (3), 277-93.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997), 'Exploring the metabolic and genetic control of gene expression on a genomic scale', *Science,* 278 (5338), 680-6.

Djermoune, E. H., Tomczak, M., and Brie, D. (2014), 'NMR Data Analysis: A Time-Domain Parametric Approach Using Adaptive Subband

Decomposition', *Oil & Gas Science and Technology-Revue D Ifp Energies Nouvelles,* 69 (2), 229-44.

Dobin, A., et al. (2013), 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics,* 29 (1), 15-21.

Dorsam, S., et al. (2017), 'Sustainable carbon sources for microbial organic acid production with filamentous fungi', *Biotechnol Biofuels,* 10, 242.

Dreyfuss, J. M., et al. (2013), 'Reconstruction and Validation of a Genome-Scale Metabolic Model for the Filamentous Fungus Neurospora crassa Using FARM', *Plos Computational Biology,* 9 (7).

Dunn, W. B., et al. (2011), 'Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry', *Nature Protocols,* 6 (7), 1060-83.

Edison, A. S., Espinoza, E., and Zachariah, C. (1999), 'Conformational ensembles: the role of neuropeptide structures in receptor binding', *J Neurosci,* 19 (15), 6318-26.

Edison, A. S., et al. (2021), 'NMR: Unique Strengths That Enhance Modern Metabolomics Research', *Analytical Chemistry,* 93 (1), 478-99.

Edson, C. M. and Brody, S. (1976), 'Biochemical and genetic studies on galactosamine metabolism in Neurospora crassa', *J Bacteriol,* 126 (2), 799-805.

Edwards, J. S., Covert, M., and Palsson, B. (2002), 'Metabolic modelling of microbes: the flux-balance approach', *Environmental Microbiology,* 4 (3), 133-40.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002), 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Research,* 30 (7), 1575-84.

Fell, D. A. and Small, J. R. (1986), 'Fat Synthesis in Adipose-Tissue - an Examination of Stoichiometric Constraints', *Biochemical Journal,* 238 (3), 781-86.

Foley, D. A., et al. (2014), 'NMR flow tube for online NMR reaction monitoring', *Anal Chem,* 86 (24), 12008-13.

Franz, M., et al. (2016), 'Cytoscape.js: a graph theory library for visualisation and analysis', *Bioinformatics,* 32 (2), 309-11.

Friebel, A., et al. (2019), 'Reaction Monitoring by Benchtop NMR Spectroscopy Using a Novel Stationary Flow Reactor Setup', *Industrial & Engineering Chemistry Research,* 58 (39), 18125-33.

Fuhrer, T., et al. (2017), 'Genomewide landscape of gene-metabolome associations in Escherichia coli', *Mol Syst Biol,* 13 (1), 907.

Gaderer, R., et al. (2017), 'N-acetylglucosamine, the building block of chitin, inhibits growth of Neurospora crassa', *Fungal Genet Biol,* 107, 1-11.

Galagan, J. E., et al. (2003), 'The genome sequence of the filamentous fungus Neurospora crassa', *Nature,* 422 (6934), 859-68.

Ghasemi, O., et al. (2011), 'Bayesian parameter estimation for nonlinear modelling of biological pathways', *BMC Syst Biol,* 5 Suppl 3, S9.

Goldberger, A. L., et al. (2000), 'PhysioBank, PhysioToolkit, and PhysioNet - Components of a new research resource for complex physiologic signals', *Circulation,* 101 (23), E215-E20.

Gonzalezmendez, R., et al. (1982), 'Continuous-Flow Nmr Culture System for Mammalian-Cells', *Biochimica Et Biophysica Acta,* 720 (3), 274-80.

Gouveia, G. J., et al. (2021), 'Long-Term Metabolomics Reference Material', *Analytical Chemistry,* 93 (26), 9193-99.

Granger, C. W. J. (1969), 'Investigating Causal Relations by Econometric Models and Cross-Spectral Methods', *Econometrica,* 37 (3), 424-38.

Greenfield, N. J., et al. (1988), 'Metabolism of D-glucose in a wall-less mutant of Neurospora crassa examined by 13C and 31P nuclear magnetic resonances: effects of insulin', *Biochemistry,* 27 (23), 8526-33.

Guijas, C., et al. (2018), 'Metabolomics activity screening for identifying metabolites that modulate phenotype', *Nature Biotechnology,* 36 (4), 316-20.

Gustavsen, JA, et al. (2019), 'RCy3: Network biology using Cytoscape from within R [version 3; peer review: 3 approved]', *F1000Research,* 8 (1774).

Hackett, S. R., et al. (2016), 'Systems-level analysis of mechanisms regulating yeast metabolic flux', *Science,* 354 (6311).

Hackett, S. R., et al. (2020), 'Learning causal networks using inducible transcription factors and transcriptome-wide time series', *Molecular Systems Biology,* 16 (3).

Hall, H., et al. (2018), 'Glucotypes reveal new patterns of glucose dysregulation', *Plos Biology,* 16 (7).

Hanahan, D. and Weinberg, R. A. (2011), 'Hallmarks of cancer: the next generation', *Cell,* 144 (5), 646-74.

Hao, J., et al. (2012), 'BATMAN-an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model', *Bioinformatics,* 28 (15), 2088-90.

Hao, J., et al. (2014a), 'Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN', *Nat Protoc,* 9 (6), 1416-27.

Hao, J., et al. (2014b), 'Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN', *Nature Protocols,* 9 (6), 1416-27.

Haralick, R. M., Watson, L. T., and Laffey, T. J. (1983), 'The Topographic Primal Sketch', *International Journal of Robotics Research,* 2 (1), 50-72.

Hattori, A., et al. (2017), 'Cancer progression by reprogrammed BCAA metabolism in myeloid leukaemia', *Nature,* 545 (7655), 500-04.

Haug, K., et al. (2020), 'MetaboLights: a resource evolving in response to the needs of its scientific community', *Nucleic Acids Research,* 48 (D1), D440-D44.

Heirendt, L., et al. (2019), 'Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0', *Nature Protocols,* 14 (3), 639-702.

Hill, E. P. and Sussman, A. S. (1964), 'Development of Trehalase and Invertase Activity in Neurospora', *J Bacteriol,* 88, 1556-66.

Hoch, J. C. (1989), 'Modern Spectrum Analysis in Nuclear Magnetic-Resonance - Alternatives to the Fourier-Transform', *Methods in Enzymology,* 176, 216-41.

Hoch, Jeffrey C. and Stern, Alan S. (1996), *NMR data processing* (New York: Wiley-Liss) xi, 196 p.

Holmes, E., et al. (2008), 'Human metabolic phenotype diversity and its association with diet and blood pressure', *Nature,* 453 (7193), 396-U50.

Horowitz, N. H., Bonner, D., and Houlahan, M. B. (1945), 'The Utilization of Choline Analogues by Cholineless Mutants of Neurospora', *Journal of Biological Chemistry,* 159 (1), 145-51.

Hwang, J. H. and Choi, C. S. (2015), 'Use of in vivo magnetic resonance spectroscopy for studying metabolic diseases', *Experimental and Molecular Medicine,* 47.

Ideker, T., et al. (2001), 'Integrated genomic and proteomic analyses of a systematically perturbed metabolic network', *Science,* 292 (5518), 929-34.

Ivanov, K., et al. (2013), 'Biotechnology in the Production of Pharmaceutical Industry Ingredients: Amino Acids', *Biotechnology & Biotechnological Equipment,* 27 (2), 3620-26.

Jiang, C., et al. (2018), 'Dynamic Human Environmental Exposome Revealed by Longitudinal Personal Monitoring', *Cell,* 175 (1), 277-+.

Johnson, C. H., Ivanisevic, J., and Siuzdak, G. (2016), 'Metabolomics: beyond biomarkers and towards mechanisms', *Nature Reviews Molecular Cell Biology,* 17 (7), 451-59.

Johnson, C. S. (1999), 'Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications', *Progress in Nuclear Magnetic Resonance Spectroscopy,* 34 (3-4), 203-56.

Joshi, M. D., Hedberg, A., and McIntosh, L. P. (1997), 'Complete measurement of the pKa values of the carboxyl and imidazole groups in Bacillus circulans xylanase', *Protein Sci,* 6 (12), 2667-70.

Judge, Michael T., et al. (2019), 'Continuous in vivo Metabolism by NMR', *Frontiers in Molecular Biosciences,* 6 (26).

Kanamori, K., et al. (1982), 'Effect of the nitrogen source on glutamine and alanine biosynthesis in Neurospora crassa. An in vivo 15N nuclear magnetic resonance study', *J Biol Chem,* 257 (23), 14168-72.

Kanehisa, M., et al. (2010), 'KEGG for representation and analysis of molecular networks involving diseases and drugs', *Nucleic Acids Research,* 38, D355-D60.

Kanehisa, M., et al. (2016), 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research,* 44 (D1), D457-D62.

Karp, P. D., Paley, S., and Romero, P. (2002), 'The Pathway Tools software', *Bioinformatics,* 18 Suppl 1, S225-32.

Khodayari, A. and Maranas, C. D. (2016), 'A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains', *Nat Commun,* 7, 13806.

Kim, J. D., et al. (2011), 'Use of 1H nuclear magnetic resonance to measure intracellular metabolite levels during growth and asexual sporulation in Neurospora crassa', *Eukaryot Cell,* 10 (6), 820-31.

Kim, S. M., et al. (2017), 'A review of parameters and heuristics for guiding metabolic pathfinding', *J Cheminform,* 9 (1), 51.

Kim, S. M., et al. (2020), 'Improving the organization and interactivity of metabolic pathfinding with precomputed pathways', *BMC Bioinformatics,* 21 (1), 13.

Klimovskaia, A., Ganscha, S., and Claassen, M. (2016), 'Sparse Regression Based Structure Learning of Stochastic Reaction Networks from Single Cell Snapshot Time Series', *PLoS Comput Biol,* 12 (12), e1005234.

Klukowski, P., et al. (2015), 'Computer vision-based automated peak picking applied to protein NMR spectra', *Bioinformatics,* 31 (18), 2981-88.

Klukowski, P., et al. (2018), 'NMRNet: a deep learning approach to automated peak picking of protein NMR spectra', *Bioinformatics,* 34 (15), 2590-97.

Koczula, K. M., et al. (2016), 'Metabolic plasticity in CLL: adaptation to the hypoxic niche', *Leukemia,* 30 (1), 65-73.

Krishnaiah, Saikumari Y., et al. (2017), 'Clock Regulation of Metabolites Reveals Coupling between Transcription and Metabolism', *Cell Metabolism,* 25 (4), 961-74.e4.

Krishnamurthy, K. (2013), 'CRAFT (complete reduction to amplitude frequency table) - robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR', *Magnetic Resonance in Chemistry,* 51 (12), 821-29.

Krumsiek, J., et al. (2011), 'Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data', *BMC Syst Biol,* 5, 21.

Kubicek, C. P., Punt, P., and Visser, J. (2010), 'Production of Organic Acids by Filamentous Fungi', *Mycota: Industrial Applications, Vol 10, Second Edition,* 10, 215-34.

Landau, David P. and Binder, Kurt (2014), *A Guide to Monte Carlo Simulations in Statistical Physics* (4 edn.; Cambridge: Cambridge University Press).

Li, B. and Dewey, C. N. (2011), 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *Bmc Bioinformatics,* 12.

Li, S., et al. (2013), 'Predicting network activity from high throughput metabolomics', *PLoS Comput Biol,* 9 (7), e1003123.

Liang, L., et al. (2020), 'Metabolic Dynamics and Prediction of Gestational Age and Time to Delivery in Pregnant Women', *Cell,* 181 (7), 1680-92 e15.

Liebeke, M., et al. (2013), 'Combining spectral ordering with peak fitting for one-dimensional NMR quantitative metabolomics', *Anal Chem,* 85 (9), 4605-12.

Link, H., Kochanowski, K., and Sauer, U. (2013), 'Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo', *Nature Biotechnology,* 31 (4), 357-+.

Link, H., Christodoulou, D., and Sauer, U. (2014), 'Advancing metabolic models with kinetic information', *Curr Opin Biotechnol,* 29, 8-14.

Link, H., et al. (2015), 'Real-time metabolome profiling of the metabolic switch between starvation and growth', *Nature Methods,* 12 (11), 1091-97.

Liu, Jun S. (2001), *Monte Carlo strategies in scientific computing* (Springer series in statistics; New York: Springer) xvi, 343 p.

Ludwig, C. and Gunther, U. L. (2011), 'MetaboLab--advanced NMR data processing and analysis for metabolomics', *BMC Bioinformatics,* 12, 366.

Maciejewski, M. W., et al. (2017), 'NMRbox: A Resource for Biomolecular NMR Computation', *Biophysical Journal,* 112 (8), 1529-34.

Macklin, D. N., et al. (2020), 'Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation', *Science,* 369 (6502), 390-+.

Magnuson, Jon K. and Lasure, Linda L. (2004), 'Organic Acid Production by Filamentous Fungi', in Jan S. Tkacz and Lene Lange (eds.), *Advances in Fungal Biotechnology for Industry, Agriculture, and Medicine* (Boston, MA: Springer US), 307-40.

Manganas, L. N., et al. (2007), 'Magnetic resonance spectroscopy identifies neural progenitor cells in the live human brain', *Science,* 318 (5852), 980-85.

Maricq, M Matti and Waugh, JS (1979), 'NMR in rotating solids', *The Journal of Chemical Physics,* 70 (7), 3300-16.

Markham, P., et al. (1993), 'Choline: its role in the growth of filamentous fungi and the regulation of mycelial morphology', *FEMS Microbiol Rev,* 10 (3-4), 287-300.

Matviychuk, Y., von Harbou, E., and Holland, D. J. (2017), 'An experimental validation of a Bayesian model for quantification in NMR spectroscopy', *Journal of Magnetic Resonance,* 285, 86-100.

Maughon, T. S., et al. (2022), 'Metabolomics and cytokine profiling of mesenchymal stromal cells identify markers predictive of T-cell suppression', *Cytotherapy,* 24 (2), 137-48.

McCluskey, K., Wiest, A., and Plamann, M. (2010), 'The Fungal Genetics Stock Center: a repository for 50 years of fungal genetics research', *J Biosci,* 35 (1), 119-26.

McNerney, M. P., Watstein, D. M., and Styczynski, M. P. (2015), 'Precision metabolic engineering: The design of responsive, selective, and controllable metabolic systems', *Metabolic Engineering,* 31, 123-31.

Mehrmohamadi, M., et al. (2014), 'Characterization of the Usage of the Serine Metabolic Network in Human Cancer', *Cell Reports,* 9 (4), 1507-19.

Meiboom, S. and Gill, D. (1958), 'Modified Spin-Echo Method for Measuring Nuclear Relaxation Times', *Review of Scientific Instruments,* 29 (8), 688-91.

Mercer, R., et al. (2021), 'Matrix Profile XXIII: Contrast Profile: A Novel Time Series Primitive that Allows Real World Classification', *2021 21st Ieee International Conference on Data Mining (Icdm 2021)*, 1240-45.

Meyer, P., et al. (2014), 'Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach', *Bmc Systems Biology,* 8.

Milewski, S., Gabriel, I., and Olchowy, J. (2006), 'Enzymes of UDP-GlcNAc biosynthesis in yeast', *Yeast,* 23 (1), 1-14.

Mobarhan, Y. L., et al. (2017a), 'Effective combined water and sideband suppression for low-speed tissue and in vivo MAS NMR', *Analytical and Bioanalytical Chemistry,* 409 (21), 5043-55.

Mobarhan, Y. L., et al. (2017b), 'Effective combined water and sideband suppression for low-speed tissue and in vivo MAS NMR', *Anal Bioanal Chem,* 409 (21), 5043-55.

Mobarhan, Y. L., et al. (2016), 'Comprehensive multiphase NMR applied to a living organism', *Chemical Science,* 7 (8), 4856-66.

Monge, M. E., et al. (2019), 'Challenges in Identifying the Dark Molecules of Life', *Annual Review of Analytical Chemistry, Vol 12,* 12, 177-99.

Montana, G., Berk, M., and Ebbels, T. (2011a), 'Modelling short time series in metabolomics: a functional data analysis approach', *Adv Exp Med Biol,* 696, 307-15.

Montana, Giovanni, Berk, Maurice, and Ebbels, Tim (2011b), 'Modelling Short Time Series in Metabolomics: A Functional Data Analysis Approach', in

Hamid R. Arabnia and Quoc-Nam Tran (eds.), *Software Tools and Algorithms for Biological Systems* (New York, NY: Springer New York), 307-15.

Moon, K. R., et al. (2020), 'Visualizing structure and transitions in high-dimensional biological data (vol 54, pg 781, 2019)', *Nature Biotechnology,* 38 (1), 108-08.

Morris, J. H., et al. (2011), 'clusterMaker: a multi-algorithm clustering plugin for Cytoscape', *BMC Bioinformatics,* 12, 436.

Muzny, D. M., et al. (2012), 'Comprehensive molecular characterization of human colon and rectal cancer', *Nature,* 487 (7407), 330-37.

Nelson, W., et al. (2019), 'To Embed or Not: Network Embedding as a Paradigm in Computational Biology', *Frontiers in Genetics,* 10.

Newman, M. E. J. and Girvan, M. (2004), 'Finding and evaluating community structure in networks', *Physical Review E,* 69 (2).

Ono, K., et al. (2015), 'CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API', *F1000Res,* 4, 478.

Park, J. O., et al. (2019), 'Near-equilibrium glycolysis supports metabolic homeostasis and energy yield', *Nat Chem Biol,* 15 (10), 1001-08.

Patil, R. S., Ghormade, V. V., and Deshpande, M. V. (2000), 'Chitinolytic enzymes: an exploration', *Enzyme Microb Technol,* 26 (7), 473-83.

Patti, G. J., Yanes, O., and Siuzdak, G. (2012), 'Metabolomics: the apogee of the omics trilogy', *Nature Reviews Molecular Cell Biology,* 13 (4), 263-69.

Paull, E. O., et al. (2013), 'Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)', *Bioinformatics,* 29 (21), 2757-64.

Peng, X. X., et al. (2018), 'Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers', *Cell Reports,* 23 (1), 255-+.

Pfister, N., Bauer, S., and Peters, J. (2019a), 'Learning stable and predictive structures in kinetic systems', *Proc Natl Acad Sci U S A,* 116 (51), 25405-11.

Pfister, N., Buehlmann, P., and Peters, J. (2019b), 'Invariant Causal Prediction for Sequential Data', *Journal of the American Statistical Association,* 114 (527), 1264-76.

Psychogios, N., et al. (2011), 'The Human Serum Metabolome', *Plos One,* 6 (2).

Radford, A. (2004), 'Metabolic highways of Neurospora crassa revisited', *Adv Genet,* 52, 165-207.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional data analysis* (2nd edn., Springer series in statistics; New York: Springer) xix, 426 p.

Ramsay, J. O., Hooker, Giles, and Graves, Spencer (2009), *Functional data analysis with R and MATLAB* (Use R!; Dordrecht ; New York: Springer) xi, 207 p.

Raue, A., et al. (2013), 'Lessons learned from quantitative dynamical modeling in systems biology', *PLoS One,* 8 (9), e74335.

Ravanbakhsh, S., et al. (2015), 'Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics (vol 10, e0124219, 2015)', *Plos One,* 10 (7).

Rhoades, Seth D., Sengupta, Arjun, and Weljie, Aalim M. (2017), 'Time is ripe: maturation of metabolomics in chronobiology', *Current Opinion in Biotechnology,* 43, 70-76.

Righi, V., et al. (2014), 'In vivo high-resolution magic angle spinning proton NMR spectroscopy of Drosophila melanogaster flies as a model system to investigate mitochondrial dysfunction in Drosophila GST2 mutants', *Int J Mol Med,* 34 (1), 327-33.

Robinette, S. L., et al. (2008), 'Web server based complex mixture analysis by NMR', *Anal Chem,* 80 (10), 3606-11.

Rose, S. M. S. F., et al. (2019), 'A longitudinal big data approach for precision health', *Nature Medicine,* 25 (5), 792-+.

Rubanova, Yulia, Chen, Ricky T. Q., and Duvenaud, David (2019), 'Latent ODEs for Irregularly-Sampled Time Series', *arXiv e-prints*. <https://ui.adsabs.harvard.edu/abs/2019arXiv190703907R>, accessed July 01, 2019.

Rubtsov, D. V. and Griffin, J. L. (2007), 'Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy', *Journal of Magnetic Resonance,* 188 (2), 367-79.

Rubtsov, Denis V., et al. (2010), 'Application of a Bayesian Deconvolution Approach for High-Resolution 1H NMR Spectra to Assessing the

Metabolic Effects of Acute Phenobarbital Exposure in Liver Tissue', *Analytical Chemistry,* 82 (11), 4479-85.

Rude, Thomas H., et al. (2002), 'Relationship of the Glyoxylate Pathway to the Pathogenesis of Cryptococcus neoformans', *Infection and Immunity,* 70 (10), 5684.

Sailani, M. R., et al. (2020), 'Deep longitudinal multiomics profiling reveals two biological seasonal patterns in California', *Nature Communications,* 11 (1).

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016), 'Probabilistic programming in Python using PyMC3', *Peerj Computer Science*.

Sarou-Kanian, V., et al. (2015), 'Metabolite localization in living drosophila using High Resolution Magic Angle Spinning NMR', *Sci Rep,* 5, 9872.

Sefer, E., Kleyman, M., and Bar-Joseph, Z. (2016), 'Tradeoffs between Dense and Replicate Sampling Strategies for High-Throughput Time Series Experiments', *Cell Systems,* 3 (1), 35-42.

Sekar, Karthik, et al. (2018), 'Synthesis and degradation of FtsZ quantitatively predict the first cell division in starved bacteria', *Molecular Systems Biology,* 14 (11), e8623.

Sekihara, K. and Ohyama, N. (1990), 'Parameter estimation for in vivo magnetic resonance spectroscopy (MRS) using simulated annealing', *Magn Reson Med,* 13 (2), 332-9.

Sengupta, Arjun, et al. (2016), 'Deciphering the Duality of Clock and Growth Metabolism in a Cell Autonomous System Using NMR Profiling of the Secretome', *Metabolites,* 6 (3), 23.

Shannon, P., et al. (2003), 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res,* 13 (11), 2498-504.

Shaver, A. O., et al. (2021), 'Culture and Assay of Large-Scale Mixed-Stage Caenorhabditis elegans Populations', *Jove-Journal of Visualized Experiments,* (171).

Show, P. L., et al. (2015), 'Overview of citric acid production from Aspergillus niger', *Frontiers in Life Science,* 8 (3), 271-83.

Sitnikov, Dmitri G., Monnin, Cian S., and Vuckovic, Dajana (2016), 'Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS', *Scientific Reports,* 6, 38885.

Slayman, C and Potapova, T (2006), 'Origin and significance of vacuolar proliferation during nutrient restriction', *Neurospora 2006 Poster Abstracts*.

Slayman, C. L. (1965), 'Electrical properties of Neurospora crassa. Respiration and the intracellular potential', *J Gen Physiol,* 49 (1), 93-116.

Slayman, C. L. and Slayman, C. W. (1968), 'Net uptake of potassium in Neurospora. Exchange for sodium and hydrogen ions', *J Gen Physiol,* 52 (3), 424-43.

Slayman, CL, Long, WS, and Lu, CY-H (1973), 'The relationship between ATP and an electrogenic pump in the plasma membrane ofNeurospora crassa', *The Journal of membrane biology,* 14 (1), 305-38.

Slayman, Clifford L, Moussatos, Vasiliana V, and Webb, WATT W (1994), 'Endosomal accumulation of pH indicator dyes delivered as acetoxymethyl esters', *Journal of Experimental Biology,* 196 (1), 419-38.

Son, J., et al. (2013), 'Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway', *Nature,* 496 (7443), 101-+.

Sousa, S. A. A., Magalhaes, A., and Ferreira, M. M. C. (2013), 'Optimized bucketing for NMR spectra: Three case studies', *Chemometrics and Intelligent Laboratory Systems,* 122, 93-102.

Steger, C. (1998), 'An unbiased detector of curvilinear structures', *Ieee Transactions on Pattern Analysis and Machine Intelligence,* 20 (2), 113-25.

Stoker, J. J. (1969), *Differential geometry* (New York: Wiley-Interscience).

Subramanian, A., et al. (2005), 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America,* 102 (43), 15545-50.

Suk, Minsoo and Bhandarkar, S. M. (1992), *Three-dimensional object recognition from range images* (Computer science workbench; Tokyo ; New York: Springer-Verlag) xxi, 308 p.

Sumner, L. W., et al. (2014), 'Proposed quantitative and alphanumeric metabolite identification metrics', *Metabolomics,* 10 (6), 1047-49.

Szakacs, Z., Hagele, G., and Tyka, R. (2004), 'H-1/P-31 NMR pH indicator series to eliminate the glass electrode in NMR spectroscopic pK(a) determinations', *Analytica Chimica Acta,* 522 (2), 247-58.

Tabatabaei Anaraki, M., Simpson, M. J., and Simpson, A. J. (2018), 'Reducing impacts of organism variability in metabolomics via time trajectory in vivo NMR', *Magn Reson Chem,* 56 (11), 1117-23.

Takis, Panteleimon G., et al. (2017), 'Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool', *Nature Communications,* 8 (1), 1662.

Tang, X., et al. (2011), 'Systems biology of the qa gene cluster in Neurospora crassa', *PLoS One,* 6 (6), e20671.

Thain, D., Tannenbaum, T., and Livny, M. (2005), 'Distributed computing in practice: the Condor experience', *Concurrency and Computation-Practice & Experience,* 17 (2-4), 323-56.

Tinevez, J. Y., et al. (2017), 'TrackMate: An open and extensible platform for single-particle tracking', *Methods,* 115, 80-90.

Tredwell, G. D., et al. (2016a), 'Modelling the acid/base 1H NMR chemical shift limits of metabolites in human urine', *Metabolomics,* 12 (10), 152.

Tredwell, G. D., et al. (2016b), 'Modelling the acid/base (1)H NMR chemical shift limits of metabolites in human urine', *Metabolomics,* 12 (10), 152.

Tredwell, G. D., et al. (2011), 'Between-Person Comparison of Metabolite Fitting for NMR-Based Quantitative Metabolomics', *Analytical Chemistry,* 83 (22), 8683-87.

Ulrich, Eldon L., et al. (2008), 'BioMagResBank', *Nucleic Acids Research,* 36 (suppl_1), D402-D08.

van Dongen, S. and Abreu-Goodger, C. (2012), 'Using MCL to Extract Clusters from Networks', *Bacterial Molecular Networks: Methods and Protocols,* 804, 281-95.

Van Geet, Anthony L (1970), 'Calibration of methanol nuclear magnetic resonance thermometer at low temperature', *Analytical Chemistry,* 42 (6), 679-80.

Vaske, C. J., et al. (2010), 'Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM', *Bioinformatics,* 26 (12), i237-i45.

Virgilio, S., et al. (2017), 'Regulation of the reserve carbohydrate metabolism by alkaline pH and calcium in Neurospora crassa reveals a possible cross-regulation of both signaling pathways', *Bmc Genomics,* 18.

Voet, Donald and Voet, Judith G. (2011), *Biochemistry* (4th edn.; Hoboken, NJ: John Wiley & Sons) xxv, 1428, 53 p.

Vrabl, P., et al. (2012), 'Organic Acid Excretion in Penicillium ochrochloron Increases with Ambient pH', *Front Microbiol,* 3, 121.

Vu, T. N. and Laukens, K. (2013a), 'Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data', *Metabolites,* 3 (2), 259-76.

Vu, T. N. and Laukens, K. (2013b), 'Getting your peaks in line: a review of alignment methods for NMR spectral data', *Metabolites,* 3 (2), 259-76.

Walejko, J. M., et al. (2018a), 'Global Metabolomics of the Placenta Reveals Distinct Metabolic Profiles between Maternal and Fetal Placental Tissues Following Delivery in Non-Labored Women', *Metabolites,* 8 (1).

Walejko, Jacquelyn M, et al. (2018b), 'Global Metabolomics of the Placenta Reveals Distinct Metabolic Profiles between Maternal and Fetal Placental Tissues Following Delivery in Non-Labored Women', *Metabolites,* 8 (1), 10.

Wang, B., et al. (2017), 'Identification and characterization of the glucose dual-affinity transport system in Neurospora crassa: pleiotropic roles in nutrient transport, signaling, and carbon catabolite repression', *Biotechnology for Biofuels,* 10.

Warburg, O. (1956), 'On the origin of cancer cells', *Science,* 123 (3191), 309-14.

Wayne, L. G. and Lin, K. Y. (1982), 'Glyoxylate metabolism and adaptation of Mycobacterium tuberculosis to survival under anaerobic conditions', *Infection and Immunity,* 37 (3), 1042.

Whiley, L., et al. (2019), 'Systematic Isolation and Structure Elucidation of Urinary Metabolites Optimized for the Analytical-Scale Molecular Profiling Laboratory', *Analytical Chemistry,* 91 (14), 8873-82.

Wishart, David S, et al. (2007), 'HMDB: the human metabolome database', *Nucleic acids research,* 35 (suppl_1), D521-D26.

Wolfinbarger, L. and Kay, W. W. (1973), 'Transport of C4-dicarboxylic acids in Neurospora crassa', *Biochimica et Biophysica Acta (BBA) - Biomembranes,* 307 (1), 243-57.

Wu, Y., et al. (2020), 'RTExtract: time-series NMR spectra quantification based on 3D surface ridge tracking', *Bioinformatics,* 36 (20), 5068-75.

Wu, Yue, et al. (2022), 'Uncovering in vivo biochemical patterns from time-series metabolic dynamics', *PLOS ONE,* 17 (5), e0268394.

Yandell, M. and Ence, D. (2012), 'A beginner's guide to eukaryotic genome annotation', *Nature Reviews Genetics,* 13 (5), 329-42.

Ye, L., De Iorio, M., and Ebbels, T. M. D. (2018), 'Bayesian estimation of the number of protonation sites for urinary metabolites from NMR spectroscopic data', *Metabolomics,* 14 (5), 56.

Yu, B., et al. (2019), 'The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies', *American Journal of Epidemiology,* 188 (6), 991-1012.

Yu, Y., et al. (2007), 'A genetic network for the clock of Neurospora crassa', *Proc Natl Acad Sci U S A,* 104 (8), 2809-14.

Zachariah, C., et al. (2001), 'Structural studies of a neuropeptide precursor protein with an RGD proteolytic site', *Biochemistry,* 40 (30), 8790-9.

Zheng, C., et al. (2011), 'Identification and quantification of metabolites in H-1 NMR spectra by Bayesian model selection', *Bioinformatics,* 27 (12), 1637-44.

APPENDICES

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Supplementary File 2.1

Tutorial for CIVM-NMR Ridge Tracing and Quantification

This is the tutorial for the ridge tracing process described in the initial CIVM manuscript. It covers the processes of tracing and quantifying peaks for time series NMR spectra. Before running any code, you need to:

2.1.1.1 Clone the Edison Lab GitHub repository locally. Corresponding functions and workflows can be found at the Edison Lab metabolomics toolbox repository: (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA). Specifically, the pieces of code central to this manuscript are in the folder: metabolomics_toolbox/code/HR_MAS

In this folder, files with names **STEP*.m** are workflows and other files are functions. This tutorial is mainly for **STEP_2_ridge_tracing.m**, **STEP_3_combining_ridges.m**, and **STEP_4_plotting.m**, and therefore assumes that the sampleData structure has been created using **STEP_1** (only necessary for several samples). Add folders under the GitHub repository to your MATLAB path before running any scripts.

2.1.1.2 Download the necessary supplement data (available at www.metabolomicsworkbench.org). Make sure folders under ‘analysis’ are in the MATLAB path and ensure MATLAB is running from the ‘analysis’ folder. The data

have been preprocessed in NMRPipe and the MATLAB script (**STEP_1_processing_combine_samples.m**). Processing steps include:

- line broadening
- Fast Fourier transform
- phasing
- end removal
- baseline correction
- solvent region removal
- normalization

as stated in the **Methods** part of the paper. Consecutive spectra (every three) have also been summed (collapsed) to improve SNR (signal to noise ratio) and the following working process is quite same for the full resolution dataset. Besides automation of quantification, we also have scripts for automatically constructing project folder (constructHRMASDirectory.m and constructHRMASDirectory_2.m).

Now you can start running the workflow.

2.1.2. Ridge tracing

This part is for the script **STEP_2_ridge_tracing.m**. Go to the script for more technical information if needed.

2.1.2.1. Load data

The data that will be worked on is a struct array containing processed NMR spectra of multiple experiments produced from the workflow (**STEP_1_processing_combine_samples.m**). Specifically, the important part

that will be used in the workflow is ppm vector (ppmR_1h1d), collapsed intensity matrix (Xcollapsed_1h1d), and collapsed time vector (timesCollapsed_1h1d). The Xcollapsed_1h1d is a matrix containing time series NMR measurement with each row for different time points and each column for different ppm.

2.1.2.2. Smoothing and Peak picking

This step will peak-pick points which can be connected in next step. Based on the specific peak that the user needs to quantify, smoothing and peak density need to be controlled by parameters in this section.

Start from this step, the region of interest (ROI) needs to be specified. The example region is [2.5 2.8] ppm, and regions with similar size should be fine in computational time. Try not to use too large or too small of a region. The former will be slow, and the latter might cause problems with peak picking.

For a selected region, the time series NMR data can be visualized as a 3D surface. On the surface, a 2D Gaussian filter (‘imgaussfilt’ in MATLAB) is used to smooth the surface. This process can be controlled by setting different sigmas for the dimensions of time and ppm. The value of the sigmas can be increased for noisy regions and decreased if the peaks are close to each other and greater resolution is needed. peakPickThreshold is the threshold for peak picking and it can be decreased to pick more peaks or increased to limit noise (some noise is okay). These parameters are fed to the function ridgeTracing_PeakPick1D for this step. In the example, we are dealing with peaks that are quite intense, so these numbers might need to be modified for smaller peaks. An example figure

(Fig. 2.1.1) produced from this step is shown here with red dots indicating the picked peaks for citrate:

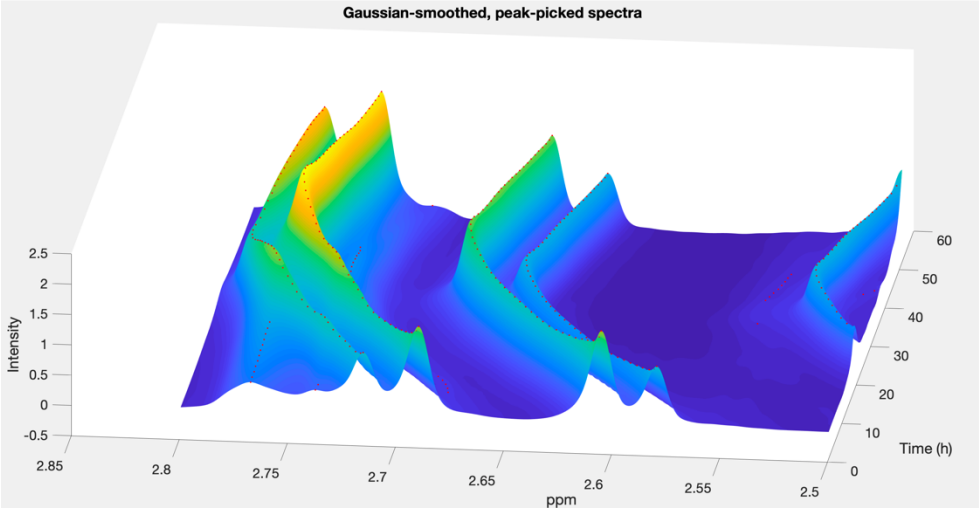


Figure 2.1.1: the peak picking result for region [2.5,2.8] with peakPickThreshold 0, sigma in gaussian filter 1 and 10.

Even though step 2.1.2.2 and step 2.1.2.3 are separated, users will often need to modify parameters on both parts to have a satisfactory result in ridge tracing.

2.1.2.3. Clustering and manual picking

This step will cluster the peaks found in last step using single linkage hierarchical agglomerative clustering. The clustering is based on distances in the 3D space of time, intensity, and ppm. For clustering, number of clusters (numberOfRidges), weight in each direction (timeWeight, ppmWeight, and intensityWeight) need to be adjusted for different conditions. For the selected region here, as the peak is shifting, a comparably larger ppmWeight is needed. If

you are working on peak that are not moving a lot and comparably crowded, then a smaller ppmWeight is needed. In general, for any dimension, increasing the weight encourages clusters to spread in that corresponding direction. Finally, the number of clusters should typically be set higher than the desired number of peaks to be traced, to allow noise to cluster. The resulting clusters are displayed as ‘ridges’ on a 3D surface plot. You will often notice noise clustering with noise.

Therefore, manual picking is also needed to select ridges with good quality. By running the function ridgeTracing_clusterPeaks_interactive_2, the peaks are clustered to different ridges with different colors. There will be a menu, from which the user should click Pick Final Clusters, after which the mouse is used to click the ridges to be kept, and then the return key is pressed. The surface plot with high-quality, traced ridges is then displayed (Fig. 2.1.2). A struct array newRidges will also be returned which contains information for generating the selected ridges and parameters.

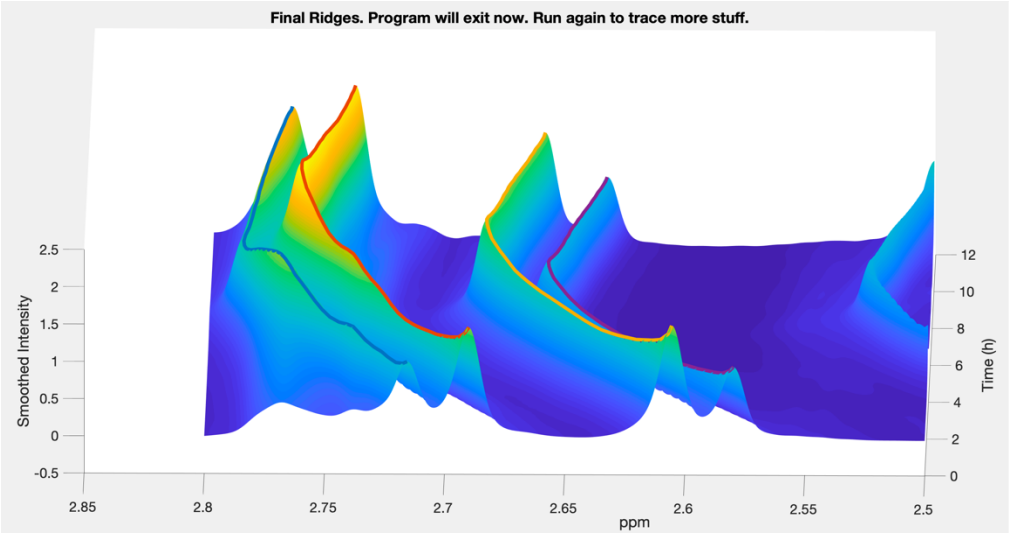


Figure 2.1.2: ridges after manual picking with numberOfRidges 13, timeWeight 1, ppmWeight 2, and intensityWeight 500.

2.1.2.4. Ridge correction

This step will refine the ridge tracing result. It will map ridges to the original (non- smoothed) matrix without smoothing at the beginning, use linear interpolation to fill gaps, and extend the ridge ends to fill the time series (in case this is required; it typically is not used). A struct array adjustedRidges will be returned containing the refined data. Figures presenting the refined result will also be shown for each ridge. Two parameters, windowWidth and viewWidth need adjustment for this step. Mapping from smoothed data to raw data is done by applying the ridge positions from the former to the data in the latter. Since the true position of a ridge is usually slightly different between the two, the maximum within the region defined by windowWidth data points to the left and right of the smoothed ridge position is taken, and its position and value become the adjusted ridge. viewWidth simply defines the width of the region plotted for inspection.

2.1.3. Combing ridge tracing information

From the data of step 2.1.3, we can summarize multiple peaks and quantify for each compound. This part is for the script STEP_3_combining_ridges.m. Go to the script for more technical information if needed.

2.1.3.1. Adjust time

There is minimal difference in time before the NMR spectral start recording and so for each dataset a time shift is added. Technical details can be found in the script.

2.1.3.2. Map ridge to compound and quantification

Quantified ridges can be annotated to specific compound by a direct mapping list. Mapping process depends on the compound annotation on NMR spectra based on extracted sample and chemical annotation experience is needed.

When multiple ridges are annotated to one compound, they are scaled such that the mean of the ridges are the same, then averaged by timepoint. This method of combining ridges facilitates comparison of trends but is not appropriate for absolute quantification.

2.1.4. Plotting

Most figures in the manuscript can be produced by the script STEP_4_plotting.m. The figure types and names are detailed indicated in the script. From the script, you can plot the compound relative concentration through time (Fig. 2.1.3). Time trajectories for single ridges and baseline can also be shown. Aside from functions mentioned in the script, you can also use the stackSpectra function to plot time series spectra as in Fig 2.1.3 in the main text.

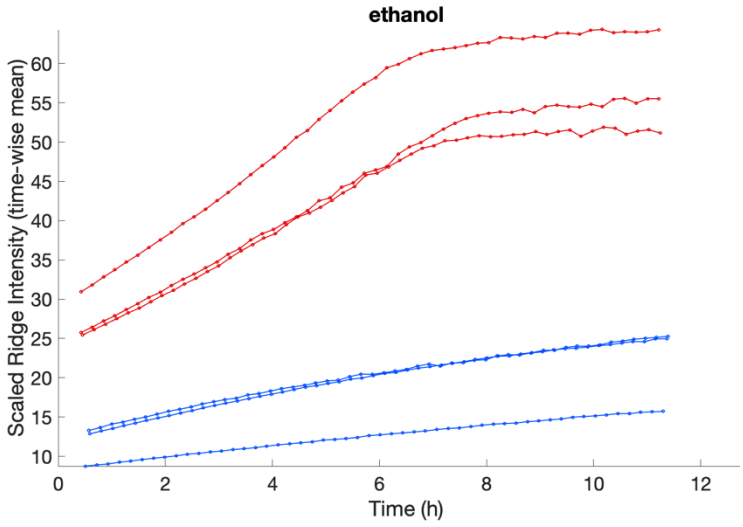
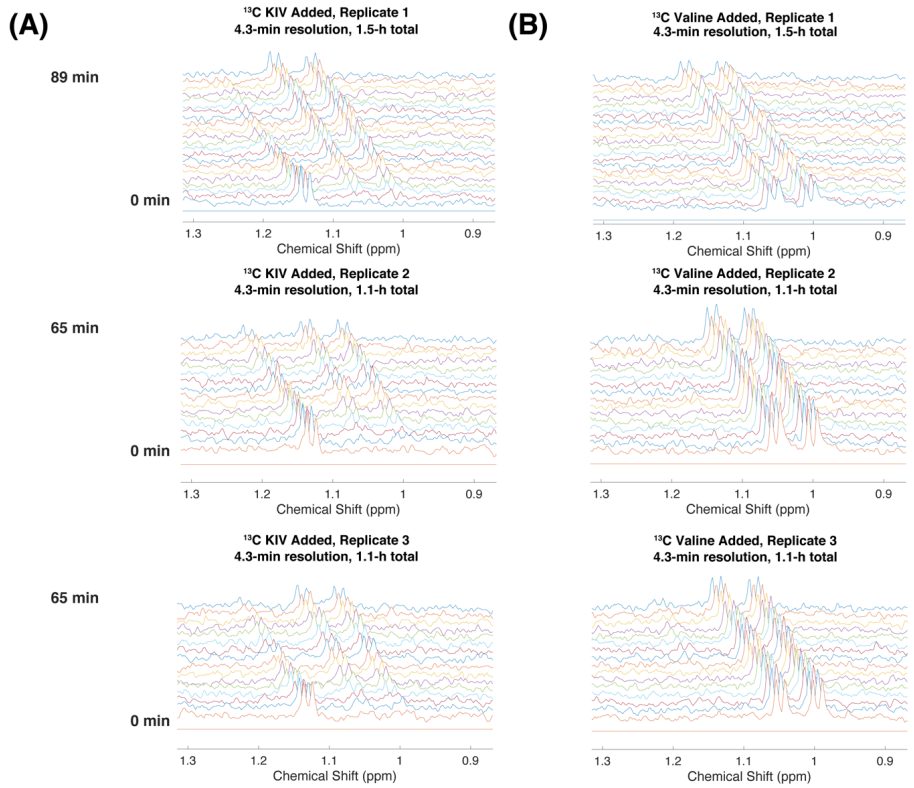
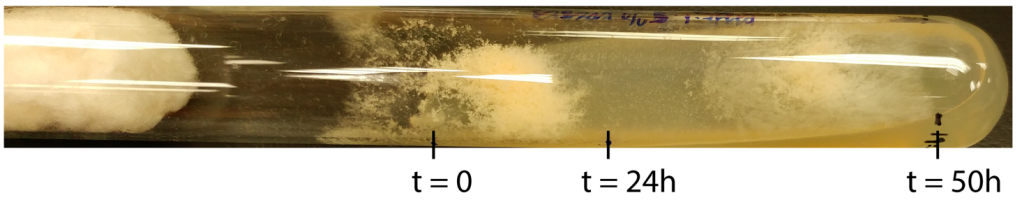


Figure 2.1.3: Change of Scaled intensity through time for replicates under different conditions. In this figure two conditions, aerobic and anaerobic are shown and for each case, there are three replicates.

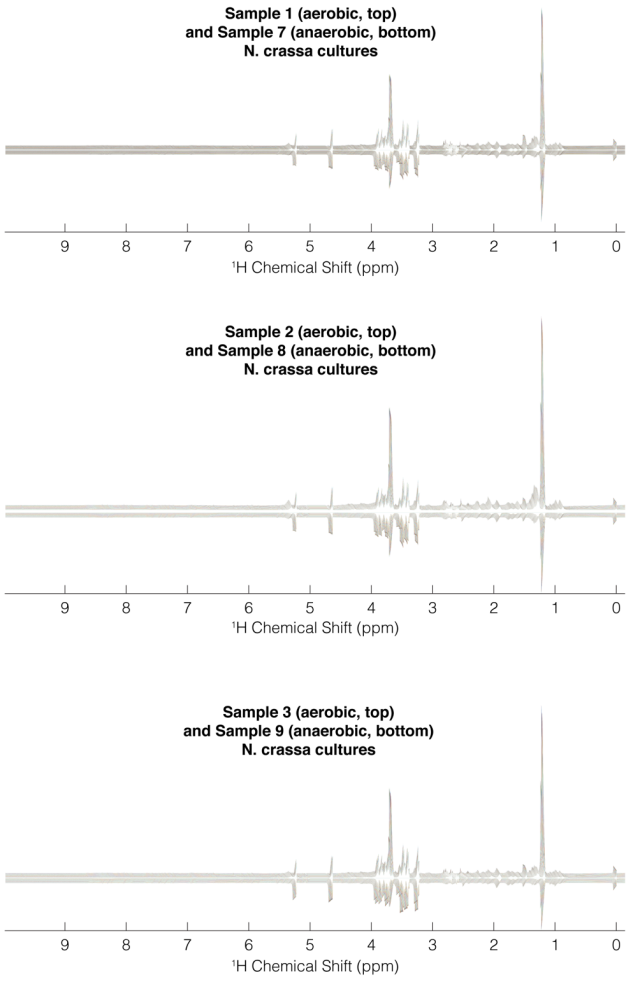
The pipelines and functions for this manuscript are extensively annotated, and if you have more technical questions, please feel free to contact M. Judge or Y. Wu. Additionally, we are working on a more efficient and extensive method for the pipeline which will be described in a later publication, and regularly update the Edison Lab Public Toolbox with new functions and workflows.



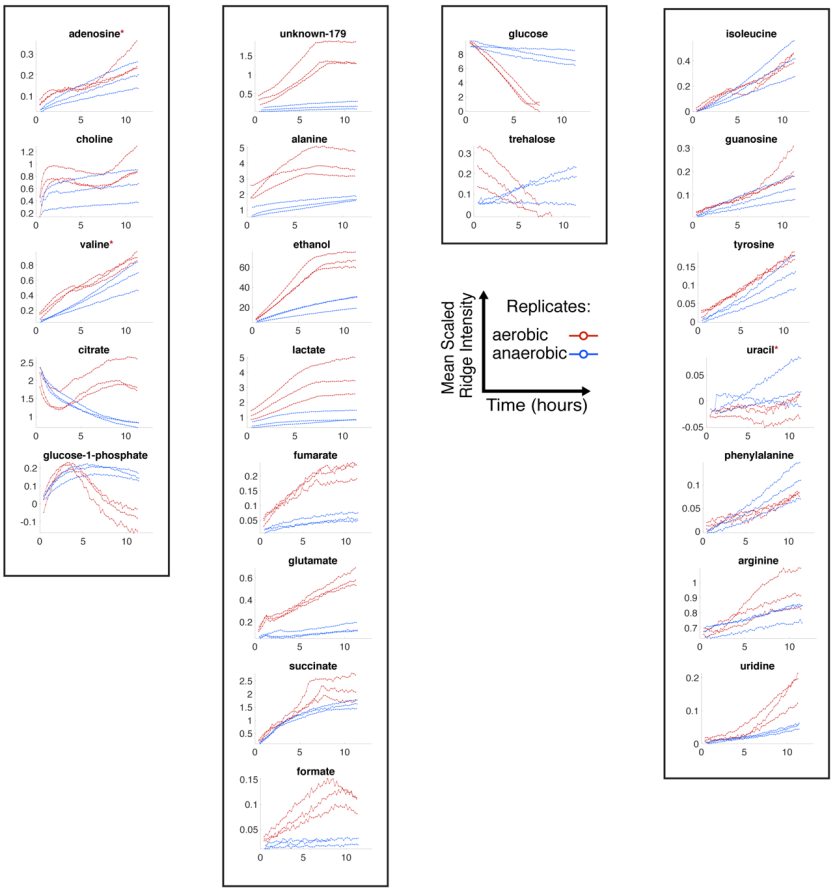
Supplementary Figure 2.1. Targeted isotopic CIVM-NMR measurements of metabolic flux in human myeloid leukemia cells are reproducible. (A) 1D ^{13}C-edited hsqc experiments were used to observe protons covalently attached to ^{13}C derived from uniformly labeled KIV in three independent replicates. ^{13}C valine accumulates as KIV is consumed. (B) The same experiments were used to observe protons covalently attached to ^{13}C derived from uniformly labeled valine in three independent replicates. ^{13}C derived from uniformly labeled valine does not accumulate to appreciable levels in KIV.



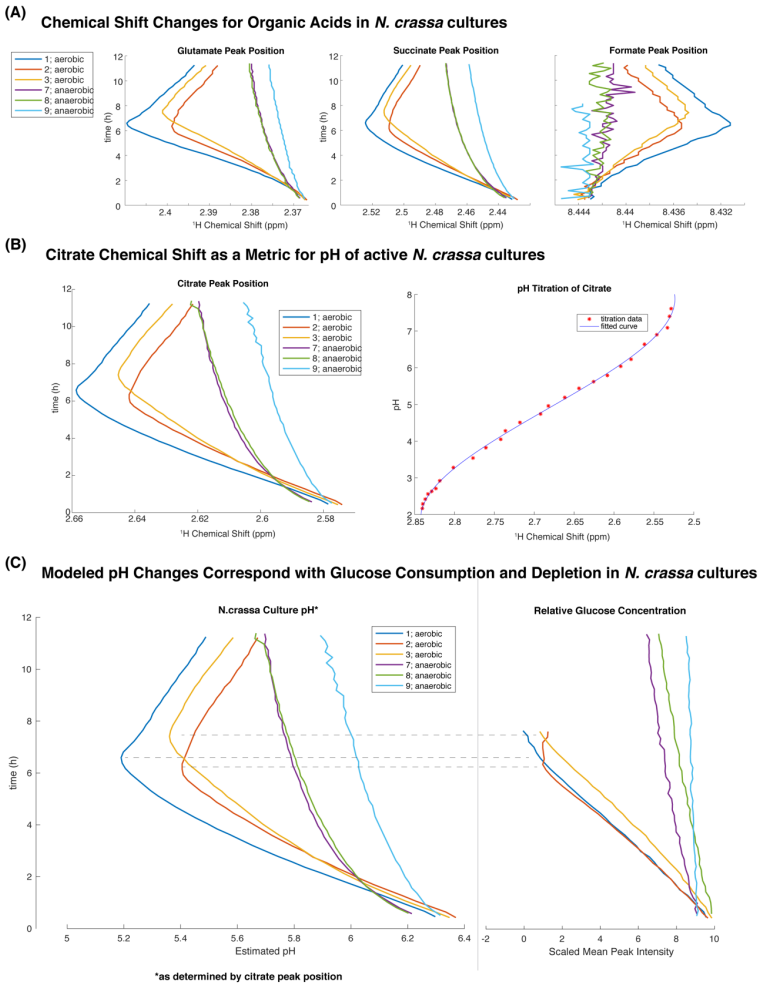
Supplementary Figure 2.2. Growth of N. crassa after a CIVM-NMR experiment. A piece of mycelium was used to inoculate a growth slant at t = 0. The culture was kept on the bench and the advancement of the growth front was marked at t = 0 h, 24 h and 50 h. Roughly circadian conidiation was observed between 0 h and 24 h, and again before the 50-h mark.



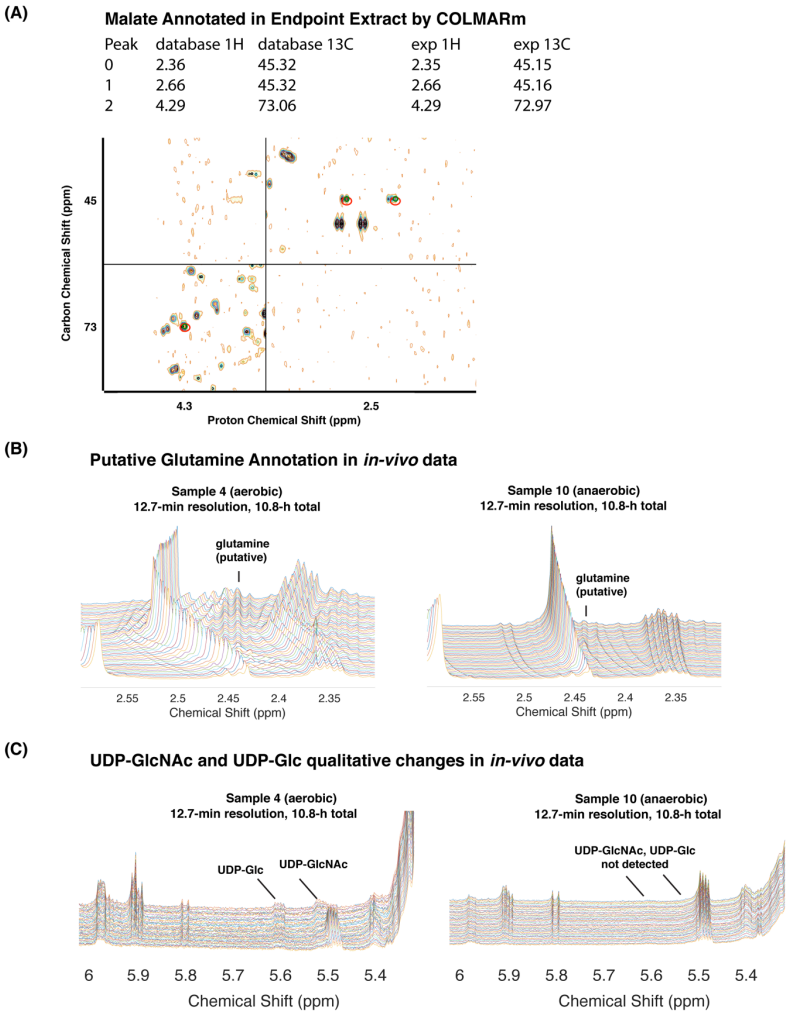
Supplementary Figure 2.3. Comparison of aerobic and anaerobic samples for three independent replicates. Replicates 1,2,3 (aerobic) are plotted against mirror images of replicates 7,8,9 (anaerobic).



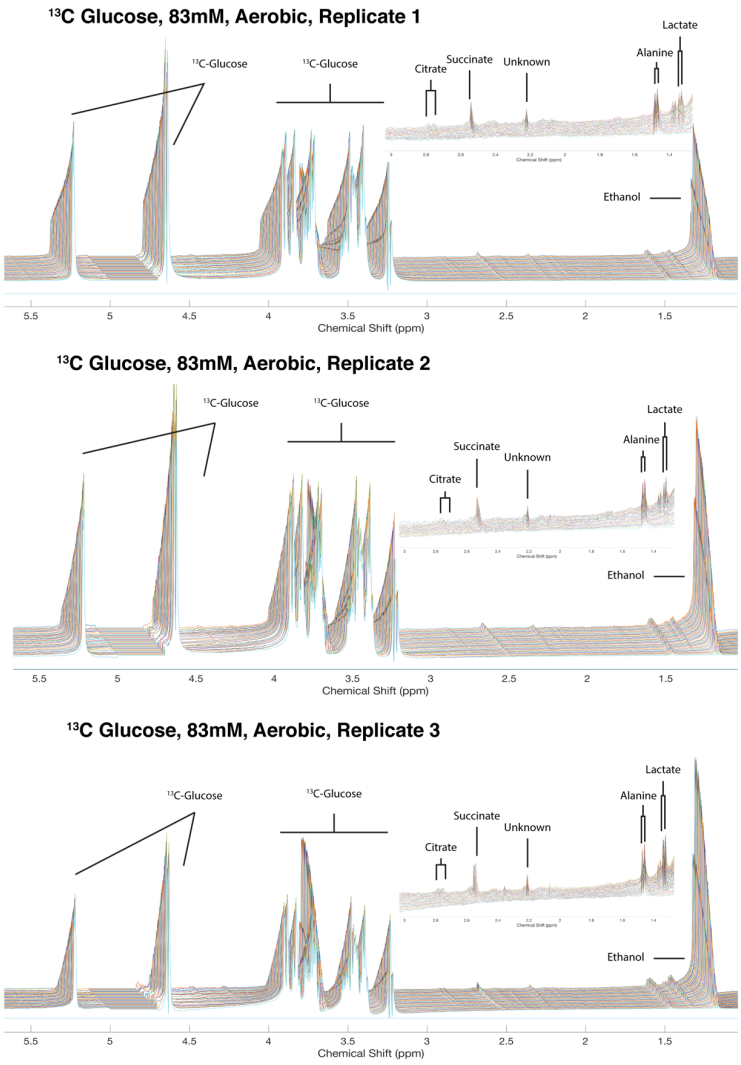
Supplementary Figure 2.4. Aerobic and anaerobic trajectories for each metabolite that was both annotated and quantified in this study. One example of an un-annotated ridge is also shown. Metabolites are grouped with those having similar profiles in one or both conditions. Red asterisks indicate compounds whose absolute peak intensities were affected by changes in baseline; these were therefore excluded from biological interpretation.



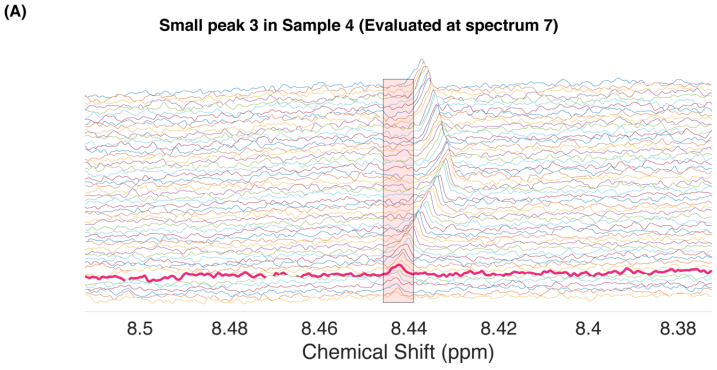
Supplementary Figure 2.5. Organic acid peak positions reflect glucose-dependent changes in pH over time. (A) Representative in-vivo peak position changes for glutamate, succinate, and formate for all samples. The formate peak position in the anaerobic samples is noisy because the peak was very low intensity. (B) Position of a representative citrate peak over time and in-house pH titration data for citrate. A 3rd-order polynomial was fit to the titration data in order to interpolate pH values at each timepoint for each sample (C). Dashed horizontal lines are used to show that glucose depletion coincided with reversal of acidification around 6-7 h in the aerobic samples.

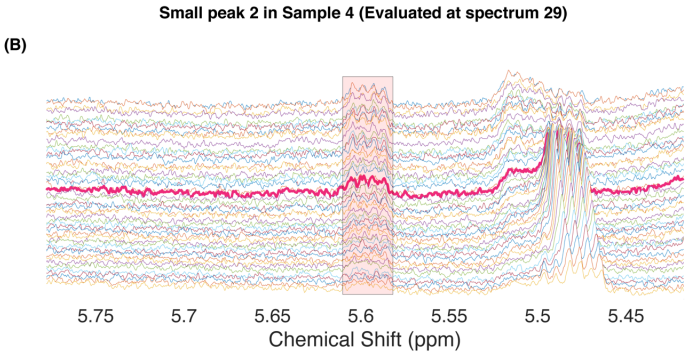


Supplementary Figure 2.6. Qualitative assessment for unquantifiable metabolites in this study. (A) Malate was annotated in the 2D data for endpoint extracts by peak matching to databases using COLMARm1. Positions of matched peaks are indicated with a red ellipse and in the peak table. (B) Peaks consistent with glutamine increased in the aerobic and may be present in the anaerobic condition. (C) UDP-GlcNAc (UDP-N-Acetyl Glucosamine; UDP-NAG) and UDP-Glc (UDPGlucose) were annotated in the aerobic condition but could not be traced reliably.



Supplementary Figure 2.7. Accumulation of 13C-labeled metabolites in three independent replicate aerobic N. crassa cultures. hmqc1d NMR experiments were used to monitor the accumulation of ^{13}Clabeled metabolites after addition of uniformly labeled ^{13}C glucose. Glucose was converted to ethanol, alanine, succinate, and lactate, citrate, and some peaks that have yet to be identified.





(C)

	Small Peak	Annotation	Integral (Peak)	Integral (DSS)	[1H] (mM)*
Individual Spectra	1	unknown	00.212	0047.526	0.036
	2	UDP-NAG (whole multiplet)	01.474	0048.040	0.246
	2	UDP-NAG (quarter multiplet)	-	-	0.062
	3	formate	00.145	0048.350	0.024
Summed Spectra	1	unknown	20.910	2466.623	0.068
	2	UDP-NAG	69.245	2466.623	0.225

* [1H] (mM) = (Peak Integral) / (DSS Peak Integral) * (0.89 mM * 9 DSS protons)

Supplementary Figure 2.8. Estimation of sensitivity for CIVM-NMR using our HR-MAS probe in an aerobic sample. (A) Formate peak in spectrum 7 (bolded in magenta, t = 84.6 min), where it became discernable from noise. (B) Similar result for a UDP-N-acetyl glucosamine peak in spectrum 29 (t = 363.8 min). (C) Table of measured intensities and calculated concentrations of 1H based on the known concentration of DSS in the sample. Integrals were assessed using the boundaries shown by the pink box. Notably, formate and an unknown peak) of similar intensity (not shown) yielded similar concentrations of 1H, and these values are within a factor of ~2 from the calculated potential sensitivity from one quarter of the UDP-N-acetylglucosamine peak.

Supplementary Table 2.1A. Annotation of NMR data using database matching and manual curation. (a) Matched compound index provided by COLMARm; (b) Compound name as provided by COLMARm1. The format is specified as CompoundName_isomerNumber_spinSystemNumber; (c) Matching ratio is an output parameter of COLMARm defined as "the ratio of the matched peaks to the total number of peaks of the metabolite"; (d) Average RMSD (root-mean-square deviation) in the 13C dimension between input and database peaks. A smaller magnitude RMSD indicates a better match; (e) Average RMSD (root-mean-square deviation) in the 1H dimension between input and database peaks. A smaller magnitude RMSD indicates a better match; (f) Uniqueness score is an output parameter of COLMARm indicating the uniqueness of the match in the COLMAR database using "the number of cross-peaks in the HSQC spectrum of the mixture that are uniquely assigned" to a metabolite; (g) User-indicated score in COLMARm. For this study, "Unknown" indicates no confidence in an assignment, "Poor" indicates having only one peak matching (even for compounds with only one peak), "Fair" indicates a low observed/expected peak ratio, and "Good" indicates a credible annotation; (h) Confidence score as described in Walejko et al.: (1) putatively characterized compound classes or annotated compounds, (2) matched to literature and/or 1D reference data such as HMDB and BMRB (3) matched to HSQC, (4) matched to HSQC and validated by HSQC–TOCSY (COLMARm), and (5) validated by spiking the authentic compound into sample; (i) "y" indicates that characteristic peaks were observed in the 1D spectra of extracts. Citrate, G1P, nicontinate, and NAD+ were observed

only in 1D spectra. * not ID'd using COLMARm; manually annotated comparing 1D data to HMDB and BMRB.

ID[a]	Name[b]	Matching Ratio[c]	13C RMSD[d]	1H RMSD[e]	Uniqueness[f]	Manual Score[g]	Confidence Score[h]	Identified in 1D?[i]
1	D_Mannitol_1	1	0.05	0.007	2/4	Good	4	y
2	N_Acetyl_L_Glutamine_1	0.6	0.14	0.014	1/3	Poor	-	n
3	gamma_Aminobutyric_acid_1	0.67	0.18	0.009	1/2	Unknown	-	n
4	Phosphoethanolamine_1	1	0.05	0.003	1/2	Unknown	-	n
5	6_Aminohexanoic_acid_1	0.6	0.18	0.01	3/3	Poor	-	n
6	L_Arabitol_1	0.6	0.11	0.008	0/3	Unknown	-	n
7	D_Aspartate_1	1	0.03	0.004	3/3	Fair	4	n
8	Acetyl_phosphate_1	1	0.07	0.027	0/1	Unknown	-	n
9	Agmatine_1	0.75	0.08	0.006	2/3	Fair	4	y
10	Allantoin_1	1	0.07	0.008	1/1	Unknown	3	y
11	alpha_Ketoglutaric_acid_1	1	0.18	0.013	1/2	Poor	-	n
12	Adenosine_1	0.75	0.03	0.009	5/6	Poor	-	n
13	L_Arginine_1	1	0.03	0.002	1/5	Good	4	n
14	L_Asparagine_1	1	0.05	0.007	3/3	Fair	4	n
15	Benzyl_alcohol_1	0.75	0.22	0.005	2/3	Poor	-	n
16	Cadaverine_1	0.67	0.13	0.003	0/2	Fair	-	n
17	Choline_1	1	0.02	0.006	3/3	Good	4	n
18	Cytidine_1	0.88	0.1	0.012	3/7	Poor	-	n
19	alpha,epsilon_Diaminopimelic_acid_1	0.67	0.2	0.009	0/2	Poor	-	n
20	Ethanolamine_1	1	0.01	0.003	1/2	Good	4	y
21	Fumaric_acid_1	1	0.04	0.001	1/1	Poor	3	y
22	D_Fructose_6_phosphate_1	0.6	0.03	0.017	2/3	Unknown	-	n
23	D_Glucosaminic_acid_1	0.67	0.17	0.012	0/4	Poor	-	n
24	D_Glucuronate_1	0.6	0.2	0.014	1/3	Poor	-	n
25	D_Glucuronate_2	0.6	0.15	0.007	0/3	Poor	-	n
26	L_Glutamine_1	1	0.12	0.002	1/3	Good	4	y
27	L_Glutathione_reduced_1	0.67	0.16	0.01	0/4	Poor	-	n
28	D_Glucose_6_phosphate_1	0.67	0.1	0.003	1/4	Fair	4	y
29	D_Glucose_6_phosphate_2	0.71	0.1	0.01	1/5	Fair	4	y
30	Glycerol_1	1	0.05	0.005	1/3	Fair	4	y
31	Glycine_1	1	0.03	0.004	1/1	Unknown	3	y
32	4_Guanidinobutyric_acid_1	0.67	0.26	0.006	2/2	Poor	-	n
33	N_Acetyl_D_glucosamine_1	1	0.11	0.008	2/7	Poor	3	n
34	N_Acetyl_D_glucosamine_2	0.75	0.08	0.006	2/6	Poor	3	n
35	D_Glucose_1	0.71	0.08	0.008	1/5	Poor	3	y
36	D_Glucose_2	1	0.13	0.008	0/7	Poor	3	y
37	L_Glutamic_acid_1	1	0.05	0.004	2/3	Good	4	y
38	DL_alpha_Glycerol_phosphate_1	0.75	0.2	0.01	0/3	Poor	-	n
39	L_Homocitrulline_1	0.6	0.17	0.021	1/3	Poor	-	n
40	Homoarginine_1	1	0.09	0.008	2/5	Fair	-	n
41	L_Isoleucine_1	1	0.02	0.005	5/6	Good	4	y
42	Lactic_acid_1	1	0.02	0.003	2/2	Good	4	y
43	Leucine_1	1	0.12	0.007	4/5	Good	4	y
44	Lysine_1	1	0.07	0.003	1/5	Good	4	y
45	Malic_acid_1	1	0.14	0.001	3/3	Fair	4	y
46	Methanol_1	1	0.06	0.002	1/1	Poor	3	y
47	5_Methyluridine_1	0.63	0.1	0.013	0/5	Poor	-	n
48	Alanine_1	1	0.02	0.002	2/2	Good	4	y
49	L_Methionine_1	0.75	0.08	0.005	2/3	Poor	3	y
50	L_Ornithine_1	1	0.13	0.01	2/4	Poor	4	y
51	6_Phosphogluconic_acid_1	0.6	0.19	0.025	2/3	Poor	-	n
52	L_Phenylalanine_1	1	0.03	0.006	6/6	Good	4	y
53	Putrescine_1	1	0.05	0.002	1/2	Unknown	-	n
54	Phenethylamine_1	0.6	0.17	0.006	2/3	Poor	-	n
55	L_Proline_1	1	0.03	0.003	6/6	Good	4	y
56	L_Serine_1	1	0.03	0.009	2/2	Good	4	y
57	D_Sorbitol_1	0.63	0.15	0.01	0/5	Poor	-	n
58	L_Tyrosine_1	1	0.16	0.012	4/5	Fair	4	y
59	L_Tartaric_acid_1	1	0.08	0.026	0/1	Unknown	-	n
60	L_Tryptophan_1	0.75	0.06	0.009	6/6	Poor	3	y
61	L_Threonine_1	1	0.05	0.006	3/3	Good	4	y
62	D_Trehalose_1	1	0.08	0.011	3/7	Poor	3	y
63	UDP_1	0.71	0.16	0.018	1/5	Poor	-	n
64	Uridine_1	1	0.05	0.007	3/8	Fair	4	y
65	UTP_1	0.71	0.14	0.014	0/5	Unknown	-	n
66	Uracil_1	1	0.11	0.003	2/2	Good	4	y
67	UDP_GlcNAc_1	1	0.06	0.008	3/16	Fair	4	y
68	UDP_glucuronate_1	0.85	0.13	0.011	0/11	Poor	-	n
69	L_Valine_1	1	0.03	0.001	4/4	Good	4	y
70	Xylitol_1	0.75	0.12	0.015	1/3	Poor	-	n
71	Maltose_1	0.64	0.15	0.009	0/9	Poor	-	n
72	Maltose_2	0.71	0.18	0.007	0/10	Poor	-	n
73	D_Ribose_2	0.83	0.04	0.016	4/5	Fair	4	y
74	Guanosine_1	1	0.13	0.003	0/5	Poor	3	y
75	Glycerophosphocholine_1	1	0.15	0.004	4/8	Poor	3	y
76	Allocystathionine_1	0.6	0.24	0.012	2/3	Poor	3	n
77	Glycyl_L_Leucine_1	0.6	0.09	0.01	3/3	Poor	-	n
78	Propionylglycine_1	0.67	0.13	0.029	1/2	Poor	-	n
79	Dimethylmalonic_acid_1	1	0.1	0.001	0/1	Unknown	-	n
80	Erythrose_1	0.6	0.13	0.01	1/3	Poor	-	n
81	Gulonic_acid_1	0.6	0.18	0.02	1/3	Poor	-	n
82	Dimethyl_sulfone_1	1	0.23	0.006	0/1	Unknown	3	y
83	L_Iditol_1	0.75	0.14	0.006	1/3	Poor	-	n
84	UDP_galactose_1	0.62	0.08	0.007	1/8	Poor	-	n
85	gamma_Glutamylcysteine_1	0.6	0.15	0.011	1/3	Poor	-	n
86	Saccharopine_1	0.88	0.09	0.008	4/7	Fair	4	n
87	UDP_glucose_1	0.93	0.16	0.007	1/13	Good	4	y
88	CDP_choline_1	0.73	0.18	0.005	2/8	Poor	-	n
89	1_Methylguanosine_1	0.75	0.17	0.016	2/6	Poor	-	n
90	Hydroxyphenyllactic_acid_1	0.6	0.13	0.01	1/3	Poor	-	n
91	D-Glucuronic_acid_1	0.6	0.16	0.008	0/3	Poor	-	n
92	Glycogen_1	0.71	0.2	0.007	0/20	Poor	-	n
93	Maltotetraose_1	0.75	0.17	0.01	0/18	Unknown	-	n
94	Maltotetraose_2	0.67	0.16	0.008	0/16	Unknown	-	n
*	citrate	-	-	-	-	-	2	y
*	glucose-1-phosphate	-	-	-	-	-	2	y
*	nicontinate (niacine)	-	-	-	-	-	2	y
*	NAD+	-	-	-	-	-	2	y

Supplementary Table 2.1B. Annotation of in vivo NMR data by mapping from extract annotations. (a) ID index from initial matching (Supplementary Table 2.1a); (b) Compound name from initial matching (Supplementary Table 2.1a); (c) Confidence score for initial matching (Supplementary Table 2.1a); (d) Quantification Score for compounds in the real-time data: 0 (unannotated), 1 (annotated only), 2 (qualitatively assessed), or 3 (relatively quantifiable); * not identified using COLMARm

ID[a]	Name[b]	Confidence Score[c]	Extract -> *in-vivo* mapping notes	Quantification Score[d]
48	Alanine_1	4	Good for quant.	3
35	D_Glucose_1	3	annot/quant.	3
36	D_Glucose_2	3	annot/quant.	3
28	D_Glucose_6_phospha	4	annot/quant.	3
29	D_Glucose_6_phospha	4	annot/quant.	3
62	D_Trehalose_1	3	OK for quant.	3
21	Fumaric_acid_1	3	Good for quant.	3
74	Guanosine_1	3	Good for quant.	3
49	L_Methionine_1	3	Good for quant.	3
52	L_Phenylalanine_1	4	Quant. possible. Use ~7.4ppm	3
58	L_Tyrosine_1	4	Good for quant. 6.9ppm	3
69	L_Valine_1	4	Good for quant.	3
87	UDP_glucose_1	4	Good for quant. upfield ~5.8ppm	3
66	Uracil_1	4	Good for quant.	3
64	Uridine_1	4	Good for quant.	3
*	citrate	2	Good for quant.	3
*	glucose-1-phosphate	2	Good for quant.	3
13	L_Arginine_1	4	Good for quant, multiplet ~1.7 ppm. Other multiplets overlapped with Lysine. ~3.24, badly o	3
*	Succinate	2	Good for quant.	3
17	Choline_1	4	Good for quant.	3
12	Adenosine_1	-	Good for quant. 8.3ppm	3
20	Ethanolamine_1	4	Good for quant., but small and noisy	2
37	L_Glutamic_acid_1	4	Overlapped with glutamine. Could quantify ~2.35ppm.	2
41	L_Isoleucine_1	4	Good for quant. Some overlap.	2
42	Lactic_acid_1	4	chemical shift changes allow discernment from Threonine	2
1	D_Mannitol_1	4	Overlapped with glutamate. Do not quantify, but qualitative OK	2
26	L_Glutamine_1	4	Overlapped with glutamate. Do not quantify, but qualitative OK	2
67	UDP_GlcNAc_1	4	Good for qualitative information about trend, but bad peak shape, cannot trace	2
73	D_Ribose_2	4	removed with water/overlapped. possible quant at 5.4ppm	1
50	L_Ornithine_1	4	Small peaks; difficult to trace	1
61	L_Threonine_1	4	Does not shift ppm, so possible to discern from lactate	1
45	Malic_acid_1	4	too low to trace	1
*	NAD+	2	not confident, but possible annotation	1
60	L_Tryptophan_1	3	Annotate only, overlapped with Uracil	1
43	Leucine_1	4	Overlapped. Do not quantify.	1
44	Lysine_1	4	All overlapped, esp. with Arginine	1
46	Methanol_1	3	cannot annotate	0
9	Agmatine_1	4	cannot annotate	0
10	Allantoin_1	3	cannot annotate	0
82	Dimethyl_sulfone_1	3	overlapping, cannot annotate	0
30	Glycerol_1	4	cannot annotate	0
75	Glycerophosphocholin	3	cannot annotate	0
31	Glycine_1	3	cannot annotate	0
55	L_Proline_1	4	cannot annotate	0
56	L_Serine_1	4	cannot annotate	0
*	nicontinate (niacine)	2	cannot annotate	0

Supplementary Table 2.1C. Annotations for in vivo data. Manually curated annotations and quantifications for *in vivo* data.

Annotated in 1D *in-vivo* data	Quantified in *in-vivo* data
alanine	adenosine
glucose	alanine
glucose-6-phosphate	arginine
trehalose	choline
fumarate	citrate
guanosine	ethanol
methionine	formate
phenylalanine	fumarate
tyrosine	glucose
valine	glucose-1-phosphate
UDP-glucose	glutamate
uracil	guanosine
uridine	isoleucine
citrate	lactate
glucose-1-phosphate	phenylalanine
arginine	succinate
succinate	trehalose
choline	tyrosine
adenosine	unknown-179
ribose	uracil
ethanolamine	uridine
glutamic acid	valine
isoleucine	
mannitol	
glutamine	
ornithine	
threonine	
lactate	
malate	
UDP-GlcNAc	
serine	
tryptophan	
leucine	
lysine	

Supplementary Table 2.2. Selected pH fits for organic acids in three timepoints representing the extremes of pH changes. pH Fits derived using AssureNMR (Bruker, Billerica, MA, USA) on spectra from timepoint 1, 89, and 157 generally agreed with interpolated pH derived from in-house citrate titration data.

Compound	Timepoint (one aerobic sample)						Peaks Used (ppm)	Notes
	1		89		157			
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound		
Citrate	5.5	6	5	5.5	5	5.5	2.8-2.5	5.6 does not make sense; ignoring
Glutamate	7.5	8	4.5	5	5	5	2.4-2.3	little movement >6
Succinate - d2O	5.6	7	--*	5.6	--	5.6	2.4-2.5	use d2O or surine?
Succinate - surine	5.5	6	4	4.5	4.5	5	2.4-2.5	use d2O or surine?
Fumarate	8	--	4.5	5	5	5.5	6.45-6.6	
Alanine	8	--	4	--	4	--	1.35-1.4	little movement 4-8ppm
*	--	not bounded						
**	nd	not detected						

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Supplementary File 3.1

Tutorial for Compound Quantification Based on Ridge Tracking

This is the tutorial for the MATLAB compound quantification workflow, and it covers the main ridge tracking process. More details can be found in the shared code and the previous publication (Judge et al. 2019). Before running any code, you need to:

3.1.1.1. Clone the Edison Lab GitHub repository locally. All relevant functions and workflows can be found in the Edison Lab metabolomics toolbox repository: (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA). Specifically, the pieces of code central to this manuscript are in the folder: metabolomics_toolbox/code/ridge_tracking.

In this folder, the directory **ridgetracing_newmeth** contains dependent functions and the directory **ridgetracing_paper_workflow** contains workflows. Each working script includes sections for tracking interesting peak ridges, storing data, and plotting figures. Make sure that this GitHub repository is added to your MATLAB path.

3.1.1.2. Download the necessary supplemental data and add the folder to MATLAB path. The data can be found on Metabolomics Workbench (Project Number: **PR000738**. /ST001103/analysis/multi_sample_data/sampleData.mat).

Details for the simulation workflow can be found in **simulate.homescript.complex.m**.

Now you can start running the workflow.

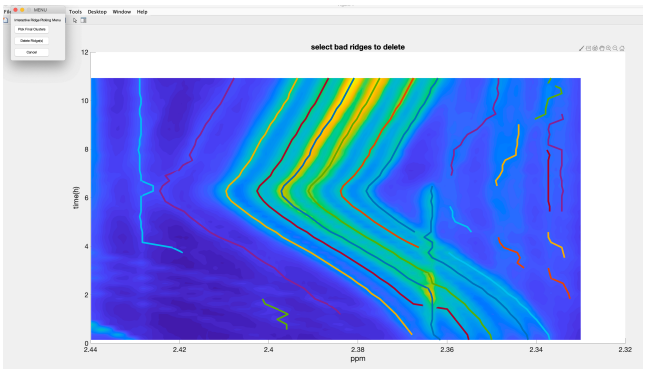
In the following section, we will use **experiment.ridtracing.manual.m** as an example. Please use this tutorial to navigate through the steps in the workflow and compare your result with figures. Alternatively, users can work on **simulation.ridtracing.manual.m**, or **repeat.experimental.tracing.pre.m**. There are parts commented out in the code. These provide options which can be ignored at first so that the user can concentrate on the main parts of the workflow. For a simplified start of the workflow, a Quick-Start Tutorial is first provided to guide the first-time user. More details are then provided to utilize the full potential of the workflow, address potential issues, and implement on other datasets.

Quick-start tutorial

3.1.2.1. Move to working folder in MATLAB and load sampleData.mat into MATLAB.

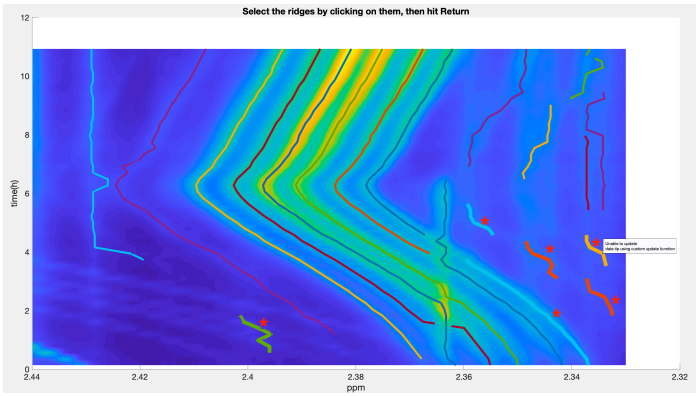
3.1.2.2. Run line 1 through 21. Have a good inspection of the spectra and close the figure.

3.1.2.3. Run from line 22 through 45. You will see the following figure. Click “Delete Ridge(s)” on the MENU.



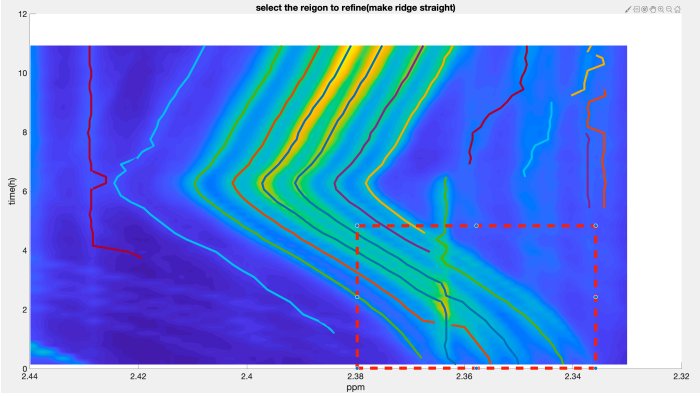
3.1.2.4. Be sure to select all the six ridges indicated by stars in the figure below (they will become thicker) for deletion and hit return/enter.

NOTE: The stars will not appear in the actual run.

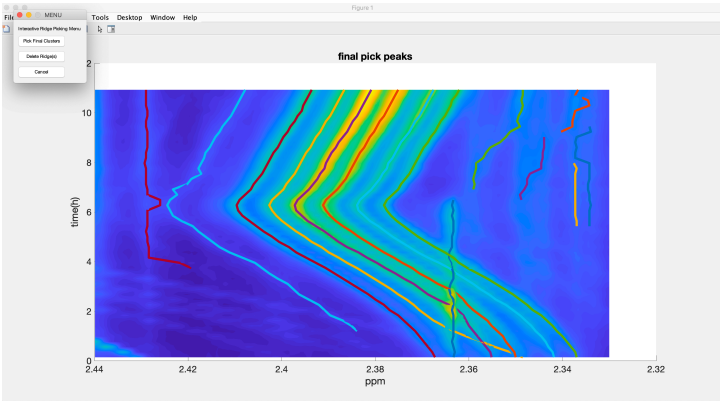


3.1.2.5. Draw a box as shown in the following figure and click mouse once anywhere in the figure.

NOTE: the box can extend slightly outside of the figure and may not be visible in the actual run.

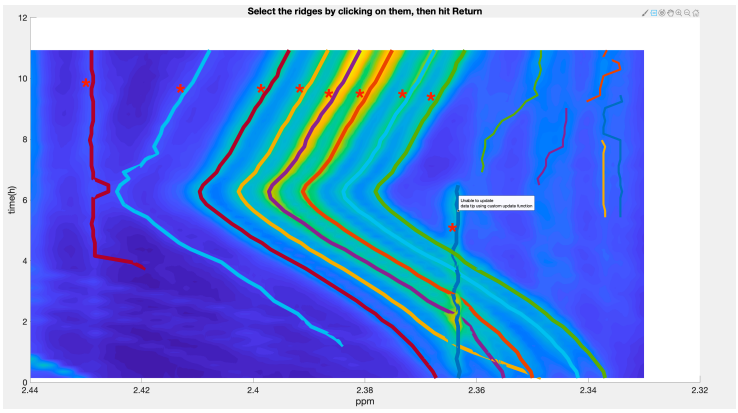


3.1.2.6. Click “Pick Final Clusters” on the MENU



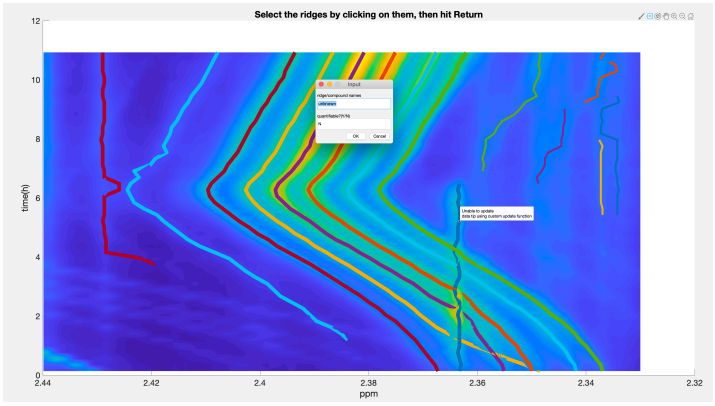
3.1.2.7. Select the nine ridges indicated by stars in the figure below and the thickness. Hit return/enter.

NOTE: this action selects the ridges which will be carried forward in the next steps.

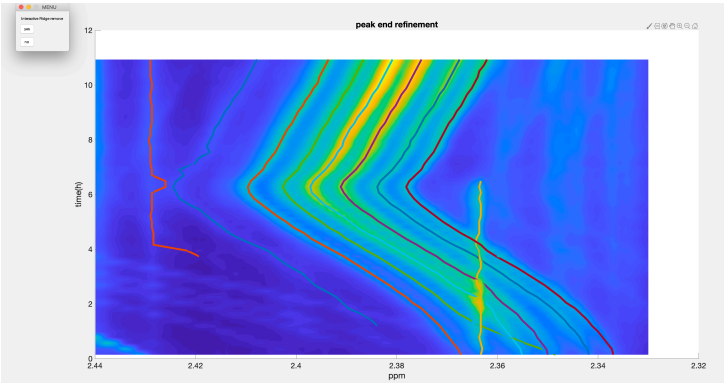


3.1.2.8. Click OK and hit space to save the information for selected ridges.

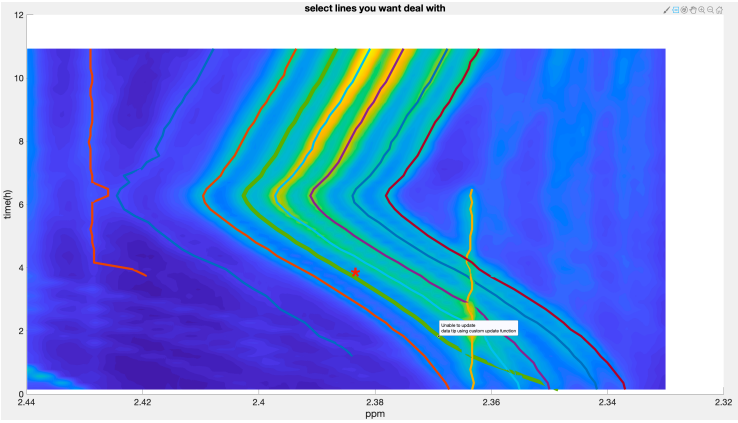
NOTE: For more details on keeping record of ridge information and tracking multiple ridges corresponding to different compounds, please refer to the detailed tutorial below (Section 4).



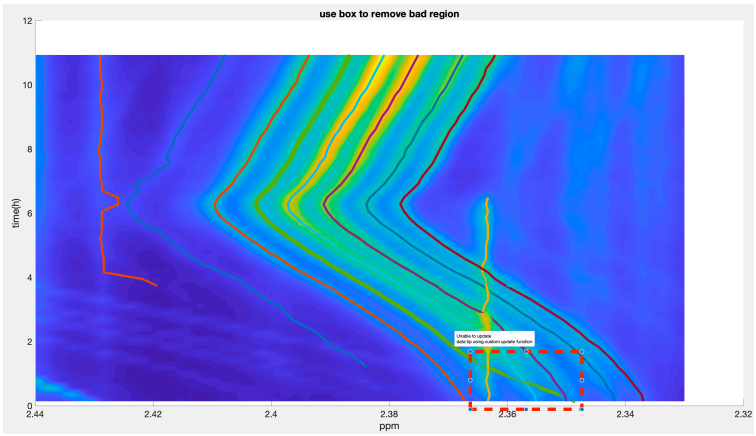
3.1.2.9. Click Yes to continue to the Ridge End Refinement step.



3.1.2.10. Click the ridge requiring refinement (indicated by the red star) to select it, then hit return/enter to continue.

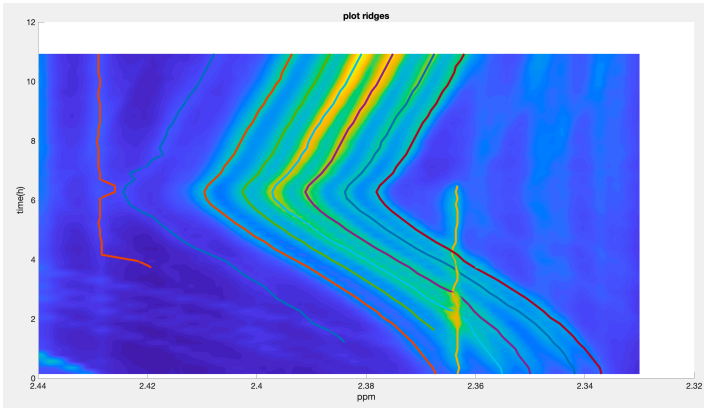


3.1.2.11. Draw a box around the region which needs to be trimmed from the selected feature as indicated by the red box on the figure below. Click mouse once and hit space to complete the process.



3.1.2.12. The final ridges will be plotted as in the figure below. These ridges and all relevant parameters are stored in the returndata structure in the MATLAB workspace.

NOTE: In this case, the leftmost ridge requires further refinement. Refer to the details below for more information on this process.



Detailed tutorial

3.1.3.1. Data input and initial visualization

Move to the working folder in MATLAB and load sampleData.mat. Instead of the directory setup (Line 8-11) and data loading (Line 12) code in the script, users can choose to move to their chosen directory and load the data ("sampleData.mat") from their own computer.

The input data is a structure array containing processed NMR spectra from multiple experiments. The important parts are ppm vector (**ppm_1h1d**), collapsed intensity matrix (**Xcollapsed_1h1d**), and collapsed time vector (**timesCollapsed_1h1d**) in the **sampleData** structure. **Xcollapsed_1h1d** is a matrix containing time-series NMR measurements with each row being a different time point and each column being a different chemical shift.

Run the script through line 21. The user can use the **stackSpectra** function (line 21) to visualize the spectra and interactively zoom in to regions of interest (Fig. 3.1.1. A more detailed view can be found in Supplementary Fig. 3.3). **horzshift** and **vertshift** can be modified to change the viewing angles. Regions of interest (ROI) can be put in the array **regionsele** (line 24), in which each row is an ROI range. The following part will focus on the ROI of chemical shift range [2.33 2.44].

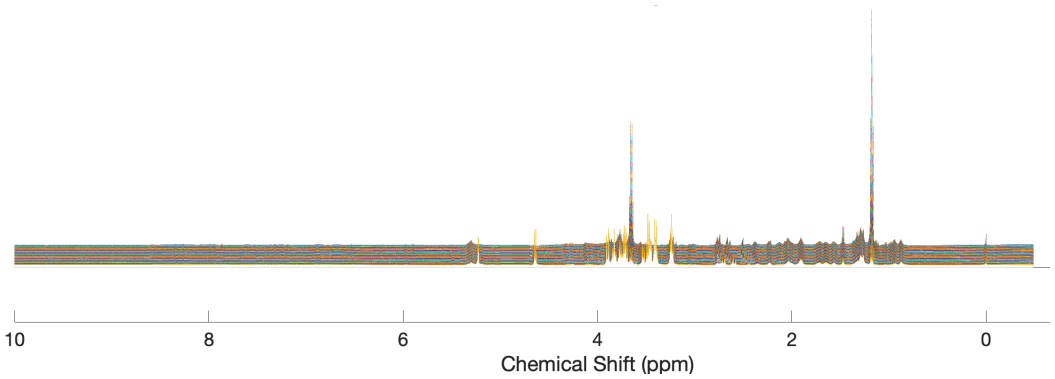


Figure 3.1.1: Stack spectral plot for experimental measurements under aerobic condition. Multiple spectra are plotted through time. The X axis indicate chemical shift.

3.1.3.2. Initial ridge tracking

Run from line 24 to 45.

After going through some intermediate steps, the user will see **ridgetrace_power2_ext** (line 45 in test run block), which is the ridge tracking function. It requires the spectral matrix, ppm vector, time vector, ROI, work directory, and tuning parameters as input. The tuning parameters include **thredseg** (L_{gap}, the maximum distance to connect a segment, in unit of indices) and **maxaddon** (N_{max}, highest local maxima to add for each time point). As L_{gap} is in unit of indices, if the user is working on spectra of resolution that is different from the example dataset`, L_{gap} can be changed according to the ratio between the resolutions. Other parameters can also be added to control function performance but are often not needed. Details in tuning parameters can be found in the main Methods and the detailed annotation of the function

ridgetrace_power2_ext. In most cases, initial tracking is enough for ridge tracking (ridge removal, retracking, and final refinement can be skipped). The example case here is relatively complex, and these steps are needed. Generally, only regions with peak shifting or overlapping need the whole workflow.

After the user executes the **ridgetrace_power2_ext** function, a window as shown in Fig. 3.1.2 will appear (another angle in Fig. 3.1.3). Most of the time, the ridge is already tracked precisely at this step. However, the example region presents a quite complex pattern with overlaps and peak shifting and requires further refinements (Fig. 3.1.2). Six peaks are shifting back and forth in the middle of the region and they overlap with a peak without much moving at around ppm 2.36. The initial ridge tracking is not perfect because of this complexity, especially in the overlapping region (Fig. 3.1.2). The user can click **Delete Ridge(s)** to refine the current result or go directly to next step by clicking **Cancel**.

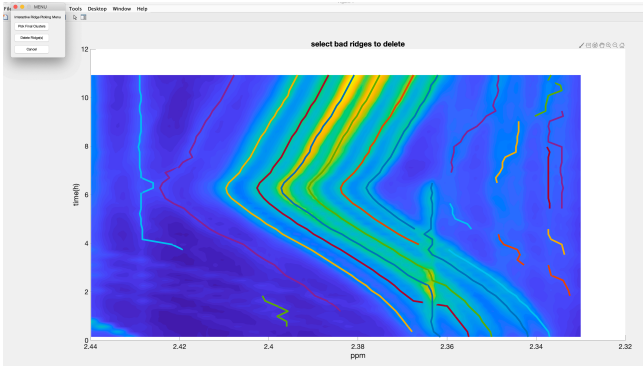


Figure 3.1.2: Initial ridge tracking result. A MENU and a surface plot will appear. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges

and different colors in surface indicate different intensity in the surface. The surface plot is also presented with another angle in Fig. 3.1.3.

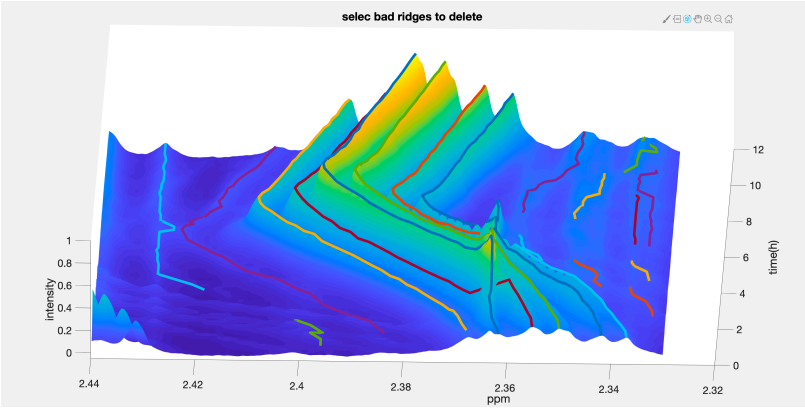


Figure 3.1.3: The surface plot with another view angle. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges. This view angle can be obtained by “Rotate 3-D” in MATLAB.

3.1.3.3. Ridge delete and retracking

In this step, the user can select ridges to delete and retrack for regions that are messy because of overlap and peak shifting. The ridge delete step will clean up the local area and help the retracking step. The area here is the rectangle region (approximately ppm [2.33, 2.38], time [0, 5]). To reproduce the result in Fig. 3.1.5, please strictly follow the details of the next two paragraphs.

If the user clicks **Delete Ridge(s)** in step 2, the mouse cursor will become a **Data Cursor** in MATLAB; in this mode ridges to be deleted can be selected. The selected ridge will become thicker and it can also be deselected by clicking again, in which case it will become thinner. There might be a message: “Unable to update data tip using custom update function”. The user can ignore this message and keep clicking until the line become thicker. In the example, we choose to delete small unwanted ridges around the selected region. Please carefully remove ridges shown by star in Figure 3.1.4. After the user selects **all** ridges to remove, they can hit **return/enter** on keyboard, after which, the surface plot will be updated (Fig. 3.1.5), and the mouse cursor will become a **cross** (retracking mode).

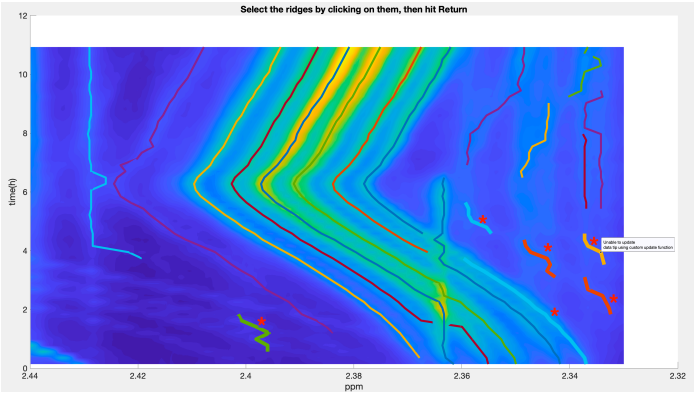


Figure 3.1.4: Remove ridges. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges. The thicker and starred ridges are selected and removed. The star will not appear when running the program.

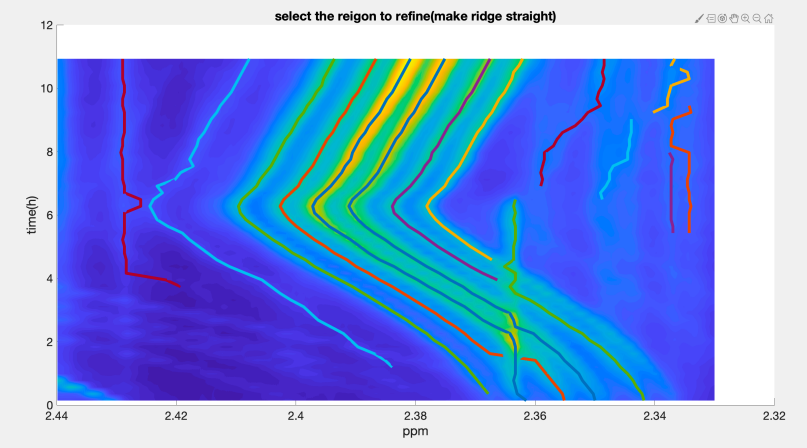


Figure 3.1.5: The ridge retracking window. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges.

In the **retracking mode**, the user can draw a box with mouse cursor (**cross**) to select regions to retrack or click any single point to finish this mode. Each ridge that enters the box will be extended in a restricted way (Method) and after each selection, the surface plot will be updated. The box we draw in this step is the approximate region (ppm [2.33, 2.38], time [0, 5], Fig. 3.1.6). Drawing the box slightly outside of the figure is fine. Please make sure a result similar to that in Fig. 3.1.7 is produced. After this step a window similar to Fig. 3.1.2 will appear (Fig. 3.1.7), and the user can click **Pick Final Clusters** to select ridges of interest. Slight problems can be seen for some ridges and will be dealt with in step 5.

A few things need to be followed in the retracking step. First, the box needs to be drawn so that all ridges just cross one boundary (i.e. not 2 or 0).

That is, the box is drawn to cover the ends of all ridges entered in the box. This is because each ridge will be retracked from where it enters the box and two entry points for one ridge can cause error. Second, keep in mind that retracking is based on linear extrapolation of earlier time points, so allowing enough time points for this estimation step (default 5) is needed. Third, retracking for noisy regions might be difficult and could benefit from a larger ROI. If you encounter any error message at this step, please check whether the three suggestions in this paragraph help.

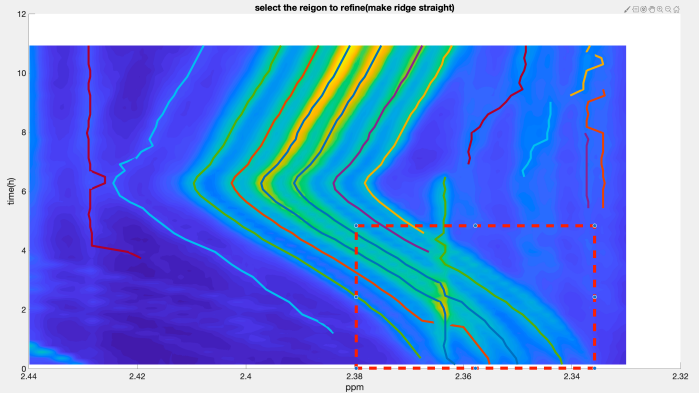


Figure 3.1.6: Ridge retracking step. The user needs to draw a box approximate to the box shown in figure to retrack the region. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges.

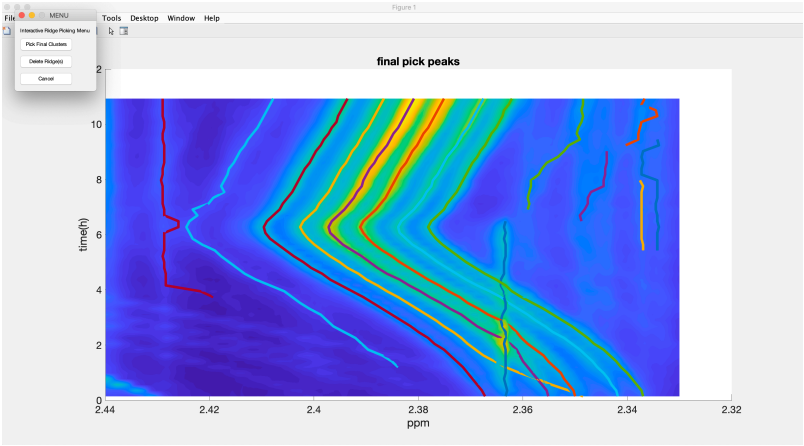


Figure 3.1.7: Result of ridge retracking. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges.

3.1.3.4. Final ridge picking

After the user finishes step 3 or clicks **Cancel** in step 2, they are presented with the new ridge tracking result and can select **Pick Final Clusters** (Fig. 3.1.7). The selection procedure is similar to the one used for ridge delete, except that after **return/enter** is hit, the user can enter more information regarding compound name and quantification (Default unknown and N. Fig. 3.1.8). These two input boxes will also reuse inputs from the last time if not modified. After OK is clicked, the user can choose to finish by pressing **return/enter** or to select more ridges iteratively by pressing other keys. A typical working process is an iterative process of selecting all the ridges for a single compound and entering the name information. This process can be repeated for different compounds. The picking process can be ended by pressing the space

key. All the selected ridges are presented in the new window (Fig. 3.1.9), from which the user can choose to do the final refinement.

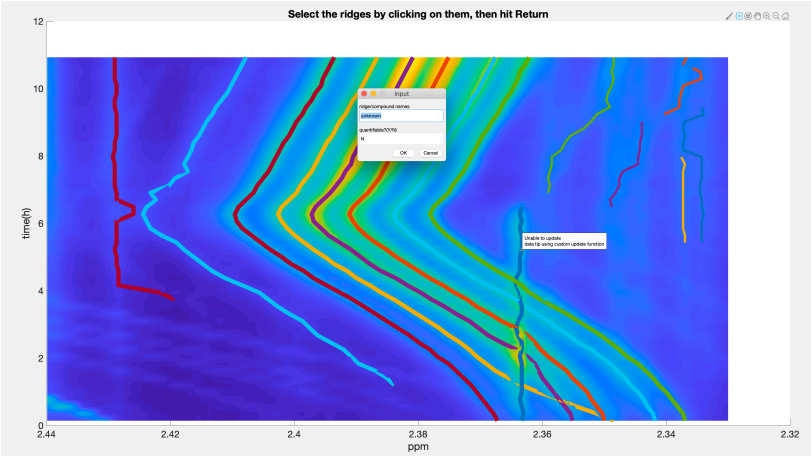


Figure 3.1.8: Information input window. In the surface plot, the X axis indicates ppm and the Y axis indicates time. Different colors in lines indicate different tracked ridges. The thicker ridges are picked.

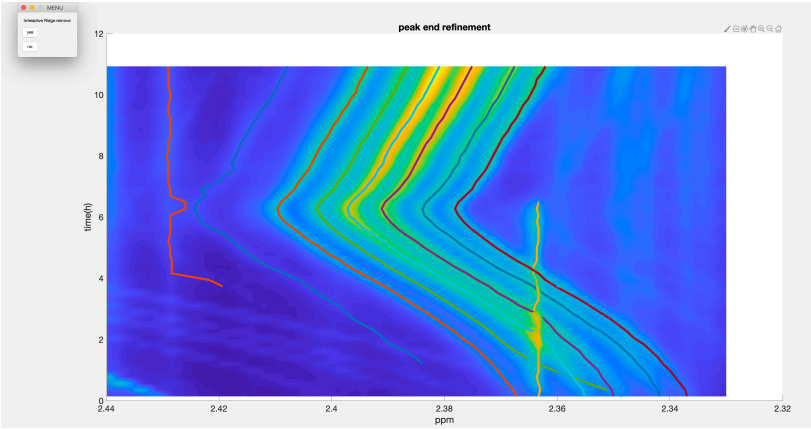


Figure 3.1.9: Final refinement window. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges.

3.1.3.5. Final refinement (Interactive Ridge remove)

If the user wants to do the final refinement (imperfect end removal), they need to click yes on the MENU. Each refinement is composed of the **ridge selection mode** and the **region selection mode**, and the process can be iterated. The selected ridge will be removed for the part of the selected region. In the **ridge selection mode**, the mouse cursor will become a **Data Cursor** and the user can select ridges to refine (selected ones will be thicker), which can be finished by pressing **return/enter**. In the **region selection mode**, the mouse cursor become a **cross** and can select regions to remove, which can be finished by clicking any single point. After **region selection mode**, the user can end the refinement by pressing the space key or start a new iteration by pressing any other key. The box you draw in region selection model might not be visible but it indeed removes the region for corresponding ridges. For refining the green ridge, we select the green ridge and draw a box covering the imperfect ridge approximately in ppm range [2.35 2.365], time [0 1.5]. Similarly, the user can choose to refine the left red ridge in a new iteration. The iterated process can be stopped by pressing the space bar after finishing region selection mode, and after this the refinement result will be updated.

For the example region, the left side red ridge and the middle green ridge were refined by end removal (Fig. 3.1.10). A structure will be returned with information on tracked ridges.

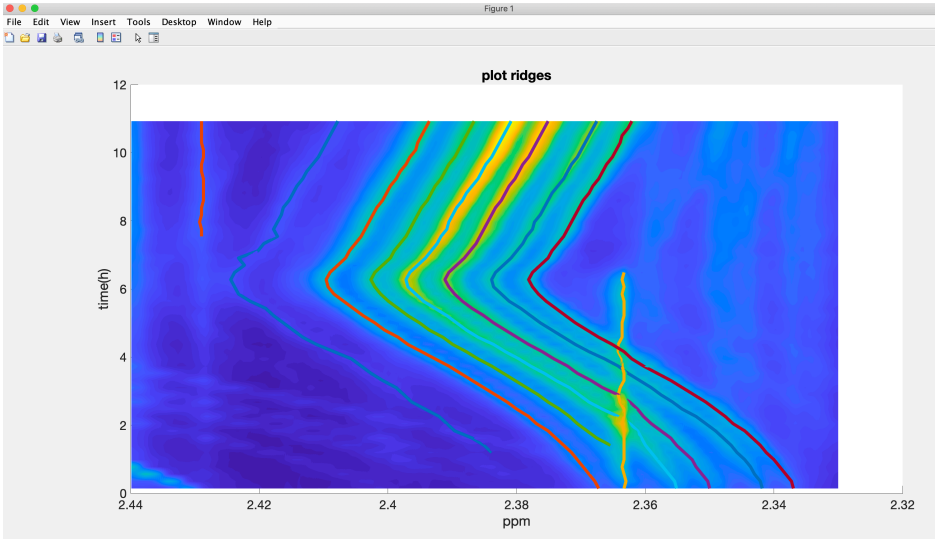


Figure 3.1.10: Ridge tracking result after final refinement. In the surface plot, the X axis indicates ppm direction and the Y axis indicates time direction. Different colors in lines indicate different tracked ridges.

In a general run, the users usually execute in order, data input (step 1 in the tutorial), the **test run** for each ROI (step 2-3) to find reasonable parameters for different ROIs, and then the **production run** for all ROI and all samples (step 2-5). The process of production run is similar to the test run. There are only two tuning parameters to optimize, **thredseg** and **maxaddon** (correspondingly **thredseglist** and **maxaddonlist** in the production run). The two parameters are

easy to guess and do not need modification for the same ROI in different replicates (Method). The other code (commented out) in the script deal with data formatting and plotting.

The region we chose for this tutorial is among the more difficult regions in the spectra. The user can also try other regions like: 1. regionhere=[1.85 1.92], thredseg=10, maxaddon=1; 2. regionhere =[6.05 6.08] thredseg=10, maxaddon=1. If errors occur, the user can rerun the same function with slight differences on parameters and manual steps until it works. To test different regions, we recommend staying within the **test run** block and change parameters, including ROI.

The pipelines and functions for this manuscript are extensively annotated, and if you have more technical questions, please feel free to contact the author.

Supplementary Methods

S3.1 NMR spectral simulation

Time-series data sets were simulated according to the known challenges that limited previous methods. The simulation includes the following procedures, and quantification of their complexity is defined in the next section (S3.2).

Reference spectra for simulation were selected from GISSMO database (http://gissmo.nmrfam.wisc.edu) because there is no noise or solvent peaks as in experimental measured spectra (Supplementary Table 3.3). This is important, especially when multiple reference spectra are summed to simulate the mixture spectra, where measurement noise can accumulate. The GISSMO database also provides for accurate simulation of NMR spectra at different field strengths, making it feasible to simulate spectra of different compounds under the same field strength as in the experimental data (600 MHZ) (Dashti et al. 2017; Judge et al. 2019).

Peak intensity was simulated to change through time by changing compound concentration linearly. Some features increased through time, some decreased, and a few (e.g. DSS as the reference) were kept constant (Supplementary Table 3.3). For each time point, compound spectra were multiplied by their concentration and summed to make the *in silico* mixture.

In addition to changes in peak intensity, we simulated pH-dependent changes in chemical shift for acetate and formate using parameters from published titration data (Ackerman et al. 1996; Tredwell et al. 2016b; Ye et al. 2018). The pH varied from 4.0 to 6.0. The simulation was based on the

Henderson-Hasselbalch equation and included parameters for pKa and chemical shift in both the protonated and unprotonated forms (Supplementary Table 3.3) (Ackerman et al. 1996; Tredwell et al. 2016b; Ye et al. 2018).

After summing all composed spectra, white noise was added (Gaussian distribution. $\mu = 0$, $\sigma = 5$). Exponential line broadening was introduced to simulate a line width similar to experimental measurements. The line broadening effect on NMR spectra (frequency domain) was implemented by multiplying the time (t) domain data with an exponential decay function $e^{\lambda t}$ ($\lambda = -0.00035$).

In addition to standard metabolite peaks, we added extra peaks (arbitrary peaks) to increase complexity and test the algorithm. More variation in chemical shift and peak intensity was introduced through these peaks. Arbitrary peaks were also introduced to overlap with shifting peaks, such as acetate.

S3.2 Metrics of spectral complexity

The complexity of experimental and simulated spectra was defined based on the following factors: SNR, overlap, change in intensity, and chemical shift.

The metrics were calculated from both the spectral matrix and the ridge intensity matrix. In the ridge intensity matrix X (N by M), each row $X(i,:)$ indicates ridge points at one time point, and each column $X(:,j)$ indicates a distinct ridge, in which $i \in 1 \dots N$ and $j \in 1 \dots M$, where N is the total number of time points and M is the total number of ridges. The vector V_{ppm} (length N_{ppm} number of points in each spectrum) denotes the whole chemical shift vector for the entire NMR spectrum, and v_{ppm}^{j} (length $N_{ridge,j}$ the length of ridge j) vector denotes the chemical shift vector for a single ridge j through existing time points.

The ridge matrix was shifted to be nonnegative (X_I) before the following calculations. A spectral region with no peaks (ppm [-0.4, -0.2]) was selected as a sample matrix (X_N, N rows) for noise level calculation.

SNR is defined here as the mean ratio of the DSS peak (at index j_{DSS}) intensity to the standard deviation (sd) in a region with no peaks (Equation [3.3.12]). The higher the noise level, the lower the SNR value. Annotation and quantification of peaks are more difficult in regions with low SNR.

$$SNR = \frac{1}{N}\sum_{i}^{N}\frac{X_I(i, j_{DSS})}{sd\big(X_N(i,:)\big)} \quad [3.12]$$

Shift complexity (C_{shift}) measures the complexity in chemical shift variation and is computed using Equations [3.3.13] and [3.3.14]. Equation [3.3.13] centered and scaled the chemical shift vector for each peak ($v'_{ppm}(j)$). Equation [3.3.14] computed an average of the normalized sum of squares (NSS, Equation [3.3.15]) of $v'_{ppm}(j)$ and was used to measure the extent of peak shift for each data set. The more the chemical shift varied for each individual peak, the larger C_{shift}. In the calculation of NSS, V is a vector, s is the index in the vector, V_s is one element in the vector, and N_L is the length of the vector.

$$v'_{ppm}(j) = \frac{v^{j}_{ppm} - \min\big(V_{ppm}\big)}{\max\big(V_{ppm}\big) - \min\big(V_{ppm}\big)} \quad [3.13]$$

$$C_{shift} = \frac{1}{M}\sum_{j}^{M} NSS\big(v'_{ppm}(j)\big) \quad [3.14]$$

$$NSS(V) = \frac{1}{N_L}\sum_{s}^{N_L}\left(V_s - \frac{1}{N_L}\sum_{s}^{N_L} V_s\right)^2 \quad [3.15]$$

Scale complexity (C_{scale}) measures the extent to which peak intensities differ from their closest neighbors (Equation [3.3.16]). N_{pair} is the number of closest neighbor ridge pairs with average differences in chemical shift less than 0.05 ppm. $N_{diff,i}$ is the number of peak pairs counted in N_{pair} and with >10-fold change in intensity for each time point. N is the total number of time points. The greater the difference in intensity between neighboring peaks, the higher the value of C_{scale}. Considerable differences in intensity between nearby peaks result in imperfect peak shapes and challenges in quantification. Sometimes, the smaller peak is only discernible in a proportion of time points because of overlap.

$$C_{scale} = \frac{1}{N(\mathrm{N_{pair}})} \sum_{i}^{N} N_{diff,i} \qquad [3.16]$$

Dynamic complexity ($C_{dynamics}$) measures the complexity in intensity variation through time and is computed with Equation [3.3.17]. For each ridge, the intensity vector was scaled by its maximum. NSS of the scaled intensity can indicate time dynamics for each ridge, and the value was averaged for all ridges. The more the intensity changes for each peak through time, the higher $C_{dynamics}$. Data sets with higher $C_{dynamics}$ have an interesting underlying variation in metabolism, such as metabolic adaption under different carbon sources (Judge et al. 2019).

$$C_{dynamics} = \frac{1}{M} \sum_{j}^{M} NSS\left(\frac{X_I(:,j)}{\max(X_I(:,j))}\right) \qquad [3.17]$$

S3.3 Parameter tuning for ridge tracking

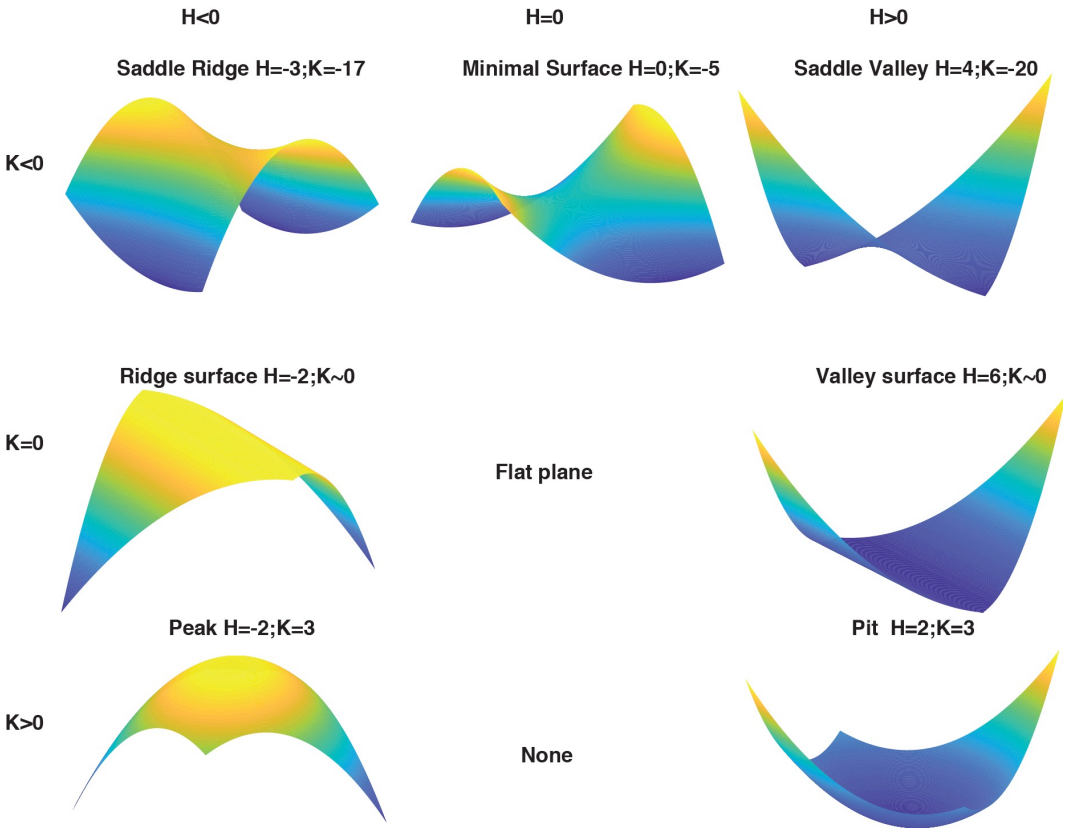
RTExtract and the previous method (Judge et al. 2019) has different parameter tuning processes. Both methods were tested by the us on the same data set and near optimal parameter sets were selected. The correctness of ridge tracking is also evaluated by figures like shown in Fig. 3.5. As the previous method need tuning of seven parameters, a good parameter set often takes longer.

Parameter tuning for RTExtract has been described in method. Following that procedure should give near optimal tuning parameter sets. We found that the results were robust to a wide-range of parameters, making this particularly simple. Additionally, to use RTExtract on spectra with different resolution, L_{gap} needs to be changed proportionally, because it corresponds with the unit of index.

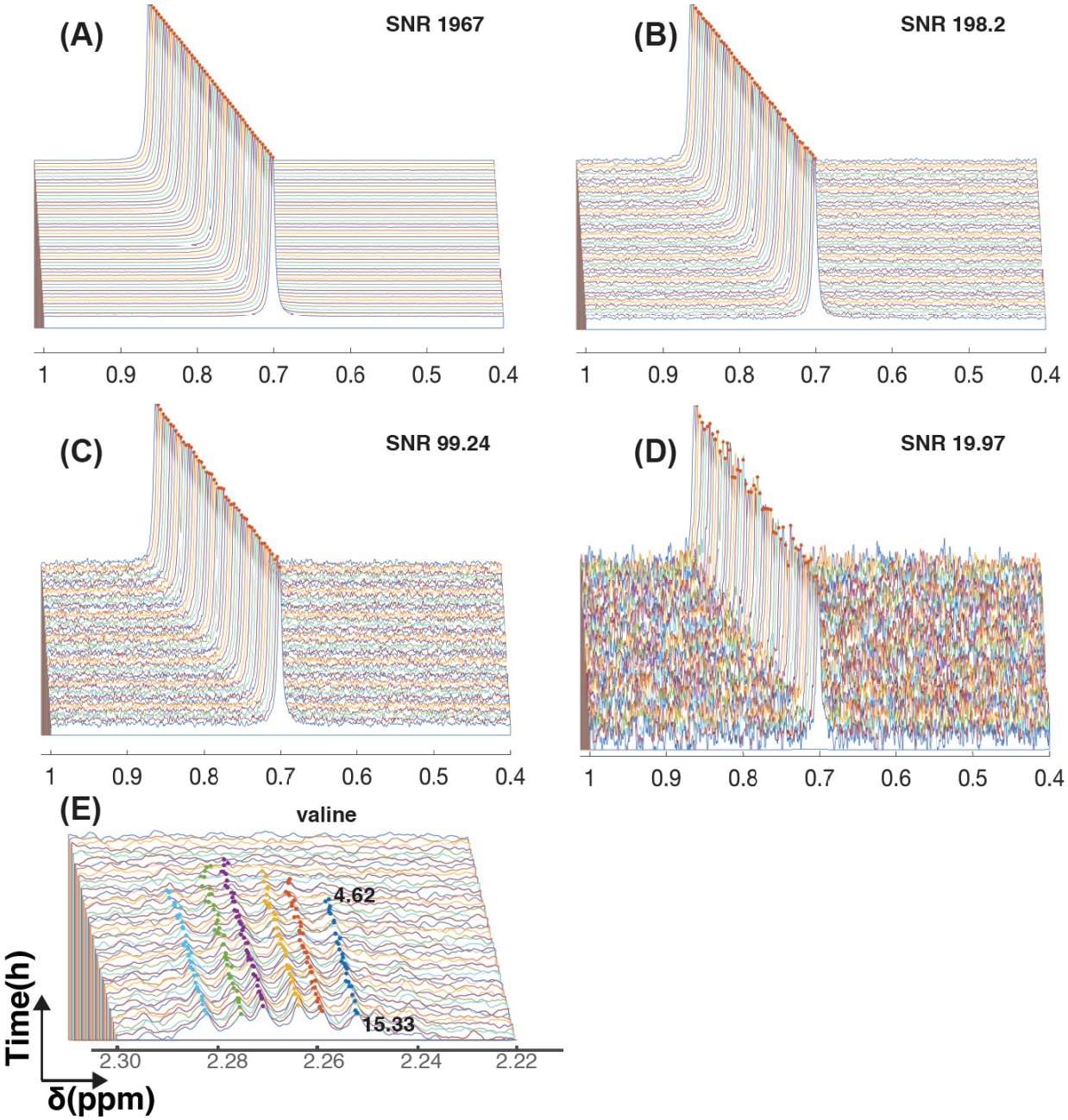
For the method in CIVM paper (Judge et al. 2019), finding working parameters requires more trial and error. Mainly seven parameters need modification: σ_{time}, σ_{ppm}, h_{peak}, N_{ridge}, W_{time}, W_{ppm}, and $W_{intensity}$. In smoothing and peak picking stage the first three parameter need tuning: σ_{time} standard deviation of the 2D Gaussian kernel at time direction; σ_{ppm} standard deviation of the 2D Gaussian kernel at ppm direction; h_{peak} threshold for peak picking. σ_{ppm} and σ_{time} need tuning for different peak density and regional SNR. h_{peak} needs tuning for different peak heights. In the clustering and connection stages, other parameters need tuning: N_{ridge} number of expected ridges; W_{time} control the weight in time dimension; W_{ppm} control the weight in ppm dimension; and

$W_{intensity}$ control the weight in intensity dimension. N_{ridge} can be initialized to $k \cdot N_{\exp}$, where most of the time, $k \in [1, 4]$, and N_{exp} is the number of expected ridges. Choices of k depends on the noise level. The three weight factors are codependent and a fixed relative ratio often gives equivalent results. Increasing of one weight will encourage the ridge to grow across the corresponding dimension. Tracking peaks of rapid chemical shift (intensity) changing will need increase of W_{ppm} ($W_{intensity}$). In the ridge correction stage, ridge points on the smoothed surface are mapped back to the real data. There are two parameters controlling this stage, but they mostly need no changes. For the presented region (Fig. 3.4 and Supplementary Fig. 3.7), the near optimal parameters are found so that peaks are correctly tracked.

Both methods need little change on parameter set among replicates. For RTExtract, the same parameter set also works for the same region of different conditions, while the previous method often need different tuning parameters for this case. All tuning paraemters are stored in data structure for reproducibility.

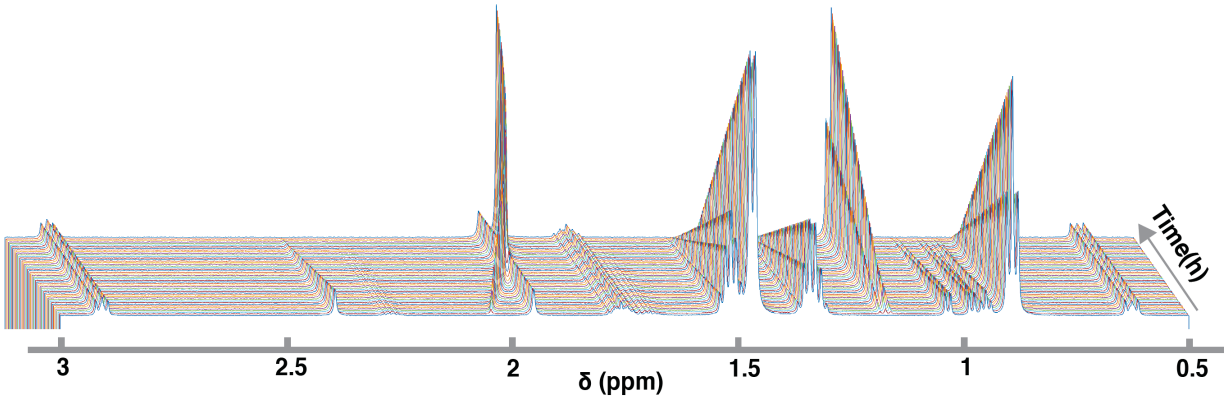


Supplementary Figure 3.1: Surface visualization for different values in H and K curvature. Different surface types were simulated by second order polynomials with interaction terms. The surfaces were classified by the value of H and K curvature values. Curvature of the central point was computed and shown with corresponding surface type name. There is no case with K>0 and H=0. The surface type of particular interest is the ridge surface ($K = 0$ and $H < 0$). $K \sim 0$ indicates that the K has a small absolute value ($< 10^{-14}$). Except for the ridge surface and valley surface, the computed H and K curvatures agreed exactly with the expected value. For ridge and valley, the expected curvature is 0 and the computed curvature was close to 0. Details in simulation can be found in Methods. A similar draft of different surfaces can also be found (Suk and Bhandarkar 1992).

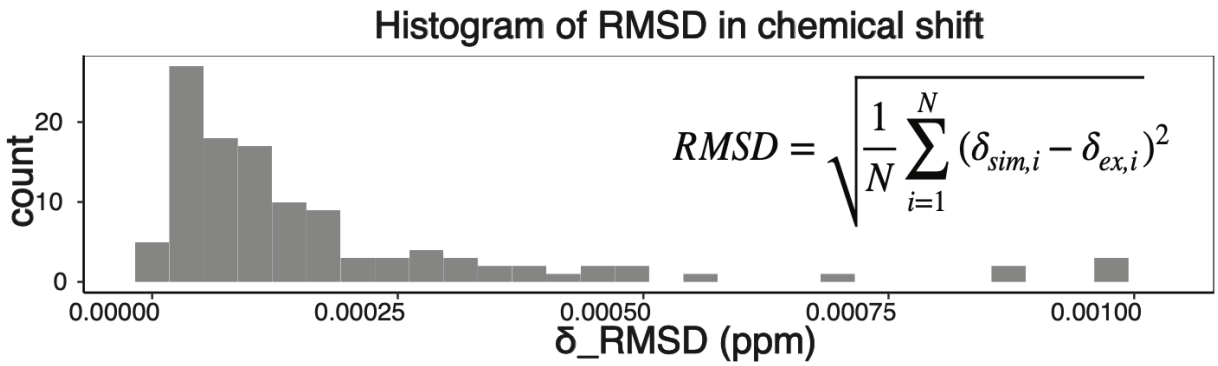


Supplementary Figure 3.2: Evaluation of noise effect on ridge tracking. (A)-(D) Spectral peaks with chemical shift variation and different SNRs were tested for evaluating performance of ridge tracking, and the red dots show the tracked peaks. The ridge can be tracked automatically for A-C, and D can be tracked with some manual refinement. (E) The peak tracking results for valine multiplet at

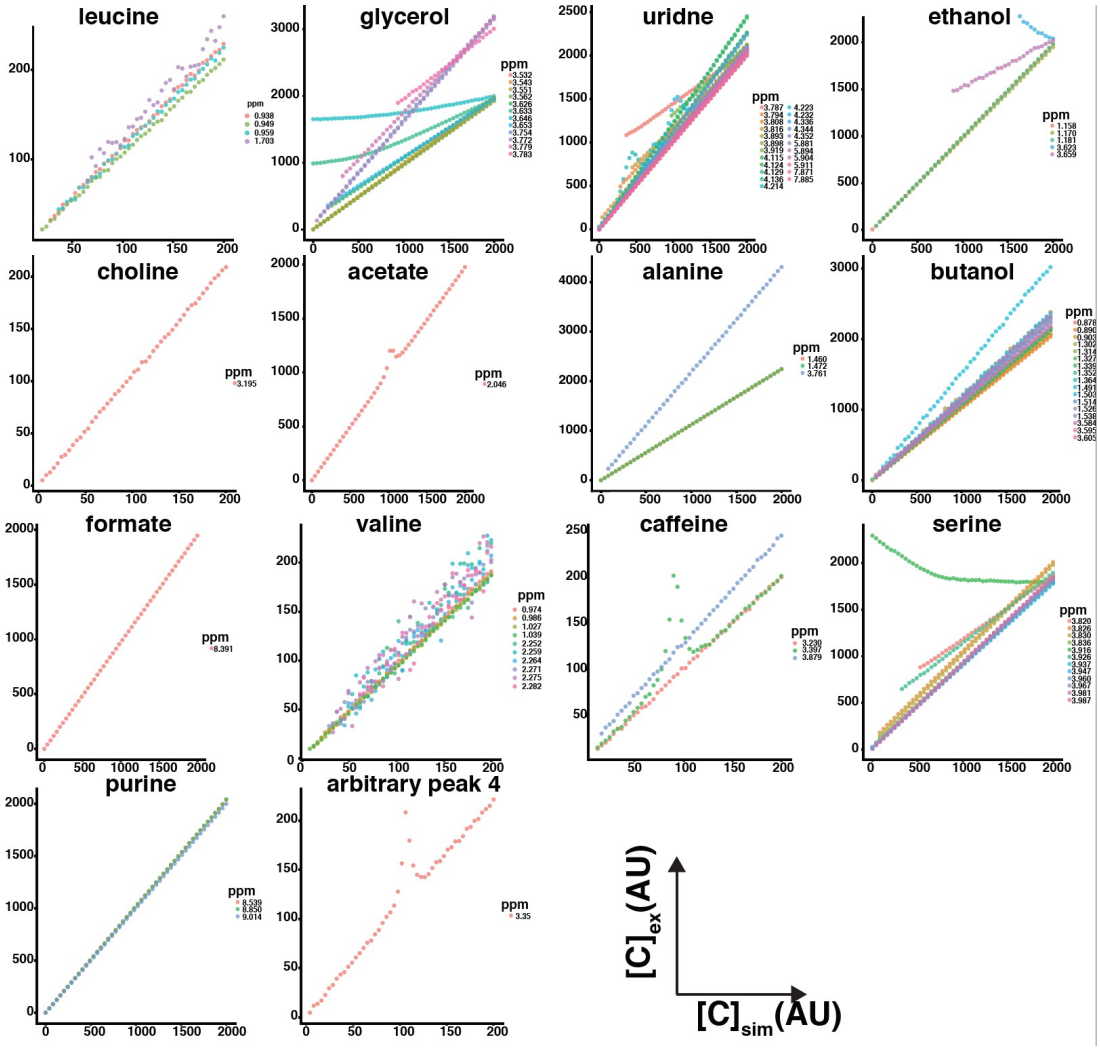
ppm [2.22, 2.30] in simulated data set. The numbers indicate the SNRs for corresponding peaks (the blue peak at time point 1 and 33), with noise regions ppm [2.05 2.2]. Spectra (X-axes chemical shift δ) were plotted against time (Y-axes), and different line colors indicated spectra at different time points. Spectral plots were shifted for visualization, and the bottom spectra corresponded to the chemical shift axes.



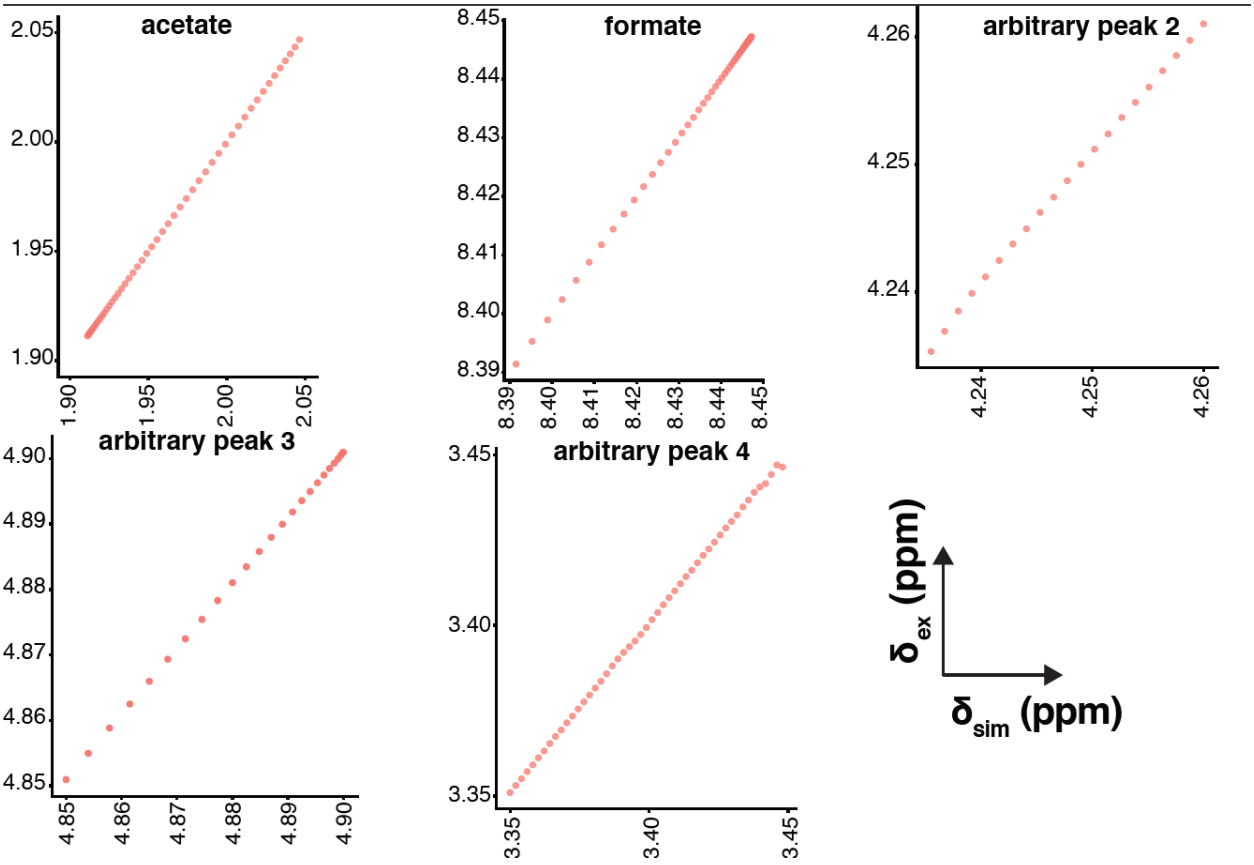
Supplementary Figure 3.3: A region of the simulated spectra. The spectra (X-axis, ppm [0.5, 3]) are presented as a stacked plot through time (Y-axis).



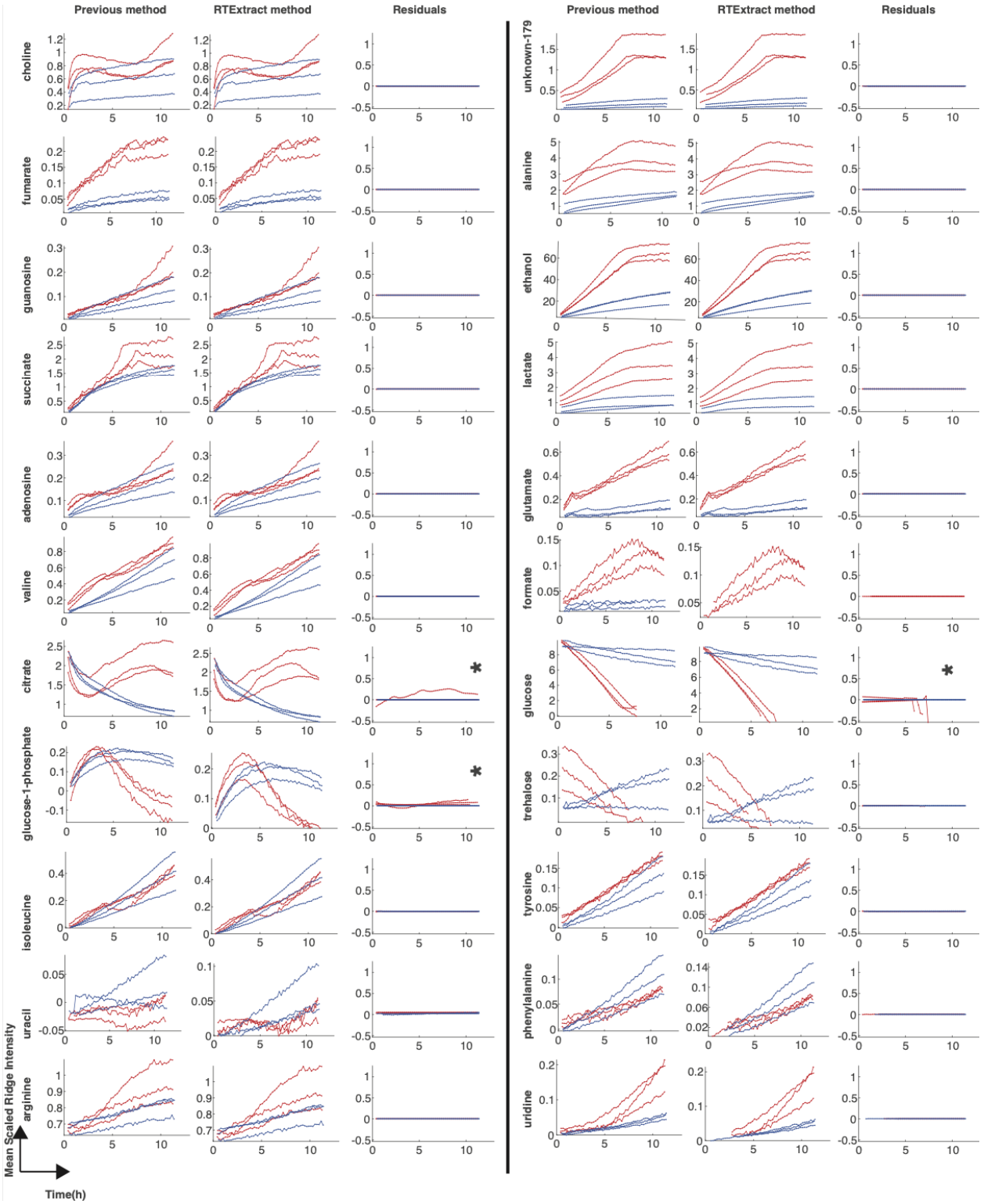
Supplementary Figure 3.4: Histogram of RMSD in chemical shift for tracked peaks by RTExtract in simulated data sets. The X axis is RMSD (Root mean square deviation) in chemical shift (δ), and the Y axis is the count number. δ_{sim}: the simulated chemical shift; δ_{ex}: the extracted chemical shift. Most tracked peaks had a small RMSD value. The equation for computing RMSD is also presented in the figure (Method). The histogram was not plotted for the previous method because much fewer ridges were tracked. As peak simulations included different compounds and line broadening, peak widths varied with a typical value of about 0.006 ppm, larger than typical RMSD values.



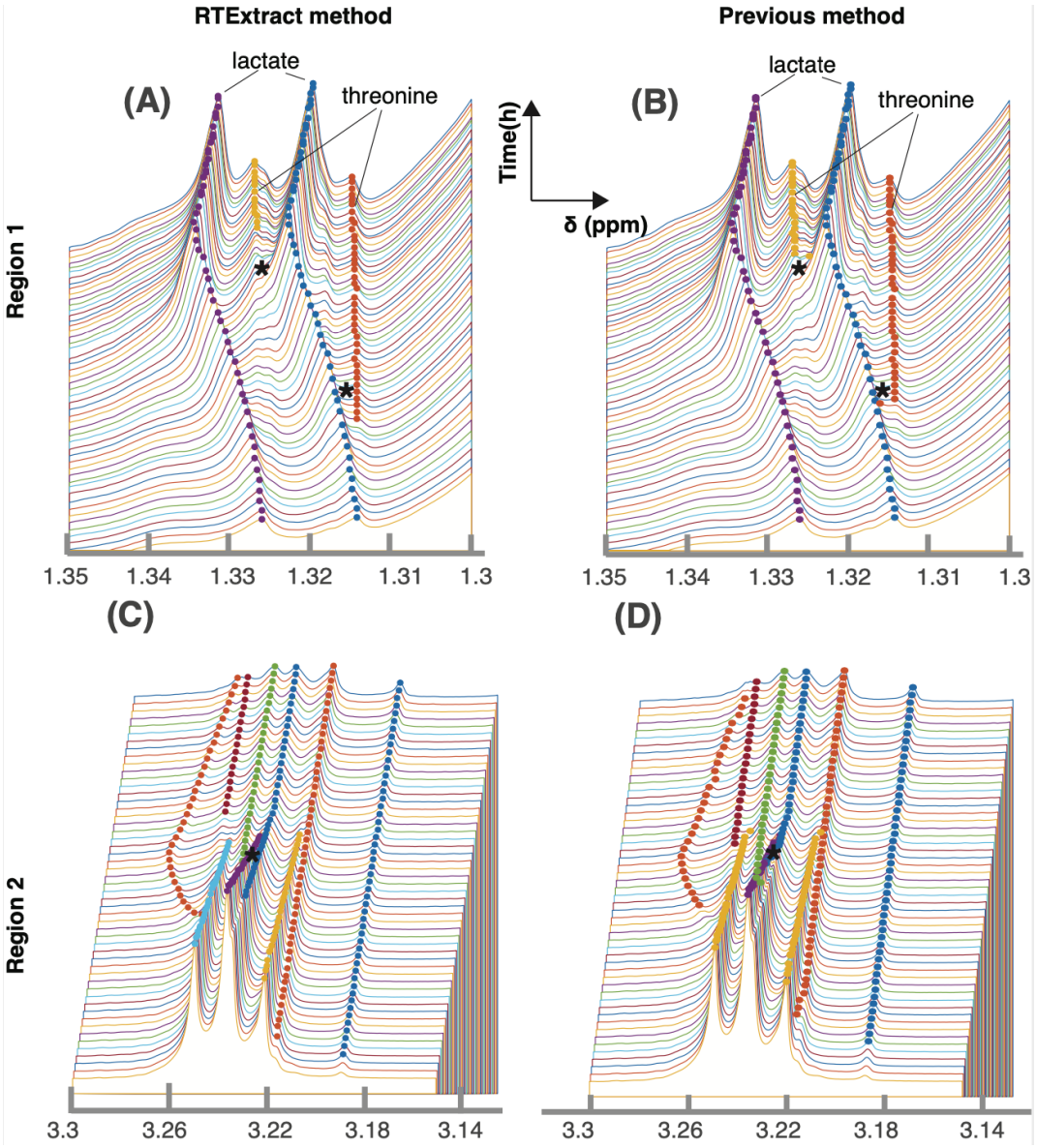
Supplementary Figure 3.5: Evaluation of intensity estimation in the simulated data set. Extracted concentration ($[C]_{ex}$) was plotted against simulated concentration ($[C]_{sim}$). Each compound and its concentrations are represented by one plot, and each peak is presented by a color. The extracted concentration was computed based on ridge intensity scaled by the DSS peak and the simulated concentration was the known true value. Extracted concentration can correctly represent simulated concentration for peaks without overlaps. “Arbitrary peak” is one added arbitrary peak to increase spectral complexity as explained in S3.1.



Supplementary Figure 3.6: Evaluation of chemical shift estimation for peaks with chemical shift variation in the simulated data set. Extracted chemical shifts (δ_{ex}) were plotted against simulated chemical shift (δ_{sim}). Each compound is presented by one plot. The X-axes indicate simulated chemical shift, and the Y-axes indicate extracted chemical shift. In the simulated spectra, five peaks changed chemical shift and the extracted chemical shifts accurately represented simulated chemical shifts for these peaks. Arbitrary peak 2, 3, and 4 are added arbitrary peaks to increase spectral complexity as explained in S3.1.

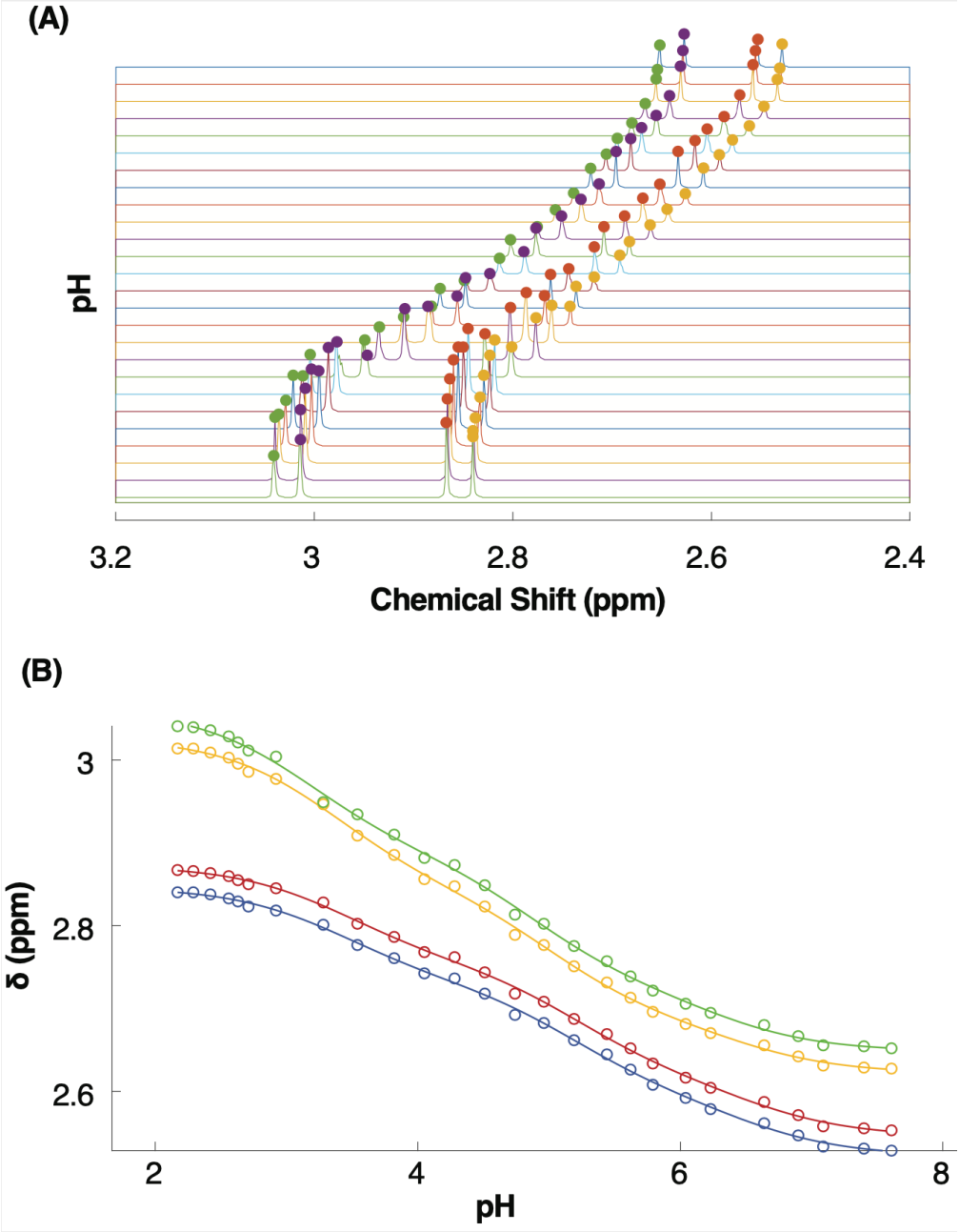

Previous method
RTExtract method
Residuals
choline
fumarate
guanosine
succinate
adenosine
valine
citrate
glucose-1-phosphate
isoleucine
uracil
arginine
unknown-179
alanine
ethanol
lactate
glutamate
formate
glucose
trehalose
tyrosine
phenylalanine
uridine
Mean Scaled Ridge Intensity
Time(h)

Supplementary Figure 3.7: Comparison of quantification results between the method presented in this paper (RTExtract method) and our previously published method (Previous method). A similar procedure in combining peaks for compound quantification was implemented as the previous method (Judge et al. 2019). In each plot, the X-axis indicates time, and the Y-axis indicates Mean Scaled Ridge Intensity. Red curves are from aerobic data sets, and blue curves are from anaerobic data sets. Details in computing Mean Scaled Ridge Intensity can be found in the previous publication (Judge et al. 2019). Each row indicates quantification for one compound, including the Previous method, RTExtract method, and Residuals (differences between the quantifications by the two methods). A value near zero in Residuals indicates agreement between results of the two methods and quantification differences are highlighted by stars. There are differences in how negative values were dealt with between the two methods. Additionally, in one experimental replicate, citrate is quantified with three ridges (rather than four for other replicates) for the previous published method, while all four ridges are used for quantification in the RTExtract method for all replicates.



Supplementary Figure 3.8: Evaluation of RTExtract and the previous method on complex overlapping regions on the experimental data sets. Two ROIs were selected as examples to compare RTExtract and the previous method. Each row shows one region and each column one method. Peaks in these ROIs can be precisely tracked by RTExtract (A and C), and the parts that are problematic in the previous method are indicated with stars (B and D). (A) is the same as Fig. 3.5 A and is repeated here for comparison. Spectra (X-axes

chemical shift δ) were plotted against time (Y-axes). Different point colors indicate different tracked peaks. Performance of the algorithm for less complex regions is in Fig. 3.4 and Supplementary Fig. 3.7.



Supplementary Figure 3.9: Ridge tracking on pH titration data of citrate. (A) The spectral plot shows tracking of the four citrate peaks through an experimental pH titration. (B) Extracted chemical shift (δ, Y-axis) is plotting with corresponding pH (X-axis). Circles are experimental measurements and solid lines are fitting results with Henderson-Hasselbalch equation with three

protonation sites. pK_a are estimated: $\widehat{pK_{a1}} = 3.34 \pm 0.23$, $\widehat{pK_{a2}} = 5.06 \pm 0.25$, and $\widehat{pK_{a3}} = 6.45 \pm 0.19$ (the confidence interval is calculated by 2 standard deviation). (Szakacs et al. 2004; Tredwell et al. 2016b). The pK_a estimation is also affected D_2O used in the measurement (Bates and Pinching 1949).

Supplement Table 3.1: Complexity metrics of spectral data sets from simulations and experimental measurements. Experimental and simulated data sets were compared for different types of complexity. SNR: signal-to-noise ratio for the DSS peak in the spectra. Shift complexity (C_{shift}): measurements of how much peaks shift in chemical shift dimension. Scale complexity (C_{scale}): measurements of how nearby peaks differ in intensity. Dynamics complexity ($C_{dynamics}$): measurements of how dynamic peak intensity changes through time. The higher the SNR, the less complex the spectral data sets. The higher the other complexity measurement, the more complex the spectral data sets. The complexity measurements were compared based on both spectra matrix and peak intensity. Details in computing these criteria can be found in S3.2.

Data source	SNR	Shift complexity	Scale complexity	Dynamics complexity
aerobic experimental data set	2.47E+02	4.53E-07	6.76E-02	5.61E-02
anaerobic experimental data set	6.49E+02	7.25E-08	3.55E-02	3.32E-02
simulated data set	1.84E+03	2.62E-07	1.67E-01	7.57E-02

Supplementary Table 3.2: Evaluation of estimation accuracy in chemical shift for simulated data sets. The accuracy was evaluated by RMSD (root mean square deviation) between extracted chemical shift and simulated chemical shift of each peak. A close to zero value indicates an accurate result. The peak column lists different peaks for different compounds with a form of "Compound_ppm". NA indicates that the corresponding peaks were not tracked and quantified. Result from both RTExtract and the previous method are presented. Because of considerable time cost and difficulty, we didn't track all ridges for the previous method.

peaks	RTExtract	the previous method	Note
ethanol_1.1578	3.50E-05	NA	
ethanol_1.1696	1.30E-05	NA	
ethanol_1.1814	2.09E-05	NA	
ethanol_3.6231	2.66E-04	NA	
ethanol_3.6349	NA	NA	
ethanol_3.6467	NA	NA	
ethanol_3.6586	2.52E-04	NA	
Valine_0.97412	5.02E-05	NA	
Valine_0.98602	6.67E-05	NA	
Valine_1.0272	5.83E-05	NA	
Valine_1.0387	1.05E-04	NA	
Valine_2.2284	NA	NA	
Valine_2.2357	NA	NA	
Valine_2.2402	NA	NA	
Valine_2.2473	NA	NA	
Valine_2.2519	2.66E-04	NA	
Valine_2.259	3.32E-04	NA	
Valine_2.2636	3.75E-04	NA	
Valine_2.2707	3.10E-04	4.21E-04	
Valine_2.2753	5.52E-04	NA	
Valine_2.2824	3.76E-04	4.44E-04	
Valine_2.287	NA	NA	
Valine_2.2942	NA	NA	
Valine_2.2986	NA	NA	
Valine_2.3058	NA	NA	
Valine_3.5954	NA	NA	
Valine_3.6024	NA	NA	
Acetate_2.0463	1.09E-04	6.21E-05	previous method track part of the ridge and RTExtract track whole ridge
Glycerol_3.5317	2.35E-05	NA	
Glycerol_3.5425	4.54E-05	NA	
Glycerol_3.5513	4.54E-05	NA	
Glycerol_3.5621	3.58E-05	NA	
Glycerol_3.6263	1.06E-04	NA	
Glycerol_3.6334	7.08E-04	NA	
Glycerol_3.6459	4.92E-04	NA	
Glycerol_3.653	1.32E-04	NA	
Glycerol_3.7536	1.04E-04	1.55E-04	
Glycerol_3.7608	NA	NA	
Glycerol_3.7645	NA	NA	
Glycerol_3.768	NA	NA	
Glycerol_3.7716	3.32E-04	3.32E-04	
Glycerol_3.7752	NA	NA	
Glycerol_3.7788	2.47E-04	2.47E-04	
Glycerol_3.7824	2.35E-04	2.35E-04	
Glycerol_3.7897	NA	NA	
Choline_3.195	1.22E-04	NA	
Choline_3.5058	NA	NA	
Choline_3.5139	NA	NA	
Choline_3.5169	NA	NA	
Choline_3.5221	NA	NA	
Choline_4.0452	NA	NA	

Choline_4.0503	NA	NA
Choline_4.0533	NA	NA
Choline_4.0613	NA	NA
alanine_1.4601	1.27E-04	NA
alanine_1.4722	1.04E-04	NA
alanine_3.7493	NA	NA
alanine_3.7613	2.10E-04	2.18E-04
alanine_3.7734	NA	NA
alanine_3.7855	NA	NA
Uridine_3.787	2.75E-04	2.75E-04
Uridine_3.7943	1.33E-04	1.60E-04
Uridine_3.8083	1.03E-04	1.03E-04
Uridine_3.8157	1.91E-04	2.46E-04
Uridine_3.8933	1.52E-04	1.52E-04
Uridine_3.8979	7.00E-05	7.00E-05
Uridine_3.9146	NA	NA
Uridine_3.9194	2.70E-04	NA
Uridine_4.1153	2.16E-04	NA
Uridine_4.1204	NA	NA
Uridine_4.1241	1.59E-04	NA
Uridine_4.1289	3.14E-04	NA
Uridine_4.1314	NA	NA
Uridine_4.1365	2.19E-04	6.90E-02
Uridine_4.2139	4.72E-04	2.47E-04
Uridine_4.223	7.10E-05	1.12E-04
Uridine_4.2321	1.06E-04	1.63E-04
Uridine_4.3359	7.20E-05	8.56E-02
Uridine_4.3439	1.74E-04	NA
Uridine_4.3523	6.94E-05	NA
Uridine_5.8807	1.94E-05	NA
Uridine_5.8942	1.00E-04	NA
Uridine_5.9039	7.54E-05	NA
Uridine_5.9114	7.11E-05	NA
Uridine_7.8711	1.00E-06	NA
Uridine_7.8846	2.00E-06	NA
Leucine_0.93771	9.90E-05	NA
Leucine_0.94855	7.47E-05	NA
Leucine_0.95941	9.89E-05	NA
Leucine_1.6244	NA	NA
Leucine_1.6328	NA	NA
Leucine_1.6473	NA	NA
Leucine_1.6569	NA	NA
Leucine_1.6628	NA	NA
Leucine_1.6667	NA	NA
Leucine_1.673	NA	NA
Leucine_1.6761	NA	NA
Leucine_1.6832	NA	NA
Leucine_1.6925	NA	NA
Leucine_1.6993	NA	NA
Leucine_1.7028	4.38E-04	NA
Leucine_1.7149	NA	NA
Leucine_1.7252	NA	NA
Leucine_1.7336	NA	NA

Leucine_1.7444	NA	NA	
Leucine_1.7626	NA	NA	
Leucine_3.7001	NA	NA	
Leucine_3.7089	NA	NA	
Leucine_3.7178	NA	3.57E-04	
Leucine_3.7262	NA	NA	
Leucine_3.7339	NA	NA	
Formate_8.3915	3.02E-05	NA	
Butanol_0.87806	3.05E-05	NA	
Butanol_0.89039	7.21E-05	NA	
Butanol_0.90272	4.72E-05	NA	
Butanol_1.3015	1.27E-04	NA	
Butanol_1.3141	4.05E-05	NA	
Butanol_1.3267	3.93E-05	NA	
Butanol_1.3393	4.57E-05	NA	
Butanol_1.3519	5.86E-05	NA	
Butanol_1.3644	1.45E-04	NA	
Butanol_1.4727	NA	NA	
Butanol_1.4913	1.31E-04	NA	
Butanol_1.5027	7.92E-05	NA	
Butanol_1.5144	3.74E-05	NA	
Butanol_1.5258	1.01E-04	NA	
Butanol_1.5375	1.84E-04	NA	
Butanol_1.559	NA	NA	
Butanol_3.5839	3.41E-05	NA	
Butanol_3.5946	4.80E-05	NA	
Butanol_3.6054	1.00E-04	NA	
Caffeine_3.2296	5.93E-05	1.42E-01	
Caffeine_3.397	1.91E-04	2.13E-04	
Caffeine_3.8786	6.98E-05	1.43E-04	
Caffeine_7.8756	NA	NA	
Serine_3.8199	1.69E-04	2.33E-04	
Serine_3.8262	7.40E-05	7.40E-05	
Serine_3.8295	9.43E-05	9.43E-05	
Serine_3.8358	3.00E-05	1.15E-04	
Serine_3.9163	8.71E-04	NA	
Serine_3.9261	1.40E-04	NA	
Serine_3.9369	5.60E-05	NA	
Serine_3.9466	2.81E-05	NA	
Serine_3.9604	3.00E-06	NA	
Serine_3.9665	4.34E-05	NA	
Serine_3.9809	3.70E-05	NA	
Serine_3.9871	3.04E-05	NA	
Purine_8.5392	6.09E-05	NA	
Purine_8.8504	7.32E-05	NA	
Purine_9.0142	1.41E-04	NA	
DSS_-7.6297e-06	3.70E-07	NA	
DSS_0.61128	4.76E-05	NA	
DSS_0.62045	NA	NA	
DSS_0.6255	1.18E-04	NA	
DSS_0.63077	NA	NA	
DSS_0.63973	3.73E-05	NA	
DSS_1.7295	1.40E-04	NA	

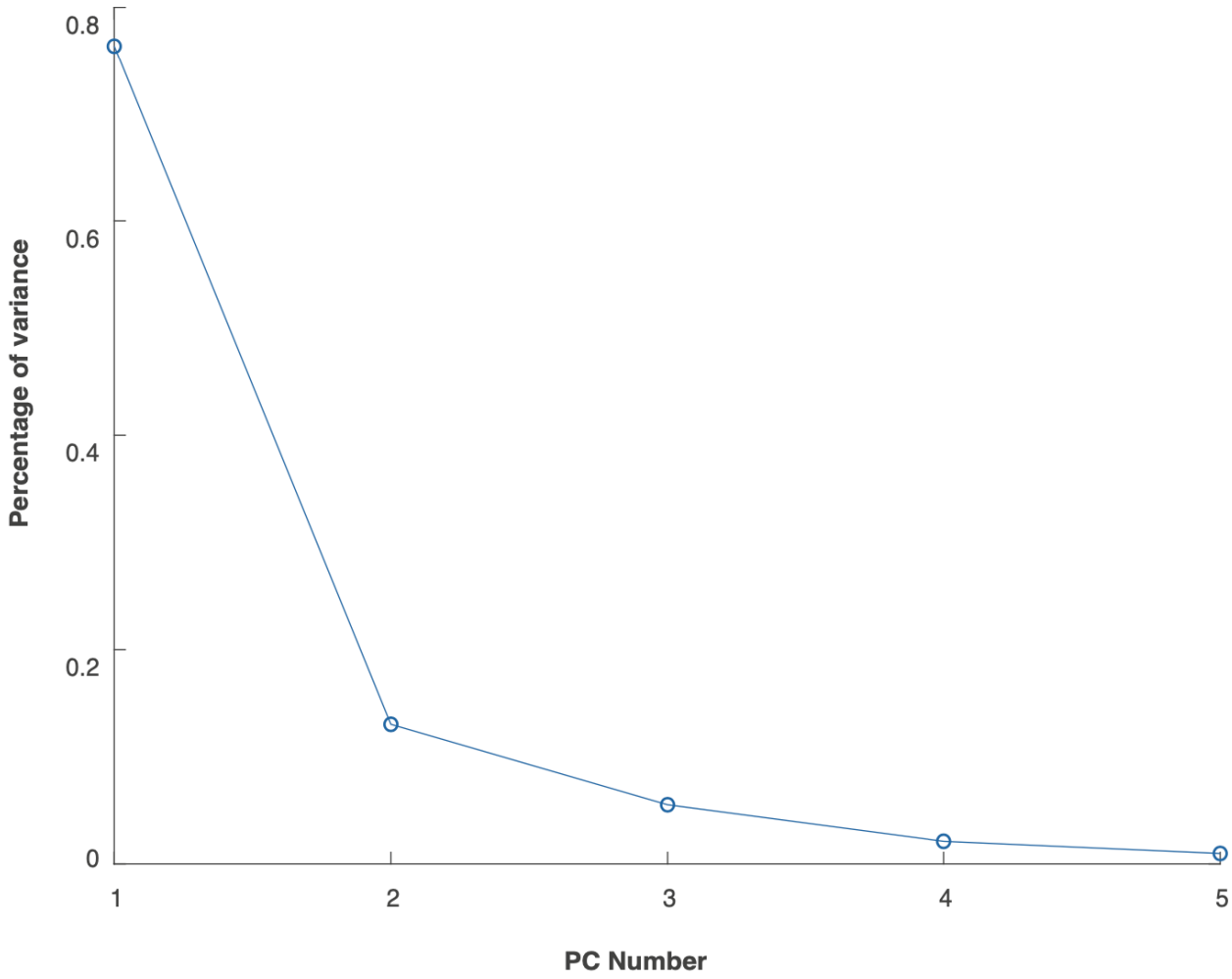
DSS_1.7385	4.24E-04	NA	
DSS_1.7429	3.17E-04	3.17E-04	
DSS_1.7495	1.80E-04	NA	
DSS_1.757	9.51E-05	9.51E-05	
DSS_1.7642	1.81E-04	1.81E-04	
DSS_1.767	NA	NA	
DSS_1.7707	1.75E-04	1.75E-04	
DSS_1.7753	4.55E-04	4.65E-04	
DSS_1.7843	1.08E-04	1.04E-04	
DSS_2.8941	4.24E-05	NA	
DSS_2.9074	5.11E-05	NA	
DSS_2.9205	4.28E-05	NA	
unknown1_1.95	9.85E-04	9.43E-04	
unknown2_4.2	8.80E-04	9.16E-04	
unknown3_4.85	9.75E-04	9.75E-04	
unknown4_3.35	9.72E-04	9.80E-04	

Supplementary Table 3.3: Detailed informtion for compound used in simulation. Each row is for a different compound. Listed information includes Compound name, BMRB Entry ID, the trend of concentration variation, concentration range, whether chemical shift variation (peak shift) from pH is introduced, pKa, chemical shift in protonated form and unprotonated form. For those compound without simulated peak shift in the pH range (pH [4.0 6.0]), no information (NA) is listed on pKa, and chemical shift in protonated form and unprotonated form. Reference spectral can be found in GISSMO database (http://gissmo.nmrfam.wisc.edu).

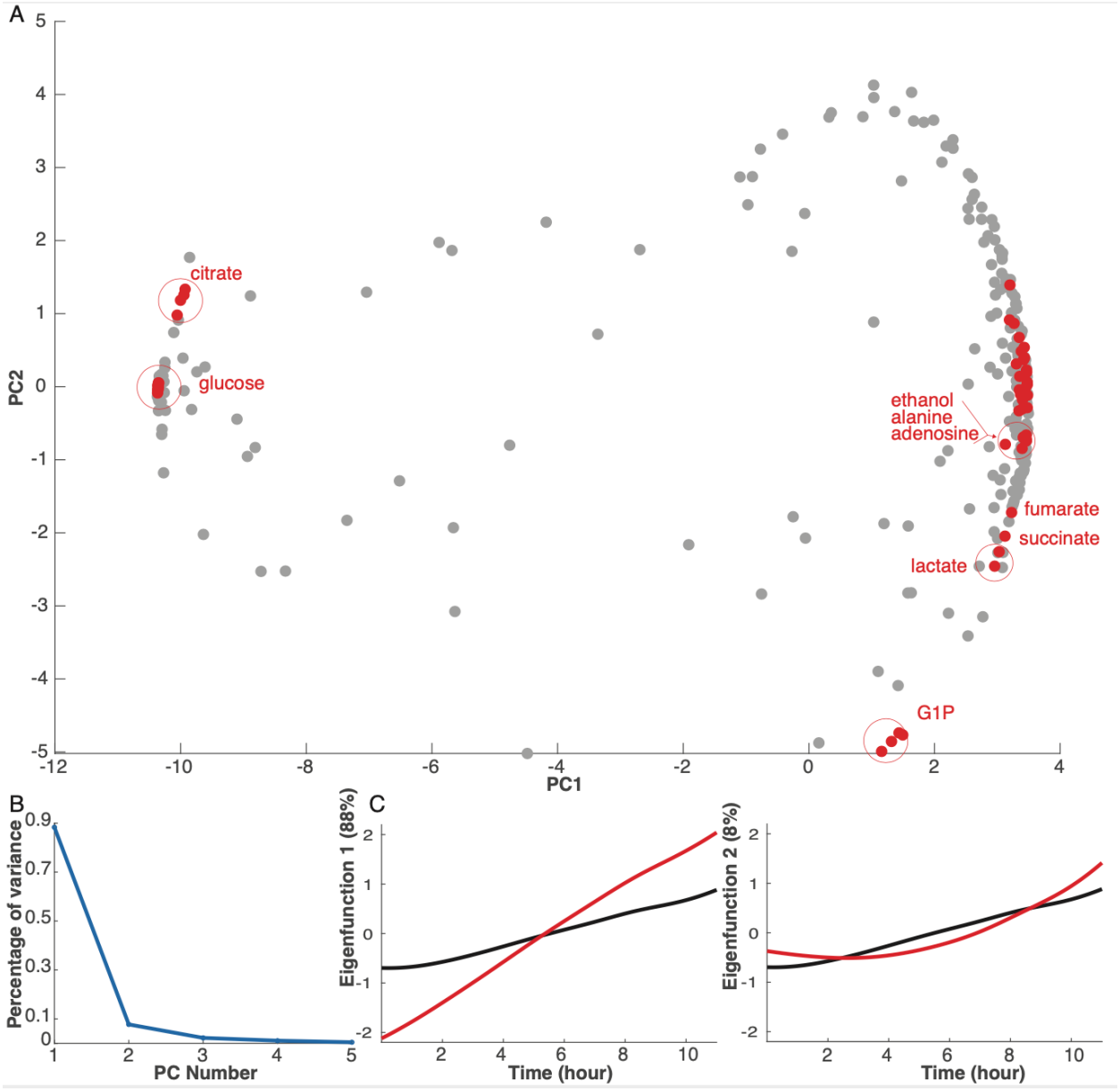
Compound	BMRB Entry ID	Concentration variation	Concentration range	Chemical shift variation	pKa	chemical shift (protonated form)	chemical shift (unprotonated form)
ethanol	bmse000297	Increase	[0 2000]	No	NA	NA	NA
valine	bmse000052	Decrease	[0 200]	No	NA	NA	NA
acetate	bmse000191	Increase	[0 2000]	Yes	4.578	2.089	1.91
glycerol	bmse000184	Decrease	[0 2000]	No	NA	NA	NA
choline	bmse000285	Increase	[0 200]	No	NA	NA	NA
alanine	bmse000028	Decrease	[0 2000]	No	NA	NA	NA
uridine	bmse000158	Increase	[0 2000]	No	NA	NA	NA
leucine	bmse000042	Decrease	[0 200]	No	NA	NA	NA
formate	bmse000203	Increase	[0 2000]	Yes	3.555	8.234	8.448
butanol	bmse000447	Decrease	[0 2000]	No	NA	NA	NA
caffeine	bmse000206	Increase	[0 200]	No	NA	NA	NA
serine	bmse000048	Decrease	[0 2000]	No	NA	NA	NA
purine	bmse000454	Increase	[0 2000]	No	NA	NA	NA
DSS	bmse000795	Constant	2000	No	NA	NA	NA

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

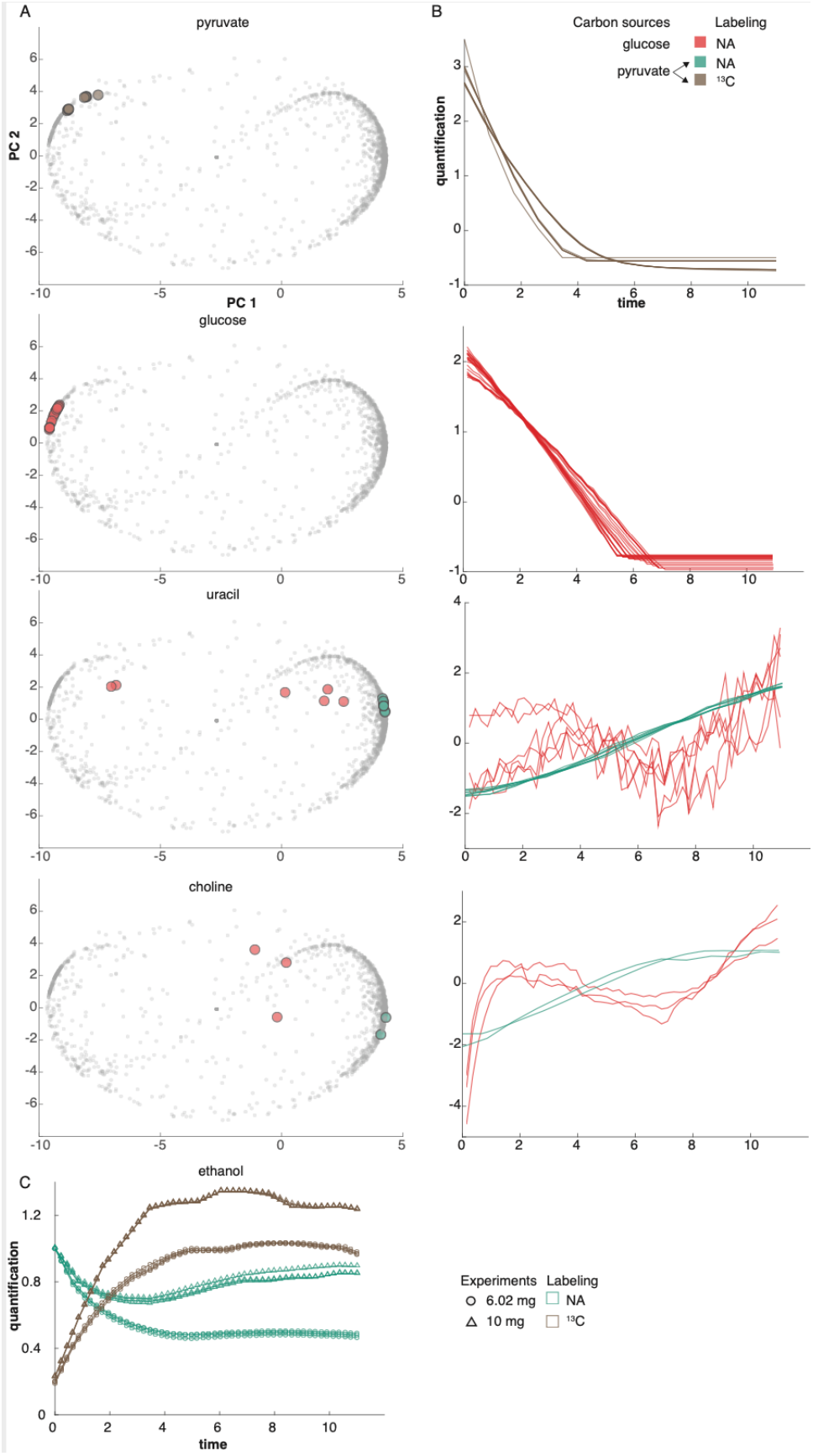


Supplementary Figure 4.1: Percentages of explained variance in FPCA in one aerobic dataset. The X (Y) axis represents the number of PC (percentage of variance). The first two PCs, especially PC1, explain the most variance. The corresponding scores plot and eigenfunction can be found in Fig. 4.2B.

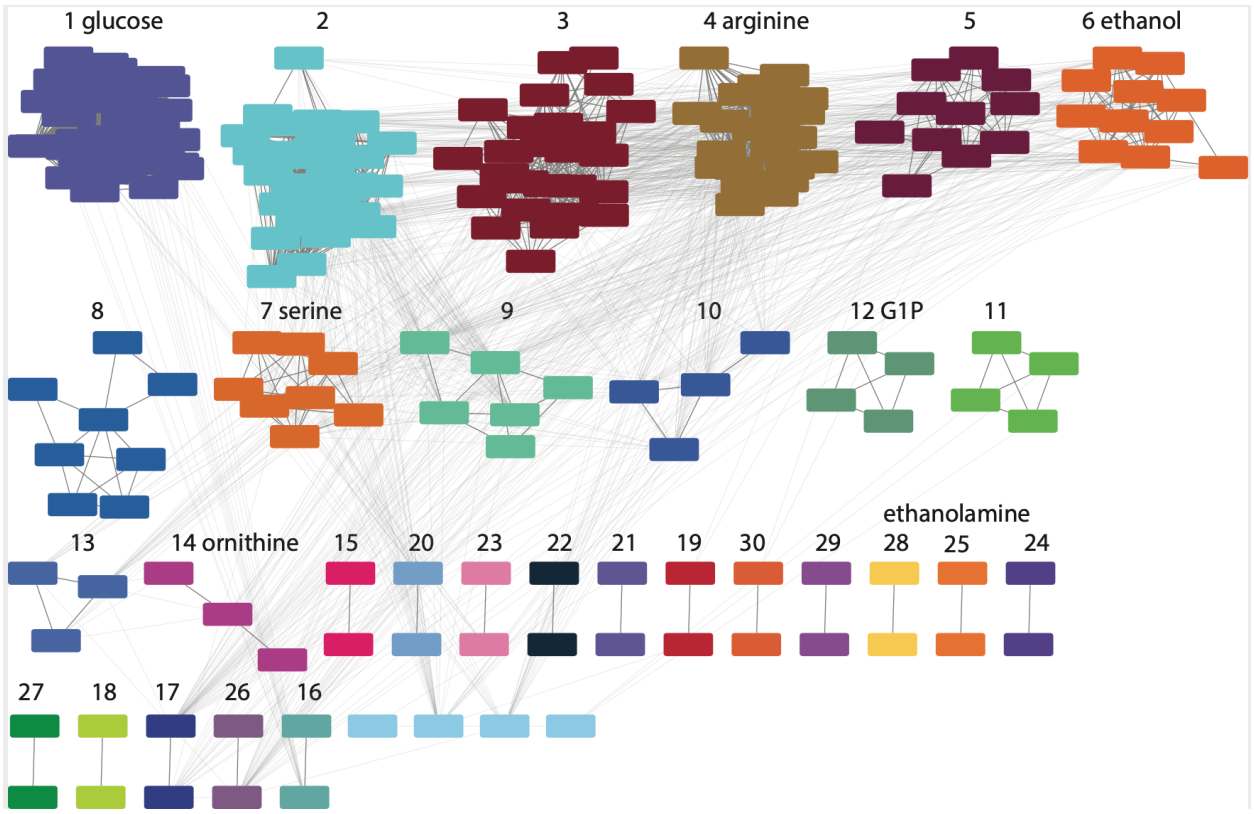


Supplementary Figure 4.2: Dimensionality reduction captures dominant variation in metabolic dynamics in one anaerobic dataset. A: The first two PC dimensions are visualized, and dominant changing patterns are presented. Each point represents the time series of one NMR feature, and some of them are highlighted with compound annotations (red). The X (Y) axis represents scores for PC 1 (2). B: Percentages of explained variance are presented for the first few

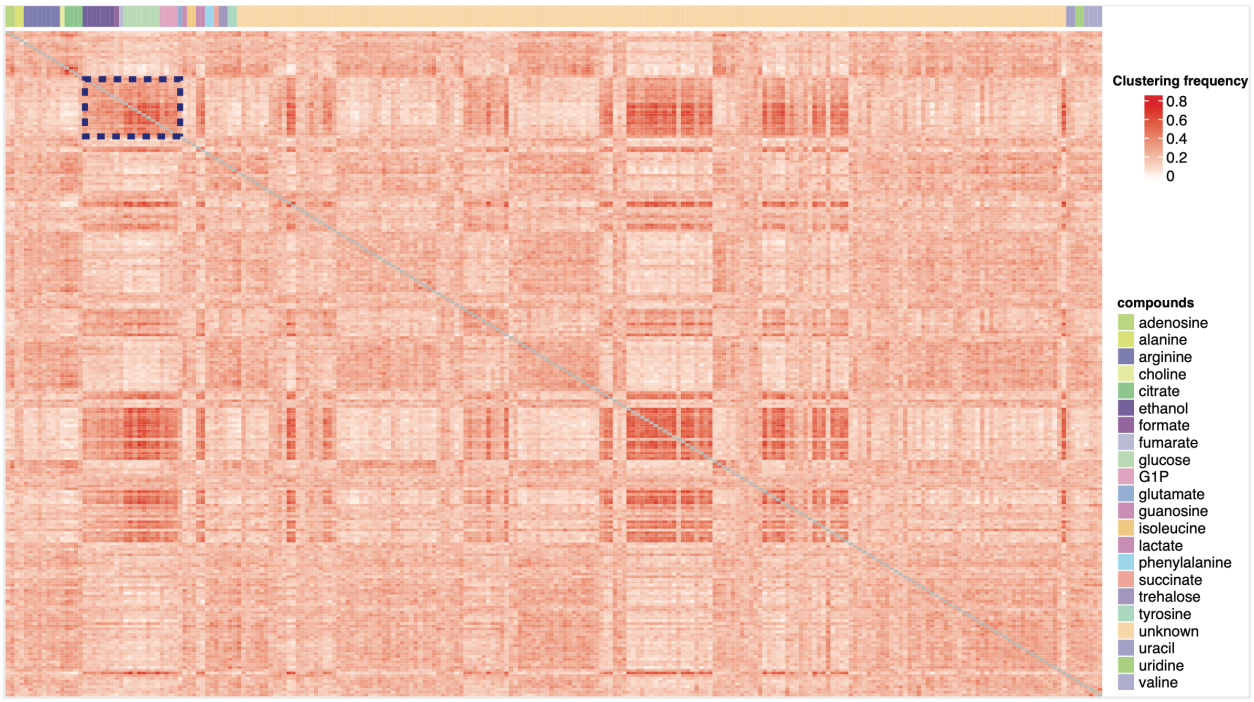
PCs. The X (Y) axis represents the number of PC (percentage of variance). The first two PCs, especially PC1, explain most variances. C: Eigenfunctions are plotted for PC1 and PC2. The middle black curve represents the mean time series; the red curve represents the effects of adding a fraction (square root of eigenvalue) of the corresponding eigenfunctions to the mean curve. The X (Y) axis represents time (value). The percentages of variance explained are presented in parentheses. NMR features were centered and scaled before the PCA analysis. Results for aerobic conditions can be found in Fig. 4.2 and Supplementary Fig. 4.1.


A
pyruvate
PC 2
PC 1
glucose
uracil
choline
ethanol
B
Carbon sources
Labeling
glucose
NA
pyruvate
NA
13C
quantification
time
C
quantification
time
Experiments
6.02 mg
10 mg
Labeling
NA
13C

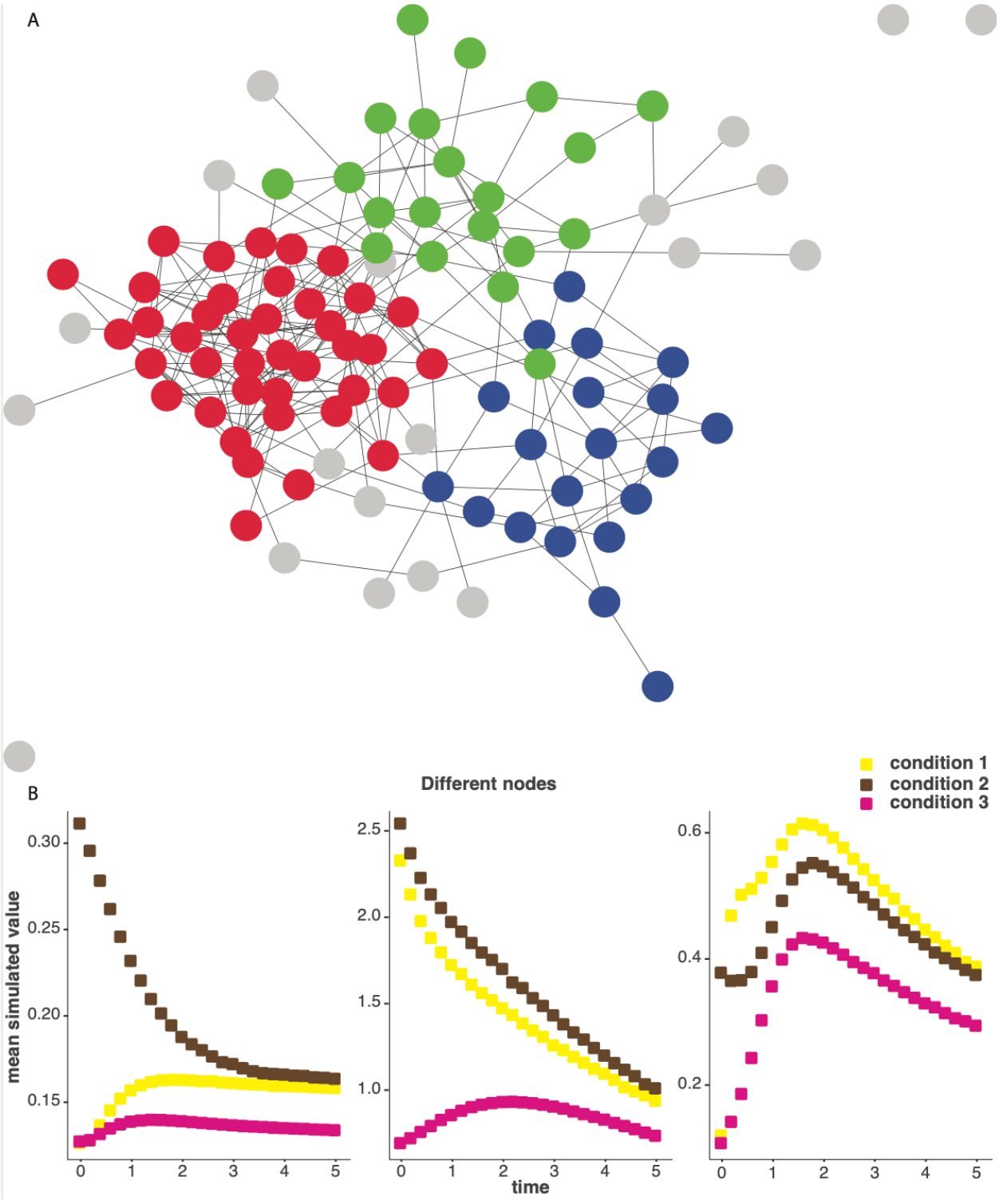
Supplementary Figure 4.3: Additional comparison of metabolic profiles under different carbon sources through FPCA. Two different carbon sources were compared: natural abundance glucose (3 experiments) and uniformly ^{13}C-labeled pyruvate (2 experiments). The glucose experiments were all done at a high density (10 mg/63 mL), and the pyruvate experiments were done at low (6mg /63 mL, solid lines) and high (10 mg/63 mL, dashed lines) densities. The chemical features from the glucose experiments are shown in red. The ^{13}C-labeled metabolites produced in the ^{13}C-pyruvate experiments are shown in brown. The unlabeled metabolites produced in the ^{13}C-pyruvate experiments are shown in green. A: FPCA score plot indicates the overall patterns of different compounds in each of these experiments. Each point represents one ridge in one sample. The small, grey points correspond to all the other ridges detected in these experiments. The X (Y) axis represents scores for PC 1 (2). The score plot was kept the same with different compounds highlighted. B: Time trajectory (hours) of the highlighted features from A. Each curve was centered and scaled for A and B. C: Detailed comparison of ethanol in the ^{13}C-pyruvate experiments. Two experiments with different amounts of organisms are visualized by different shapes. ^{13}C-labeled and unlabeled ethanol are distinguished by colors. Normalization was applied to make them comparable. Ridges of triplets around 1.1 ppm were used to quantify ethanol in C, and all tracked ridges were used in Fig. 4.3A-B.



Supplementary Figure 4.4: Clustering of the correlation network helps compound annotation. A correlation network was built upon time-series features in the glucose feeding experiments, and clusters were found (More details in Methods). Each cluster is highlighted by a different color and assigned one number. Single nodes are in blue. Details of specific clusters are shown in Fig. 4.4.

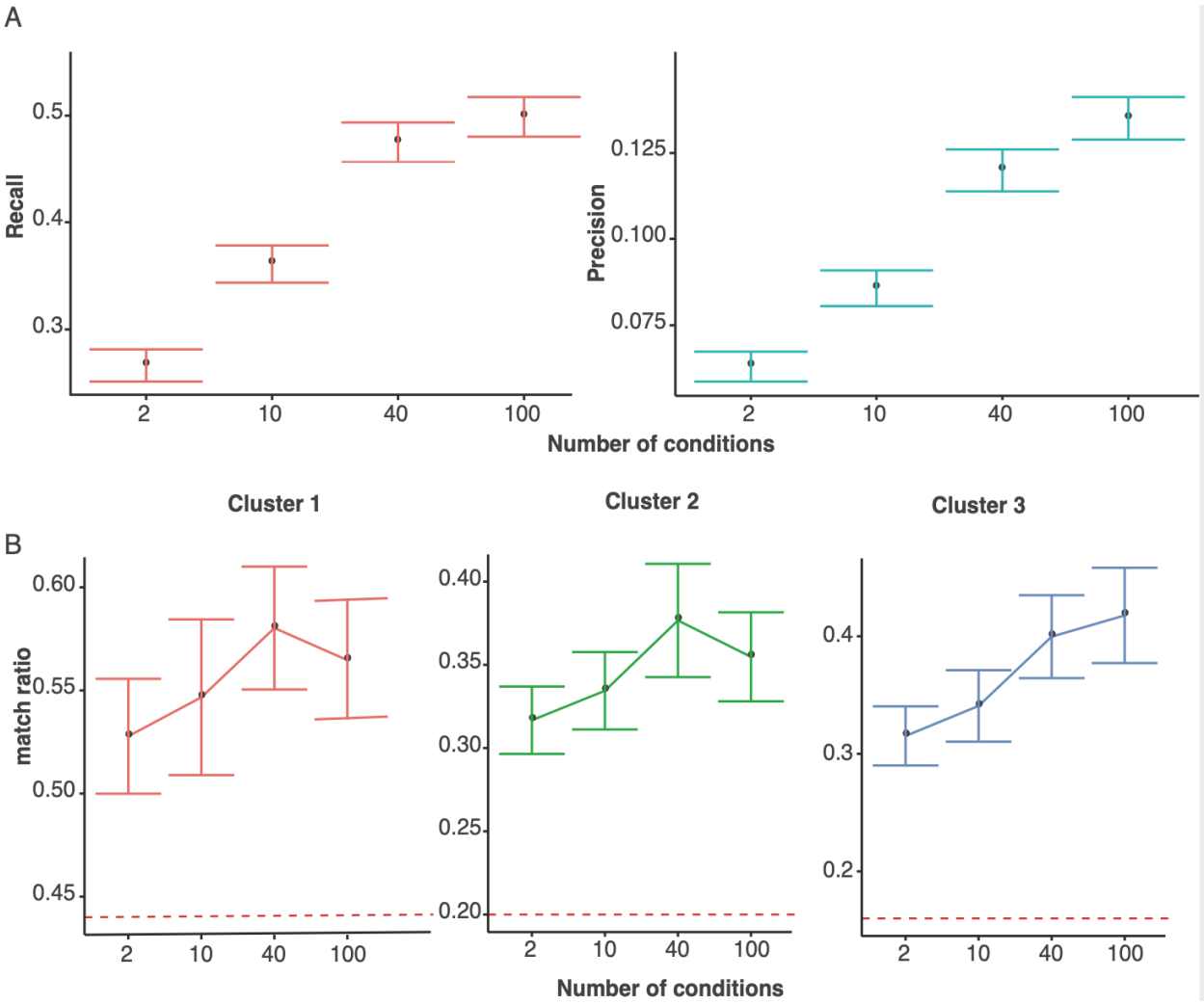


Supplementary Figure 4.5: Consistency of functional networks was evaluated by clustering frequency. Colors in the heatmap represent relative frequencies that two NMR features share the same clusters in bootstrapping. Each row or column represents one feature. Red (white) indicates more (less) co-occurrences of the two features. The top bar indicates different compounds by colors. The dashed box highlights nodes of glucose, ethanol and G1P that are frequently presented in the same cluster. Diagonal values are not presented, as the same feature will always be in the same cluster. The bootstrapping results are also visualized in the network in Fig. 4.5B.



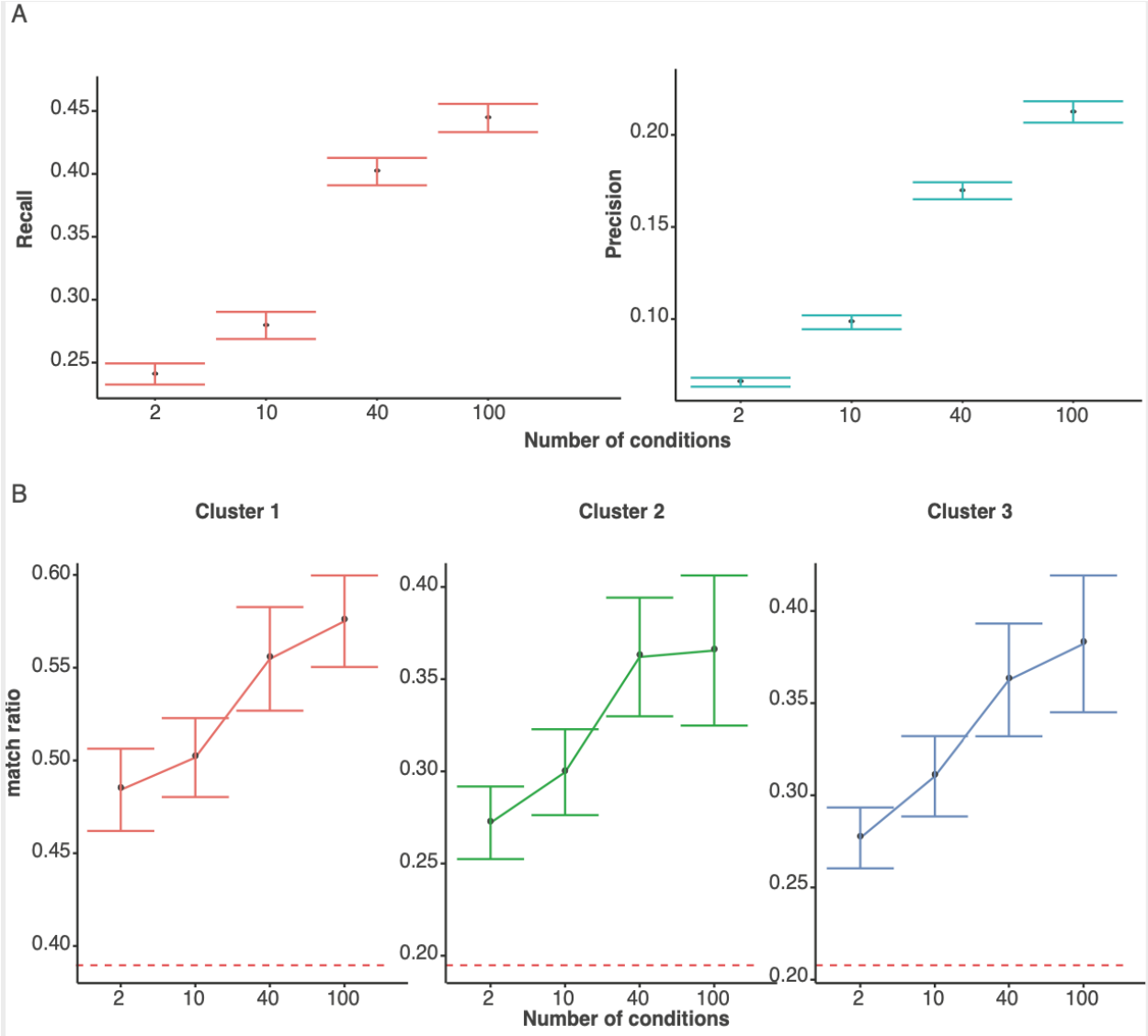
Supplementary Figure 4.6: Time-series dynamics were simulated from random networks. One example random network and corresponding dynamics are presented. A: The random network (with 100 nodes) was simulated with

clusters. Red, green and blue nodes indicate three simulated clusters where internal links are denser than inter-cluster links. Gray nodes do not belong to any clusters. B: Time-series dynamics were simulated for each node under different initial conditions. The X (Y) axis indicates time (mean simulated value). Trajectories of three different nodes under three different conditions are presented. The mean value was calculated from three different replicates with the same nodes and conditions but different random noise. More detail on random networks and dynamics simulation can be found in Methods and Supplementary File 4.1.



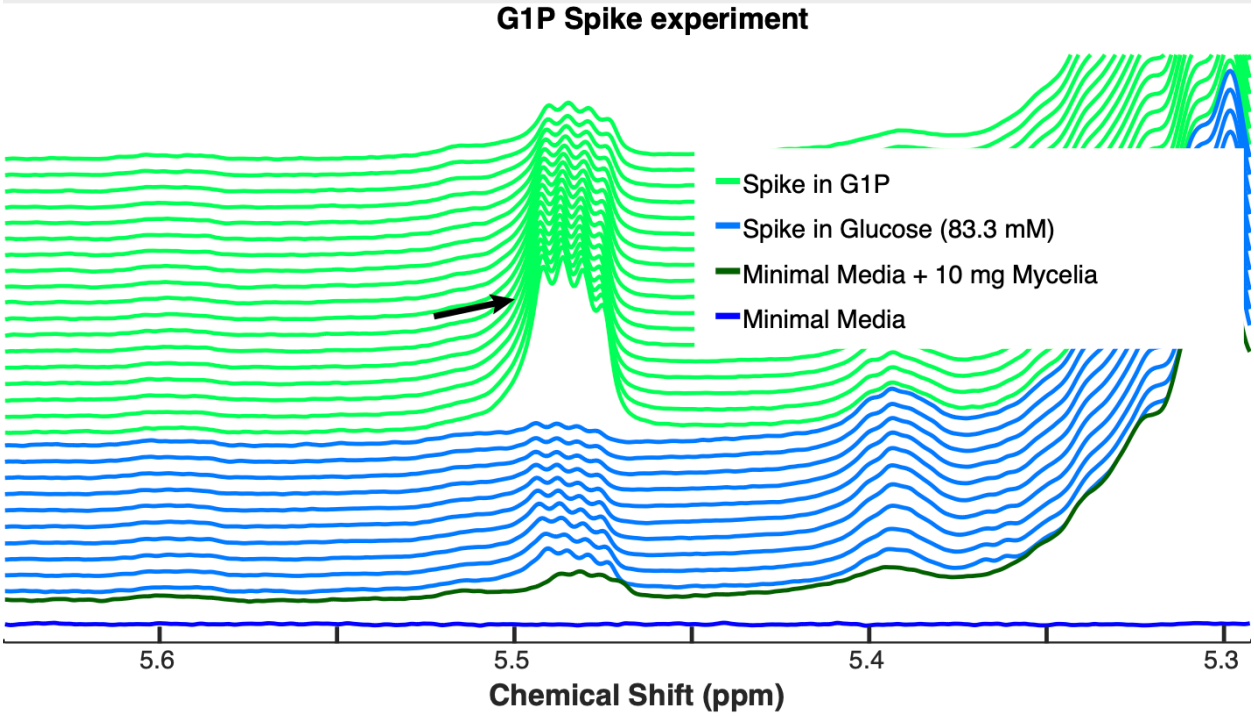
Supplementary Figure 4.7: Model performances were benchmarked on a simulated dataset with partial observation. Our workflow was evaluated on a simulated benchmark dataset with partial observation. In the random network, each node (100 in total) represents one compound, and each cluster is a set of compounds with denser inner cluster connections. Time dynamics were simulated based on the networks through ODEs under different initial conditions. A subset of the time-series features was observable and clustered (More details in Methods). The performance in estimating edge and recovering clusters was evaluated. A: Recall and precision (Y-axis) for edge estimation are presented

under different numbers of initial conditions (X-axis). Edge estimation improves with more conditions. B: The performance in cluster recovering is presented for the three clusters under a different number of conditions. The X (Y) axis represents the number of conditions (match ratio). The match ratio is the proportion of nodes from the matched real cluster in the best-recovered cluster (More details in Methods). The red dotted lines indicate the match ratio of a random group of nodes (baseline). The estimated clusters can recover more nodes from real clusters, and the performance improves with more conditions. The error bars represent two standard errors calculated from the simulation and reconstruction of 59 random networks. Simulated random networks and example time series can be found in Supplementary Fig. 4.6. Simulation-based evaluation with redundant signals can be found in Supplementary Fig. 4.8. The performance on an experimental dataset can be found in Fig. 4.5B and Supplementary Fig. 4.5.



Supplementary Figure 4.8: Model performances were benchmarked on a simulated dataset with partial observation and redundant signals. Our workflow was evaluated on a simulated benchmark dataset with partial observation and redundant signals. In the random network, each node (100 in total) represents one compound, and each cluster is a set of compounds with denser inner cluster connections. Time dynamics were simulated based on the networks through ODEs under different initial conditions. Time-series features were also expanded and scaled by random factors as an analogy of multiple

peaks corresponding to the same compound in NMR. A subset of the time-series features was observable and clustered (More details in Methods). The performance in estimating edge and recovering clusters was evaluated. A: Recall and precision (Y-axis) for edge estimation are presented under different numbers of initial conditions (X-axis). Edge estimation improves with more conditions. B: The performance in cluster recovering is presented for the three clusters under a different number of conditions. The X (Y) axis represents the number of conditions (match ratio). The match ratio is the proportion of nodes from the matched real cluster in the best-recovered cluster (More details in Methods). The red dotted lines indicate the match ratio of a random group of nodes (baseline). The estimated clusters can recover more nodes from real clusters, and the performance improves with more conditions. The error bars represent two standard errors calculated from the simulation and reconstruction of 59 random networks. Simulated random networks and example time series can be found in Supplementary Fig. 4.6. Simulation-based evaluation with no redundant signals can be found in in Supplementary Fig. 4.7. The performance on the experimental dataset can be found in Fig. 4.5B and Supplementary Fig. 4.5.



Supplementary Figure 4.9: The G1P Spiking experiment showed a level 5 annotation. Multiple contrasting experiments were collected through time: minimal media (Judge et al. 2019), minimal media and 10 mg mycelia, spiking glucose in the culture, and spiking G1P in the culture. The X-axis represents chemical shift. The arrow shows the G1P peaks, which increased after G1P spiking.

Supplementary Table 4.1: Neighbors of annotated peaks in central energy metabolism. Neighbors of selected driver nodes in the CausalKinetiX network were listed (Fig. 4.5B). The searching starts from glucose, G1P, ethanol, alanine, and lactate. Columns include driver (the searching start node), annotation (the annotation of the neighbor), ppm (chemical shifts of the neighbor nodes), edge_support (whether the edge is supported by bootstrapping), and cluster (which cluster the neighbor node belongs to).

driver	annotation	ppm	edge_support	cluster
glucose	unknown-179	2.2169	TRUE	4
glucose	unknown	4.0237	FALSE	4
glucose	serine (overlapped)	3.9754	FALSE	2
glucose	serine (overlapped)	3.9478	FALSE	3
glucose	unknown	3.815	FALSE	3
glucose	glucose (overlapped)	3.4817	TRUE	4
glucose	glucose (overlapped)	3.2211	TRUE	4
glucose	unknown	2.4064	TRUE	4
glucose	unknown	1.29	FALSE	1
glucose	unknown	1.1077	TRUE	4
glucose	lactate	1.318	FALSE	4
glucose	glucose-1-phosphate	5.4781	TRUE	4
glucose	glucose-1-phosphate	5.4723	TRUE	4
glucose	unknown	1.1291	FALSE	3
glucose	glucose-1-phosphate	5.484	TRUE	4
glucose	alanine	1.4763	FALSE	4
glucose	alanine	1.4643	FALSE	4
glucose	unknown	4.1984	FALSE	4
glucose	unknown	4.3159	FALSE	2
glucose	unknown	4.2278	FALSE	2
glucose	glucose (overlapped)	3.7066	TRUE	4
glucose	serine (overlapped)	3.9572	FALSE	1
glucose	glucose (overlapped)	3.5389	FALSE	4
glucose	unknown	3.3187	FALSE	4
glucose	phenylalanine	7.4206	FALSE	3
glucose	ethanolamine (overlapped)	3.1479	FALSE	3
glucose	unknown	1.0803	FALSE	1
glucose	unknown	3.5658	FALSE	1
glucose	uracil	5.8016	FALSE	3
glucose	serine (overlapped)	3.9692	FALSE	3
glucose	glucose (overlapped)	3.745	FALSE	3
glucose	glucose (overlapped)	3.5225	TRUE	4
glucose	glucose (overlapped)	3.4057	TRUE	1
glucose	unknown	2.8307	FALSE	4
glucose	unknown	1.9916	FALSE	4
glucose	unknown	1.0685	FALSE	4
glucose	citrate	2.7226	FALSE	1
g1p	unknown	5.0796	FALSE	3
g1p	adenosine (overlapped)	6.0616	FALSE	3
g1p	unknown	4.1984	FALSE	4
g1p	unknown	3.815	FALSE	3
g1p	glucose (overlapped)	3.7131	TRUE	4
g1p	unknown	3.6027	FALSE	4
g1p	unknown	3.5508	FALSE	3
g1p	glucose (overlapped)	3.5326	TRUE	4
g1p	glucose (overlapped)	3.4969	TRUE	3
g1p	glucose (overlapped)	3.4664	TRUE	4
g1p	glucose (overlapped)	3.4508	TRUE	4
g1p	glucose (overlapped)	3.4817	TRUE	4
g1p	glucose (overlapped)	3.4109	TRUE	4
g1p	unknown	2.9159	FALSE	4
g1p	unknown	2.8307	FALSE	4
g1p	unknown	2.4064	FALSE	4
g1p	unknown	2.0986	FALSE	1
g1p	arginine (overlapped)	1.904	FALSE	3
g1p	unknown	1.2782	FALSE	1
g1p	unknown	1.1195	TRUE	2
g1p	unknown	1.1077	TRUE	4
g1p	unknown	1.0803	FALSE	1
g1p	unknown	0.87654	FALSE	1
g1p	succinate	2.4735	FALSE	3
g1p	ethanol	3.6571	FALSE	4
g1p	fumarate	6.5246	FALSE	3
g1p	citrate	2.615	FALSE	1
g1p	glucose-1-phosphate	5.4723	FALSE	4
g1p	glucose	4.6318	TRUE	4
g1p	glucose	3.8853	TRUE	4
g1p	ethanol	3.6453	FALSE	4
g1p	ethanol	3.6335	FALSE	4
g1p	alanine	1.4763	FALSE	4
g1p	ethanol	1.175	FALSE	4
g1p	adenosine	8.2584	FALSE	3
g1p	unknown-179	2.2169	TRUE	4
g1p	unknown	5.2998	FALSE	1
g1p	unknown	5.3953	FALSE	1

g1p	glucose (overlapped)	3.8333	FALSE	4
g1p	glucose (overlapped)	3.8194	TRUE	4
g1p	glucose (overlapped)	3.811	TRUE	4
g1p	glucose (overlapped)	3.8074	TRUE	4
g1p	glucose (overlapped)	3.8282	TRUE	4
g1p	unknown	3.7885	FALSE	3
g1p	glucose (overlapped)	3.7738	TRUE	4
g1p	glucose (overlapped)	3.7535	TRUE	4
g1p	glucose (overlapped)	3.745	FALSE	3
g1p	glucose (overlapped)	3.7336	TRUE	4
g1p	glucose (overlapped)	3.7066	TRUE	4
g1p	glucose (overlapped)	3.7036	FALSE	2
g1p	glucose (overlapped)	3.7238	TRUE	4
g1p	unknown	3.6786	FALSE	4
g1p	unknown	3.5594	FALSE	1
g1p	glucose (overlapped)	3.5225	FALSE	4
g1p	glucose (overlapped)	3.4611	TRUE	4
g1p	glucose (overlapped)	3.4545	TRUE	4
g1p	glucose (overlapped)	3.4448	TRUE	4
g1p	glucose (overlapped)	3.4413	TRUE	4
g1p	glucose (overlapped)	3.4212	TRUE	3
g1p	glucose (overlapped)	3.4057	TRUE	1
g1p	glucose (overlapped)	3.3894	FALSE	1
g1p	glucose (overlapped)	3.3796	TRUE	4
g1p	glucose (overlapped)	3.3957	TRUE	4
g1p	unknown	3.3384	FALSE	2
g1p	glucose (overlapped)	3.2499	FALSE	3
g1p	glucose (overlapped)	3.2366	FALSE	4
g1p	ethanolamine (overlapped)	3.1392	FALSE	3
g1p	unknown	2.9621	FALSE	1
g1p	unknown	2.2371	FALSE	1
g1p	unknown	1.8582	FALSE	3
g1p	unknown	1.5669	FALSE	1
g1p	unknown	1.3748	FALSE	2
g1p	unknown	1.1488	FALSE	3
g1p	lactate	1.3296	FALSE	4
g1p	glucose-1-phosphate	5.484	FALSE	4
g1p	glucose	5.2299	TRUE	4
g1p	glucose	5.2237	TRUE	4
g1p	ethanol	1.1868	FALSE	4
g1p	glucose	4.645	FALSE	4
g1p	glucose	3.9057	TRUE	4
g1p	glucose	3.9022	FALSE	4
g1p	glucose	3.8817	TRUE	4
g1p	ethanol	1.1632	FALSE	4
g1p	alanine	1.4643	FALSE	4
g1p	uracil	5.7888	FALSE	1
g1p	uracil (overlapped)	7.5193	FALSE	3
g1p	unknown	5.894	FALSE	3
g1p	unknown	4.0355	FALSE	3
g1p	glucose (overlapped)	3.6908	FALSE	3
g1p	glucose (overlapped)	3.5162	TRUE	4
g1p	unknown	3.2894	FALSE	3
g1p	ethanolamine (overlapped)	3.1305	FALSE	2
g1p	ethanolamine (overlapped)	3.1479	FALSE	3
g1p	unknown	2.0447	FALSE	1
g1p	unknown	1.9378	FALSE	3
g1p	arginine (overlapped)	1.7537	FALSE	4
g1p	arginine (overlapped)	1.7092	FALSE	3
g1p	unknown	1.3628	FALSE	1
g1p	trehalose	5.1815	FALSE	2
g1p	lactate	1.318	FALSE	4
g1p	arginine	1.648	FALSE	1
ethanol	unknown	5.8865	FALSE	3
ethanol	unknown	4.1984	FALSE	4
ethanol	glucose (overlapped)	3.4508	FALSE	4
ethanol	glucose (overlapped)	3.3796	FALSE	4
ethanol	unknown	2.8307	FALSE	4
ethanol	unknown	2.4289	FALSE	3
ethanol	unknown	1.9916	FALSE	4
ethanol	unknown-179	2.2169	TRUE	4
ethanol	unknown	5.2998	FALSE	1
ethanol	glucose (overlapped)	3.2211	FALSE	4
ethanol	unknown	2.4064	FALSE	4
ethanol	glucose-1-phosphate	5.484	FALSE	4
ethanol	glucose-1-phosphate	5.4723	FALSE	4

ethanol	adenosine	8.3425	FALSE	3
ethanol	glucose (overlapped)	3.7738	FALSE	4
ethanol	unknown	2.3621	FALSE	3
ethanol	arginine (overlapped)	1.8925	FALSE	3
ethanol	fumarate	6.5246	FALSE	3
ethanol	alanine	1.4643	FALSE	4
ethanol	glucose (overlapped)	3.4969	FALSE	3
ethanol	glucose (overlapped)	3.4664	FALSE	4
ethanol	unknown	2.0033	FALSE	4
ethanol	citrate	2.6404	FALSE	1
ethanol	arginine	1.6314	FALSE	1
ethanol	unknown	4.0237	FALSE	4
ethanol	glucose (overlapped)	3.4109	TRUE	4
ethanol	glucose-1-phosphate	5.4781	FALSE	4
alanine	uracil	5.7888	FALSE	1
alanine	unknown	5.9765	FALSE	2
alanine	unknown	5.8865	FALSE	3
alanine	unknown	4.0609	FALSE	2
alanine	glucose (overlapped)	3.8463	FALSE	2
alanine	glucose (overlapped)	3.8194	FALSE	4
alanine	glucose (overlapped)	3.8074	FALSE	4
alanine	glucose (overlapped)	3.7131	FALSE	4
alanine	glucose (overlapped)	3.7066	FALSE	4
alanine	unknown	3.6151	FALSE	1
alanine	unknown	3.6027	FALSE	4
alanine	unknown	3.5911	FALSE	1
alanine	glucose (overlapped)	3.5326	FALSE	4
alanine	glucose (overlapped)	3.5162	FALSE	4
alanine	glucose (overlapped)	3.4969	FALSE	3
alanine	glucose (overlapped)	3.4664	FALSE	4
alanine	glucose (overlapped)	3.4545	FALSE	4
alanine	glucose (overlapped)	3.4413	FALSE	4
alanine	glucose (overlapped)	3.4817	FALSE	4
alanine	glucose (overlapped)	3.4212	FALSE	3
alanine	glucose (overlapped)	3.4109	FALSE	4
alanine	glucose (overlapped)	3.3957	FALSE	4
alanine	glucose (overlapped)	3.2366	FALSE	4
alanine	glucose (overlapped)	3.2211	FALSE	4
alanine	unknown	2.4064	FALSE	4
alanine	unknown	1.1423	FALSE	1
alanine	unknown	1.1077	FALSE	4
alanine	unknown	1.1291	FALSE	3
alanine	unknown	1.0685	FALSE	4
alanine	trehalose	5.1815	FALSE	2
alanine	glucose-1-phosphate	5.4723	FALSE	4
alanine	glucose	4.6318	FALSE	4
alanine	glucose	3.8853	FALSE	4
alanine	glucose	3.8817	FALSE	4
alanine	ethanol	3.6453	FALSE	4
alanine	ethanol	3.6335	FALSE	4
alanine	ethanol	1.1632	FALSE	4
alanine	choline	3.1895	FALSE	3
alanine	alanine	1.4763	FALSE	4
alanine	uridine	7.8552	FALSE	1
alanine	uracil	5.8016	FALSE	3
alanine	glucose (overlapped)	3.8333	FALSE	4
alanine	glucose (overlapped)	3.7336	FALSE	4
alanine	glucose (overlapped)	3.7238	FALSE	4
alanine	unknown	3.4571	FALSE	3
alanine	glucose (overlapped)	3.2499	FALSE	3
alanine	unknown	2.8307	FALSE	4
alanine	unknown	2.3687	FALSE	4
alanine	glucose-1-phosphate	5.484	FALSE	4
alanine	glucose	3.9057	FALSE	4
alanine	glucose	3.9022	FALSE	4
alanine	alanine	1.4643	FALSE	4
lactate	unknown-179	2.2169	TRUE	4
lactate	unknown	5.9765	FALSE	2
lactate	unknown	2.3687	FALSE	4
lactate	unknown	2.4064	TRUE	4
lactate	unknown	1.2782	FALSE	1
lactate	unknown	1.0803	FALSE	1
lactate	trehalose	5.1815	FALSE	2
lactate	glucose	5.2237	FALSE	4
lactate	glucose-1-phosphate	5.4781	FALSE	4
lactate	unknown	5.2998	FALSE	1

lactate	unknown	2.8307	FALSE	4
lactate	unknown	1.9378	FALSE	3
lactate	glucose-1-phosphate	5.4723	FALSE	4

Supplementary File 4.1: Performance Evaluation Based on Benchmarking Dataset

Background of the simulation test

Our work also adds to topology estimation based on time-series data (A. Al-Omari et al. 2018; Meyer et al. 2014). Different approaches in topology (edge) and parameter estimation in biological dynamical systems have been compared in the DREAM7 competition (Meyer et al. 2014). This early benchmarking study focused on networks with few hidden connections and limited topological diversity. The experimental dataset provides a useful but limited evaluation of network construction and cluster. Collecting time series data with multiple perturbation conditions is also expensive.

Here, we simulated benchmarking datasets with a considerable number of unknown connections, extensive partial observation and signal duplication, which better represents exploratory metabolomic experiments. Partial observation represents the absence of data for many metabolites in time-series profiling experiments (Judge et al. 2019; Koczula et al. 2016; Link et al. 2015). Duplicated signals arise in NMR spectra, where the same compound produces multiple peaks.

Performance evaluation based on benchmarking dataset

Metabolites (nodes) are connected through reactions and regulation (edges), and there are multiple functional clusters in biochemical pathways. To resemble the pathway architecture, we simulated random networks with clusters, and nodes were more frequently connected within clusters (Supplementary Fig.

4.6A; More details in Methods). The three clusters represent groups of nodes related to three different functions. From the random network, time-series dynamics were simulated based on different initial conditions (Supplementary Fig. 4.6B), and they were the benchmark input.

Edge estimation improves with more experimental conditions (Supplementary Figs 4.7A and 4.8A). In the simulation with partial observation and redundant signals (Supplementary Fig. 4.8A), precision and recall improve monotonically with more conditions. For dynamics generated from unknown network topology, recall (precision) can achieve around 45% (20%) with 100 conditions. Even with a low number of conditions, about 25% of real edges can be recovered though precision is low. As a comparison, in totally random cases, recall and precision should be about 2%. For the simulation with no redundant signals (Supplementary Fig. 4.7A), there are similar patterns in precision and recall though the improvement seems to plateau at around 100 conditions. Experimental measured metabolic data (e.g., NMR) have redundant signals for the same compounds so should still improve with more conditions.

The performance of recovering functional clusters also improves with more conditions (Supplementary Fig. 4.7B and 4.8B). The match ratio represents the proportion of nodes from the matched real cluster in the best recovered cluster. The ratios improve with more conditions for both simulations, even though the variance is still high with our number of simulation samples (59). The pattern is consistent for different cluster sizes, and they are all significantly higher than the random guess (Wilcox t-test $p<0.01$). For a cluster with 20% of nodes (total 100

nodes) and 100 conditions, on average more than 35% nodes of the best cluster are from the real cluster (Supplementary Fig. 4.8B), and those recovered real nodes represent around 40% of the real cluster.