

**Decomposition and Feature Selection of Comprehensive 2-Dimensional
Gas Chromatography - Time-of-Flight Mass Spectrometry
(GC×GC-TOFMS) Data**

by

Michael Drew Sorochan Armstrong

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry
University of Alberta

© Michael Drew Sorochan Armstrong, 2021

Abstract

Comprehensive Two-Dimensional Gas Chromatography - Time-of-Flight Mass Spectrometry (GC \times GC-TOFMS) is an advanced instrumental technique that separates complex mixtures along two chromatographic dimensions, followed by multivariate detection that collects mass spectral information at a high acquisition rate. GC \times GC-TOFMS improves upon the sensitivity and selectivity of traditional Gas Chromatography - Mass Spectrometry (GC-MS), and as such many more chemicals can be identified and quantified within a much shorter span of time.

Current commercial offerings, and some academic works have largely focused on capitalising upon the sensitivity and selectivity of GC \times GC-TOFMS in order to find more chemical components per chromatogram, often achieved by removing interfering noise from the signal and digging far into the Signal-to-Noise Ratio (SNR). For experiments where it is necessary to correlate some observable characteristic of the samples being analysed with the chemical information available in the GC \times GC-TOFMS chromatograms, this usually creates far more features than samples. This is a common problem in the practice of chemometrics, and there are a number of feature selection routines and rank-deficient solutions to the inverse least squares problem that can correct for this inequality of variables to samples. However, a problem arises when these features are poorly integrated and/or associated across multiple samples. This has been a persistent and known problem within the chromatography community for years, and while it remains an active area of research, little has been done to develop an algorithm to properly quantify and identify these chemical components without excessive programmatic steps that are prone to failure.

The main issue surrounding this problem is the fact that chemical components often drift between runs along both their first- and second-dimension retention modes. Although chemometricians have been using Parallel Factor Analysis 2 (PARAFAC2) to model chromatographic drift along one mode for decades, thus far, no algorithm has been developed to handle drift in two modes using a similarly mathematically satisfying way.

In this work, I present improvements to the Feature Selection by Cluster Resolution (FS-CR) algorithm that enables high quality information to be extracted from peak tables with a number of integration artefacts such that many more combinations of data can be analysed in a much shorter span of time; generally improving upon the feature selection routine. This algorithm was tested upon a number of datasets, most of which were created during the course of this research. Following this, a parsimonious solution for the analysis of GC \times GC-TOFMS data with drift in two modes will be proposed, named PARAFAC2 \times 2. Within a particular region of the chromatogram, this algorithm appears capable of deconvolving components with drift that varies across each sample independently, under close to the worst conditions possible. To the end of creating a parameter-free pre-processing routine for entire chromatograms, a novel method for predicting the chemical rank of a matrix will be proposed. This may enable automated, parameter-free processing of raw GC \times GC-TOFMS data sometime in the near future.

Preface

Chapter 2 has been published as: M. S. Armstrong, A. P. de la Mata, and J. J. Harynuk, “An efficient and accurate numerical determination of the cluster resolution metric in two dimensions” *Journal of Chemometrics*, e3346

Chapter 3 has been submitted for publication as: M. S. Armstrong, O. R. Arredondo Campos, C. C. Bannon, A. P. de la Mata, R. J. Case, and J. J. Harynuk, “Global metabolome analysis of *Dunaliella tertiolecta*, *Phaeobacter italicus* R11 co-cultures using thermal desorption - comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry” *Journal of Phytochemistry*

Chapter 5 has been submitted for publication as: M. S. Armstrong, J. L. Henrich, O. R. Arredondo Campos, A. P. de la Mata and J. J. Harynuk, “PARAFAC2×N: coupled decomposition of multi-modal data with drift in N modes” *Chemometrics and Intelligent Laboratory Systems*. A derivation used in this work was previously published online by Michael SoroChan Armstrong [1].

Chapter 6 has been submitted for publication as: M. S. Armstrong, R. J. Abel, O. R. Arredondo Campos, A. P. de la Mata, and J. J. Harynuk “*A priori* prediction of chemical rank in gas chromatography - mass spectrometry data with projection pursuit analysis”, *Analytical Chemistry*

Michael SoroChan Armstrong was responsible for the conceptualisation and actualisation of all algorithms and analyses of the data within this thesis. He was also responsible for the planning and execution of the acquisition of most data, including the operation of the instruments. All experiments requiring derivatisation were planned and executed by Michael SoroChan Armstrong, frequently with the assistance

of O. René Arredondo Campos.

Data for Chapter 2 was collected as part of an earlier study, credited to A. Paulina de la Mata [2]. Some data acquired for Chapter 6 is credited to Robin Abel, who collected and excised a number of chromatographic regions for analysis.

Catherine Bannon is credited with the planning, and culturing of all algae, bacterial, and co-culture samples used in Chapter 3. Rebecca Case is also credited for her input for the experimental design, along with Catherine Bannon. Both are credited for their input for the draft of the submitted paper.

Lu Deng is credited for providing the urine samples prepared and analysed as part of Chapter 4, in addition to her input regarding the experimental design.

Jesper Henrich, and Rasmus Bro are credited for their feedback and input regarding the function of the PARAFAC2×2 algorithm in Chapter 5, in addition to the helpful comments regarding notation.

Quite generally, the familiar, just because it is familiar, is not cognitively understood. The commonest way in which we deceive either ourselves or others about understanding is by assuming something as familiar, and accepting it on that account; with all its pros and cons, such knowing never gets anywhere, and it knows not why. Subject and object, God, Nature, Understanding, sensibility, and so on, are uncritically taken for granted as familiar, established as valid and made into fixed points for starting and stopping. While these remain unmoved, the knowing activity goes back and forth between them, thus moving only on their surface.

-Phenomenology of Spirit (1807) §31, Georg Wilhelm Friedrich Hegel (translated by A.V. Miller)

Dedicated to my late grandfather, Frank Stephan Sorochan.

*With gratitude to my close family: Neil Armstrong, Brenda Sorochan, Sarah
Armstrong, and Taylor Hurdle*

Acknowledgements

This thesis was made possible by the combined efforts of a great many people. I would like to thank first and foremost the support staff at the University of Alberta Department of Chemistry. The administrators Kelly Fowler, Esther Moibi, Anuar Riviero, and Laura Pham. The chemical store staff Andrew Yeung, Matthew Kingston, Ryan Lewis, and Michael Barteski. The graduate student service coordinator Anita Weiler, and the Assistant Chair of student services Dr. Christie McDermott. Broderick Wood from the University of Alberta's Internet Service and Technology has also been an incredible help at various stages of research.

I would like to thank my undergraduate research supervisor Dr. Michael Serpe for introducing me to research, and providing me with so many excellent opportunities to further my personal and professional development. Dr. Hong Chen and Dr. Zhonglin Lyu from the MacBio Research Group at Soochow Univeristy, Dr. Rasmus Bro, his students Dr. Jesper Henrich and Dr. Dillen Augustin from the University of Copenhagen have all be instrumental in this regard as well. The undergraduate lab coordinators, Dr. Norman Gee, Dr. Yoram Appelblat, and Dr. Gregory Kiema taught me a great deal about time management and interpersonal skills, and I am fortunate to have been able to work with them.

A few people have worked exceptionally hard to make this thesis possible. I'd like to thank our group's massively over-qualified and wonderful research associate, Dr. A. Paulina de la Mata, and the exceptional young researcher O. Rene Arredondo Campos for the hours of help I have outright failed to return in kind. I would also like to thank my supervisor, Dr. James Harynuk for his patience - not only allowing

me to change my research direction, but for offering additional support and teaching relief to ensure I had enough time to write up my somewhat unorthodox ideas for data analysis. On a similar note, I would like to thank my committee - Dr. Liang Li and Dr. Juli Gibbs for their willingness to support the new avenues for research as they became available to me.

The International School of Chemometrics at the University of Copenhagen 2019 was a truly transformative experience, and I would like to thank Dr. Jose Amigo and all of the professors who work hard to make it happen every year.

It has been a privilege to have learned so much from the people passing through the Harynuk group. I'd especially like to thank the senior students for offering their hard-won wisdom and support: Dr. Lawrence Adutwum, Dr. Keisean Stevenson, Dr. Robin Abel, and Dr. Seo Lin Nam. While writing this thesis, I became even more aware of the high level of stress involved with original research, especially when faced with the very real possibility of failure. This has made all of the help they were willing to offer all the more selfless, and I cannot understate how thankful I am to have grown in the company of such excellent researchers.

This is not to say I am not grateful to the students (Brittany Reib, Trevor Johnson, Ryan Dias, Kieran Tarazona, and Sheri Schmidt) who joined the group later than myself. They have all become experts within their own specialties, and it has been great to learn from them regarding things like chemistry, how to operate specialised instrumentation, or even making sure that the eye wash stations are cleaned regularly. Thanks as well to the undergraduate and contract research assistants who have passed through the lab, it has been good to have gotten to know you all as well.

I am extremely grateful to my mathematics tutor, Gerry Leenders. I would also like to mention my fifth grade teacher Mr. Mark Goos for inspiring a love of science, along with Dr. Elizabeth Hill and Mr. Robin Hill for their help later in my academic career; especially during those instances where science didn't love me back. A very big thank you to all of my friends not otherwise mentioned for their support.

Table of Contents

1	General Introduction	1
1.1	Motivation	1
1.2	Analytical Instrumentation	3
1.2.1	Gas Chromatography	3
1.2.2	Mass Spectrometry	5
1.2.3	Comprehensive Two-Dimensional Gas Chromatography	7
1.2.4	Comprehensive Two-Dimensional Gas Chromatography - Time- of-Flight Mass Spectrometry	8
1.3	Chemometrics	9
1.3.1	Unsupervised learning: Matrix Decomposition	10
1.3.2	Tensor Decomposition	16
1.3.3	Supervised Learning: Regression and Discriminant-Type Anal- yses	24
1.3.4	Feature Selection	26
1.4	Thesis Objectives	32
1.5	Thesis Outline	33
2	An Efficient and Accurate Numerical Determination of the Cluster Resolution Metric in Two Dimensions	35
2.1	Theory	35
2.1.1	Cluster Resolution	35
2.1.2	Mathematical Description of Confidence Ellipses	36

2.1.3	Derivation of a Numerical Solution	37
2.1.4	Practical Considerations	38
2.2	Materials and Methods	39
2.2.1	Implementation	39
2.2.2	Experimental Data	40
2.3	Results and Discussion	41
2.3.1	Calculation of the cluster resolution metric for N clusters . . .	43
2.3.2	Comparison of Predictive Capabilities	44
2.4	Conclusions	46
3	Global Metabolome Analysis of <i>Dunaliella tertiolecta</i>, <i>Phaeobacter italicus</i> R11 Co-cultures using Thermal Desorption - Comprehensive Two-dimensional Gas Chromatography - Time-of-Flight Mass Spectrometry (TD-GC×GC-TOFMS)	49
3.1	Introduction	49
3.2	Materials and Methods	52
3.2.1	Growth and maintenance of algal and bacterial strains	52
3.2.2	Preparation of samples	52
3.2.3	Collection of culture samples	53
3.2.4	Sample preparation and derivatisation	54
3.2.5	Sample introduction and operating conditions	55
3.2.6	Thermal desorption, sample introduction	55
3.2.7	GC×GC-TOFMS method	57
3.2.8	GC×GC-TOFMS data pre-processing	57
3.2.9	Data analysis	59
3.2.10	Sample normalisation	59
3.2.11	Feature selection, cross-validation	59
3.3	Results	60

3.4	Discussion	62
3.5	Conclusion	67
4	Application of FS-CR for Urinary Metabolite Profiling of Human Colorectal Cancer using GC×GC-TOFMS: Limitations of Feature Selection	69
4.1	Introduction	69
4.2	Materials and Methods	71
4.2.1	Reagents	71
4.2.2	Urine Samples	72
4.2.3	GC×GC-TOFMS Method	74
4.2.4	Data Pre-processing Method	75
4.2.5	Normalisation	76
4.2.6	Data Analysis	76
4.3	Results	77
4.3.1	Discussion	80
4.4	Conclusion	84
5	PARAFAC2×N: Coupled Decomposition of Multi-modal Data with Drift in N Modes	86
5.1	Background	86
5.1.1	GC×GC-TOFMS Data Structure	89
5.1.2	PARAFAC Modelling of GC×GC-TOFMS Data	90
5.2	PARAFAC2 modelling for 4-way data unfolded as: $\mathcal{X} \in \mathbb{R}^{I \times K \times J \times L}$. .	92
5.2.1	A Flexible Coupling Approach for Non-negative PARAFAC2 .	93
5.3	PARAFAC2 modelling of 4-way data unfolded as: $\mathcal{X} \in \mathbb{R}^{I \times J \times K \times L}$ or $\mathcal{X} \in \mathbb{R}^{K \times J \times I \times L}$	95
5.4	The PARAFAC2×2 algorithm	96
5.4.1	Analysis of Synthetic Data using PARAFAC2×2	99

5.4.2	Analysis of Calibration Data using PARAFAC2 \times 2	104
5.5	Extension to Multidimensional Separations Data	105
5.6	Conclusions	108
6	<i>A priori</i> prediction of chemical rank in gas chromatography - mass spectrometry data with projection pursuit analysis	110
6.1	Introduction	110
6.2	Motivation	113
6.3	Materials and Methods	116
6.4	Results and Discussion	118
6.4.1	GC-MS Results	118
6.4.2	GC-TOFMS Results	120
6.4.3	Considerations for GC \times GC-TOFMS Data	122
6.4.4	Limitations of the approach	122
6.5	Conclusion	123
7	Conclusions and Future Work	124
7.1	Conclusions	124
7.2	Future Work	125
7.2.1	Future work to do with FS-CR	125
7.2.2	Future work to do with PARAFAC2 \times 2	126
7.2.3	Future work on Applications	127
7.2.4	Future work on automated k estimation	128
7.2.5	Region of Interest Selection	128
7.3	Outlook	129
Appendix A: Sample Calculations for “An Efficient and Accurate Numerical Determination of the Cluster Resolution Metric in 2 Dimensions”		145

Appendix B: Supporting Information for “Global Metabolome Analysis of <i>Dunaliella tertiolecta</i>, <i>Rueger iaitalica</i> Co-cultures using Thermal Desorption - Comprehensive 2-Dimensional Gas Chromatography - Time-of-Flight Mass Spectrometry (TD-GC×GC-TOFMS)”	156
B.1 Example Instrument Blank, Reagent Blank	157
B.2 Overview of Extracted Features	158
Appendix C: Derivation of Non-trivial Expressions used in Flexible Coupling PARAFAC2 - ALS, and Coupled PARAFAC2×2 - ALS	175
C.1 Derivation of an Expression to Solve for A_{kl} , and A_{il}	175
C.2 Derivation of an Expression to Solve for B_k, B_{kl} , and B_{il}	177
Appendix D: Supporting Information for: “<i>A-priori</i> prediction of chemical rank in gas-chromatography mass spectrometry data with projection pursuit analysis”	179
D.1 Library of Chemical Components used for Synthetic Data	179
D.2 Summary Analyses of GC-qMS Data	200
D.3 Summary Analysis of GC-TOFMS Data	204
D.4 Summary Analysis of GC×GC-TOFMS Data	205
D.5 Summary Analyses of Synthetic Data	206
D.5.1 1 Factor Synthetic Data	206
D.5.2 2 Factor Synthetic Data	215
D.5.3 3 Factor Synthetic Data	225
D.5.4 4 Factor Synthetic Data	235
D.5.5 5 Factor Synthetic Data	245

List of Tables

2.1	Comparison of variables selected using two FS-CR routines	47
3.1	Summary of identification levels for the significant metabolites	66
4.1	Demographic information for the study participants	74

List of Figures

1.1	A graphical representation of a 4 th -order tensor	16
2.1	FS-CR computation time comparison	41
2.2	Numerical FS-CR example calculation	41
2.3	Numerical vs. dynamic programming CR determinations	42
2.4	<i>N</i> -class cluster resolution, computation time comparison	44
2.5	Summary of classification results, NM-FS-CR vs. DP-FS-CR	46
3.1	Total ion current chromatograms for each class	61
3.2	Principal component analyses of the algae dataset	63
3.3	Graphical representation of cross-validation results	63
4.1	Quality control sample projections	78
4.2	Comparison of normalisation techniques	79
4.3	Graphical representation of cross-validation results	80
4.4	Graphical overview of relative peak areas for the internal standard from various samples at $m/z = 251$. The data was acquired and visualised using ChromaTOF [®] , and each black square represents the apex of a feature that was integrated by the software. Drift along the first mode affects the distribution of the relative peak area across multiple modulations, and rules out an assessment of the raw chromatographic signal, without first visualising it as a contour plot.	81
4.5	Sparsity of features, colorectal cancer dataset	83

4.6	Sparsity of selected features, colorectal cancer dataset	84
5.1	Simulated two-component model	102
5.2	Simulated 3-component model	103
5.3	Summary analyses of the calibration data	106
6.1	Summary of automated analysis, GC-MS data	119
6.2	Summary of automated analysis, GC-TOFMS data	121
B.1	Reagent blank	157
B.2	Instrument blank	157
B.3	Summary analyses, Analyte 5083	159
B.4	Summary analyses, Analyte 8351	160
B.5	Summary analyses, Analyte 13706	161
B.6	Summary analyses, analyte: “Hexadecane”	162
B.7	Summary analyses, Analyte 16951	163
B.8	Summary analyses, Analyte 17909	164
B.9	Summary analyses, Analyte 19404	165
B.10	Summary analyses, Analyte 21239	166
B.11	Summary analyses, Analyte 21318	167
B.12	Summary analyses, Analyte 22226	168
B.13	Summary analyses, Analyte 23041	169
B.14	Summary analyses, Analyte 23397	170
B.15	Summary analyses, Analyte 24829	171
B.16	Summary analyses, Analyte 27584	172
B.17	Summary analyses, Analyte 27833	173
B.18	Summary analyses, Analyte 33374	174
D.1	Automated MCR analysis, GC-qMS 1	200
D.2	Automated MCR analysis, GC-qMS 2	201

D.3 Automated MCR analysis, GC-qMS 3	202
D.4 Automated MCR analysis, GC-qMS 2	203
D.5 Automated MCR analysis, GC-TOFMS 1	204
D.6 Automated PARAFAC analysis, GC×GC-TOFMS 1	205
D.7 Automated MCR analysis, synthetic 1-factor data 1	206
D.8 Automated MCR analysis, synthetic 1-factor data 2	207
D.9 Automated MCR analysis, synthetic 1-factor data 3	208
D.10 Automated MCR analysis, synthetic 1-factor data 4	209
D.11 Automated MCR analysis, synthetic 1-factor data 5	210
D.12 Automated MCR analysis, synthetic 1-factor data 6	211
D.13 Automated MCR analysis, synthetic 1-factor data 7	212
D.14 Automated MCR analysis, synthetic 1-factor data 8	213
D.15 Automated MCR analysis, synthetic 1-factor data 9	214
D.16 Automated MCR analysis, synthetic 2-factor data 1	215
D.17 Automated MCR analysis, synthetic 2-factor data 2	216
D.18 Automated MCR analysis, synthetic 2-factor data 3	217
D.19 Automated MCR analysis, synthetic 2-factor data 4	218
D.20 Automated MCR analysis, synthetic 2-factor data 5	219
D.21 Automated MCR analysis, synthetic 2-factor data 6	220
D.22 Automated MCR analysis, synthetic 2-factor data 7	221
D.23 Automated MCR analysis, synthetic 2-factor data 8	222
D.24 Automated MCR analysis, synthetic 2-factor data 9	223
D.25 Automated MCR analysis, synthetic 2-factor data 10	224
D.26 Automated MCR analysis, synthetic 3-factor data 1	225
D.27 Automated MCR analysis, synthetic 3-factor data 2	226
D.28 Automated MCR analysis, synthetic 3-factor data 3	227
D.29 Automated MCR analysis, synthetic 3-factor data 4	228
D.30 Automated MCR analysis, synthetic 3-factor data 5	229

D.31 Automated MCR analysis, synthetic 3-factor data 6	230
D.32 Automated MCR analysis, synthetic 3-factor data 7	231
D.33 Automated MCR analysis, synthetic 3-factor data 8	232
D.34 Automated MCR analysis, synthetic 3-factor data 9	233
D.35 Automated MCR analysis, synthetic 3-factor data 10	234
D.36 Automated MCR analysis, synthetic 4-factor data 1	235
D.37 Automated MCR analysis, synthetic 4-factor data 2	236
D.38 Automated MCR analysis, synthetic 4-factor data 3	237
D.39 Automated MCR analysis, synthetic 4-factor data 4	238
D.40 Automated MCR analysis, synthetic 4-factor data 5	239
D.41 Automated MCR analysis, synthetic 4-factor data 6	240
D.42 Automated MCR analysis, synthetic 4-factor data 7	241
D.43 Automated MCR analysis, synthetic 4-factor data 8	242
D.44 Automated MCR analysis, synthetic 4-factor data 9	243
D.45 Automated MCR analysis, synthetic 4-factor data 10	244
D.46 Automated MCR analysis, synthetic 5-factor data 1	245
D.47 Automated MCR analysis, synthetic 5-factor data 2	246
D.48 Automated MCR analysis, synthetic 5-factor data 3	247
D.49 Automated MCR analysis, synthetic 5-factor data 4	248
D.50 Automated MCR analysis, synthetic 5-factor data 5	249
D.51 Automated MCR analysis, synthetic 5-factor data 6	250
D.52 Automated MCR analysis, synthetic 5-factor data 7	251
D.53 Automated MCR analysis, synthetic 5-factor data 8	252
D.54 Automated MCR analysis, synthetic 5-factor data 9	253
D.55 Automated MCR analysis, synthetic 5-factor data 10	254

List of Symbols

Latin

\mathcal{X}	N^{th} order tensor of observations
${}_nC_k$	Binomial coefficient for a pair of integers, n and k
A	Eddy diffusion coefficient
B	Longitudinal diffusion coefficient
B^*	$R \times R$ matrix, Flexible Coupling PARAFAC2
B_k	$I \times R$ Non-negative peak profiles, Flexible Coupling PARAFAC2
C_m	Resistance to mass transfer (mobile phase)
C_n	Quantitative score vector for the n^{th} column of a C score matrix, for Multivariate Curve Resolution
C_s	Resistance to mass transfer (stationary phase)
D_k	$R \times R \times K$ tensor of diagonal matrices, that describe the relative expression of components in a PARAFAC or PARAFAC2 model
E	Error matrix of residuals
F	$I \times R$ Matrix of scores (PARAFAC), or $R \times R$ square matrix (PARAFAC2)
H	Height Equivalent to a Theoretical Plate (HETP)
K	Kurtosis
k_n	Retention factor for the n^{th} peak
P_k	$I \times R$ matrix of orthogonal peak profiles

P_M	Modulation period
Q_{left}	Left eigenvectors
Q_{right}	Right eigenvectors
R_s	Rotation matrix for a confidence ellipsis
T_n	Quantitative score vector for the n^{th} column of a T score matrix, for either Principal Component Analysis or Projection Pursuit Analysis
V	Loadings matrix (Principal Component Analysis, Singular Value Decomposition)
X	$m \times n$ matrix of observations
Y	Matrix of characteristics
A	$J \times R$ matrix of mass spectral loadings (PARAFAC and PARAFAC2)

Greek

α	Selectivity ratio
χ^2	Chi squared statistic
$\Gamma(\nu)$	Gamma function for ν degrees of freedom
Λ	Major eigenvalue for a confidence ellipsis
λ	Minor eigenvalue for a confidence ellipsis
μ_A	Mass spectral coupling constant for PARAFAC2 \times 2
μ_k	Coupling constant for non-negative PARAFAC2
μ_{il}	Coupling constant, first dimension retention mode
μ_{kl}	Coupling constant, second dimension retention mode
ϕ	Angle of a confidence ellipsis
Σ	Matrix of singular values
σ_{new}	Sum of squared residuals from the current iteration

σ_{old}	Sum of squared residuals from the previous iteration
θ_n	Angle of a line drawn from the centre of the n^{th} confidence ellipsis
Ξ	Cluster resolution (3 or more classes)
ξ	Cluster resolution (2 classes)

Abbreviations

DBSCAN Density-based spatial clustering of applications with noise.

FS-CR Feature Selection by Cluster Resolution.

GC Gas Chromatography.

GC×GC Comprehensive 2-Dimensional Gas Chromatography.

GC×GC-TOFMS Comprehensive 2-Dimensional Gas Chromatography - Time-of-Flight Mass Spectrometry.

ICA Independent Component Analysis.

LC Liquid Chromatography.

MCR Multivariate Curve Resolution.

MS Mass Spectrometry.

MSTFA N-Methyl-N-trimethylsilyltrifluoroacetamide.

PARAFAC Parallel Factor Analysis.

PARAFAC2 Parallel Factor Analysis 2.

PARAFAC2×2 Parallel Factor Analysis 2×2.

PCA Principal Component Analysis.

PLS Partial Least Squares.

PLS-DA Partial Least Squares - Discriminant Analysis.

PPA Projection-Pursuit Analysis.

SPME Solid Phase Micro-Extraction.

SVD Singular Value Decomposition.

TD Thermal Desorption.

TMCS trimethylchlorosilane.

TOFMS Time-of-Flight Mass Spectrometry.

Glossary of Terms

Decomposition Reduction of a matrix or tensor into a series of informative linear components.

Kurtosis The fourth standardised statistical moment, which for larger values describes the tendency of a series of observations to trend towards a higher number of outliers with more extreme values. Observations that are bimodally distributed, or present fewer and less extreme outliers trend towards lower values of kurtosis.

Latent Variable A low-rank representation of a matrix, manifest as a linear combination of the original variables.

Loadings A matrix containing descriptors of all latent variables.

Parsimony Principle of simplicity of an analysis or algorithm; mathematically viable, and free of excess pre-treatment or programmatic steps.

Partial Least Squares A rank-deficient method for solving the least squares problem $Y = Xb$, such that the covariance between Y and X is maximised.

Scores Projections of the original data within latent variable subspace.

Tensor An ordered series of numbers of N orders, where N is the length of the vector required to reference any position within the tensor itself. A 1st-order tensor for example is a vector, and a 3rd-order tensor is a cube of numbers..

Utilitarian Antonym of parsimony: a description of an algorithm or analysis with no well-defined theory, or too much pre-treatment of the data and excess programmatic steps.

Chapter 1

General Introduction

1.1 Motivation

Modern analytical instruments have allowed researchers to investigate the chemical characteristics of complex samples in unprecedented detail. In most cases very little of this information is of practical use, but it is possible to correlate some quantifiable metric of an interesting batch of samples with a useful, (typically small) subset of their observed chemical characteristics. This is a type of “soft” modelling, wherein the goal of an experiment is not to prove a hypothesis, but rather to generate one where none exists. This type of analysis is particularly useful for samples whose characteristics are difficult to measure conventionally; so much so, that unconventional analytical measurements become a viable alternative. Many diseases are difficult to diagnose using the tools currently available to physicians, and correlating the endogenous chemical composition of bio-fluids (such as blood or urine) with diseases states represents a very promising research avenue for which modern instrumentation is well-suited [3].

Soft modelling is a departure from the practice of “hard” modelling more widespread in the practice of chemistry, or any other physical science. Hard modelling involves a relationship between experimental data, and a parametric equation that can be derived from, or related back to, first principles. Hard models have the advantage of being consistent with much broader theories of the physical sciences, but rely on

a comprehensive understanding of the system being studied. This may be inspired by, or derived from relationships observed in experimental data, but for insight into complicated, indirectly observed phenomena this may not always be possible for a human theorist.

Soft models are named for their ability to adapt to the information provided, independent of human intervention [4]. For either the flexible soft models calculated by computers, or the inflexible hard models made by humans, the predictive ability of a model must be evaluated by determining the accuracy of predictions made on previously unconsidered data, or compared against a known solution for which an automated determination is sought. Since soft modelling has begun to enjoy considerable attention in recent years within the relatively new disciplines of “Artificial Intelligence” (AI), and “Machine Learning” (ML), there have been many prominent studies that have flirted with disregarding model validation altogether [5]. This practice is roughly analogous to deriving an expression incorrectly, and refusing to apply it to the experimental data it was designed to predict.

The study of the algorithms and statistics that generate new hypotheses from an over-abundance of chemical information, generally falls within the scope of “chemometrics”[6]. Application of similar principles to problems of biological interest, fall within one of the newer disciplines with the “-omics” suffix: “Genomics” for the analysis of genetic data, “Proteomics” for the analysis of protein data, and “Metabolomics” for the analysis of small-molecule metabolites [7], etc. The chemical diversity for each type of data increases as a limited number of genes transcribe a greater number of proteins, and as proteins are modified post-translationally and catalyze or affect a currently unknowable number of different chemical reactions in a biological system[8]. The potential for insight into disease appears to be much greater for metabolite information, as metabolic expression can vary due to a biochemical dysfunction at any step. However, the challenge for analysing complex mixtures with a high degree of chemical diversity is also considerable and necessitates sophisticated analytical in-

strumentation[9].

1.2 Analytical Instrumentation

1.2.1 Gas Chromatography

The practice of chromatography encompasses different methods for separating complex mixtures into (ideally) pure components by exploiting differences in their chemical properties. Unlike spectrometric techniques, which analyse mixtures directly, a carefully optimised chromatographic separation can yield quantitative and qualitative information of several distinct chemical species, by physically separating them and passing them sequentially to a detection device. Chromatographic separations typically vary by their choice of stationary phase and mobile phase. Movement of analytes of interest between the stationary phase and mobile phase typically follows a partition coefficient K , analogous to a reversible chemical reaction at equilibrium. Liquid Chromatography (LC) exploits the relative affinities for molecules soluble in a liquid mobile phase, versus a solid stationary phase. Reverse-phase liquid chromatography is a common technique, that utilises a hydrophobic stationary phase (typically functionalised silica) versus a relatively polar liquid mobile phase that is usually comprised of water mixed with either methanol or acetonitrile[10].

Gas Chromatography (GC) separates mixtures in the gas phase, according to the partition coefficient between the analyte and stationary phase which affects the distribution of a particular analyte between the mobile phase (where components are able to move through the column), and the stationary phase where the molecules are stationary. Chemicals with a higher partition coefficient spend longer in the stationary phase, and less time in the mobile phase which increases the amount of time required to reach the detector. The time a particular analyte spends in the column, from the moment it is injected, until the moment it reaches the detector is defined as the retention time, or t_R . For the mobile phase, a sufficiently inert, low viscosity

gas such as helium, hydrogen, or occasionally nitrogen, is typically used. Since there is no significant interaction between the analytes and the mobile phase, the choice of stationary phase is an especially important consideration for GC.

Ideally, the stationary phase should be able to exploit minor differences in chemical characteristics of the analytes of interest while maintaining a reasonable degree of affinity for the analytes. An overzealous stationary phase can slow down the analysis time, with diminishing returns for the quality of the separation, while a poorly retentive phase will yield no separation. Most popular stationary phases are based on films of modified polydimethylsiloxane (PDMS), due to their stability at high temperatures and predictable interactions with the analytes of interest. However there is also growing interest in Ionic Liquid (IL) stationary phases, which are typically reserved for mixtures that are unable to be resolved using the more affordable and robust PDMS-based types.

Following a GC separation, a univariate detector at the end of the column can record the analytical signal as a time-series measurement. Since the behaviour of chemical species in the column is not discrete, the retention of a chemical component is (under idealised circumstances) distributed about an average retention time as a function of its variance and magnitude. Optimisation of GC separations seeks to minimise the variance (peak width), and maximise the resolution (peak separation) of each component, within a time-frame that is as short as possible. Since closely eluting chemical components can be difficult to quantify using a univariate detector, a carefully optimised separation also has consequences for the accuracy of an analysis. The efficiency of a separation is related to the well-known Van-Deemter Equation[11]:

$$H = A + \frac{B}{u} + C_s u + C_m u \quad (1.1)$$

H is the height equivalent to a theoretical plate; a lower theoretical plate height suggests a greater number of theoretical plates within a given length of column, L , such that the number of plates is given by $N = L/H$. u is the average linear

velocity of the carrier gas in cm/s. The A , B , C_s , and C_m terms are the Eddy Diffusion, Longitudinal Diffusion, resistance to mass-transfer in the stationary phase, and resistance to mass transfer in the mobile phase terms, respectively. The A term refers to diffusion brought about by the collision of analyte molecules with small particles within the column (which is common-place for HPLC). Since GC columns are almost exclusively open-tubular, the A term is zero. Within the practice of GC using open-tubular columns, depending on the linear velocity of the carrier gas, either the B or C terms dominate, depending on whether the value for u is less than or greater than where its derivative $dH/du = 0$.

The resolution of two closely eluting analytes is dictated by the Purnell Equation[11]:

$$R_s = \left(\frac{\sqrt{N}}{4} \right) \left(\frac{\alpha - 1}{\alpha} \right) \left(\frac{k_2}{k_2 + 1} \right) \quad (1.2)$$

Where N is the number of theoretical plates (i.e. the efficiency of the separation), k_2 is defined as the retention factor ($k = \frac{t_R - t_m}{t_m}$) of the later eluting component, and α is the selectivity ratio ($\alpha = \frac{k_2}{k_1} = \frac{K_2}{K_1}$) The Purnell Equation illustrates that optimising resolution between peaks is a balancing act involving the column geometry (N), the retention it offers (k_2), and the relative differences between component interactions with the column (α).

GC separations can be optimised via a wealth of knowledge on the topic, based on parameters such as the oven temperature program, the column geometry, stationary phase chemistry, and the choice of carrier gas [12][13][14].

1.2.2 Mass Spectrometry

Mass spectrometry can be used to analyse molecules based on their spectra of mass-to-charge ratios (m/z) following fragmentation via a high energy ionisation state. Unlike LC, it is readily coupled to GC systems and there are no significant consequences regarding ion suppression as is frequently encountered when using electrospray ionisa-

tion (ESI) [15]. This is because the high volatility and minimal volumes of gas required for a GC separation are much more favourable, and the more straightforward method of Electron Impact Ionisation (EI) is possible. Electron impact ionisation bombards neutral analytes leaving the separation step with high energy electrons generated via thermionic emission of a high-temperature filament with an electrical current applied. A plate situated opposite to the filament is maintained at a high positive voltage [16], which draw the electrons across the ionisation source through electrostatic forces. When the neutral molecules are bombarded with an electron of sufficient energy, an electron is removed to generate a radical cation. Molecules that contain sufficient kinetic energy from the impact will fragment through loss of neutral radicals into smaller cations that can be detected using a mass analyser.

Time-of-Flight (TOF) Mass Spectrometers and quadrupole (qMS) Mass Spectrometers are popular choices for mass analysers. Briefly, TOF instruments separate mass-to-charge ratios through measurement of the length of time spent in a flight tube due to the action of a constant kinetic energy pulse applied to the ions perpendicular to their trajectory leaving the ion source. Quadrupole mass analysers utilise at least four metal rods: two opposite to each other held at a constant voltage, and two whose voltage varies across a continuum of radio-frequencies to select for stable ion trajectories [17]. The length of time required to scan through all interesting m/z ratios is not trivial, and quadrupoles are generally capable of much lower acquisition rates. The quality of the mass spectra extracted can also suffer, as the magnitude of the peak varies during the time-scale of the quadrupole scan[18]. TOF instruments on the other hand are able to handle a much higher acquisition rate and are suitable for short, high-efficiency separations with lower peak widths, but generally suffer from a reduction in dynamic range [19].

Closely eluting peaks, when analysed using GC-MS can be resolved using multivariate deconvolution techniques [20][21]. This decreases the reliance on a perfect chromatographic separation, as imperfectly resolved components can still be quan-

tified and identified using a library search on each components' deconvolved mass spectra. The practice of assigning more or less significance to particular ions is an active area of research, and affects what hits are listed in a library search [22][23]. The retention index, defined as the retention of an analyte relative to a series of n -linear alkanes is also useful for this purpose [24].

1.2.3 Comprehensive Two-Dimensional Gas Chromatography

Mixtures that challenge the practical limitations of a single chromatographic separation may benefit from an additional chromatographic dimension that can differentiate poorly resolved components using a complementary selection criteria - evading the limitations of Equation 1.2 for two unresolved components. Improvements to the selectivity of the separation for one pair of closely-eluting analytes can negatively affect the resolution of other analytes, and increasing column performance will not help if the limitation is k_2 or α in Equation 1.2. This issue scales with the complexity of the samples, since the number of poorly-resolved pairs predictably increases as more components are introduced within the limited time-frame of a chromatographic separation, as is commonly encountered with petroleum, natural products, and biological samples. For these mixtures, even using hyphenated techniques such as GC-MS, deconvolution of complex co-elutions may not always be sufficient.

For analytes that are unresolved along one chromatographic dimension, it is possible to couple the first separation to further chromatographic separations to better resolve closely-eluting components. This can be done for specific regions of a single chromatogram, where there are a number of poorly resolved components collected and separated using a different mechanism, or for an entire chromatographic separation for eluent fractions collected at regular intervals. Comprehensive two-dimensional chromatography belongs to the latter of the two categories, and separates an entire chromatographic run along two dimensions.

The most mature comprehensive two-dimensional separation technology is compre-

hensive two-dimensional gas chromatography, owing to the relative ease of modulating fractions of eluent from the first dimension onto the second at regular intervals, and the ability to affect complementary separations using only a change in stationary phase chemistry without the need to change or modify the composition of the mobile phase [25].

1.2.4 Comprehensive Two-Dimensional Gas Chromatography - Time-of-Flight Mass Spectrometry

The modulation period, P_M , in the practice of GC \times GC is the amount of time between injections of collected fractions of first-dimension effluent onto the second-dimension column. Typically this period lasts only a few seconds, allowing for adequate sampling of first dimension elution profiles, and sufficient time for the second-dimension separation to proceed. As such, peaks that approach the detector following a short second-dimension separation are typically quite narrow. This necessitates the use of a high speed detection system, which is not a problem for widespread univariate detectors such as flame-ionization detector (FID) systems. However, for applications requiring multivariate detection it does necessitate the use of a high-throughput mass spectrometer. Most hyphenated GC \times GC systems are coupled with a high-speed TOFMS usually operating between 100-200 spectra/s; however fast quadrupole systems may also be used, provided they can acquire at least 20 spectra/s without significant reduction in spectral quality.

GC \times GC-TOFMS out-paces traditional GC-MS in every performance-based metric [26][27]. Utilising two complementary separations offers far more peak capacity (the maximum theoretical limit of how many components can be well-resolved during a chromatographic run), and sensitivity (since the action of the modulator tends to increase the Signal-to-Noise ratio). Despite these advantages, and despite widespread use of GC-MS, GC \times GC-TOFMS has not yet reached widespread acceptance in commercial laboratories. There are many hypotheses as to why this may be the case: the

relatively high cost of a GC×GC-TOFMS instrument, the need for expert personnel to operate a GC×GC-TOFMS, and the generally poor output of commercial data analysis software used to identify and quantify interesting signals [28].

Over the last few years, GC×GC-TOFMS instrumentation has become much more robust and affordable. There are now several vendors who provide a variety of instrumentation alternatives. With this proliferation of the technique, users are demanding more from the software, and are becoming more interested in processing suites of dozens, hundreds, or even thousands of samples for discovery-type problems (e.g. biomarker discovery in metabolomics data sets). In these situations, the traditional GC×GC data processing tools are inadequate, leaving users with complex data sets that require days or weeks of expert user intervention for curation of robust peak tables. Addressing these needs is the focus of this thesis.

1.3 Chemometrics

Chemometrics is the study of mathematical, statistical, and computational methods to analyse chemical data, or to design optimal experimental or measurement conditions within either a practical or theoretical framework. It is roughly analogous to similar disciplines in Psychology, and Economics (Psychometrics, and Econometrics) that encountered a need for such work much earlier in their histories. Chemometrics developed alongside the use of computers in the laboratory, as some of the very first digital instruments became commercially available in the 1970s. This is likely one of the earliest points where analytical chemists were commonly faced with the task of analysing data with many more variables than observations, since experimental observations acquired through conventional chemical analyses had hitherto been prohibitively expensive and time-consuming.

A distinguishing characteristic of chemometrics is its focus on algorithmic, rather than purely statistical methodologies. For example, variants of the NIPALS algorithm (Nonlinear Iterative Partial Least Squares) [29] are incredibly widespread, and

have been extensively developed by chemometricians. This has led to the distinction between chemometrics and disciplines such as machine learning (more closely associated with computer science), being somewhat blurry. However, there is typically a bias towards linear models within chemometrics as these models tend to be more interpretable and correspond better to the latent chemical phenomena being studied. This is different than more typical machine learning tools which often rely on highly non-linear techniques such as Artificial Neural Networks (ANNs). Although pervasive and well-characterised, neural networks are generally poorly interpretable in a chemical sense. Consequently, they are generally avoided as a principal means of analysing chemical data.

1.3.1 Unsupervised learning: Matrix Decomposition

Unsupervised learning extracts useful information from data (typically a matrix, X), without any additional user-supplied information such as a panel of observable characteristics (typically a vector, or matrix Y of length equal to the number of observations in X). Examples of unsupervised learning are clustering algorithms such as k -means or k -nearest neighbour, and DBSCAN, but encompass certain matrix decomposition techniques as well.

Matrix decomposition reduces the dimensionality, or rank, of an $m \times n$ matrix X via projections of the original variables to latent variable structures calculated as linear combinations of observable variables that maximise or minimise a desirable characteristic of the data. The latent variables can be used to recover informative trends within the data, reduce the amount of memory required to store the information encoded in the matrix, or the scores themselves can serve as a simplified, lower-dimensional representation of the data as pre-processing for further analysis.

Different techniques for matrix decomposition can be classified by what characteristics of the data are being sought, and the amount and manner in which each latent variable influences the determination of others. Considerations for numerical stability

are also important, since convergence to a global optimum is not guaranteed for all matrix decompositions [30].

Principal Component Analysis

Principal Component Analysis (PCA), is one of the most ubiquitous matrix decomposition techniques. PCA decomposes an $m \times n$ matrix, X into an $m \times R$ column-wise orthogonal score matrix (T) projected within an $n \times R$ column-wise orthonormal basis (V):

$$X = TV^T + E \quad (1.3)$$

Where R is the number of latent variables. TV^T can be used reconstruct the original matrix, X , using the components that maximise the variance explained by the model. E is the $m \times n$ matrix of residual errors not accounted for by the model. T is commonly referred to as the principal component “scores”, and V the “loadings”. Either can be calculated from the right and left eigenvalues of matrix X , via:

$$XX^T = Q_{left}\Lambda Q_{left}^{-1} \quad (1.4)$$

$$X^TX = Q_{right}\Lambda Q_{right}^{-1} \quad (1.5)$$

Due to the fact that eigendecompositions can only operate on square matrices the scores matrix, T , is calculated from the left eigenvectors and the square root of the $R \times R$ matrix of eigenvalues:

$$T = Q_{left}\sqrt{\Lambda} \quad (1.6)$$

Singular value decomposition is a more straightforward way of performing PCA,

decomposing the rectangular matrix, X directly via:

$$X = U\Sigma V^T \quad (1.7)$$

Where the $m \times n$ is a diagonal, rectangular singular value matrix contains the square root of the eigenvalues. This relationship can be proven using the well-known identities:

$$XX^T = U\Sigma V^T V\Sigma^T U^T = U(\Sigma\Sigma^T)U^T \quad (1.8)$$

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V(\Sigma^T \Sigma)V^T \quad (1.9)$$

Where it is evident that $XX^T = Q_{left}\Lambda Q_{left}^{-1} = U(\Sigma\Sigma^T)U^T$ and $X^T X = Q_{right}\Lambda Q_{right}^{-1} = V\Sigma^T U^T U\Sigma V^T = V(\Sigma^T \Sigma)V^T$, and since Σ is a diagonal matrix, the eigenvalue matrix, Λ corresponds to the square of the singular values along the diagonal. Analogous to Equation 1.7, the principal component scores of matrix X can be calculated as:

$$T = U\Sigma \quad (1.10)$$

Variables in the matrix, X are typically centred and scaled to ensure that the most interesting aspects of the data are being considered. Centring such that the mean of each column = 0, is commonly referred to as “mean-centring”, and scaling such that the standard deviation, σ , = 1 is commonly referred to as “autoscaling” when used in conjunction with mean-centring. If the column-wise standard deviations and means are known, then the original data can be recovered without any loss of information.

Multivariate Curve Resolution

Multivariate Curve Resolution (MCR) is another common matrix decomposition technique applied to matrices of chromatographic regions with hyphenated, multivariate

detectors such as mass spectrometers or Diode Array Detectors (DAD). Any such matrix can be decomposed using MCR:

$$X = CS^T + E \quad (1.11)$$

Where C is an $m \times R$ matrix of parallel vectors that contain the elution profiles of the calculated components, and are typically quantitative. S is an $n \times R$ matrix that contains the spectral information, or the resultant deconvolved multivariate signals of each component. Unlike PCA, there are no constraints for how the calculation of one component affects the other (i.e. orthogonality), and because of this there is no guarantee for the uniqueness of the resultant solution. This is commonly known as “rotational ambiguity”, and can be demonstrated with the following expression:

$$CS^T = C(P P^{-1})S^T \quad (1.12)$$

Where any non-singular, arbitrary $R \times R$ rotation matrix, P , can be applied to either the “scores” or “loadings” in Equation 1.11 such that: $C = CP$ and $S = S(P^{-1})^T$ to explain exactly the same amount of variance in matrix X . This ambiguity is equally applicable to PCA, but since the rotation of the matrix itself is fixed relative to the orientation of the axes of most significant variance, the rotation matrix, P , in effect can only affect the signage of the principal components (i.e. a rotation of π about some axis).

While PCA can be calculated via eigendecomposition, or singular value decomposition there is no convenient mathematical method for calculating MCR. As such, the Alternating Least Squares (ALS) algorithm is the most widely used method of calculating an MCR model:

Algorithm 1: Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS)[31]

Result: $C, S = \text{MCR-ALS}(X, R, \epsilon)$
 $C \in \mathbb{R}^{n \times R}$, $S \in \mathbb{R}^{m \times R}$, $X \in \mathbb{R}^{m \times n}$
 $\sigma_{old} = 1e9$ %%Arbitrarily large number
while $\frac{\sigma_{old} - \sigma_{new}}{\sigma_{old}} > \epsilon \sigma_{old}$ **do**
 $\sigma_{old} = \sigma_{new}$
 $C = X S (S^T S)^{-1}$
 $S = X^T C (C^T C)^{-1}$
 $\sigma_{new} = \|X - C S^T\|_F^2$
end

At each iteration, each column of C or S can be normalised to its Euclidean norm, depending on what mode is desired to contain the quantitative information. PCA can also be calculated using a minor variation of Algorithm 1, known as Nonlinear Iterative Partial Least Squares (NIPALS) [32], by adding a matrix deflation step following the calculation of each latent variable and its scores:

Algorithm 2: Principal Component Analysis - Nonlinear Iterative Partial Least Squares (PCA-NIPALS)[33]

Result: $T, V = \text{PCA-NIPALS}(X, R, \epsilon)$
 $T \in \mathbb{R}^{n \times R}$, $V \in \mathbb{R}^{m \times R}$, $X \in \mathbb{R}^{m \times n}$
 $\sigma_{old} = 1e9$ %%Arbitrarily large number
for $r \in [1, R]$ **do**
 while $\frac{\sigma_{old} - \sigma_{new}}{\sigma_{old}} > \epsilon \sigma_{old}$ **do**
 $\sigma_{old} = \sigma_{new}$
 $T = X v_r (v_r^T v_r)^{-1}$
 $V = X^T t_r (t_r^T t_r)^{-1}$
 $\sigma_{new} = \|X - t_r v_r^T\|_F^2$
 end
 $E = X - t_r v_r^T$
 $X = E$
 $T(:, r) = t_r$
 $V(:, r) = v_r$
end

PCA-NIPALS is more numerically stable compared with methods utilising eigendecomposition or SVD, but the same advantages afforded to it as an iterative algorithm, also necessitate much more computation time. The matrix deflation step helps to ensure that the calculation of each subsequent principal component is orthogonal to the

previous loadings and scores matrices, but due to rounding errors this relationship is not guaranteed[34]. As with MCR-ALS, one mode is normalised at each iteration within the ALS step to constrain the other mode to express the relative expression of the components being studied in the other mode.

Projection Pursuit Analysis

A more general theory of matrix decomposition is encompassed by Projection Pursuit Analysis (PPA). Projection Pursuit Analysis (PPA) is a technique first proposed by Friedman and Tukey [35] that seeks to find “interesting” projections based on the pursuit of a particular projection index. This is a general enough description to encompass other common linear decomposition techniques: Independent Component Analysis (ICA) maximises a projection index of statistical independence, and PCA maximises an index of the explained variance of the data. All projection indices make assumptions about what characteristics of the data are most interesting for the analyst, save for those instances where the projection index is selected manually. Hou and Wentzell in 2011 [30] first described the minimisation of kurtosis as a projection index, to reveal resultant clustering of the data. This was motivated by the observation that highly resolved score clusters present a very low kurtosis, K , or tendency for the data to feature a relatively low number and extremity of outliers as described by the following equation:

$$K = \frac{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^4}{\left(\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 \right)^2} \quad (1.13)$$

Where z_i refer to the score of each sample, as projected along a vector, \mathbf{v} such that $z_i = \mathbf{x}_i^T \mathbf{v}$. Calculation of the projection vector is tantamount to minimising:

$$K = \frac{n \sum_{i=1}^n (\mathbf{v}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v})^2}{(\mathbf{v}^T X^T X \mathbf{v})^2} \quad (1.14)$$

A drawback of PPA is that for novel projection indices there is seldom a convenient mathematical method for calculating the model, so sophisticated algorithms

are required to find a solution. Oftentimes, these algorithms require several different parallel initialisations to ensure convergence to a global minimum.

1.3.2 Tensor Decomposition

A tensor is a multidimensional array of numbers of order N . They are a natural extension of vectors from scalars, and matrices from vectors. Conversely, vectors are commonly referred to as 1st-order tensors, and matrices as 2nd-order tensors, since both are indexed by an $N \times 1$ vector.

4th-order tensors are frequently encountered when discussing decompositions of GC×GC-TOFMS data. For a particular region of the chromatogram, a tensor, $\mathcal{X} \in \mathbb{R}^{I \times J \times K \times L}$ presents as I mass spectral acquisitions, J mass-to-charge ratios, K modulations, and L samples as illustrated below:

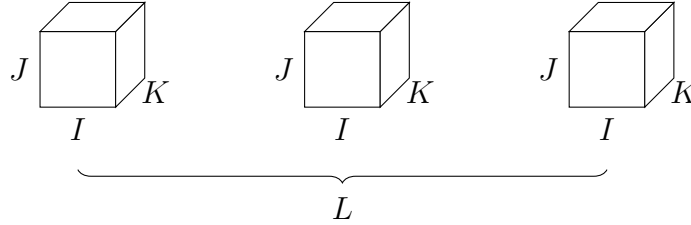


Figure 1.1: A graphical representation of a 4th-order tensor

As for matrices, it is possible to decompose a tensor of any order into a series of informative parallel vectors. There are a number of different techniques for tensor decomposition but for the purposes of this thesis, only Parallel Factor Analysis (PARAFAC) and the related technique PARAFAC2 will be discussed in detail.

Parallel Factor Analysis (PARAFAC)

PARAllel FACtor analysis (PARAFAC) is one extension of PCA to higher-order tensors, and was proposed independently by several authors. Hitchcock first proposed the idea of representing a tensor as a finite series of rank-1 tensors [36], but the idea of a PARAFAC model was proposed by Harshman in 1970 [37]. At close to the same

time two other researchers (Carroll and Chang) published a similar model under the name of CANDECOMP (CANonical DECOMPosition) in the sphere of psychometrics [38].

A 3rd-order tensor, $\mathcal{X} \in \mathbb{R}^{I \times J \times K} = X_k$ can be decomposed using a PARAFAC model, using matrix notation:

$$X_k = F D_k A^T + E_k \quad (1.15)$$

Where F is an $I \times R$ matrix of scores, D_k is series of diagonal matrices of $R \times R \times K$ dimensions, and A is a matrix of $J \times R$ loadings. Typically, F and A are normalised column-wise to their Euclidean norm such that the diagonal matrices D_k contain the quantitative information. E_k are the error matrices, as slabs. Rather than unfolding X_k as an $I * K \times J$ matrix, PARAFAC models do not suffer from the rotational ambiguity of bilinear models of matrices. If it were possible to rotate the components of a PARAFAC model as in Equation 1.15, such that the same data is explained, albeit with variations of the existing components then the following identity would hold:

$$F D_k A^T = F N N^{-1} D_k M M^{-1} A^T \quad (1.16)$$

Where N and M are $R \times R$ rotation matrices, similar to P in Equation 1.12. Since $N^{-1} D_k M$ would need be a diagonal matrix for equivalency, the only possible matrices that would satisfy Equation 1.16 are permutation (where the components could be represented in a different order) or scaling matrices (where the quantitative information would be distributed across either A or F in addition to D_k). The resultant indeterminacies can be categorised as to do with the order of the components, or the treatment of non-quantitative loading matrices, and are therefore considered to be trivial [39]. For most practical considerations, the PARAFAC model is considered to have a unique solution, and is preferable in most cases to decompositions that

operate on the unfolded matrix.

Equation 1.15 is a convenient representation of a 3rd-order tensor using matrix notation, however the model is typically calculated using the Khatri-Rao (KR) product. The KR product is defined as the column-wise Kronecker product for two matrices, A and B with an equal number of columns, R , that correspond to the number of chemical factors:

$$A = [a_1, a_2, \dots, a_R] \quad (1.17)$$

$$B = [b_1, b_2, \dots, b_R] \quad (1.18)$$

$$A \odot B = [a_1 \otimes b_1, a_2 \otimes b_2, \dots, a_R \otimes b_R] \quad (1.19)$$

Using the KR product, a PARAFAC model can be calculated (using a variant of Harshman's tensor notation for brevity):

$$\mathcal{X}^{I \times J \times K} = F(A \odot D)^T \quad (1.20)$$

Using the Alternating Least Squares Algorithm [39] for a 3rd-order tensor, a PARAFAC model that minimises $\|\mathcal{X}^{I \times J \times K} - F(A \odot D)^T\|_F^2$ can be found:

Algorithm 3: Parallel Factor Analysis - Alternating Least Squares (PARAFAC-ALS)[39]

Result: $F, D, A = \text{PARAFAC-ALS}(\mathcal{X}, R, \epsilon)$
 $F \in \mathbb{R}^{I \times R}$, $D \in \mathbb{R}^{K \times R}$, $A \in \mathbb{R}^{J \times R}$, $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$
 $\sigma_{old} = 1e9$ %Arbitrarily large number
while $\frac{\sigma_{old} - \sigma_{new}}{\sigma_{old}} > \epsilon$ **do**
 $\sigma_{old} = \sigma_{new}$
 $Z = (A \odot D)$
 $A = \mathcal{X}^{J \times I * K} Z (Z^T Z)^{-1}$
 $Z = (A \odot F)$
 $D = \mathcal{X}^{K \times J * K} Z (Z^T Z)^{-1}$
 $Z = (D \odot F)$
 $F = \mathcal{X}^{I \times J * K} Z (Z^T Z)^{-1}$
 $\sigma_{new} = \|\mathcal{X}^{I \times J * K} - F(A \odot D)^T\|_F^2$
end

Where D is not solved for as a series of diagonal matrices as in Equation 1.15, but as a matrix of parallel vectors, similar to a bilinear decomposition. Implementation of Algorithm 3 is highly vectorisable, computationally efficient, and can be extended to higher orders without very much difficulty. A PARAFAC model for a 4th-order tensor can be described as the following:

$$\mathcal{X}^{I \times J * K * M} = F(B \odot A \odot D)^T \quad (1.21)$$

Where B is an additional $M \times R$ series of parallel vectors, as-yet undefined for the M^{th} mode. It could easily be considered as an additional chromatographic mode, in which case we can describe Equation 1.21 using more familiar terms relevant to GC×GC-TOFMS data analysis. Let $F_1 \in \mathbb{R}^{K \times R}$ be the first-dimension modulations, $F_2 \in \mathbb{R}^{I \times R}$ be the second-dimension acquisitions, $D_l \in \mathbb{R}^{R \times R \times L}$ of quantitative loadings per sample, and $A \in \mathbb{R}^{J \times R}$ of mass-to-charge ratios:

$$\mathcal{X}^{I * K \times J \times L} = (F_2 \odot F_1)_l D_l A^T \quad (1.22)$$

Equation 1.22 is displayed primarily to demonstrate that the unfolded elution times of the first- and second-dimensions can be described as the KR product of F_1 and F_2 .

If a 4th-order PARAFAC model for a GC×GC-TOFMS region were being constructed, it would be more straightforward to operate on a variation of Equation 1.21.

Modelling chromatographic data using tensor decomposition techniques, such as PARAFAC present many benefits over commercial software offerings. Namely, that the quantity of each component is solved for as a regression step in the calculation of the model such that low-abundance peaks that are close to the background noise can still be reasonably quantified as a very small number as opposed to an outright zero. The presence of too many zeros in a $m \times n$ matrix of chemical characteristics and samples can make modelling the observable characteristics of the samples extremely difficult, especially in those cases where the number of chemical factors far exceeds the number of samples. Tensor methods are generally seen as being more reliable for the identification and quantification of chemical factors in chromatographic data; however, they do require some expert user intervention in order to determine an appropriate number of components and for the selection of regions of interest (ROIs), that best describe the latent chemical phenomena being studied.

PARAFAC2

A fundamental assumption of any PARAFAC model is that across all slices of the data, similar information is indexed in the same locations. Very small drifts do not have disastrous impacts on model performance; however, the percent of variance explained by the model will be lower. This is somewhat at odds with what is observed in chromatographic data, as typically retention times will drift over the course of an experiment; more-so when significant numbers of samples have been run, or when a significant period of time elapses from the collection of the first sample to the last. Since retention time is not as readily quantised in the same way as mass-to-charge ratios, the retention of a band of analytes exhibits significant Gaussian behaviour exiting the column as an average of the equilibrium conditions between the stationary and mobile phases along the length of the column. Over time the retention of the

analytes may become markedly different as the column conditions change, primarily due to the degradation of the column. Since each column is somewhat unique due to minor imperfections in the manufacturing process, correcting retention time drift is not as simple as installing a new column.

PARAFAC2 is a variation of a standard PARAFAC model that is equipped to handle drift in one mode, and typically operates on 3rd-order tensors, although extensions to higher orders for instrumentation such as GC-MS/MS are possible. The basic concept for PARAFAC2 was first described by R.A. Harshman in 1972 [40]. H.A.L Kiers published an "indirect" fitting of PARAFAC2 in 1993 [41], and published a much simpler version of the algorithm in 1998 [42] along with R. Bro [43]. The simpler, "direct fitting" approach calculates P_k orthogonal, unique peak profiles per each k matrix slab along X_k , and calculates the remaining terms via one iteration of PARAFAC-ALS on $X_k P_k$:

$$X_k = P_k F D_k A^T \quad (1.23)$$

P_k is calculated using a simplification of the following expression via SVD [42]:

$$P_k = X_k A D_k F^T (F D_k A^T X_k^T X_k A D_k F)^{-1/2} \quad (1.24)$$

$$F D_k A^T X_k^T = U_k \Sigma_k V_k^T \quad (1.25)$$

$$P_k = V_k U_k^T \quad (1.26)$$

Algorithm 4: Parallel Factor Analysis 2 - Alternating Least Squares (PARAFAC2-ALS)[42]

Result: $P_k, F, D_k, A = \text{PARAFAC2-ALS}(X_k, R, \epsilon)$
 $P_k \in \mathbb{R}^{I \times R}, F \in \mathbb{R}^{R \times R}, D \in \mathbb{R}^{K \times R}, A \in \mathbb{R}^{J \times R}, \mathcal{X}_{\parallel} \in \mathbb{R}^{I \times J \times K}$
 $\sigma_{old} = 1e9$ %% Arbitrarily large number
while $\frac{\sigma_{old} - \sigma_{new}}{\sigma_{old}} > \epsilon$ **do**
 $\sigma_{old} = \sigma_{new}$
 for $k \in [1, K]$ **do**
 $U_k \Sigma_k V_k^T = \text{SVD}(F D_k A^T X_k^T, R)$
 $P_k = V_k U_k^T$
 end
 %% PARAFAC-ALS
 $Z = (A \odot D)$
 $F = X_k^T P_k^{R \times J \times K} Z (Z^T Z)^{-1}$
 $Z = (A \odot F)$
 $D = X_k^T P_k^{K \times J \times R} Z (Z^T Z)^{-1}$
 $Z = (D \odot F)$
 $A = X_k^T P_k^{J \times R \times K} Z (Z^T Z)^{-1}$
 $\sigma_{new} = \sum_{k=1}^K \|X_k - P_k F D_k A^T\|_F^2$
end

In Algorithm 4, the quantitative information is typically stored in either the F or D_k components. The direct fitting solution for PARAFAC2 has been shown to be unique under relatively mild conditions [42]. PARAFAC2 does not function exclusively on tensors, since the dimension of each frontal slab can vary along the mode that is free to vary. As such, it is not necessarily a tensor decomposition technique, but is generally mentioned in the context of other tensor decomposition tools [44].

Flexible Coupling PARAFAC2

A drawback of PARAFAC2 is the reliance on SVD to calculate the elution profiles directly, such that the profile scores within P_k are orthogonal. This is a severe constraint for relatively closely eluting components that can be corrected by F , but only assuming that the components do not drift independently (See Chapter 5 for a more detailed treatment of this). Recently [45], a flexible coupling approach has enabled

the calculation of the elution profiles via a coupled expression that minimises:

$$X_k = \operatorname{argmin} \|X_k - B_k D_k A^T\|_F^2 + \mu_k \|B_k - P_k B^*\|_F^2 \quad (1.27)$$

Note that P_k in this case is calculated via: $P_k = \operatorname{SVD}(B_k B^*) \rightarrow UV^T$, which enables for a flexible descent as the model iterates, to a non-negative determination of B_k using a non-negative least-squares solver:

Algorithm 5: Flexible Coupling PARAFAC2-ALS [46]

Result: $B_k, D_k, A = \operatorname{PARAFAC2} \times 2(X_k, R)$
 $B_k \in \mathbb{R}^{I \times R}$, $D_k \in \mathbb{R}^{R \times R \times L}$, $A \in \mathbb{R}^{J \times R}$, and $X \in \mathbb{R}^{I \times J \times K}$
 Initialise randomly;
while $\frac{\sigma_{old} - \sigma_{new}}{\sigma_{old}} > \epsilon \sigma_{old}$ **do**
 for $\forall k \in [1, K]$ **do**
 $[U, \Sigma, V] = \operatorname{SVD}(B_k * B^*, R)$
 $P_k = UV^T$
 end
 $B^* = \|\sum_{k=1}^K \mu_k P_k^T B_k\|_R$
 $A = \|\sum_{k=1}^K \frac{X_k^T B_k D_k}{D_k B_k^T B_k D_k}\|_R$
 $B_k = \|\frac{X_k A_k D_k + \mu_k P_k B_k^*}{D_k A_k^T A_k D_k + \mu_k I_R}\|_R$ %% See Appendix C.2
 for $\forall k \in [1, K]$ **do**
 $D_k = \frac{B_k^T X_k A_k}{(B_k^T B_k)(A_k^T A_k)}$
 end
 if $i = 1$ **then**
 for $\forall k \in [1, K]$ **do**
 $\Sigma = \operatorname{SVD}(X_k, 2)$
 $SNR \approx \Sigma_1 / \Sigma_2$
 $\mu_k = 10^{-SNR/10} \frac{\|X_k - B_k D_k A_k^T\|_F^2}{\|B_k - P_k B_k^*\|_F^2}$
 end
 else
 if $i < 10$ **then**
 for $\forall k \in [1, K]$ **do**
 $\mu_k = \mu_k * 1.05$
 end
 end
 end
 $i = i + 1$
end

Note that in Algorithm 5, the determination for many components is expressed

as a normalised sum of slice-wise calculations, as opposed to previously where these components were determined via the KR product.

Certain features of Algorithm 5 are original work for this thesis, in particular the determination of an appropriate SNR via SVD. Further details will be provided in Chapter 5. While uniqueness for this approach has not been proven for the flexible coupling approach, it is guaranteed to improve the fit with each iteration.

1.3.3 Supervised Learning: Regression and Discriminant-Type Analyses

The goal of the aforementioned matrix and tensor decomposition techniques serve as a method for extracting useful chemical information from the data in an unsupervised way. With enough extracted chemical factors, it is possible to correlate their relative abundances across several samples with a vector or matrix of the samples’ observable qualities:

$$Y = Xb \tag{1.28}$$

For a simple case, where each sample has one observation associated with it in the Y block, Y is an $m \times 1$ vector. X is an $m \times n$ matrix of samples and variables or features. The Y block can contain quantitative information (for regression-type problems), or class information for discriminant-type problems. Discriminant-type problems typically encode class information as a vector or matrix of 1s and 0s, and a value of 0.5 at the line across \hat{Y} is typically used as a decision boundary to assess the class membership of the predicted scores according to $\hat{Y} = Xb$. In either case, the method by which the model is calculated is virtually identical. The regression vector, b , can be solved in the least-squares sense:

$$b = (X^T X)^{-1} X^T Y \tag{1.29}$$

For cases where there are many more variables than observations $n \gg m$, the

nature of the inverse in Equation 1.29 marks the difference between Multivariate Linear Regression (MLR) for a classical least squares inverse, a Principal Component Regression (PCR) for a rank-deficient pseudo-inverse, or Partial Least Squares Regression (PLSR). Partial Least Squares is a widely used inverse least squares solver in chemometrics, since it uses the characteristics of Y to inform the calculation of the latent variable structures in X such that the variables most correlated with Y in X inform the model more strongly. Conversely, poorly informative variables in X do not influence the predicted scores $\hat{Y} = Xb$, and in turn do little to influence the model. PLSR is usually calculated using a variation of the NIPALS algorithm, or more recently SIMPLS [47]. There is a wealth of literature regarding PLSR and its equivalent for discriminant-type problems, Partial Least Squares - Discriminant Analysis (PLS-DA) [29] for the interested reader. As with any supervised learning technique, model validation is critical since the ability for PLSR to correctly model the training data is almost a surety given a matrix with enough variables.

Least-Squares Solvers

Least-squares solvers, such as those used to calculate Equation 1.29, or any latent variable structures as part of the ALS algorithm can be constrained such that the solution meets some user-defined conditions. Typically, this condition is informed by a user’s domain-specific knowledge of the data. While calculating the quantitative scores for a chromatographic peak for example, it may be useful to constrain each score vector unimodally (i.e., such that there is a single peak for each parallel vector). The use of constraints on least-squares solvers is motivated by the idea that the latent variable structures ought to correspond to physio-chemical phenomena that are not at odds with the more general theories of chemistry. An unconstrained least squares solver may just as easily solve for two negative components, whose product amounts to the same positive values observed in the data. While the two components may accurately represent the data when considered together, based on our current under-

standing of physics, there exists no negative value for a possible quantity of chemical. The components themselves may be poorly interpretable in this scenario, since an entire component is not likely to be entirely negative or entirely positive. Constraints are usually applied to data where the data is not typically centred prior to analysis (i.e. MCR-ALS and PARAFAC), and a direct representation of the latent chemical phenomena is required. Negative values are to be expected with most centring, and enjoy usage whenever the interpretation of the relative magnitude of the resultant scores and loadings is required.

The analyst must balance the needs of representing the latent chemical phenomena accurately, versus constraining the model unnecessarily. Generally speaking, unimodality is the harshest constraint that can be applied to chromatographic data; oftentimes bimodal factors can be decomposed into further components, provided that the loadings are sufficiently resolved along the spectral mode. For most applications in chromatographic data analysis, non-negativity constraints are applied in (typically) one mode to gently guide the algorithm to converge to a non-negative solution in the remaining modes.

Non-negative least squares solvers have been historically quite slow, but recent developments by Bro [48] and Van Benthem [49] have enabled much faster non-negative matrix factorisations. Bro also published one of the first unimodally constrained least-squares solver in 1998 [50]. The combinatorial approach proposed by Van Benthem was used in the vast majority of cases in this work.

1.3.4 Feature Selection

Feature selection seeks to identify a useful subset of variables in the data for inclusion, while excluding variables that do not contribute to model performance. Ideally, feature selection will identify a subset of variables that yield a robust model that is resistant to noise and easy to interpret. Feature selection is usually motivated by instances where there are many more variables than samples, in order to simplify the

regression step of the analysis, and to minimise the risk of over-fitting the data.

A number of variable selection techniques exist, which can be broadly classified as belonging to filter, wrapper, or embedded-type methods. All of these techniques have relative advantages and disadvantages [51] [52]. Briefly, filter methods are easy to apply to many different types of data, but require careful optimisation of thresholds. Selection based on a Fisher-ratio threshold is a common example of a filter method [53] [2] [54]. The approach is fast, but since it evaluates each variable independently, it cannot account for relationships between correlated variables. Selectivity ratios and Variable Importance in Projection (VIP) scores are metrics for feature selection which do consider each variable in the context of others - either by examining the weighted variable correlation with the vector of observed values, y , within its projection to the latent variable space (in the case of VIP scores) or the ratio of variance explained to residual variance (in the case of Selectivity Ratios) [55]. Wrapper-type methods reduce user intervention through automated model quality assessment as different combinations of variables and samples are tested and validated, but may still require carefully optimised user parameters. These also require large numbers of iterations in order to find an optimal variable subset. Recursive weighted Partial Least Squares (rPLS) [56] is a modern implementation of a wrapper-type method [57], but methods such as Genetic Algorithms (GA) [58] and Random Forests (RF) [59] have been used as variable selection routines within the framework of wrapper methods as well. Embedded methods incorporate an extra step to make decisions about variable selection during model calculation, independent of model quality assessment. Powered partial least squares discriminant analysis is an example of this technique [60], but embedded methods based on classification by Support Vector Machines (SVM) are also applied [61]. Ideally, this approach makes objective decisions about variable selection, and reduces the dependency on extensive cross-validation, but the extra step increases the computation time required for these techniques. Embedded methods are not as extensively used, perhaps because they assess variable significance based

on optimisation criteria instead of a statistical measure of performance [62].

Hybrid variable selection routines incorporate elements of two or more classes of variable selection, most commonly to reach a compromise between the simplicity of threshold methods, and the reliability of wrapper methods [63] [64]. An initial subset of variables exceeding a particular threshold are considered, and model performance is evaluated once per iteration or multiple times either by adding (forward-selection) or removing (backwards-elimination) high-ranking variables until the predictive ability of the model is no longer improved. Variables that are consistently retained across multiple cross-validation sets are retained, and others are discarded.

PCA can be employed as a data reduction technique that can be used to reduce the complexity of the feature selection problem, and a qualitative tool often used by investigators to examine the effect of variable selection on the most significant axes of variance in the data. Classes well-separated along their most significant principal components typically perform well using a targeted discrimination technique such as Linear Discriminant Analysis (LDA), Partial Least Squares - Discriminant Analysis (PLS-DA), or Support Vector Machines (SVM). As such, it is often convenient to examine variable subsets within their principal component space [65] [66].

Feature Selction by Cluster Resolution (FS-CR)

Feature selection by cluster resolution (FS-CR) is a supervised learning technique that returns a subset of variables whose linear combination provides the best possible separation between two or more sample classes in principal component (PCA) space as determined by the Cluster Resolution (CR) metric. There are two basic assumptions for the operation of the algorithm: That a useful, discriminating subset of variables is of relatively low rank, and can be adequately described using only a few principal components, and that classes resolved along relatively few principal components are trivial to separate using supervised classification methods.

The latest implementation of FS-CR operates within the framework of a hybrid

filter/wrapper method with a combination of backwards-elimination and forward-selection to consider individual variables within the context of others. Variables are first ranked, typically through the application of either Fisher ratio [67] or selectivity ratio [68] [69], and the algorithm evaluates candidate variables via a hybrid backwards-elimination / forward-selection routine [70]. The initial population of variables to be included in the preliminary model is determined through analysis of the true and null distributions of ranking metric values [66]. Backwards-elimination proceeds by sequentially removing variables beginning with the lowest-ranked (based on ranking metric) variable and working towards the highest-ranked. If CR improves when a variable is removed, that variable is permanently discarded; otherwise it is returned and permanently retained. This proceeds until the entire initial population of variables has been tested. Forward-selection is then performed, testing as-yet unconsidered variables to see if their inclusion improves the model based on the variables that survived the backwards-elimination step. Forward-selection proceeds from the highest-ranked variable that was not included in the initial population being considered until a stop condition is met [66].

The FS-CR algorithm has several advantages, including the fact that the utility of a variable is evaluated in the context of the information provided by other variables (unlike in methods such as a Fisher-ratio cut-off threshold). Unlike feature selection methods based on a partial least-squares regression, variables are considered in an unsupervised projection to their principal component space. This helps to reduce the risk of over-fitting because the model is not seeking to impose a favourable projection on the data, but is seeking to find a group of variables that naturally lend themselves to favourable projections via PCA [71]. The results can also be thoroughly cross-validated by redistributing the samples among the training (data which is used to calculate the principal components), optimization (data for which CR is calculated within the previously calculated principal component space), and validation (data that measures the predictive accuracy of the variables) sets through multiple

iterations of the algorithm and retaining only those variables that survive in a given fraction of all iterations (typically 75-90%). Other hybrid feature selection routines are structurally similar to the FS-CR algorithm in this regard, but a metric describing the accuracy of cross-validation results is more commonly employed to assess variable subsets.

FS-CR employs the CR metric to assess variable subsets, which is defined as the maximum confidence interval over which two or more classes can be separated when projected into the principal component space of the candidate feature subset. This is more sensitive to favourable orientations of sample scores, and is much less granular than cross-validation results alone.

FS-CR has been successfully deployed for datasets where the number of features greatly exceeds the number of samples, and in cases where there are many pre-processing artefacts or spurious signals. These situations often arise with weak signals close to the detection limits of analytical instrumentation or in data derived from highly variable populations of samples (often encountered in natural products, petroleum, and metabolomics samples). Two challenges for the application of FS-CR include the need for a relatively large number of samples (30 per class is a typical minimum) so that they can be properly partitioned into training, optimisation, and validation sets, and the relative slowness of the cluster resolution calculation which is performed nC_2 times for each variable being tested where n is the number of classes in the optimisation problem. Calculation of CR scales poorly for multi-class problems: for a three-class problem, CR is calculated three times for each variable considered, and in datasets with seven classes, CR is calculated 21 times per variable [69], as the overall CR for the iteration is calculated as the product of the individual pair-wise binary combinations of different classes. This puts a practical limitation on the number of classes that can be analysed within a reasonable time-frame using this technique.

As mentioned previously, GC×GC-TOFMS data presents thousands of unique chemical features per sample, owing to the high specificity and sensitivity of the

instrument. In most cases, where the chemical characteristics of the samples are complex, not nearly all of these features are identified within every sample. Making this problem worse is the fact that there is no surety that the features identified in individual samples are reliable. Commercial software, which grew out of traditional one-dimensional chromatography software, utilises simple programmatic steps to try and find similar features across multiple samples, provided that these features fall within a threshold of the maximum retention times and exceed a threshold for a mass spectral match factor. Chemical features that do not meet the selected criteria are either excluded from the final peak table, or if there is a sufficient number of features that are consistently different enough in terms of their retention time and extracted mass spectra, they may be assigned into their own column. Within a matrix of samples and observations, each column of chemical features is analogous to an entirely separate dimension within an n dimensional space spanned by features of the data; without reduction of the feature space in some way, it is difficult to associate improperly separated variables with each other.

The raw output of chromatographic instruments are not especially useful. Each observation is a multivariate signal that does not correspond to the presence of a single chemical component, but one of several observations of many potential chemical components in the case of co-elutions. Only once the chemical components have been deconvolved from interfering signals, quantified, and found across multiple samples can a matrix of m samples and n chemical characteristics be generated.

The aforementioned matrix of chemical characteristics is especially important for the interpretation of the resultant model. Knowledge of what chemicals are correlated with an observable outcome offers better insight into the system being studied, than by analysis of the raw signal alone. Additionally, the same signals may drift between the collection of each sample due to minor variations in the operating conditions of the instrument. Correlating the same component across multiple samples within multidimensional separations is even more difficult, as variation is possible in each

chromatographic mode. This necessitates the development of tools that can identify identical chemical factors across multiple samples, despite the very likely outcome that these factors may not always be found in exactly the same place.

Principal component analysis decomposes the matrix into components that best inform the axes of the most variance within the data. Data that is highly correlated with the principal component axes score higher, than data that is not highly correlated. In this way, principal component analysis can mitigate the effect of poorly integrated features on the overall structure of the data. For example, if two features are incorrectly identified as different peaks due to failure of the pre-processing and integration software, they may both score highly along the principal components, even as different features. In effect, it is easier to observe either variables' influence on the model, despite being part of an entirely different dimension in the original data.

1.4 Thesis Objectives

A single GC×GC-TOFMS chromatogram can present thousands of unique chemical features. Of these features, it is likely that only a few hundred will be found across multiple samples. There are currently extremely low standards for commercial GC×GC-TOFMS data analysis platforms. Few of these platforms provide detailed explanations of the algorithms used in their data processing, making objective scrutiny of the mechanism by which raw data are transformed into a peak table difficult. Additionally, data is often stored in proprietary file formats and exporting the data to a generic format is not easily automated, potentially slow, and not practical for routine users. This was the impetus for the development of software to permit the rapid translation of Leco ChromaTOF[®](v4.xx) “.peg” data files into generic formats or to permit them to be read directly into MATLAB[72]. This has made comparison of different software platforms an extremely difficult task [73]; after all, comparing platforms that do not disclose their methodologies is ultimately just a comparison of different brands and trademarks, with no insight to truly explain the reasons for the

observed differences. Additionally, there are dozens of user parameters needed to extract and align features from GC \times GC-TOFMS experiments. This makes processing of data highly subjective, based on the experience and opinions of the user. With few detailed explanations of commercial algorithms, it is difficult to choose optimal parameters for analysis. For example, a common challenge is that a set of parameters that identify and quantify small chromatographic peaks very well, are unlikely to quantify large peaks well. In some software offerings, large peaks quantified with parameters optimised for smaller peaks are likely to be split into several smaller peaks that do not adequately represent the underlying chemical information.

1.5 Thesis Outline

In Chapter Two of this thesis, improvements to the Feature Selection by Cluster Resolution (FS-CR) algorithm will be presented. These improvements are primarily achieved through the development and implementation of a numerical solution to the calculation of the cluster resolution metric. This replaces the previous dynamic programming approach and improves upon the computational efficiency of this calculation by almost 70-fold.

Subsequently, In Chapters Three and Four, applications of the new feature selection routine will be shown, demonstrating its utility and its power. Difficulties associated with generating a reliable, aligned series of peak tables in these studies demonstrate the need for a more powerful, robust, and objective approach to extracting qualitative and quantitative data from a set of raw GC \times GC-TOFMS data, providing the motivation for Chapters Five and Six.

In Chapter Five, an algorithm that can deconvolve multiple coeluting features in a region of GC \times GC-TOFMS data excised from multiple samples in a straightforward and parsimonious way is introduced. This algorithm is all that is needed for a targeted analysis of GC \times GC-TOFMS data, and can be used to generate calibration information for metabolite standards. However it is difficult to generalise the use of

this algorithm to multiple regions of interest without human intervention to select an appropriate number of components to deconvolve. This number of chemical components, denoted as either k or R , is related to a concept known as the chemical rank of a matrix. A novel method for estimating the chemical rank of a matrix is presented in Chapter Six. This may one day enable the automated analysis of series of entire chromatograms without any human intervention.

Chapter 2

An Efficient and Accurate Numerical Determination of the Cluster Resolution Metric in Two Dimensions

2.1 Theory

2.1.1 Cluster Resolution

Cluster resolution (CR) is defined as the maximum confidence interval over which two confidence ellipses (drawn around the scores of their corresponding sample classes) can be separated within a linear subspace. This linear subspace is typically comprised of the first and second principal components. As it currently stands, CR is calculated by increasing and decreasing the value for the confidence interval until a point is reached where the two ellipses are just “touching”. However in addition to being slow, dynamic programming is mathematically unsatisfying. The method works by calculating a number of points along confidence ellipses projected within two or three principal components, and uses graphical methods to determine whether or not they intersect. Improvements to the efficiency of this method have been made by “hopping”[70] between intervals where the ellipses do intersect, and where the ellipses do not intersect. This reduces the number of iterations of the algorithm, and solves some issues regarding the granularity of confidence ellipse calculations. However there is still the need to determine the coordinates of many points multiple times for

each calculation of cluster resolution, and the accuracy scales with the computational workload required for a properly representative graphical determination.

2.1.2 Mathematical Description of Confidence Ellipses

For two uncorrelated score vectors in principal component space, confidence ellipses for one class follow the form [74]:

$$\left(\frac{\mathbf{T}_1}{\sqrt{S_1}}\right)^2 + \left(\frac{\mathbf{T}_2}{\sqrt{S_2}}\right)^2 = \chi^2 \quad (2.1)$$

Where T_1 , and T_2 are vectors containing the first and second principal component scores of the confidence ellipsis, S_1 and S_2 are the variances associated with each principal component, and χ^2 corresponds to the size of the ellipse for a given confidence interval, as defined by the χ^2 distribution. Equation 2.1 can be rewritten in more general parametric form:

$$\mathbf{T}_1 = T_1^0 + \sqrt{\Lambda\chi^2} \cos(\theta) \quad (2.2)$$

$$\mathbf{T}_2 = T_2^0 + \sqrt{\lambda\chi^2} \sin(\theta) \quad (2.3)$$

In Equation 2.2, T_1^0 and T_2^0 refer to the mean of each cluster along the first and second principal components. Λ , and λ describe the major and minor eigenvalues, which correspond to the variance of the data, and θ encompasses the angle associated for a given point along the ellipse. For the majority of cases where T_1 , and T_2 are not completely uncorrelated, the angular components of Equations 2.2 are multiplied by a rotation matrix, R_s , as a function of an angle ϕ . ϕ is calculated as the angle between the major eigenvector of the ellipse, \mathbf{v}_1 , relative to the first principal component of the data:

$$\phi = \arctan \frac{\mathbf{v}_1(2)}{\mathbf{v}_1(1)} \quad (2.4)$$

Where R_s for a two-dimensional case follows:

$$R_s = \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix} \quad (2.5)$$

Previously, CR was calculated by increasing the confidence statistic by the same factor for each of the two confidence ellipses until the point where a collision between the two ellipses was detected (or decreased similarly until a collision no longer occurred). This requires several hundred points around the ellipse to be calculated for each increment. Accuracy is improved by increasing the granularity of the expansions/contractions and/or the number of points along the ellipse, at the expense of computation time. While this is a reliable method of determining cluster resolution, it is computationally costly. Consequently, in this work a numerical solution through minimisation of some cost function is sought.

2.1.3 Derivation of a Numerical Solution

The intersection of two confidence ellipses can be described as the intersection of two lines, for a pair of angles that stem from the centre of each confidence ellipse.

$$\begin{bmatrix} T_1^1 \\ T_2^1 \end{bmatrix} + \sqrt{\chi_1^2} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} T_1^2 \\ T_2^2 \end{bmatrix} + \sqrt{\chi_2^2} \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} \quad (2.6)$$

Where T_j^i refers to the i^{th} confidence ellipse centre for the j^{th} principal component scores. Vector components u, v are shorthand for the following expansions from Equations 2.2 and 2.5:

$$u_1 = \sqrt{\Lambda_1} \cos \theta_1 \cos \phi_1 - \sqrt{\lambda_1} \sin \theta_1 \sin \phi_1 \quad (2.7)$$

$$u_2 = \sqrt{\Lambda_2} \cos \theta_2 \cos \phi_2 - \sqrt{\lambda_2} \sin \theta_2 \sin \phi_2 \quad (2.8)$$

$$v_1 = \sqrt{\Lambda_1} \cos \theta_1 \sin \phi_1 + \sqrt{\lambda_1} \sin \theta_1 \cos \phi_1 \quad (2.9)$$

$$v_2 = \sqrt{\Lambda_2} \cos \theta_2 \sin \phi_2 + \sqrt{\lambda_2} \sin \theta_2 \cos \phi_2 \quad (2.10)$$

For almost any pair of θ_1, θ_2 , depending on their positions relative to the angle of the ellipses, ϕ_1, ϕ_2, χ_1^2 , and χ_2^2 can be solved for by rearranging Equation 2.6.

$$\frac{1}{-(u_1 v_2) + u_2 v_1} \begin{bmatrix} -v_2 & u_2 \\ -v_1 & u_1 \end{bmatrix} \begin{bmatrix} T_1^2 - T_1^1 \\ T_2^2 - T_2^1 \end{bmatrix} = \begin{bmatrix} \sqrt{\chi_1^2} \\ \sqrt{\chi_2^2} \end{bmatrix} \quad (2.11)$$

The euclidean norm of Equation 2.11 can be used to constrain the problem as the minimization of a cost function:

$$\min f(\theta_1, \theta_2) = \sqrt{\left(\sqrt{\chi_1^2}\right)^2 + \left(\sqrt{\chi_2^2}\right)^2} \quad (2.12)$$

2.1.4 Practical Considerations

For an accurate numerical solution to the cluster resolution problem, at the minimum of Equation 2.11, χ_1^2 ought to be equal to χ_2^2 . By minimising Equation 2.12, the results often approach this equality. Ideally, a solution is found when $\partial/\partial\theta_1 = 0$ and $\partial/\partial\theta_2 = 0$, such that the intersection of two lines at the minimum of Equation 2.12 becomes:

$$\sqrt{\chi_1^2} \begin{bmatrix} \partial u_1 / \partial \theta_1 \\ \partial v_1 / \partial \theta_1 \end{bmatrix} = 0 \quad (2.13)$$

$$\sqrt{\chi_2^2} \begin{bmatrix} \partial u_2 / \partial \theta_2 \\ \partial v_2 / \partial \theta_2 \end{bmatrix} = 0 \quad (2.14)$$

Setting Equations 2.13 and 2.14 equal to each other and expanding the differentials yields:

$$\sqrt{\chi_1^2} \begin{bmatrix} -\sqrt{\Lambda_1} \sin \theta_1 \cos \phi_1 - \sqrt{\lambda_1} \cos \theta_1 \sin \phi_1 \\ -\sqrt{\Lambda_1} \sin \theta_1 \sin \phi_1 + \sqrt{\lambda_1} \cos \theta_1 \cos \phi_1 \end{bmatrix} = \sqrt{\chi_2^2} \begin{bmatrix} -\sqrt{\Lambda_2} \sin \theta_2 \cos \phi_2 - \sqrt{\lambda_2} \cos \theta_2 \sin \phi_2 \\ -\sqrt{\Lambda_2} \sin \theta_2 \sin \phi_2 + \sqrt{\lambda_2} \cos \theta_2 \cos \phi_2 \end{bmatrix} \quad (2.15)$$

It is clear that a solution for θ_1, θ_2 can be found that satisfies Equation 2.15 as a system of nonlinear equations. However, there is no guarantee that a solution to this

system of equations would minimise Equation 2.12 nor would a minimum of Equation 2.12 necessarily satisfy Equation 2.15. It is possible to find a minimum subject to the constraints of 2.13 and 2.14 via Lagrange’s method; however, it was shown to be computationally inefficient, and unstable given the complexity of the equations involved, and the potential for undifferentiable points on the optimisation surface (i.e. for two lines parallel to each other such that the χ^2 at which they converge is undefined). The accuracy of the algorithm is therefore somewhat limited, but it will be shown that minimising Equation 2.12 yields a workable approximation by calculating an intermediate of the upper and lower bounds of χ^2 via the mean:

$$\chi_{mean}^2 = \min f^2/2 \quad (2.16)$$

The confidence interval (defined as cluster resolution, ξ , for this particular problem) is calculated from the cumulative χ^2 distribution function with two degrees of freedom (DOF) using the *chi2cdf* function in the MATLAB[®] Statistics and Machine Learning Toolbox [75]. Where:

$$\xi = F(x|\nu) = \int_0^x \frac{t^{(\nu-2)/2} e^{-t/2}}{2^{\nu/2} \Gamma(\nu/2)} dt \quad (2.17)$$

In Equation 2.17, ν refers to DOF, Γ is the Gamma function, and x is the input χ_{mean}^2 value from Equation 2.16.

2.2 Materials and Methods

2.2.1 Implementation

Equation 2.12 was minimised using an implementation of the Nedler-Mead Simplex algorithm available as *fminsearch* in MATLAB[®] 2018b (64 bit)[76]. This algorithm outperformed its equivalent quasi-Newtonian counterparts, both in terms of the reliability and speed of its convergence rate, due in part to its ability to operate without the need for an analytical determination of the gradient at each iteration. The tolerance

for convergence was set at 2.5×10^{-6} for the numerical experiments, and 1×10^{-8} for the classification data. Randomly generated two-dimensional data, simulating scores in the first and second principal components for a balanced dataset were generated (See A). The original (dynamic programming) and new numerical approach to determination of cluster resolution were applied to the data. All computations were performed on a Lenovo ThinkCentre M700 running Ubuntu 18.04 LTS "Bionic Beaver" with 8 Gb RAM, and an Intel i3-6100T CPU @ 3.20 GHz.

The most recent implementation of the FS-CR algorithm was used for selecting discriminating features in the experimental data using both the current dynamic programming, and proposed numerical method for determining CR. Variables were ranked using Fisher Ratios and populations for backwards elimination and forward selection were calculated using experimental true and null distributions of significant features [66]. The numerical implementation of the CR algorithm has been made freely available online: [10.5281/zenodo.4064280](https://doi.org/10.5281/zenodo.4064280).

2.2.2 Experimental Data

A dataset comprising the volatile organic chemical signatures of 162 samples of cotton and polyester fabrics recovered from a wear trial, wherein participants each wore bi-symmetrical shirts comprised of one-half cotton, and one-half polyester fabric was used to compare the algorithms with real data. Details of the wear trial can be found elsewhere [77], but the stated goal of the analysis was to find discriminating chemical signatures between the cotton and polyester samples indicating which compounds were particularly well-retained on the different fabrics following multiple wear-wash cycles. Both washed and unwashed samples were categorised only as belonging to either the cotton or polyester classes. The sampling was performed using Solid Phase Micro-Extraction (SPME) fibres with a "tri-mode" divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) extraction phase (SUPELCO, Bellefonte, PA). Extractions were performed on the headspace of 2.0×2.0 (± 0.2 cm) samples of fab-

ric, sealed within 10 mL crimp-top vials at 30 °C for 21 h. The potentially large in-class variation makes for a somewhat challenging dataset for classification.

2.3 Results and Discussion

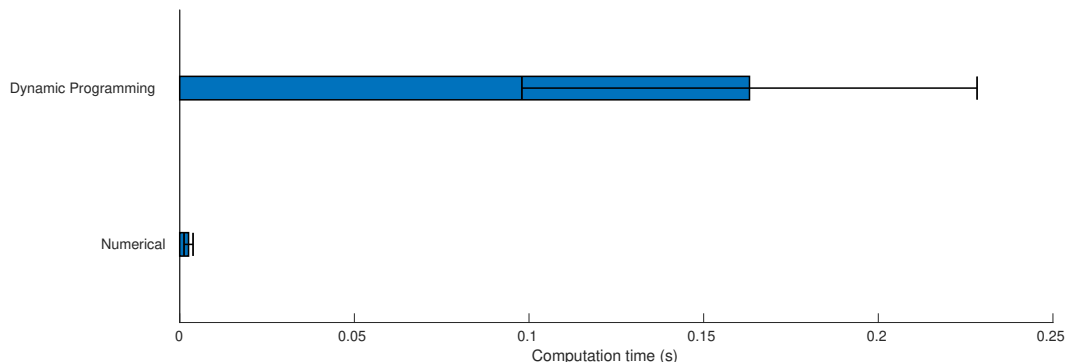


Figure 2.1: Average computation times for numerical and dynamic programming determinations of cluster resolution for two clusters. Error bars indicate $\pm s$.

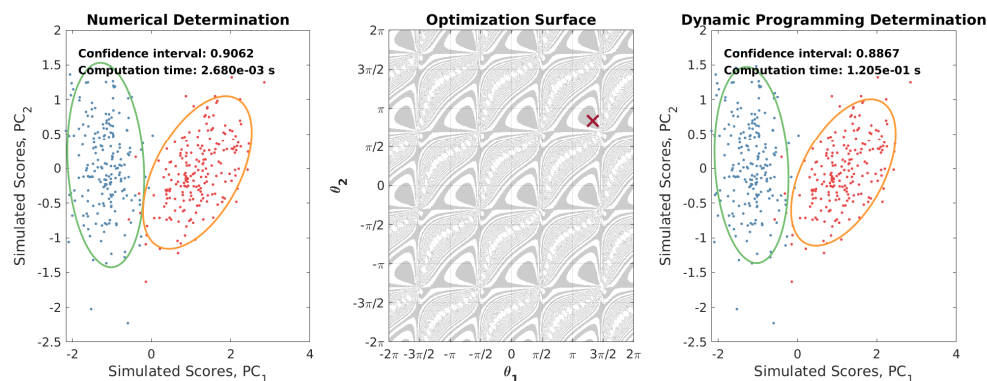


Figure 2.2: Example calculation, comparing the numerical and dynamic programming implementation of the cluster resolution metric, with the corresponding optimisation surface. “X” is the location of the optimum values for θ_1 , and θ_2 found by the Nedler-Mead Simplex algorithm via the minimisation of Equation 2.12.

In the absence of an analytical solution to the cluster resolution metric, a numerical experiment consisting of 200 randomly generated data sets was used to compare the performance of the numerical implementation of the algorithm with the current version. The dynamic programming implementation requires an initial guess for CR,

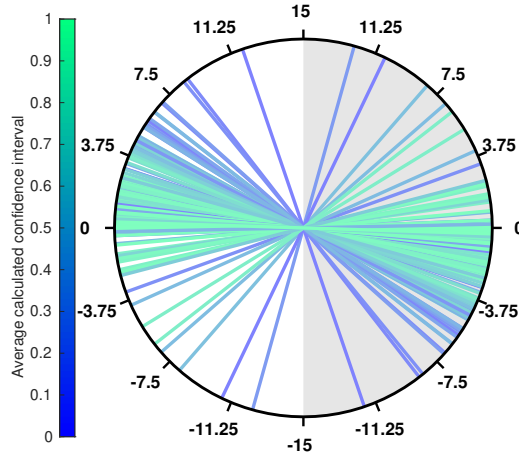


Figure 2.3: Secant plot comparing the relative percent differences in the determination of cluster resolution using the numerical (left, white hemisphere) vs. dynamic programming (right, grey hemisphere) approach for data with sample classes. All calculations agreed within a 15% relative error. Lines intersecting with the edge of the circle at the left side of the circle indicate the % relative difference of the numerical method for the determination of cluster resolution, for an averaged value for cluster resolution indicated by the colour of the line - similar to the intersection with the right hand side of the circle, expect with respect to the % relative difference for the dynamic programming approach.

and 0.75 was used, as this is a typical initial guess used in practice. The numerical solution does not require an initial guess for CR. An example solution and the corresponding optimisation surface for the numerical method is shown in Figure 2.2. For $N = 200$ sets of randomly generated data, the average time required to calculate cluster resolution was 2 ± 1 ms using the numerical method, and 163 ± 65 ms using the current dynamic programming approach. For two clusters in two-dimensional space, the numerical method is on average 65 times faster (Figure 2.1). The two methods provided similar results that agreed within 15% (2.3) and the ellipses do not appear to significantly overlap in any of the solutions from the numerical method (See: Supporting Information 1). Comparing the two methods with a secant plot (Figure 2.3) shows the tendency for the the dynamic programming approach to underestimate cluster resolution vs. the new, numerical approach.

2.3.1 Calculation of the cluster resolution metric for N clusters

The number of times cluster resolution is calculated per evaluated variable depends on the binomial coefficient, ${}_nC_k$ where $k = 2$ and n is the number of sample classes. This is due to the fact that it is necessary to calculate the cluster resolution between each pair of classes to evaluate the overall cluster resolution for the model (Ξ). Consequently, computation time scales poorly for variable selection problems with more than two classes. In general, the overall cluster resolution is calculated as the product of the individual cluster resolutions for each possible combination of clusters:

$$\Xi = \prod_{n=2} {}^nC_2 \xi \binom{n}{2} \quad (2.18)$$

To compare the compounded improvement for multi-class problems offered by the numerical approach vs. the current approach, randomly generated data sets simulating n -class problems were generated as before ($2 < n < 7$). 30 data sets were simulated per value of n . Results are summarized in Figure (2.4). For the 7-class problem, a single Ξ calculation required 3.3 ± 1 s using the dynamic programming approach, and 0.040 ± 0.005 s using the numerical approach, an 82-fold improvement.

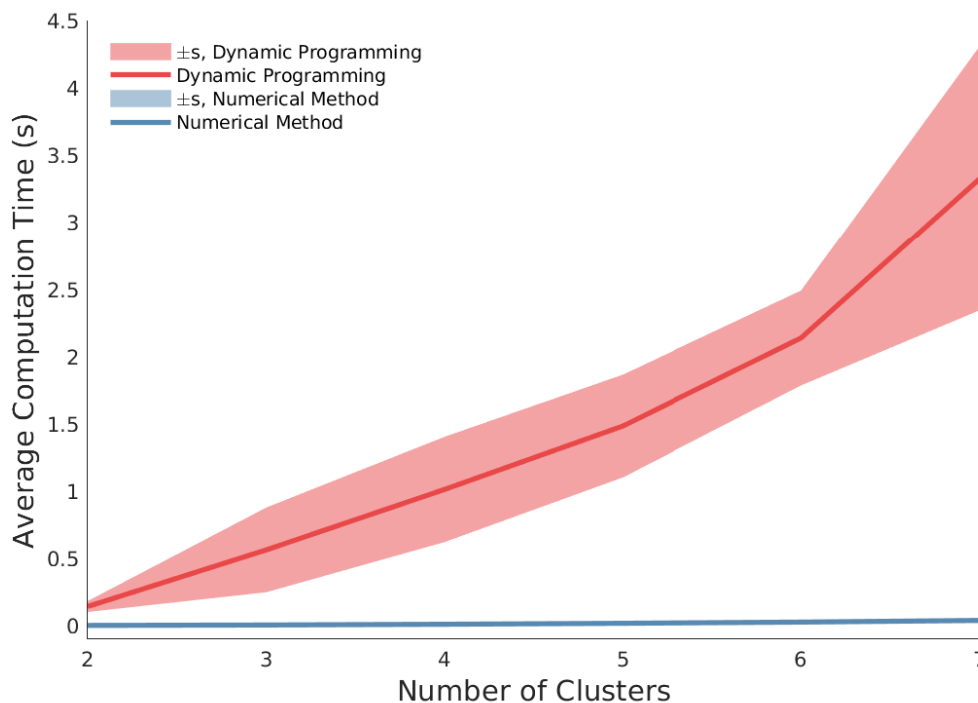


Figure 2.4: Average computation time for Ξ in N -class problems for $2 < N < 7$ for numerical and dynamic programming methods. s refers to sample standard deviation for the dynamic and numerical computation times.

2.3.2 Comparison of Predictive Capabilities

Feature Selection by Cluster Resolution has proven to be extremely useful for selecting useful subsets of sparse datasets, typical of peak tables generated GC \times GC-TOFMS data, and so one such dataset was used from a previous study [78]. The dataset is available at: <https://doi.org/10.7939/DVN/RLMSRW>.

Partial Least Squares Discriminant Analysis (PLS-DA) was used to generate a classification model; the discrimination threshold for predicted Y scores was generated using a Bayesian technique [79]. The external validation set was used to evaluate prediction results, following strict class membership assignment designated by the aforementioned threshold. Results for predictions were made using PLS-DA without feature selection (Figure 2.5, row 1), with the current dynamic programming (DP-FS-CR) implementation (Figure 2.5, row 2), and the numerical (NM-FS-CR) imple-

mentation (Figure 2.5, row 3), and summarized using predicted Receiver-Operator Characteristics (ROC) and prediction accuracy, where prediction accuracy is defined as the ratio of the sum of the true positive rate (TP) and true negative rate (TN) over the sum of all prediction rates including the false positive (FP) and false negative (FN) rates:

$$Accuracy = (TP + TN) / (TP + FP + TN + FN) \quad (2.19)$$

The dataset contains a high number of replicates, and was divided evenly between training and validation sets for a critical analysis of each models' predictive ability. External validation samples were centred and scaled according to values calculated in the training set. All results are presented with respect to Class 1 (cotton samples) versus Class 0 (polyester samples). Within the training set, 200 combinations of training and optimization sets were generated and used for both DP-FS-CR and NM-FS-CR routines, with variables selected at least 90% of the time across all sample combinations being included in the final feature subset. In the training set, there were a total of 81 samples: 63 class 1, and 18 class 0. In the validation set there were also a total of 81 samples, with 61 class 1 and 20 class 0 samples.

Results from the confusion matrices and predicted ROC curves suggest that the variables selected using NM-FS-CR perform better than the much slower DP-FS-CR algorithm in terms of predictive ability. DP-FS-CR selected a total of 32 variables, and NM-FS-CR selected a total of 46, with total computation times of 14760 and 1468.3 seconds respectively. All but two variables that were selected using DP-FS-CR were also selected using NM-FS-CR (Table 2.1), in addition to 16 variables that were unique to the NM-FS-CR method. It was previously shown that using a simple thresholding method for feature selection with this dataset was unsuccessful [2].

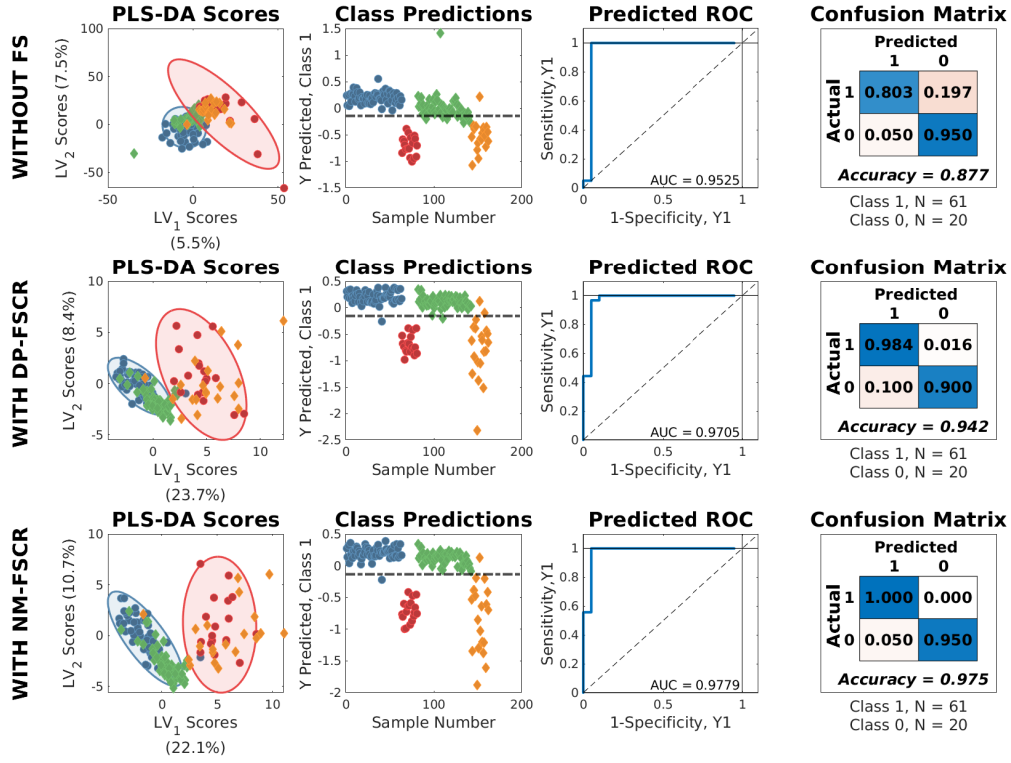


Figure 2.5: Summary of classification results using PLS-DA, DP-FS-CR, and NM-FS-CR. Confidence ellipses are displayed for a confidence interval of 95%

2.4 Conclusions

Cluster resolution is a useful metric for evaluating model quality and guiding variable selection routines. Cluster resolution permits consideration of favourable changes to the relative positions and orientations of score clusters representing the distribution of sample classes in principal component space, without relying solely on cross-validation results as is done with other methods. Previously, there existed no mathematical formalization of the cluster resolution metric, and its determination relied on dynamic programming. The speed and accuracy of the dynamic programming method depends primarily on the number of points used in the confidence ellipse projections, and a reasonable initial guess for the cluster resolution. The numerical solution to

DP-FS-CR	NM-FS-CR	Common
32	46	30
1	1	1
3	3	3
4	4	4
	5	
9	9	9
11	11	11
21	21	21
22	22	22
23	23	23
28		
30	30	30
35	35	35
	54	
	56	
	69	
75	75	75
76	76	76
78	78	78
79	79	79
84	84	84
85	85	85
	109	
123	123	123
141	141	141
148	148	148
	165	
228		
236	236	236
	260	
	280	
308	308	308
336	336	336
458	458	458
483	483	483
610	610	610
	662	
	806	
	912	
1022	1022	1022
1342	1342	1342
1573	1573	1573
	1614	
	1763	
1842	1842	1842
2230	2230	2230
	2531	
	2708	
	2766	

Table 2.1: Variables selected using the DP-FS-CR and NM-FS-CR feature selection routines. NM-FS-CR identified all but two variables identified using DP-FS-CR in addition to 16 variables that were not identified using DP-FS-CR.

the calculation of cluster resolution presented herein has demonstrated a substantial improvement in computation speed, while generally maintaining or improving the accuracy of the calculation. Preliminary results show that the variables selected using the new numerical determination largely encompass the variables selected using the dynamic programming approach, while also identifying additional useful variables. Including these previously hidden variables has been shown to improve the predicted ROC and prediction accuracy of the model.

The improvements in speed make it feasible to analyse many more combinations of training and optimisation sets, and a greater number of classes within a reasonable time-frame. As with any hybrid feature selection method, this extensive cross-validation is generally considered to improve the robustness and predictive accuracy of the model.

Although employed here as a feature selection routine, CR is a generally useful metric for model quality that can be used in conjunction with validation and residual analysis in any linear space to describe the expected utility of classification models. It is the authors' hope that the described mathematical formalization and freely-available code will enable its use in a variety of different fields where multivariate classification problems are encountered. Further studies are necessary to derive cost functions for the resolution of N -Dimensional confidence ellipses, and validate the applicability of this method for determining CR in higher dimensional PCA space.

An online tutorial for constructing confidence ellipses was instrumental in making this work possible [80]. Automated colour palette generation made use of *linspecer.m* [81]. Figure 2.4 also utilised code for visualising error bars as a translucent background [82].

Chapter 3

Global Metabolome Analysis of *Dunaliella tertiolecta*, *Phaeobacter italicus* R11 Co-cultures using Thermal Desorption - Comprehensive Two-dimensional Gas Chromatography - Time-of-Flight Mass Spectrometry (TD-GC×GC-TOFMS)

3.1 Introduction

Microalgae oils represent a source of energy-rich fatty acids and lipids that have long been considered as a carbon-neutral source of biologically-derived fuels [83] [84]. Limiting the commercial applicability of microalgae oil are the long times required for algal growth and the limited yields of oil. In addition to optimisation of growth conditions, bacterial co-cultures have been explored as a way of improving yields of microalgae oils [85]. These co-cultures have improved the rate of growth for microalgae colonies via directly observable mechanisms such as the synergistic exchange of oxygen and carbon dioxide for aerobic bacteria[86], or the release of extracellular compounds into the growth media [87]. For many co-cultures, although a change in the microalgal

growth is apparent, there is no clear mechanism that can be inferred through analysis of the growth media alone. In these cases, it may be useful to examine the biomass directly.

Metabolomics has been used to great effect to correlate differences in small-molecule metabolite expression with macroscopically observable phenomena. Algae metabolomics in particular has seen some interest [88] [89], and has been used to gain insight into chemical responses of microalgae to changes in their environment. The relative expression of these chemicals can be used to deduce the ecology of the bacterial-algal relationships as either mutualistic [90] [91] or antagonistic [92], and such insight may be used to further improve upon the microalgae oil yield.

Comprehensive two-dimensional gas chromatography - time-of-flight mass spectrometry (GC×GC-TOFMS) is a powerful analytical tool for the non-target examination of the chemical diversity of samples of volatile and semi-volatile organic compounds owing to its improved resolution, sensitivity, and identification capabilities over traditional one-dimensional gas chromatography - mass spectrometry (GC-MS). For non-volatile species and those which chromatograph poorly, such as lipids, fatty acids, and amino acids, extra derivatisation steps are required. These steps hydrolyze bonds in large molecules (e.g. triacyl glycerides) and substitute groups such as methyl-, ethyl-, or trimethylsilyl- for labile protons, enabling subsequent analysis in the gas-phase. Techniques specific to a particular class of chemical compound, usually fatty acids, are common in algae metabolomics since only a limited number of chemical classes dominate the composition of microalgae oil. Analysis of Fatty Acid Methyl Esters (FAMEs) is a common way of determining fatty acid expression by GC-MS, for example. Targeted analysis of a limited number of chemical species often fails to explain the observed phenomena however. Global metabolomic profiling aims to encompass the broadest possible scope of all small-molecule metabolites, for the similar aim of correlating changes in abundances of some small number of metabolites with the different classifications of populations comprising the study data set (e.g.:

healthy/diseased or different cultures of microalgae).

GC×GC-TOFMS often reveals several thousands of unique chemicals in a metabolomics study. For a properly optimised analysis, most of these chemicals originate from the biomass itself. However contamination at some stage of the sample preparation is inevitable, especially for samples requiring derivatisation, as the reagents involved (being reactive) are particularly difficult to purify. Even for chemical features that are biological in origin, typically only a small subset of these features contain useful or discriminating information. As such, for a limited number of samples, a feature selection step is necessary to create informative models that are easy to interpret.

This chapter presents a workflow for preparing microalgae samples for global metabolomic analyses that includes a novel technique for introducing the derivatised sample matrix into the gas chromatograph. Using thermal desorption (TD), a concentrated sample of microalgae extract can be introduced directly into the instrument without a prior filtration step. The injected sample is deposited into a small insert in the thermal desorption unit, and heated to transfer the volatile and semi-volatile components to a cryogenically cooled inlet. All low-volatility components that would otherwise be deposited into the inlet itself remain within the insert inside of the TD unit. Subsequent pyrolysis of heavy biomass residues between sample injections is avoided in this way, and the analyses are relatively free of interference and contamination. 70 samples of *Dunaliella tertiolecta*, *Phaeobacter italicus* R11 [93] [94], and co-cultures of the two species were cultured and filtered for analysis. A useful sample normalisation and feature selection routine is demonstrated on this dataset, and the authors present a short list of candidate metabolites that may be biologically interesting for future study.

3.2 Materials and Methods

3.2.1 Growth and maintenance of algal and bacterial strains

The *Dunaliella tertiolecta* CCMP 1320 strain was obtained from the Provasoli-Guillard National Centre for Marine Algae and Microbiota (NCMA). The chlorophyte was maintained in L1-Si media made with artificial seawater (35 g/L of Instant Ocean, Blacksburg, VA, USA), at 18 °C with a diurnal incubator cycle (12:12 hour dark-light cycle). Samples from the cultures of microalgae were examined microscopically to rule out bacterial contamination before experimental use. These samples were also inoculated onto marine agar plates, and incubated at 28°C for three days to identify any colony forming units (CFUs). (18.7 g of Difco Marine Broth 2216 with 9 g NaCl and 15 g Difco agar in 1 L) . Experimental use of the algal cultures proceeded once a cell concentration of 10^4 cells/mL was reached.

Samples of *Phaeobacter italicus* R11 were acquired from Botany Bay, Australia [93]. The bacterial cultures were maintained at 28 °C on the aforementioned marine agar plates, then transferred to 5 mL 50% dilute marine broth media (2216 Marine Broth, Difco) where it was grown until reaching a stationary phase for 24 hours before the experiments. Cell concentration for stationary phase *Phaeobacter italicus* R11 was similar to the cell concentration of the algae cultures, at 10^4 cells/mL.

3.2.2 Preparation of samples

Algal-bacterial co-cultivation experiments were performed as described by Bramucci et al. [95] in 12 well plates (Standard TC Growth Surface, Bacto (Oakville, Canada)). Briefly, stationary phase bacterial colonies of *Phaeobacter italicus* R11 were washed twice by centrifugation and re-suspended in L1-Si medium before dilution to the target cell concentration 10^4 colony-forming units (CFU) /mL. For co-culture samples, *D. tertiolecta* and *Phaeobacter italicus* R11 were mixed in a 1:1 ratio by volume at equivalent cell concentrations in L1-Si medium made with artificial seawater. Mono-

culture controls of both *D. tertiolecta* and *Phaeobacter italicus* R11 were inoculated in 1:1 (v:v) ratio with L1-Si made with artificial seawater. Mono- and co-cultures were aliquoted in 6 mL volumes into 12-well plates and grown in a diurnal incubator with a 12 hr dark-light cycle at 18°C. During the mid-point of each dark cycle, 20 μ L of each sample were plated onto a 1.5% agar, 1/2 marine broth plates and incubated for 24 h at 28°C to confirm the absence of bacterial in mono-culture samples, and enumerate the bacteria in co-culture samples. All samples were cultured until the stationary phase, after 18 days.

3.2.3 Collection of culture samples

On day 18 of each samples' incubation, each sample was collected by vacuum filtration onto pre-weighed glass fibre filters (0.22 μ m). Each sample was rinsed three times with 1 M PBS buffer to wash away the growth media. Filters were placed into clean, glass vials covered with Kimwipes and stored at -80 ° before being lyophilised for 24 hours. After the drying step, the filters were weighed once again to record the biomass. The biomass was recorded to (\pm 0.001 g), which was insufficient to accurately determine the mass of either bacterial culture samples, or the algae and culture samples. The bacterial samples were recorded at 0.000 ± 0.002 g. Co-culture, and *D. tertiolecta* samples presented an average biomass of $0.005 \text{ g} \pm 0.002 \text{ g}$ and $0.005 \text{ g} \pm 0.002 \text{ g}$ respectively.

20 samples of *D. tertiolecta* + 1 replicate, quality control sample, 23 samples of *Phaeobacter italicus* R11 + 1 replicate, quality control sample, and 27 co-cultured samples + 3 replicate quality control samples were prepared and analysed for the experiment. One quality control sample, part of the co-cultured samples, indicated poor agreement with its corresponding sample, and indicated a change in the tightly-controlled analytical conditions of the instrument. This sample was excluded from the final dataset, bringing the total number of samples (including quality controls) to 74. Careful inspection of data for samples before and after this particular sample

indicated that this one anomalous result was the result of an isolated incident that did not affect other samples.

3.2.4 Sample preparation and derivatisation

Liquid measurements were performed using either a 100- μ L or 1000- μ L Microman positive displacement pipette and disposable pipette tips. For measurements of pure reagent, a reusable positive displacement syringe system with a digital readout was used (SGE eVolTM handheld automated analytical syringes).

The sample preparation procedure is divided into two main steps. During the first step, an extraction protocol based on the widely-known Bligh and Dyer [96] [97] method for the analysis of fatty acids was used to extract and separate the macromolecular plant material from the plant metabolites. Use of the chloroform extract appeared to be a good choice of extraction solvent and offered decent coverage of a wide variety of different chemical classes. During the second step, trimethylsilyl derivatives of the extracts were generated based on the protocol of Chan et al. [98].

The glass fibre filter papers (GFFPs) used to collect each sample were submerged in 7-mL volumes of HPLC-grade methanol (>99.9%, Millipore-Sigma Canada) in 20-mL scintillation vials (Chromatographic Specialties Inc., Oakville, ON, Canada) using a clean spoonula. To each vial, 7 mL of HPLC-grade chloroform (>99.8%, Millipore-Sigma Canada, Oakville, ON, Canada) were added, and sonication proceeded for an additional hour. 3.5 mL of 18.2 M Ω deionised water (Elga PURELAB flex 3 system, VWR International, Edmonton, AB, Canada) was then added to effect a separation into a polar methanol/water layer and a chloroform layer, which was allowed to rest overnight at 3 °C prior to extraction. 1.8 mL of the bottom (chloroform) layer was extracted and transferred into 2-mL glass GC vials. For replicate measurements, an additional 1.8 mL was transferred to an additional vial. The extracts were blown down with nitrogen at 40 °C until there was no visible chloroform remaining, approximately two hours. 100 μ L of toluene was added to the dry residue, which was

then vortexed briefly before being evaporated under nitrogen at 40 °C once again. The extra evaporation step using toluene ensures there were no remnants of moisture in the vials as even traces of moisture interfere with the subsequent derivatisation steps. 50 μ L of methoxyamine (Millipore Sigma, Oakville, ON, Canada) in HPLC-grade pyridine (Millipore Sigma, Oakville, ON, Canada) was added via a digital positive displacement syringe, and the solution was incubated at 60 °C for two hours. Following this, using another positive displacement syringe, 100 μ L of N-Methyl-N-(trimethylsilyl)trifluoroacetamide + 1% trimethylchlorosilane (MSTFA + 1% TMCS, Fisher Scientific, Ottawa, ON, Canada) was added and vials were incubated at 60 °C for an additional hour. 100 μ L of the resultant solution was transferred to 1.8 mL vials with fused glass 300- μ L inserts. These vials were stored at 3 °C for up to 48 hours prior to analysis.

3.2.5 Sample introduction and operating conditions

3.2.6 Thermal desorption, sample introduction

The chloroform extracts were not centrifuged to remove non-volatile components from the extract solution. Being a relatively non-polar solvent, exposure of chloroform to plastic centrifuge tubes would present a significant risk of leeching contaminants from the vials into the samples. To avoid this, the authors opted to inject the unfiltered, derivatised extract directly into TDU insert tubes, with subsequent evaporation of (semi-)volatile components directly from the insert tube, leaving the non-volatile components in the insert. Inserts were replaced for every analysis.

A Gerstel MPS autosampler and sample preparation robot equipped with a 10- μ L Gerstel TriStar Liquid Syringe, a Thermal desorption Unit (TDU 2), and a programmed temperature vaporization inlet (CIS4; Gerstel US - 701 Digital Drive, Suite K, Linthicum, MD 21090). The MPS system was programmed in Automated TDU-Liner Exchange (ATEX) mode, where 9- μ L aliquots of sample were injected into clean TDU tubes containing a disposable microvial insert. TDU tubes (straight tubes

with notch), were rinsed with high-purity (99.9%) toluene (Millipore Sigma Canada, Oakville, Ontario) and baked in an oven at 400 °C for 1 hour between runs, with new microvial inserts baked inside of the TDU tubes. Both the microvial inserts and TDU tubes were allowed to cool inside of the oven before being fitted with transport adapters for liquid injections. The transport adapter seals the microvial inside of the TDU tube, maintaining cleanliness. A Teflon-coated septum in the adapter maintains carrier gas pressure before and after the liquid injections. The liquid syringe was washed 6 times with each of 10 μ L of 1:1 (v/v) acetone:hexane and HPLC grade methanol (Fisher Scientific Co, Edmonton Alberta) both before and after the liquid injection.

The TDU was operated in solvent vent mode, followed by a splitless injection from the TDU to the inlet, and then splitless injection from the inlet to the GC \times GC-TOFMS. During the solvent vent step, the TDU was kept at an initial temperature of 128 °C, and the TDU split vent was open for 5 min to vent the pyridine and MSTFA solvent mixture. The TDU was fed with a constant flow supply of ultra-high purity helium carrier gas (Linde Canada (formerly Praxair), Edmonton Alberta, Canada) at 50 mL/min with the permanent split vent from the TDU set to 2.5 mL/min. The remaining flow during the solvent vent step exited the system via the split line from the CIS. Following solvent venting, the temperature of the TDU was raised to 280 °C and injected via a splitless injection into the CIS at a flow rate of 50 mL/min for an additional 5 min. During this step, the CIS was maintained at a temperature of 30 °C with the split valve open, until the splitless injection from the TDU was completed. 30 °C was the lowest practical temperature that could be reliably maintained with the cryogenic cooling system during the run. In the next step, the CIS ramped to 300 °C and injected with the split vent closed into the GC \times GC-TOFMS. During this time, the TDU ramped to 300 °C to clean the system, with the TDU split vent open. The CIS operated in splitless mode for 150 s at 300 °C, until the split vent was opened, at a total gas flow rate of 252 mL/min for a constant flow rate of 2 mL/min

delivered to the head of the column. The inlet liner was baffled (Gerstel, US) to trap volatile components, while allowing for an effective purging step between each run, and system cleanliness was monitored via instrument blanks (one fast blank after every sample, and one instrument blank under identical operating conditions to those used for the samples twice per batch of 14 samples).

3.2.7 GC×GC-TOFMS method

The samples were separated on a LECO Pegasus 4D system (LECO, St. Joseph, MI, USA) outfitted with a quad-jet dual-stage cryogenic modulator. The column set featured a 60 m × 0.25 mm internal diameter; 0.25 μ m film thickness Rxi-5SilMS in the first dimension, and a 1.4 m × 0.25 mm internal diameter; 0.25 μ m film thickness Rtx-200MS second dimension column (Chromatographic Specialities, Brockville, ON, Canada). The initial oven temperature was set to 80 °C, held for 4 min, and ramped at 3.5 °C/min to a maximum oven temperature of 315 °C with a 10 min final hold. The temperature program and flow rate of the method were directed by considerations for speed optimised flow (SOF) [14] and optimal heating rate (OHR) [13] derived from the column geometry and dead-time, respectively. The secondary oven offset was set at +10 °C relative to the primary oven temperature and the modulator temperature offset was set at +15 °C relative to the secondary oven temperature. The modulation period (P_M) was 2.50 s, with a hot pulse time of 0.60 s and a cool time of 0.65 s between stages. The mass spectrometer collected spectra at 200 Hz, from 40 to 800 (m/z). The electron impact ionisation energy was -70 eV. The ion source temperature was 200 °C, with a transfer line temperature of 300 °C. An acquisition delay of 650 s was used to ensure residual solvent did not damage the filament.

3.2.8 GC×GC-TOFMS data pre-processing

The data were pre-processed using LECO ChromaTOF® version 4.72. Baseline offset for peak detection was set as a factor of 1.2 above the estimated noise level. Antici-

pated peak widths were determined through a survey of 10 different peaks, both large and small. The authors opted to direct the pre-processing parameter optimisation towards detecting smaller peaks, so the peak size parameters were as follows: 10 s for first-dimension peak width, and 0.1 s for second-dimension peak widths. The second-dimension sub-peaks were combined if their deconvolved mass spectra met a match factor of 650, and sub-peaks were only integrated if their signal-to-noise ratio (SNR) was greater than a value of 6.

Peaks were integrated into the final peak table, if the SNR for the base peak was greater than a value of 15, with 5 or more apexing masses. Peaks across multiple samples were aligned if they were within two modulation periods of each other in the first-dimension, or within 0.2 s of each other in the second-dimension. Quality control samples were included in the alignment procedure, bringing the total number of samples to 75 (including one sample that was later discarded). Before being included in the final table, an analyte must have populated at least 33% of the samples (25 samples), otherwise that analyte was discarded. This parameter was selected to minimise the very common phenomenon of peak dropout in processed GC×GC-TOFMS peak tables, and was not based on the population of unique classes of sample. Since the number of variables for typical GC×GC-TOFMS experiments is much higher than the number of samples, spurious class separations for targeted classification are very common for poorly optimised peak tables. However considering the class makeup of the dataset, chemical components unique to *D. tertiolecta* and co-cultured samples, *Phaeobacter italicus* R11 and co-culture samples, as well as the co-culture class itself, are not disallowed from the final peak table under these conditions.

Peak searching for analytes not recognised in the sample-wise pre-processing were searched for down to a SNR ratio of 12.5. Initial library searching for all peaks was performed for all samples using the National Institute of Standards and Technology (NIST) mass spectral database (2017).

3.2.9 Data analysis

The final peak table was exported from ChromaTOF[®], and into MATLAB[®] 2020b. Some functionalities of PLS Toolbox (R8.5.2; Eigenvector Research Inc.) were used for the receiver-operator characteristics (ROC), as was the MATLAB[®] Statistics and Machine Learning Toolbox. All data used in this work are available online at <https://doi.org/10.20383/102.0510>.

3.2.10 Sample normalisation

Samples were normalised according to a class-based total useful peak area (cTUPA) criterion, based on the absence of a reliable biomass measurement, and the need to account for variability in the quantity of sample that eventually reached the instrument. Using this method, peak areas were divided by the total peak area of analytes detected in each sample of a given class. This allowed for a reasonable, and simple method for comparing highly dissimilar samples (i.e. comparing a small bacterial biomass against a much larger algal biomass) and extracting a useful discriminating variable subset that best describes their differences. A similar approach, agnostic to class labels, was reported for a study conducted on GC×GC-TOFMS using human urine [99].

3.2.11 Feature selection, cross-validation

The normalised data were analysed using feature selection by cluster resolution (FS-CR), a hybrid wrapper/threshold method for selecting features based on favourable projections within a principal component subspace. This technique has shown to be effective on a number of GC×GC-TOFMS datasets, and includes a robust cross-validation routine that trains, optimises, and validates the model during each iteration. To demonstrate the effectiveness of this tool, a series of ROC curves were generated by reshuffling the data for each iteration, utilising a variable subset that survived a certain ratio of previous combination of the data. A PLS-DA model with

two latent variables was trained using half of the reshuffled samples, and validated using the other half of the reshuffled samples. Further details are presented in the Results section.

The FS-CR algorithm was operated for projection into two-dimensional principal component subspaces, using autoscaled data (i.e. data that was mean-centred and scaled via each variable’s standard deviation). The most recent version of the algorithm was selected, utilising a numerical determination of the cluster resolution metric[100], which allowed for 200 iterations of the algorithm to complete in about 30 min. Variables that survived 90% of all iterations were ultimately selected for the final model.

Mass spectra and retention indices for analytes of interest were recovered from the raw data and a library search using the Golm Metabolome Database [101] was performed.

3.3 Results

Each of the resultant chromatograms are rich in chemical information, but at first glance they are visually quite similar. The colour axes of Figures 1-3 are scaled to the same maximum TIC of 7.5×10^5 . In spite of the high sample load, instrument blanks were clean between runs, but despite the authors’ best efforts, the reagent blank itself presents a great deal of interfering chemical information (See: Appendix B). This is likely unavoidable, as the purity of MSTFA used for derivatisation is generally quite poor (98%), and the pre-concentration step during the sample introduction likely exacerbated this problem.

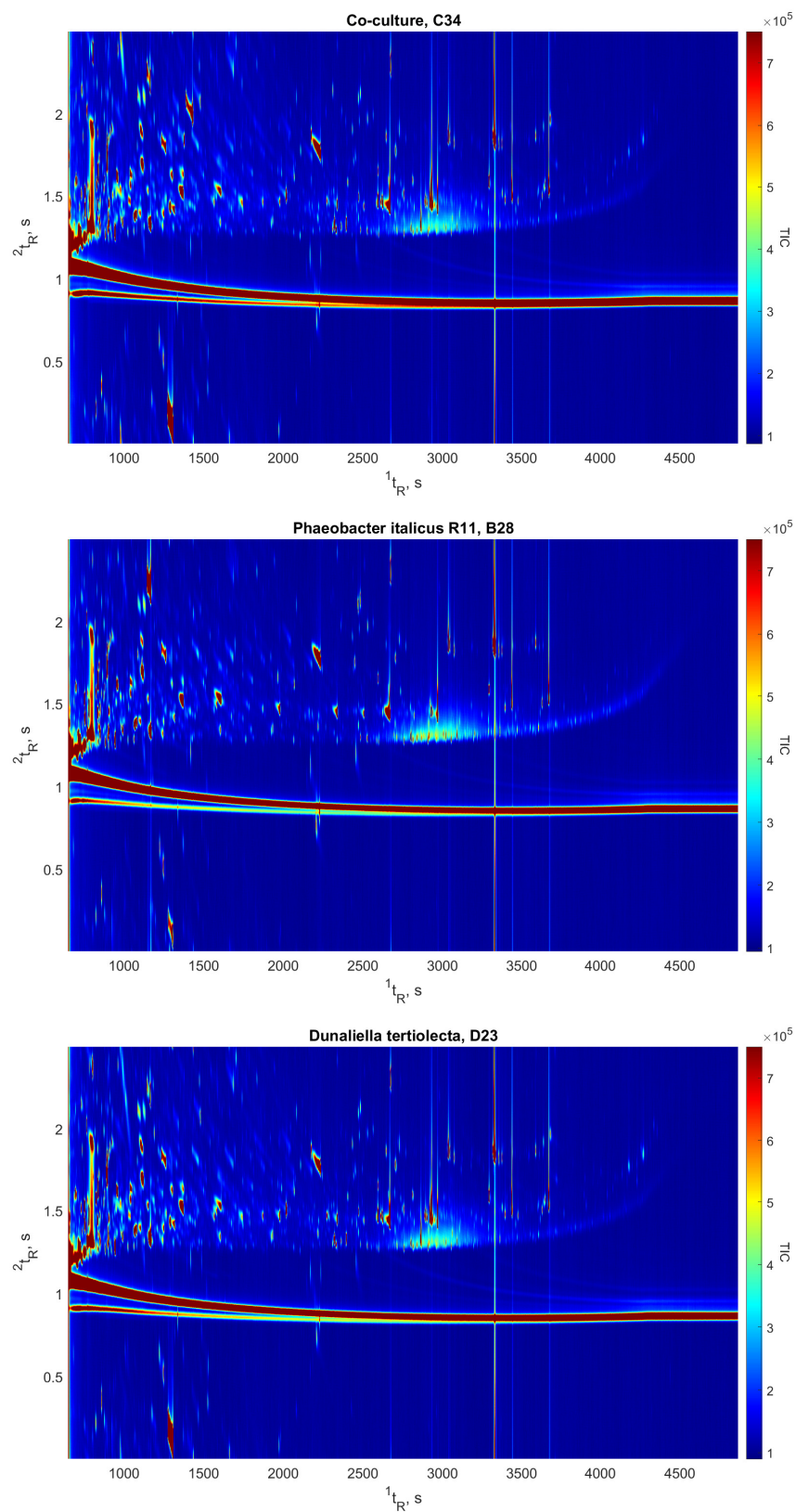


Figure 3.1: Example Total Ion Current (TIC) chromatograms from each sample class

Prior to analysis, all data was autoscaled. For the ROC curves, validation data was centred and scaled according to the values obtained in the training set for a more critical assessment of the model performance. The principal component analysis of the raw data revealed some bimodality, likely due to the number of interfering chemicals present in the reagent blank. Following cTUPA, the severity of the bimodality was reduced, and the cluster of bacterial samples was well resolved from the chemically similar co-culture and mono-culture microalgae samples (Figure 3.2). This aligns well with our initial expectations of the data.

Using the selected features, the samples appear to be normally distributed about the two axes of variance within each cluster. This suggests that the extracted features are robust against interfering chemical information, and that the previously observed bimodality of the data may have been a chemical, or instrumental artefact of the analysis that did not significantly affect projections using the selected features (Figure 3.2).

In-class variance within the co-cultured microalgae samples appears to be much higher than in the mono-culture samples, according to the projection within the subspace of the selected features. Conversely, in-class variation for the bacterial and mono-culture classes is relatively low, suggesting the selected metabolites are dysregulated within co-cultured samples (Figure 3.2).

The cross-validated receiver operator characteristics support the utility of this feature selection routine, and the extracted chemical characteristics. Although external validation was not used, regardless of the samples chosen, the calculated area under the curve is very close to ideal for all 200 combinations of the data that were independently selected relative to the feature selection routine (Figure 3.3).

3.4 Discussion

16 analytes of interest were selected using the feature selection routine. In the supporting information, there is a summary of each extracted analyte with the most

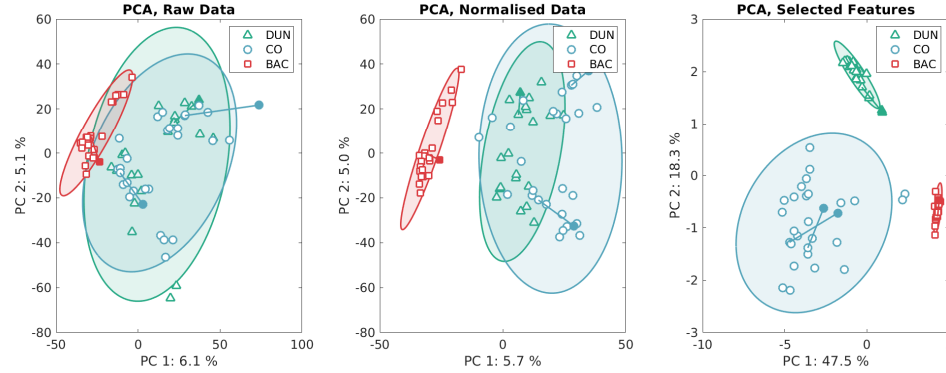


Figure 3.2: From left to right: results of principal component analysis of the raw data (autoscaled), similarly scaled data normalised to class-specific TUPA, and the normalised, scaled data using the selected features from the FS-CR routine. Quality control samples were not included in the feature selection routine, and are displayed as filled icons connected to their corresponding replicate with a straight line, following projection into the optimised principal component space.

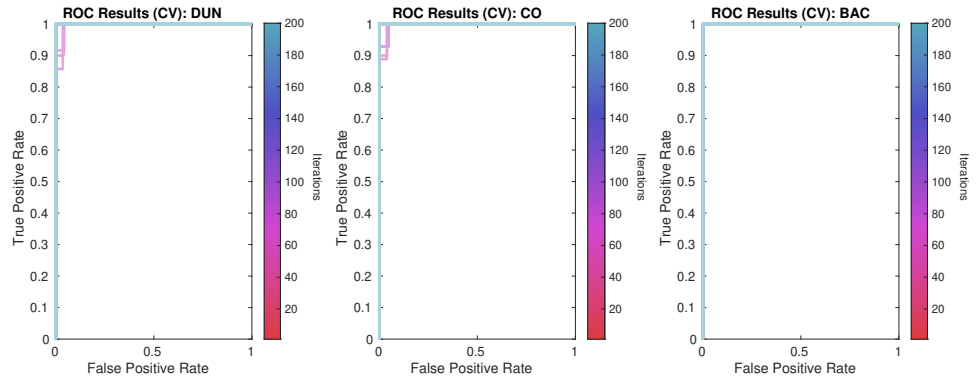


Figure 3.3: Cross validated receiver operator characteristics suggest that the calculated model is robust, and improves with more iterations. Light blue lines show the results of further iterations, and red lines show the results of fewer iterations. Each line is semi-transparent, but the AUC is close to 1 in all cases. Calculation of the classification scores was based off the named class (Class 1), versus everything else (Class 0).

relevant information presented. Using each analyte’s quantification ion listed in the output of ChromaTOF[®], it appears that the majority of the selected features are not spurious signals, given their roughly Gaussian, uni-modal peak profiles in both dimensions. However, the quality of the deconvolution appears to be poor in some features that returned no probable library hit. This can be observed by the lack of isotopic mass distributions for many prominent peaks in the mass spectra. It’s unlikely that these features can be identified based solely on the extracted mass spectra, although that isn’t to say that the features are not reliable.

Some library hits from the Golm Metabolome Database scored reasonably well on the mass spectral dissimilarity score (1-dot product), but did not account for a number of prominent peaks in the observed mass spectra. It’s possible that for these standards, no good library spectra exist. For Analytes 5083 and 22226, a very prominent peak at $m/z = 143$ is observed. This is a prominent peak in the mass spectra for alkyl-quinolones [102], a class of antibiotics. Analyte 27584 appears to bear some similarity to an unidentified compound uploaded to the Golm database as part of an earlier study [103].

Representative mass spectra were extracted from individual samples where that compound was identified. A drawback of ChromaTOF[®] software is that each sample’s features present their own mass spectra, and these mass spectra are associated across multiple samples provided that a certain similarity in mass spectral characteristics and retention time is reached. This means that the mass spectra from each sample may differ somewhat, but the decision was made not to average the mass spectra across multiple samples, as doing so could lower the precision of the extracted mass spectra or bias the match score of the library match.

Further study will be needed to confirm the identities of these analytes, since few were identified at level two or higher according to the Metabolomics Standards Initiative [104]. Below is a table summarising the quality of each of the mass spectral hits, illustrating the reasoning behind each hit’s identification level. Compounds that

were found in the reagent blank in addition to the samples were assumed to be less reliable features, but may nonetheless still be identified.

Analyte Name	Not found in Reagent Blank	Reasonable Mass Spectrum	Library hit accounts for significant peaks	Retention index match	Identification level
Analyte 5083	TRUE	TRUE	FALSE	TRUE	4
Analyte 8351	FALSE	TRUE	FALSE	FALSE	4
Analyte 13706	TRUE	FALSE	FALSE	FALSE	4
Hexadecane	FALSE	TRUE	TRUE	TRUE	2
Analyte 16951	TRUE	TRUE	FALSE	TRUE	4
Analyte 17909	TRUE	TRUE	FALSE	TRUE	4
Analyte 19404	TRUE	TRUE	FALSE	FALSE	4
Analyte 21239	TRUE	TRUE	FALSE	TRUE	4
Analyte 21318	TRUE	FALSE	FALSE	FALSE	4
Analyte 22226	TRUE	TRUE	FALSE	TRUE	4
Analyte 23041	TRUE	TRUE	FALSE	TRUE	4
Analyte 23397	TRUE	TRUE	TRUE	TRUE	2
Analyte 24829	TRUE	TRUE	FALSE	FALSE	4
Analyte 27584	TRUE	TRUE	TRUE	TRUE	4
Analyte 27833	TRUE	TRUE	FALSE	TRUE	4
Analyte 33374	TRUE	TRUE	FALSE	TRUE	4

Table 3.1: Identification levels for the significant features in the dataset, summarising the factors that went into ascribing MSI Identification levels two (punative identification) or four (unknown metabolite). Further details are available in the supporting information.

Several of listed metabolites appear to be biologically active based on their relative expressions. These metabolites may affect cell-cell signalling, hormone regulation, or be utilised as antibiotics in co-culture samples. These are of particular interest, since bioactive compounds are known to play a role in bacterial-algal interactions. Of particular interest is the homoserine lactone (HSL) which is produced by *Dunaliella tertiolecta*, however it is absent from *Phaeobacter italicus* R11 B28 and the co-culture sample C34 (Appendix B). HSLs are only known to be produced by the bacterial group Proteobacteria [105] and so further structural elucidation is necessary to assert whether or not HSL has actually been produced by a eukaryote. Previously, Schaffer et al have identified a plant metabolite, coumaric acid, that can replace the HSL tail to form a hybrid material-host signal [106], and so an analogous system with the HSL ring structure being host derived warrants further investigation. Alternatively, the *Dunaliella tertiolecta* metabolite could be an HSL antagonist as *Phaeobacter italicus* R11 is known to produce HSLs which are antagonised by the algal metabolites, furanones [93] [107] [108].

3.5 Conclusion

Despite considerable interference from chemical and pre-processing artefacts, using advanced instrumentation and data analysis methods it is possible to gain unprecedented insight into the metabolome of commercially interesting microalgae samples. A complete workflow for profiling the global metabolome of microalgae samples has been proposed, that may guide selection of bacterial inoculations for microalgae cultures to improve the yield of cultivated, carbon-neutral biofuels in the future. This study has also demonstrated the feasibility of using thermal desorption as a sample introduction technique that can allow larger-than-normal aliquots of sample to be introduced with effective pre-concentration and clean-up of dirty samples, without the need for a centrifugation step. Additionally, the utility of cTUPA as a normalisation strategy for highly dissimilar sample classes was demonstrated.

Further work is needed to interpret, and test resultant theories of the relationship between *D. tertiolecta* and *Phaeobacter italicus* *R11* cultures. Doing so may enable cultivation techniques that exploit this relationship for dividends in sustainable fuel development. Presented here is a complete workflow, with some preliminary results, that can serve as a benchmark for future experiments.

Chapter 4

Application of FS-CR for Urinary Metabolite Profiling of Human Colorectal Cancer using GC \times GC-TOFMS: Limitations of Feature Selection

4.1 Introduction

Colorectal cancer is the 2nd most common form of cancer in Canada, accounting for 13% of all new cancer diagnoses in 2017; an estimated 26,800 individuals. It is critical for all forms of cancer to be detected at an early stage to maximize the efficacy of treatment, but it is especially important for colorectal cancer: the 5-year survival rate for Stage I is estimated to be 92%, compared with only 11% for Stage IV. Colorectal cancer is typically diagnosed when a patient is at Stage III (29.1% of diagnoses), while Stage IV accounts for 19.9% of diagnoses [109]. Symptoms for the disease often do not manifest until the later stages, when the cancer has spread to other organs, and early symptoms can be easily attributed to other diseases of the gastrointestinal tract [110][111]. Diagnosis of colorectal cancer is confirmed by tissue biopsy; however it is recommended for even asymptomatic individuals at a low risk for developing the disease to start regular screening at age of 50 with a bi-annual guaiac fecal occult blood test (gFOBT), and a sigmoidoscopy every five years [112].

gFOBT tests for traces of blood within a fecal sample - often failing to diagnose colorectal cancer at a sufficiently early stage, and with relatively low patient compliance [113]. Fecal Immunochemical Tests (FIT) are a more recent development, and significantly improve upon the sensitivity of gFOBT tests, however the poor specificity of this test highlights the need for improvement [114]. Sigmoidoscopies are an invasive method of testing, whereby a physician examines the bowel up to the sigmoid using an endoscope. Effective screening has been shown to reduce patient mortality, but constraints due to invasiveness, resources, and patient compliance reduce its efficacy, especially within younger demographics not indicated to be at risk for the disease [115]. Biological samples used for screening must therefore be easy to obtain to improve patient compliance. Samples must also possess sufficient chemical information to indicate the presence of the disease. Urine is a particularly attractive medium for metabolomics studies because it is easy to obtain in large volumes, is relatively safe for technicians to handle, and features a high degree of chemical complexity[116].

Metabolomics has been used to discover small molecule markers of colorectal cancer within human biofluids. Many different instrumental techniques have been employed to build profiles of the disease[117][118][119], and recently a high-throughput, fully validated LC-MS method has been published, further indicating the utility of metabolomics for colorectal cancer screening [120]. Gas Chromatography-Mass Spectrometry (GC-MS) is a commonly used instrument for the analysis of biofluid metabolites; separations performed on a capillary gas chromatograph are relatively fast, efficient, and the sensitivity and identification capabilities of the mass spectrometer are good enough for discovery-based applications[116] [98] [121] [122].

Comprehensive two-dimensional gas chromatography – Time of flight mass spectrometry (GC \times GC-TOFMS) has been demonstrated to outperform traditional GC-MS on several different samples, including urine [123][124]. In this instrument, two capillary columns of different chemical selectivities are coupled by means of a modulator. Poorly resolved compounds eluting from the first column undergo further

separation in the secondary column. The reduction of signal noise coupled with a cryogenic modulator has been demonstrated to improve the sensitivity of detection [27] [125].

4.2 Materials and Methods

4.2.1 Reagents

HPLC Grade Methanol (>99.9%) was purchased from Millipore-Sigma Canada, HPLC Grade Toluene (Millipore Sigma Canada), was dried with anhydrous sodium sulfate (Millipore Sigma Canada) prior to use. 1 mL ampoules of N-Methyl-N-trimethylsilyltrifluoroacetamide + 1 % chlorotrimethylsilane (MSTFA + 1% TMCS) were purchased from Fisher Scientific, Canada. Urease suspensions of approximately 160 mg mL⁻¹ of water were prepared the day of derivatization using urease from Millipore-Sigma Canada, and 18.2 MΩ deionized MilliQ water (Elga PURELAB flex 3 system, VWR International Edmonton). 2 mL Safe-Lock amber centrifuge tubes were purchased from Eppendorf Canada Ltd. 2 mL GC vials, 300 μL GC vials with inserts, and GC vial caps were all purchased from Chromatographic Specialties Inc (Canada). Liquid handling was performed using 20 μL, 200 μL, and 1000 μL Rainin XLS Digital pipettes (Mettler Toledo Inc., Canada) with filter tips (Froglab Inc. Canada) for aqueous and urine samples, and 100 μL and 1000 μL Microman Pipettes and pipette tips for transferring organic solvents and various stages of the derivatisation process. Samples were treated with heat and nitrogen using a 099A EV2412S Glas-Col Heated Analytical Evaporator (Cole-Parmer Canada) using pre-purified nitrogen (Praxair Canada Inc., Edmonton). Two quality control mixtures were used in the study: one, a standard mixture (QC1) of adipic, azelaic and succinic acids (Millipore Sigma Canada), and the second (QC2) was a fatty acid methyl ester mixture in dichloromethane (SUPELCO 37 Component FAME Mix, Millipore Sigma Canada). The internal standard used was d₄-succinic acid dissolved in a 20 mM solution of

sodium bicarbonate buffer at a concentration of 78.6 mg/L.

4.2.2 Urine Samples

Using a similar population from a previous study [126], urine samples were collected from patients from the Grey Nuns Hospital, the Misericordia Hospital, the University of Alberta Hospital, and the Royal Alexandra Hospital from October 2008-2010 in Edmonton, Alberta. Patients who had been diagnosed with colorectal cancer but had not previously undergone any treatment for the disease were eligible to participate in the study. Information such as age, gender, and smoker status were collected during recruitment, and for each patient the cancer was staged based off a review of the pathology reports (Table 4.1). Within 1 hour of collection, urine samples were transferred to 1 mL vials that were labelled and frozen at -80 °C. Frozen urine was shipped on dry ice in an insulated Styrofoam container, and immediately transferred to a -80°C freezer at the University of Alberta prior to analysis. Samples from a healthy population were collected through the Stop COlorectal cancer through Prevention and Education (SCOPE®) program [127]. Here, study participants of average or increased risk of colorectal cancer provided midstream urine samples and demographic information and were verified not to be suffering from colorectal cancer through a colonoscopy performed 2 – 6 weeks after their urine was collected. The Health Research Ethics Boards at the University of Alberta provided ethics approval for the study.

Sixty-one samples were used in the study, with three randomly chosen replicate samples. QC1 was run at the start and end of each day of analysis, and QC2 was run towards the end to assess derivatisation and instrument variation respectively. Reagent blanks were also run each day, in order to screen for derivatisation artefacts in samples, and instrument blanks were run halfway through each day to ensure system cleanliness.

Urine samples were prepared according to a modified protocol for the global derivati-

sation of urine metabolites[98]. Samples were thawed on ice for 1 h, then vortexed for 1 min. 200 μ L of urine were collected and transferred into 2 mL centrifuge tubes. Here, 1 μ L of the internal standard was added along with 20 μ L urease in water (equivalent to 100 units). Samples were vortexed for one minute before incubating at 37°C for 1 h. 1.7 mL of methanol was added to each of the samples, and they were vortexed again for 1 min to precipitate the urease enzyme and extract the metabolites. Samples were then centrifuged for 10 min at 10,000 g and 4 °C. 1 mL of the supernatant was then transferred to a 2 mL vial. Vials were dried carefully under nitrogen at 60°C (approx. 2 h). Following the drying step samples were stored at -80°C for up to 1 week prior to analysis. Frozen samples were thawed at room temperature for 1 h, 100 μ L of dry toluene was added, vials were vortexed for 1 min, and dried under nitrogen at 60°C to remove the toluene and residual moisture. All samples were dried within 30 minutes. To the dried metabolite extracts, 50 μ L of 20 mg mL⁻¹ methoxyamine in pyridine solution was added, and the samples were incubated at 60°C for 2 h. 100 μ L of MSTFA was added to each sample and the samples were incubated again at 60°C for 1 h. Vials were then cooled at room temperature for 20 min. For each sample, 100 μ L of the derivatised metabolite extract was transferred to GC vials with 300 μ L inserts using the disposable positive displacement pipet tips for analysis by GC \times GC-TOFMS.

Label	Diagnosis	Age	Sex	Smoker
Control $N = 26$	Healthy $N = 26$	$\mu = 63.2$ $s = 5.8$	Female = 11 Male = 15	Yes = 5 Ex-Smoker = 12 No = 9
Case $N = 32$	Stage I $N = 4$ Stage II $N = 6$ Stage III $N = 17$ Stage IV $N = 5$	$\mu = 62.0$ $s = 5.5$	Female = 12 Male = 20	Yes = 4 Ex-Smoker = 10 No = 18

Table 4.1: Demographic information for the study participants

4.2.3 GC×GC-TOFMS Method

Samples were analyzed on a Leco Pegasus 4D GC×GC-TOFMS (Leco Instruments, St. Joseph, MI). The column used for the first dimension was a 60 m \times 0.25 mm; 0.25 μ m Rxi-5SilMS, and for the second dimension a 1.2 m \times 0.25 mm; 0.25 μ m Rtx-200MS (Chromatographic Specialties). Ultra-pure helium (5.0 grade; Praxair Canada Inc., Edmonton) was used as the carrier gas, with a constant flow rate of 2.0 mL min⁻¹. Injection was splitless, using a Restek Topaz split/splitless liner (Chromatographic Specialties), and an injection volume of 0.2 μ L. Inlet temperature was kept constant at 250 °C for all runs. The temperature program of the primary oven began at 70 °C (1 min hold) followed first by a ramp of 1°C min⁻¹ to 76°C, followed immediately by a second temperature ramp of 6.10 °C min⁻¹ to a final temperature of 300°C which was held for 7 minutes. The secondary oven and modulator temperature offset were constant at +5 °C and +15 °C respectively. The modulation period (P_M) was 2.5 s.

Mass spectra were collected at an acquisition rate of 200 Hz over a mass range

between 50 and 660 m/z. The detector voltage was 1700 V with an electron impact energy of -70 eV. The ion source temperature was 225 °C with a transfer line temperature of 225 °C. Total analysis time for each run was 52.36 min excluding cool-down time.

4.2.4 Data Pre-processing Method

GC×GC-TOFMS data were processed using ChromaTOF[®] (v.4.43; Leco). Baseline offset was set to 0.7 above the middle of the noise. The minimum SNR for base- and sub-peaks were set at 6, and the mass spectral match required for the subpeaks to be combined was set at 750. Expected peak widths throughout the entire chromatographic run were expected to be approximately 8 seconds in the first dimension and 0.16 seconds in the second dimension. A region of each chromatogram from $0.5 \text{ s} \leq t_R \leq 0.75 \text{ s}$, comprised largely of siloxanes (column degradation artefacts), was excluded from data processing.

The statistical compare feature of ChromaTOF[®] was used to align the peak table based on the parameters of retention times (in the first- and second-dimensions) and mass spectral match scores as described in previous literature[2]. Tolerances for retention time shifts were ± 1 modulation period ($P_M = 2.5 \text{ s}$) in the first dimension, and tolerances for the second dimension separation were set to 0 s (default parameters, used in previous studies). Mass spectra were associated across samples if they matched with a score of 700 or greater for all m/z values with intensities greater than 10% of the most abundant peak.

Statistical compare was constrained to only accept analytes found in at least five different samples. The peak table containing the raw data was then exported as a .csv file for analysis in MATLAB[®] R2017a, Windows 64-bit version (The Mathworks Inc., Natick, MA, USA), with multivariate statistical analysis performed using PLS Toolbox (R8.5.2; Eigenvector Research Inc., Wenatchee, WA, USA).

4.2.5 Normalisation

In order to select an informative subset of metabolites to distinguish healthy patients from those diagnosed with colorectal cancer, a proper normalisation was sought to correct for difference in hydration levels that would effect the relative quantification of each metabolite. It is common to normalise urine samples to the expression of creatinine, assuming that creatinine is only affected by hydration levels and not metabolic dysregulation. For the normalisation strategy utilising creatinine, all features were normalized to a targeted creatinine peak at $m/z = 329$, using the calibration feature of ChromaTOF[®].

Total Useful Peak Area (TUPA) was also employed for the sake of comparison. Using TUPA, all features were normalised according to the total peak area of features identified in all other samples. This attempts to correct for hydration level by considering the relative expression of many features, and may guard against the model being biased towards the integration of spurious features.

4.2.6 Data Analysis

The FS-CR algorithm was used to select a useful subset of features. Using different samples as training, optimization, and validation sets for the construction of the model, the algorithm utilized the aforementioned hybrid backward-elimination/forward selection (BE/FS) mechanism to maximize cluster resolution (CR). 100 different set combinations were used to select the features used in the model. More detailed information about this algorithm can be found in other literature [66] [71] [128] [100].

Prior to analysis, all input was split (80:20) into training and validation sets. The validation data was mean-centred and scaled according to the values calculated in the training set for a critical analysis of the predictive ability of the models, and to mimic circumstances in which this test would be employed as a diagnostic tool, where feature-wise variance and averages would not be available for individual samples.

The training set was cross-validated using a visualisation of the Receiver-Operator

Characteristics (ROC) of the features that exceeded the selected survival rate of 90%. This was to assess the utility of cross-validation metrics for GC \times GC-TOFMS data, in addition to the external validation.

PLS-DA was used to classify the training and validation sets using the selected features, and predicted scores in the validation set were used to generate the ROC curves.

4.3 Results

Following feature selection the results using three normalisation strategies were compared: no normalisation, normalisation to creatinine, and normalisation to TUPA. Feature selection was performed on each of the resultant datasets, and the predictive ability of each model, as assessed by cross-validation and the indication of samples external to the training set were assessed.

Also considered are the agreement of the samples to the three randomly-selected quality control samples, both before and after the feature selection routine:

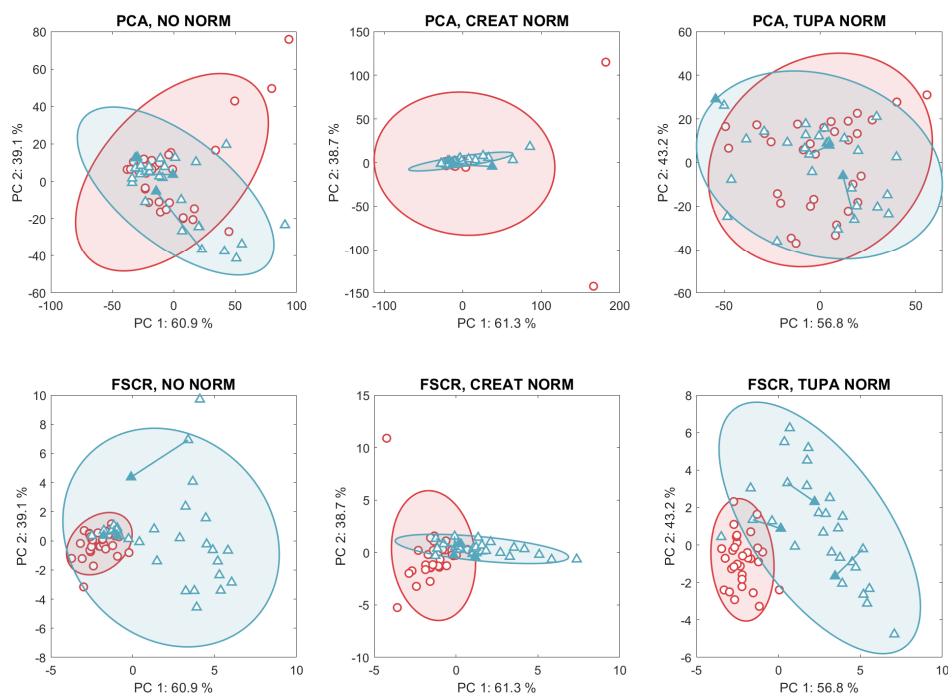


Figure 4.1: Projection of quality control samples into PCA space. Confidence ellipses represent a confidence interval of 95%.

In Figure 4.1, the data was projected into PCA space using the selected features. The solid shapes represent the QC samples, and are connected to their corresponding replicates by a line. Replicates that are in good agreement present a relatively short line, while replicates that are not in good agreement present longer lines. Note that this diagram does not indicate classification accuracy, but rather the consistency of the data insofar as the replicates agree with each other. Due to an instrumental failure, and the fact that the replicates were chosen randomly, replicates in class 1 (case samples - red circles) were not accessible due to shift along the first dimension retention axis that could not be accounted for by the commercial software being used. Control samples are indicated by the blue triangles (class 0).

A PLS-DA model was constructed using the selected features to see how well each normalisation strategy correctly indicated the external set.

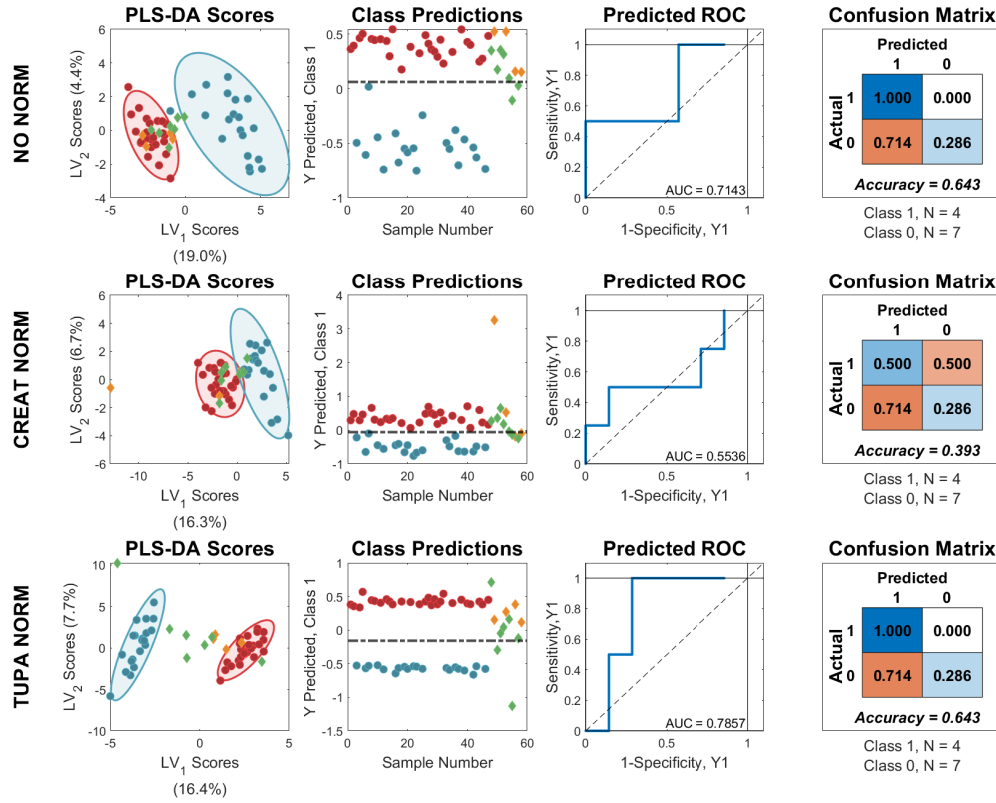


Figure 4.2: Comparison of different normalisation techniques

Figure 4.2 summarises the results of the PLS-DA classifier on the selected features. A decision boundary was determined using a bayesian method [79], and the ROC was calculated based on the scores of the validation set.

For further insight into what appears to be a problem with correctly indicating the external set, the training samples were cross-validated by re-shuffling the data into a 1:1 test:validation set using the method described in Chapter 3.

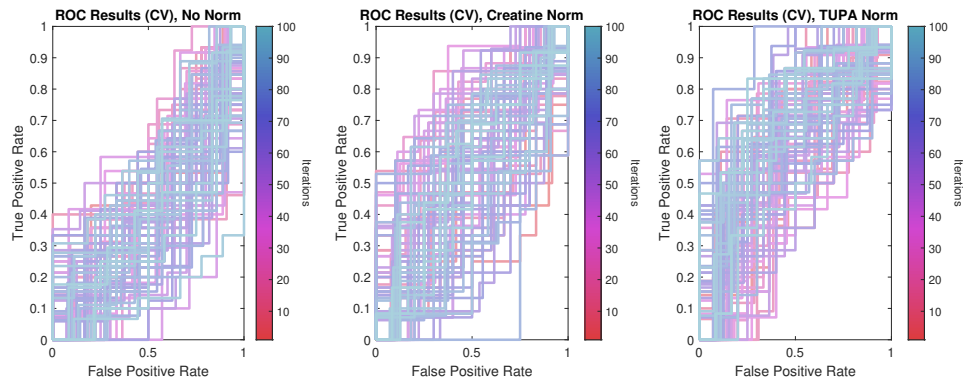


Figure 4.3: Cross-validation results at each iteration of the FS-CR algorithm

4.3.1 Discussion

Despite the appearance of sufficient cluster resolution for the TUPA normalised data in Figure 4.1, when a PLS-DA classifier is utilised, the samples external to the model are not correctly indicated. As such, the model is not useful from either a practical or theoretical standpoint, since any model used to gain insight into a biological system such as colorectal cancer should be able to indicate samples not originally considered in the analysis. The results of the cross-validation study also indicates that the model has failed to function as intended, since despite extensive combinations of data fed into the FS-CR algorithm, it is clear that the AUC for the reshuffled data does not improve significantly.

As with any supervised learning method, validation is the most important step of the analysis. Despite favourable projections of the entire dataset, the variables judged as most significant by the PLS-DA model do not correspond to the same features present in the validation set. It has been shown, despite being able to overcome some challenges regarding GC \times GC-TOFMS data, that the FS-CR algorithm can suffer from similar drawbacks of over-fitting, as does any supervised method. This may be due to either the content of the data itself, or the quality of the integrated features.

Error and uncertainty to do with the quality of the raw data is propagated at various stages throughout the sample preparation and introduction steps. There is error

associated with any measurement, but similar protocols that involve the trimethylsilylation of human biofluids have been used routinely for critical applications[129]. While it is straightforward to determine the measurement uncertainty for targeted analyses, measurements of several thousand different analytes in an untargeted analysis as presented above is not trivial. However, owing to the profound reduction of matrix effects and negligible ion-suppression during the acquisition of the mass spectra, GC-MS and GC \times GC-TOFMS are widely regarded as being as highly reproducible techniques. To highlight this, and to further implicate the failure of the proprietary ChromaTOF[®] software, shown below are replicate injections of the d4-succinic acid internal standard; each sample was extracted, derivatised, and analysed on different days.

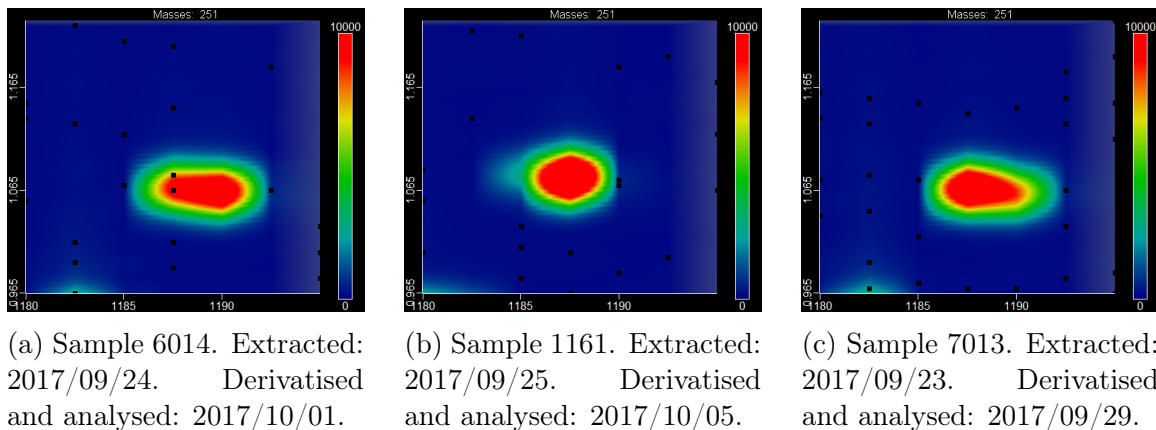


Figure 4.4: Graphical overview of relative peak areas for the internal standard from various samples at $m/z = 251$. The data was acquired and visualised using ChromaTOF[®], and each black square represents the apex of a feature that was integrated by the software. Drift along the first mode affects the distribution of the relative peak area across multiple modulations, and rules out an assessment of the raw chromatographic signal, without first visualising it as a contour plot.

This is far from an exhaustive overview of the reproducibility of each feature that was analysed, but the results do suggest that the peak areas for the internal standard appear to be consistent despite a significant overlap with the un-deuterated formulation of the same metabolite, which is present in relatively high levels in human urine. It would not be unreasonable however to extrapolate the performance of this partic-

ular analyte to many more analytes, given that ion-suppression is not an issue. Since the deuterated standard was not integrated in any of the samples ChromaTOF[®] in any of the samples, the objectivity and utility of the software is called into question once more. Although it may be that the samples as they were prepared and analysed do not contain the necessary information to discriminate between case and control samples, it is still worth pursuing further work to try and rule out the role of the data analysis software itself.

Shown below is a graphical representation of the population of the features within the entire colorectal cancer dataset:

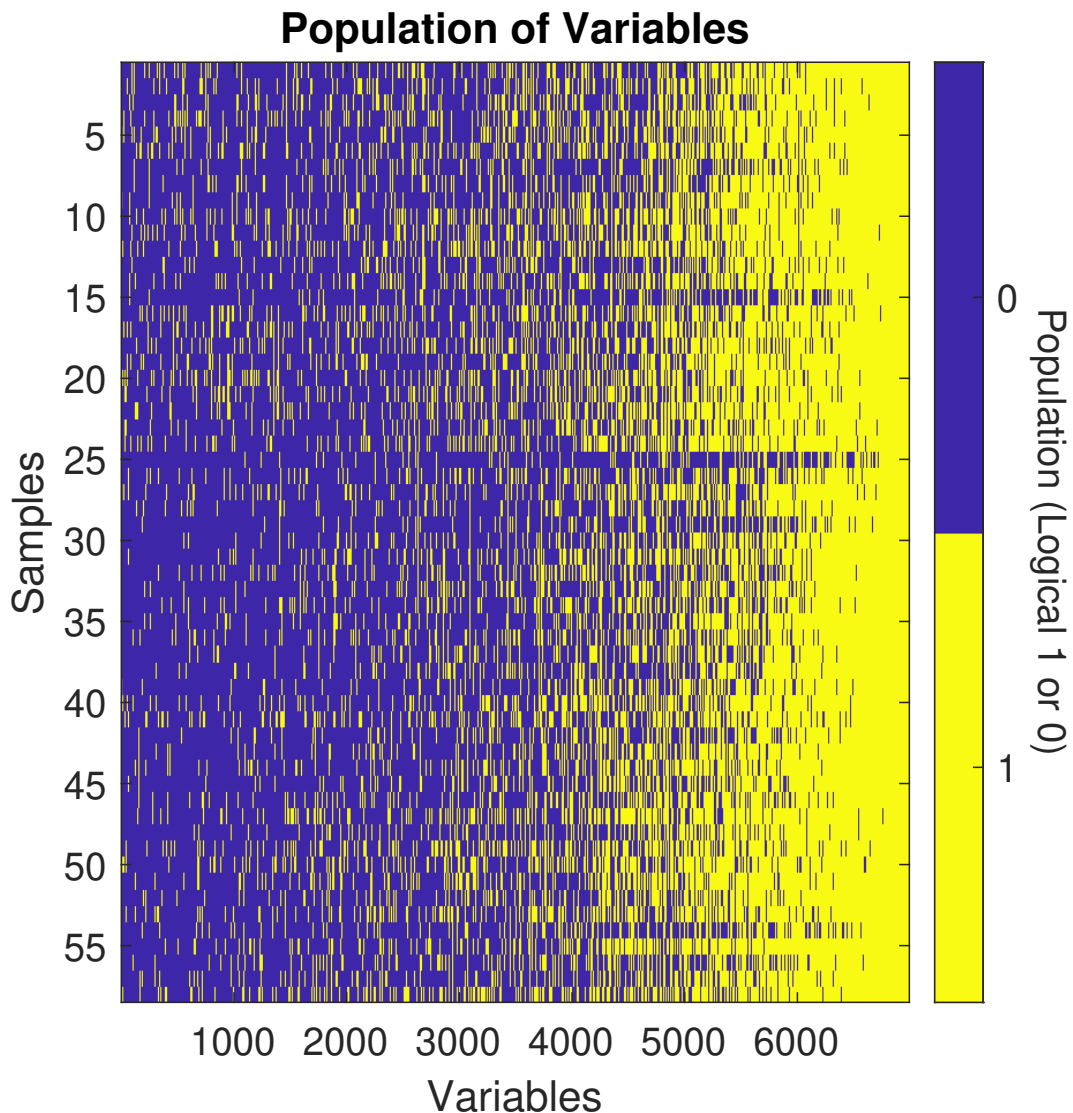


Figure 4.5: Sparsity of the colorectal cancer dataset

It appears that there is a very high degree of “zero pollution” in the data. This is present because the commercial software algorithms identified many of the features in at least five samples, but in no other samples. This high degree of sparsity, at 38.74% zeros, has a similar effect to a very large, random matrix. That is, it is well-known in chemometrics that for a sufficiently large number of random variables relative to the number of samples, a solution that discriminates between two classes is all but assured. This commonly held dogma cautions against disregarding model validation,

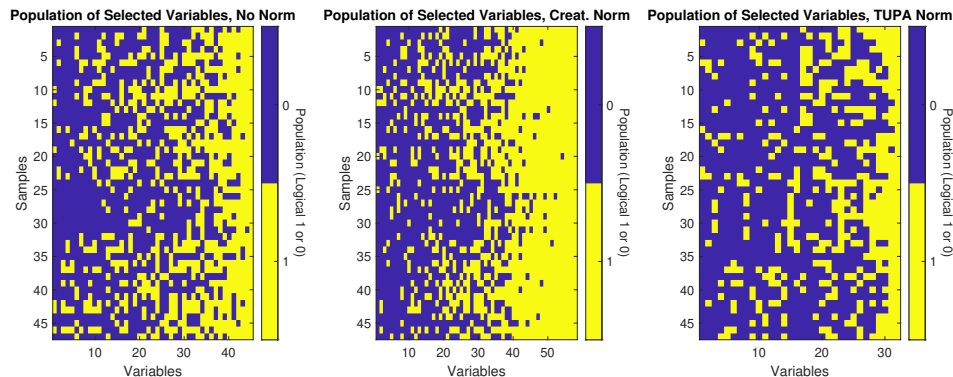


Figure 4.6: Sparsity of selected features

despite how attractive the training set may appear to the analyst.

The FS-CR algorithm does not appear to reduce the degree of sparsity under all normalisation routines. Despite the fact that the TUPA normalised data appeared to present the greatest agreement between replicate and QC samples in Figure 4.1, the sparsity of matrix for the selected features increased (Figure 4.6).

The percent of populated features following feature selection was 41.09% for the un-normalised data, 52.79% for the data normalised to creatinine, and 31.58% for the TUPA-normalised data.

4.4 Conclusion

Under any of the normalisation schemes utilised, there appears to be no model that can correctly indicate the external validation set using the data pre-processing and exploration tools that have been proposed in this chapter. As such, the proposed models are poorly diagnostic, and are likely a poor representation of the underlying biological phenomena. There are two possible reasons for the unsatisfying conclusion to this research. One possible reason is that there is no usable information in the dataset that can correctly indicate case versus control samples. While this is certainly possible, recent studies have proven that it is possible to detect colorectal cancer polyps at a very early stage using a metabolomic test using LC-MS [126].

Although differences in instrumentation can account for the failure of this study, another possibility is that the information is still present, but obscured by sub-optimal data pre-processing methods. It can certainly be argued that the parameters that were used are sub-optimal. As mentioned previously, optimisation of the chromatographic processing method is highly subjective, and many combinations of settings generate data that looks acceptable for future analysis, and it is only through many iterations of processing the entire data set through both the chromatographic and subsequent chemometric processing that a “correct” solution is found.

This iterative, subjective process is incredibly slow and a waste of analysts’ valuable time. This motivated the remainder of this thesis, where robust, objective methods for analysing GC×GC-TOFMS data without the need for many subjective inputs are sought.

Chapter 5

PARAFAC2 \times N: Coupled Decomposition of Multi-modal Data with Drift in N Modes

5.1 Background

Multidimensional chromatographic separations are becoming more widespread, thanks to advances in modulator technology that have enjoyed considerable interest over the past two decades. The most mature of these technologies is Comprehensive Two-Dimensional Gas Chromatography (GC \times GC) which is frequently hyphenated together with a Time-of-Flight Mass Spectrometer. GC \times GC-TOFMS is more sensitive and selective than traditional gas chromatography - mass spectrometry, but despite its considerable advantages, and many innovations that have reduced the analysis cost for GC \times GC-TOFMS separations, the technology suffers from challenges surrounding data analysis that hinder its widespread deployment. Currently, few software packages offer a transparent and mathematically satisfying way of handling data from multiple non-target analyses such as those frequently encountered in forensics and metabolomics. Much of the challenge arises due to the fact that chemical components are free to shift independently along the first and second chromatographic modes between runs. This is the major drawback encountered when performing a separation utilising multiple chromatographic modes, as opposed to tandem mass spectrometric

detectors, which due to regular and thorough mass calibrations do not suffer from mass-to-charge ratio (m/z) drift between runs.

A number of proposals for the analysis of GC \times GC-TOFMS data, based on the well-understood theories of Multivariate Curve Resolution (MCR) [130], Parallel Factor Analysis (PARAFAC) [131] and PARAFAC2 [132] have been presented in the literature. Models utilising linear rank-deficient solutions have proven to do well to extract meaningful information that is robust against interfering chemical and/or electronic noise. The drawback of these techniques is that skilled user intervention is necessary to determine the chemical rank of the data, and identify regions of interest. While PARAFAC2 has shown to be a useful, parsimonious approach to model drifting chromatographic data with multi-channel detectors such as mass spectrometers, it is limited in that it allows for drift in only one mode.

A number of practical solutions to handling GC \times GC-TOFMS data have been proposed, but these typically lean heavily on the dynamic programming aspect of data analysis. Rather than modelling the data, programmatic solutions find, analyse and associate regions of interest across multiple samples and correlate the chemical information for inclusion into a peak table that describes similar chemical characteristics of different samples. This is often done as part of a commercial software solution, or as an additional piece of software designed to work on the peak tables for each individual sample as exported by other software packages. A major issue with dynamic programmatic solutions is that failure of the software at any step can result in misalignment of analytes across multiple samples, or as is more commonly observed, splitting misidentified peaks into separate columns. In either case, further analysis of imperfect peak tables may lead to erroneous conclusions for untargeted analyses.

There are typically a number of different parameters that require optimisation using the software currently available. Since there is no objective measure for the performance of different data analysis parameters, results that best align with the analysts' expectations are usually assumed to be correct. While the intuition of an

experienced analyst is certainly useful, reliance on subjective measures for model performance is far from an ideal solution. And for complex mixtures, there is often not an ideal set of parameters that can handle the entire dataset in such a way that matches the expectations of the analyst. For instance, parameters that integrate and align large peaks handily, may miss smaller peaks which fall below integration thresholds. This then obviates the purported advantages of GC \times GC-TOFMS in terms of sensitivity.

The primary motivation for this work has been the frustrating and time-consuming chore of curating peak tables for large (i.e. > 200 sample) studies prior to performing multivariate analyses of the data. Curation of peak tables or feature lists currently requires the use of a variety of imperfect commercial and home-built software tools, at the hands of subjective analysts with varying levels of skill and expertise.

For the subsequent multivariate analyses of peak tables, optimisation of the processing parameters is a significant time-sink, where the processing parameters for the peak table are modified several times in order to find a set of parameters that yields a set of peak tables that appear to be of sufficient quality. Even worse, for the careless analyst, it is relatively easy to generate peak tables that when processed using multivariate tools, appear to offer illuminating results, when in fact the model may do little more than pick up on spurious signals leading to a meaningless result.

We propose a new modelling technique that exploits the high degree of redundancy in GC \times GC-TOFMS, as replicate samples via a direct decomposition of the 4-way data. This approach is based off of the flexible coupling method for 3-way PARAFAC2, with an additional coupling constraint that restricts the descent of the extracted mass spectra calculated from models that describe the first- and second-dimension retention drifts. While extremely useful for analysing GC \times GC-TOFMS data, this technique can offer a general theory for modelling multi-dimensional chromatographic data with drift in N modes, and may also be extensible to hyperspectral imaging datasets. Much like PARAFAC2, the proposed algorithm we are calling

PARAFAC2×2, requires only the number of components and a region of interest in order to work. This greatly simplifies the task of analysing GC×GC-TOFMS data, removing the long lists of parameters to be optimised and subjectivity in data analysis.

5.1.1 GC×GC-TOFMS Data Structure

A univariate detector such as a flame-induction detector (FID) performs a series of regular measurements at regular intervals as chemical components enter the detector from the GC column. Certain regions of a single chromatogram with one dimension of separation can be excised to analyse the chromatographic peak in question for quantitative purposes. An excised region is a vector of length I , where I is the number of acquisitions in the region of interest. The relative abundance of peaks in this region can be obtained by the euclidean norm of the vector, assuming no interfering analytes are present within the region, or by a non-linear, parametric fit of several idealised Gaussian or modified Gaussian [133] functions to deconvolve the signals of chemical interference.

When the separation is coupled to a multivariate detector, such as a Mass-Spectrometer or a Vacuum UV detector, the number of different variables encompassed by the detector can span an additional mode, denoted as J . An excised region encompasses the number of acquisitions, I by the number of individual “detectors” (e.g. mass-to-charge ratios, or m/z , for mass spectral detectors), J . Closely co-eluting factors can be deconvolved using multivariate methods such as MCR or ICA, both of which decompose the resultant $I \times J$ matrix.

A single GC×GC-TOFMS chromatogram presents itself as a 3rd order tensor, while a series of GC×GC-TOFMS chromatograms presents itself as a 4th order tensor comprising $I \times J \times K \times L$ modes of mass spectral acquisitions, mass-to-charge ratios (mass channels), modulations, and samples. The multidimensional separation is generated by capturing fractions of effluent from the first dimension and injecting them

at regular intervals onto the second-dimension column via the modulator. The action of the modulator creates slices of second-order information along the first chromatographic dimension, such that for an individual GC×GC-TOFMS sample the data is a 3rd order tensor. The L^{th} mode describes multiple samples extracted from either the same region of the chromatogram, or entire chromatograms depending on what is being considered.

5.1.2 PARAFAC Modelling of GC×GC-TOFMS Data

A PARAFAC model can be constructed that describes a 4th order tensor, $\mathcal{X} \in \mathbb{R}^{I \times J \times K \times L}$, using the Khatri-Rao (KR) product:

$$\mathcal{X} = F_2(D_L \odot F_1 \odot A)^T \quad (5.1)$$

Where F_2 is an $I \times R$ matrix, A is a $J \times R$ matrix, F_1 is a $K \times R$ matrix, and D_L is an $L \times R$ matrix that correspond to the characteristics of the data mentioned previously in addition to the R chemical factors that best represent the characteristics of the data being analysed. The KR product is commonly used as the tensor product, owing to the simplicity by which the PARAFAC model can be optimised using the Alternating Least Squares (ALS) algorithm that is analogous to the way in which bilinear models are traditionally optimised.

A trilinear decomposition of the unfolded tensor $\mathcal{X} \in \mathbb{R}^{I * K \times J \times L}$ of the data can be made, observing that the KR product of the second-dimension elution profiles ($I \times R$) and the modulation matrices ($K \times R$) are equal to a single unfolded retention mode, of dimension $I * K \times R$:

$$\mathcal{X} = (F_2 \odot F_1) D_L A^T \quad (5.2)$$

Which is structurally similar to the trilinear PARAFAC1 model as described by

Kiers and Bro, substituting F for $(F_2 \odot F_1)$:

$$X_k = F D_k A^T \quad (5.3)$$

X_k here describes a 3rd order tensor, $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ as a series of matrices, $X \in \mathbb{R}^{I \times J}$. This notation is used to keep the notation consistent with the original direct fitting algorithm for PARAFAC2 [43].

It is relatively straightforward to frame the problem as a PARAFAC2 model with multiple samples' first- and second-dimensions unfolded as one, with L samples:

$$X_l = (F_2 \odot F_1)_l D_l A^T \quad (5.4)$$

Using the direct-fitting method, there are l unique, orthogonal peak profiles of $I * K \times R$, P_l , and a non-singular $R \times R$ matrix, F :

$$X_l = P_l F D_l A^T \quad (5.5)$$

In addition to unfolding the $\mathcal{X} \in \mathbb{R}^{I \times K \times J \times L}$ tensor along the first and second retention modes to effect an $\mathcal{X} \in \mathbb{R}^{I * K \times J \times L}$ 3rd order tensor, tensors of similar orders can be made by “stacking” second-dimension retention profiles for an $\mathcal{X} \in \mathbb{R}^{I \times J \times K * L}$, or the first-dimension retention times' equivalent as: $\mathcal{X} \in \mathbb{R}^{K \times J \times I * L}$. In all cases, it is possible to construct a PARAFAC2 model on the resultant trilinear data. In the first case, the $X_{I * K \times J \times L}$ appears to avoid the problem of drift in two modes, by artificially reducing the problem to drift along one combined retention mode. This method appears to have an additional benefit, wherein the quantities of each component are solved for directly. It is not possible to solve for the relative expression of each component, per sample directly using the unfolded data in the other two cases.

5.2 PARAFAC2 modelling for 4-way data unfolded as: $\mathcal{X} \in \mathbb{R}^{I \times K \times J \times L}$

A PARAFAC2 model for data unfolded as: $\mathcal{X} \in \mathbb{R}^{I \times K \times J \times L}$ may appear to account for drift in two modes; however, there are practical limitations to the PARAFAC2 model, such that this method can only properly account for drift in one mode which in this case is the mode containing the second-dimension acquisitions. PARAFAC2 can only account for small variations in retention time drift, based on the assumption that the inner-product matrices: $F_l^T F_l$ (unfolded scores matrices from Equation 5.4) are consistent across all samples. In the method for direct fitting of PARAFAC2, F_l is defined as $P_l F$, where P_l are the orthonormal scores matrices that are free to vary across each sample, calculated as:

$$P_l = X_l A D_l F^T (F D_l A^T X_l A D_l F)^{-1/2} \quad (5.6)$$

Through the singular value decomposition of:

$$F D_l A^T X_l^T = U_k \Sigma_l V_l^T \quad (5.7)$$

$$P_l = V_l U_l^T \quad (5.8)$$

Because $F_l^T F_l$ is calculated as $F^T P_l^T P_l F$, and because P_l is orthonormal such that $P_l^T P_l = I_R \in \forall l$, $F^T P_l^T P_l F = F^T F$. Consequently, F itself is assumed to be constant across all samples, and is calculated as an average for those samples where it differs.

F is calculated from the PARAFAC model of $P_l^T X_l$ which minimises:

$$P_l^T X_l = ||P_l^T X_l - F D_l A^T||_F^2 \quad (5.9)$$

$$F = \sum_{l=1}^L P_l^T X_l A D_l (D_l A^T A D_l)^{-1} \quad (5.10)$$

The $P_l^T X_l$ term presents mass spectral, or second mode loading information that is also proportional to the relative abundance of each chemical factor. Across l samples, this information will be relatively consistent, as long as P_l is describing the latent chemical phenomena in the same way. And for small variations in retention time, this is not usually a problem. Small retention time drifts of each component relative to each other may not be significant, and small modelling errors are averaged out via the calculation of F in Equation 5.10. For unfolded data however, small drifts across the first retention mode are in practice large drifts across the combined first-and second-dimension modes. This problem can be mitigated using the flexible coupling approach for non-negative PARAFAC2 by Cohen and Bro, which does not rely on the intermediate calculation of orthogonal peak profiles, and permits modelling on more substantial retention drift relative to the different chemical factors thanks to softer constraints on modelling the data.

5.2.1 A Flexible Coupling Approach for Non-negative PARAFAC2

Cohen and Bro proposed a flexible coupling method for modelling non-negative scores along the mode that is allowed to vary in the PARAFAC2 model. Using this technique, on a 3rd order tensor, $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ the non-negatives scores, B_k , are calculated as the minimisation of:

$$X_k = \underset{X_k}{\operatorname{argmin}} \sum_{k=1}^K ||X_k - B_k D_k A^T||_F^2 + \mu_k ||B_k - P_k B^*||_F^2 \quad (5.11)$$

Where non-negativity can be enforced for any term with any non-negative least squares solver. The P_k and B^* terms in Equation 5.11 are the orthonormal scores of B_k via SVD, and an $R \times R$ latent coupling factor, which together minimize the second term proportional to the coupling factor, μ_k . The flexible coupling approach for calculating non-negative PARAFAC2 can be implemented using the ALS algorithm, although the numerical stability depends on an appropriate estimate for the coupling constants, μ_k . As the solution approaches the optimum, it's reasonable to increase

the coupling constants, and to tighten restriction of the coupled terms. The distance from the optimum can be estimated using:

$$\mu_k^1 = 10^{-SNR/10} \frac{\|X_k - B_k^1 D_k^1 A^{1T}\|_F^2}{\|B_k^1 - P_k^1 B^{1*}\|_F^2} \quad (5.12)$$

After the first iteration of the algorithm, where SNR is the estimated Signal-to-Noise Ratio for each chromatographic slice. A convenient estimate of the SNR can be used by calculating the ratio of the first singular value to the second singular value for non-centred data. The first singular value can be thought of as the distance along the axis of greatest variance within the data from zero, and the second singular value as an estimate for the noise. This assumption breaks down where there is significant instrumental noise and the offset of the data is far from zero. However, an advantage of GC×GC-TOFMS data is that thanks to the action of the modulator there is relatively little chemical noise, and the aforementioned assumptions are usually sufficient.

It is possible to solve for the unfolded scores, and sample-wise relative abundances using the flexible coupling approach for data unfolded along one retention mode as $\mathcal{X} \in \mathbb{R}^{I \times K \times J \times L}$

$$X_l = \underset{l=1}{\operatorname{argmin}} \sum_{l=1}^L \|X_l - B_l D_l A^T\|_F^2 + \mu_l \|B_l - P_l B^*\|_F^2 \quad (5.13)$$

This helps to avoid the issue of inconsistent cross-products that limit the accuracy of the model where there is significant drift of the chemical components relative to one another. However the scores matrix B_l , is of a relatively high dimensionality at $I * K$ unique indices. This introduces a high number of degrees of freedom, at the expense of the high number of replicates it is possible to achieve by unpacking GC×GC-TOFMS data in a different fashion. It is well known that PARAFAC models benefit from relatively high numbers of replicates, and lower degrees of freedom when compared with 2nd order modelling of the similar, unfolded data. This exploits the rotational determinacy of the PARAFAC model versus equivalent matrix

decomposition techniques.

5.3 PARAFAC2 modelling of 4-way data unfolded

as: $\mathcal{X} \in \mathbb{R}^{I \times J \times K * L}$ or $\mathcal{X} \in \mathbb{R}^{K \times J \times I * L}$

A GC×GC-TOFMS dataset comprised of multiple samples can also be unfolded into two third-order tensors as either consecutive slabs of second-dimension retention slices, or first-dimension retention slices ($X_{kl} \in \mathbb{R}^{I \times J \times K * L} = \mathcal{X}_{:, :, k, l}$ or $X_{il} \in \mathbb{R}^{K \times J \times I * L} = \mathcal{X}_{i, :, l}$). The matrices B_{kl} and B_{il} are score matrices for the hl^{th} unfolding of the tensor \mathcal{X}_{ijkl} , where h is one of k or i . The same notation applies to all of the other matrices with similar designations.

$$X_{kl} = \underset{B_{kl}}{\operatorname{argmin}} \sum_{kl=1}^{KL} \|X_{kl} - B_{kl} D_{kl} A^T\|_F^2 + \mu_{kl} \|B_{kl} - P_{kl} B_{kl}^*\|_F^2 \quad (5.14)$$

$$X_{il} = \underset{B_{il}}{\operatorname{argmin}} \sum_{il=1}^{IL} \|X_{il} - B_{il} D_{il} A^T\|_F^2 + \mu_{il} \|B_{il} - P_{il} B_{il}^*\|_F^2 \quad (5.15)$$

The advantage with these two methods for unfolding, is that there are either $K * L$ or $I * L$ numbers of replicates. The problem of inconsistent cross-product matrices is not eliminated; however, as in either case peaks invariably disappear and reappear for properly selected regions of interest. As mentioned earlier, the problem of inconsistent cross-product matrices is mitigated through the use of the flexible coupling approach for non-negative PARAFAC2.

The advantage of unfolding the data as a series of second- or first-dimension elution profiles is that it exploits the high degree of redundancy of GC×GC-TOFMS data. This plays into the advantages of PARAFAC over second-order modelling. However, further manipulation of the resultant scores are required in order to solve for the sample-wise relative abundances. This is simple to do - the scores of either B_{kl} or B_{il} can be unfolded for each sample as an $I * K \times R$ matrix and the data matrix itself unfolded as an $I * K \times R$ matrix. The sample-wise abundances can be solved for in

the least-squares sense, where by unfolding $B \in \mathbb{R}^{I \times K \times R \times L}$, it is possible to solve for each l^{th} slice of the tensor D_l with the similarly unfolded tensor $\mathcal{X} \in \mathbb{R}^{I \times K \times J \times L}$:

$$D_l = (B^T B^{-1}) B \mathcal{X} A (A^T A)^{-1} \quad (5.16)$$

Where A is the $J \times R$ matrix of the extracted mass spectra, common to the entire dataset.

There are two further advantages of unfolding the data as first- or second-dimension retention slices: unimodality constraints can be applied using an appropriate least-squares solving algorithm, and modelling one sample is highly extensible to further samples. Applying unimodality constraints may help to resolve co-elutions whose mass spectra are highly similar, but would make it difficult to factor out the noise in a straightforward manner. Applying the calculated model to new data would make it easy to quantify analytes of interest across different analysis conditions (e.g. analysing a sample using a low split ratio to find analytes of interest, and then searching for those analytes in the same region of interest using a high split ratio).

5.4 The PARAFAC2×2 algorithm

While it is clear that a GC×GC-TOFMS dataset can be decomposed in a manner that preserves the high number of replicates, the question remains as to which retention mode to model. In theory, the retention mode to model should have the highest resolution between closely co-eluting chemical factors. However this information is difficult to predict, without first calculating the model itself.

In order to minimise the reliance on dynamic programming to select an appropriate method for unfolding the data, and with an eye towards creating the most general solution possible for the deconvolution and quantification of GC×GC-TOFMS features, we propose a method that models both models simultaneously (using Equations 5.14 and 5.15), and at convergence averages the elution scores and corresponding mass

spectra for each of the modelled retention modes to solve for the sample-wise loadings. The mass spectra for each model should be allowed to vary slightly, since while the data is the same for both models, the observed chemical environment may differ somewhat between models. The unified model can be described as the minimisation of the following expression:

$$\begin{aligned} \mathcal{X}_{ijkl} = \operatorname{argmin} \sum_{kl=1}^{KL} ||X_{kl} - B_{kl}D_{kl}A^T||_F^2 + \mu_{kl}||B_{kl} - P_{kl}B_{kl}^{*T}||_F^2 + \\ \sum_{il=1}^{IL} ||X_{il} - B_{il}D_{il}A^T||_F^2 + \mu_{il}||B_{il} - P_{il}B_{il}^{*T}||_F^2 + \\ \mu_A||A_{kl} - A_{il}||_F^2 \quad (5.17) \end{aligned}$$

If the scores and mass spectra are in good agreement with each other, the average of the results do not differ significantly from the results of the individual models themselves. The descent that minimises the sum of residual squares for each model informs the learning of the other via the mass spectral coupling constant, μ_A . As long as an appropriate value for μ_A is selected, this method is readily able to converge to a usable solution in relatively few iterations. However if this coupling constant is too small the two models may begin to diverge, and if it is too large then the coupling constant may limit the descent of the mass spectra, and converge to a sub-optimal solution. It is also important to be cognisant of risk of converging to sub-optimal solutions, depending on the initialisation of the algorithm. For this reason, as is commonly done with PARAFAC2, 10 random initialisations may be utilised and the sum of residual squares was measured after 80 iterations. The model with the lowest sum of residual squares is selected as the “correct” initialisation, and is allowed to continue to convergence.

Shown below is the description of the algorithm in its current implementation. For each least-squares step, constraints such as non-negativity or unimodality can be applied depending on what is deemed appropriate for the analyst.

Algorithm 6: Coupled PARAFAC2×2 ALS

Result: $\mathcal{F}, D_l, A = \text{PARAFAC2} \times 2(\mathcal{X}, R)$
 $\mathcal{F} \in \mathbb{R}^{I \times R \times K \times L}$, $D_l \in \mathbb{R}^{R \times R \times L}$, $A \in \mathbb{R}^{J \times R}$, and $\mathcal{X} \in \mathbb{R}^{I \times J \times K \times L}$
 initialization: $B_{kl}^0 = \|\text{rand}(I, R, K * L)\|_R$, $B_{il}^0 = \|\text{rand}(K, R, I * L)\|_R$
 $A_{il}^0 = A_{kl}^0 = \|\text{rand}(J, R)\|_R$, $B_{kl}^{*0} = B_{il}^{*0} = \|\text{rand}(R, R)\|_R$
 $D_{kl}^0 = I_R, \forall kl \in [1, K * L]$, $D_{il}^0 = I_R, \forall il \in [1, I * L]$
 $\mu_{kl}^0 = \frac{\|B_{kl} D_{kl} A^T\|_F^2}{\|B_{kl}\|_F^2}$, $\mu_{il}^0 = \frac{\|B_{il} D_{il} A^T\|_F^2}{\|B_{il}\|_F^2}$, $\mu_A^0 = 10^\omega \frac{\|X_{kl} - B_{kl}^0 D_{kl}^0 A^{0T}\|_F^2 + \|X_{il} - B_{il}^0 D_{il}^0 A^{0T}\|_F^2}{\|A_{kl}\|_F^2}$
 $i = 1$
while $\frac{\sigma_{old} - \sigma_{new}}{\sigma_{old}} > \epsilon \sigma_{old}$ **do**
 for $h, H \in k, i$ & I, K **do**
 for $\forall hl \in [1, H * L]$ **do**
 $[U, \Sigma, V] = \text{SVD}(B_{hl} * B_{hl}^*, R)$
 $P_{hl} = UV^T$
 end
 $B_{hl}^* = \|\sum_{hl=1}^{H*L} \mu_{hl} P_{hl}^T B_{hl}\|_R$
 $A_{hl} = \|\sum_{hl=1}^{H*L} \left(\frac{\mu_A A_{hl} + X_{hl}^T B_{hl} D_{hl}}{D_{hl} B_{hl}^T B_{hl} D_{hl} + \mu_A I_R} \right)\|_R$ %% See Appendix C.1
 $B_{hl} = \|\frac{X_{hl} A_{hl} D_{hl} + \mu_{hl} P_{hl} B_{hl}^*}{D_{hl} A_{hl}^T A_{hl} D_{hl} + \mu_{kl} I_R}\|_R$ %% See Appendix C.2
 for $\forall hl \in [1, H * L]$ **do**
 $D_{hl} = \frac{B_{hl}^T X_{hl} A_{hl}}{(B_{hl}^T B_{hl})(A_{hl}^T A_{hl})}$
 end
 if $i = 1$ **then**
 for $\forall hl \in [1, H * L]$ **do**
 $\Sigma = \text{SVD}(X_{kl}, 2)$
 $\text{SNR} \approx \Sigma_1 / \Sigma_2$
 $\mu_{hl} = 10^{-\text{SNR}/10} \frac{\|X_{hl} - B_{hl} D_{hl} A_{hl}^T\|_F^2}{\|B_{hl} - P_{hl} B_{hl}^*\|_F^2}$
 end
 else
 if $i < 10$ **then**
 for $\forall hl \in [1, H * L]$ **do**
 $\mu_{hl} = \mu_{hl} * 1.05$
 end
 end
 end
 $i = i + 1$
 end
end
 $\mathcal{F} = \|B_{kl} D_{kl} + B_{il} D_{il}\|_F \forall l \in [1, L]$
 $D_l = \frac{F_{I*K \times R}^T X_{I*K \times J \times L} A_{J \times R}^T}{(F_{I*K \times R}^T F_{I*K \times R})(A_{J \times R}^T A_{J \times R})} \forall l \in [1, L]$
 $A = \|A_{kl} + A_{il}\|_R$

Note the convention: $\|M\|_R$ in Algorithm 6 is used to denote that each factor (R) is normalised to its Euclidean norm. For the initial mass spectral coupling constant, μ_A , an additional exponential term, ω , is added to control the initial descent of the two modes with respect to each other. That is, typically a value of 2 or 3 is used so that the two calculated mass spectra do not diverge at the outset. This constant becomes more important for higher component numbers, but an exact value is not critical for the proper functioning of the algorithm. Future work may highlight a better way to estimate this parameter at the outset.

5.4.1 Analysis of Synthetic Data using PARAFAC2×2

Synthetic data, mimicking replicate samples of GC×GC-TOFMS data was generated in MATLAB[®] 2021a to evaluate the performance of the algorithm. Random independent drift in both the first- and second-dimension retention modes across three samples was chosen. For the first-dimension retention time, each peak was allowed to vary ± 1.5 modulations, and for the second-dimension retention time each peak apex was allowed to vary randomly ± 25 acquisitions. Relatively few replicate samples were chosen for this data set to make the results easier to display, and to demonstrate this algorithm’s utility despite handling relatively few samples. Synthetic mass spectra were generated from random distributions of 45 “peaks” representing isotopic mass distributions, and each spectrum was normalised to its Euclidean norm. Nominally, the SNR was set to 500. White noise was added across every acquisition, modulation, and mass channel relative to the maximum score value out of all components, per unit of SNR. An additional offset of 6 times the maximum score value out of all components per unit of SNR was added to ensure that all data was positive. The distribution for each peak along the first-dimension retention time was set at 1.5 modulations, and along the second-dimension retention time 20 acquisitions. The magnitude of the data was multiplied by a factor of 10^4 to simulate ion counts typically encountered during GC×GC-TOFMS experiments. The synthetic data was then inspected to ensure a

visual similarity to real data, by examining the retention profiles across all channels. The code used to generate the synthetic data, as well as the PARAFAC2 \times 2 algorithm is available online at <https://github.com/mdarmstr>.

All computations were performed on a computer equipped with an Intel[®] Core[™] i7-6700K CPU @ 4.00 GHz with 16.0 Gb of installed memory (RAM) using a 64-bit Windows 10 operating system.

Hyperparameters and convergence criteria for the PARAFAC2 \times 2 algorithm as applied for the synthetic experiments were as follows: ϵ was set at 2.5×10^{-6} , and the number of iterations during which the coupling constants μ_{kl} , μ_{il} , and μ_A the coupling constants increased was 10. The non-negative solver used was the Fast Combinatorial Non-negative Least Squares algorithm [49]. A constant of 3 was used as the value for ω . In both cases, the algorithm converged to a solution in fewer than 30 iterations, which took less than 30 seconds in total. Initialisation estimates were not replicated to critically assess the utility performance of the algorithm, but several random initialisations are possible in practise.

Because of the way the data is generated, the expected peak intensities depend on the maximum score of each of the input factors. This maximum score varies, depending on the first-dimension retention times (i.e. a Gaussian along the first dimension may be modulated close to its apex in one sample, and further away from its apex in another sample), but the recovered abundances are in relatively good agreement with our expectations despite this limitation in precision.

The cosine correlation coefficient ($\cos(\theta)$) was used to measure the agreement between the calculated and synthetic scores and loadings (ν_{SYN} , and ν_{OBS}) using the unfolded GC \times GC scores and mass spectral loadings. This was calculated as the inner product of each pair of matrices, with each column normalised to its Euclidean norm

such that for each pair of entries:

$$\cos(\theta) = \frac{\nu_{SYN} \cdot \nu_{OBS}}{||\nu_{SYN}|| \cdot ||\nu_{OBS}||} \quad (5.18)$$

The percent variance explained was calculated using the formula from Bro et al. [43]

$$\%VAR = 100 \times \left(1 - \frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L (X_{ijkl} - F D_l A^T)^2}{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L (X_{ijkl})^2} \right) \quad (5.19)$$

Shown below are the results of analysing a synthetic two-component elution. For ease of comparison, an extra component was not used to model the noise.

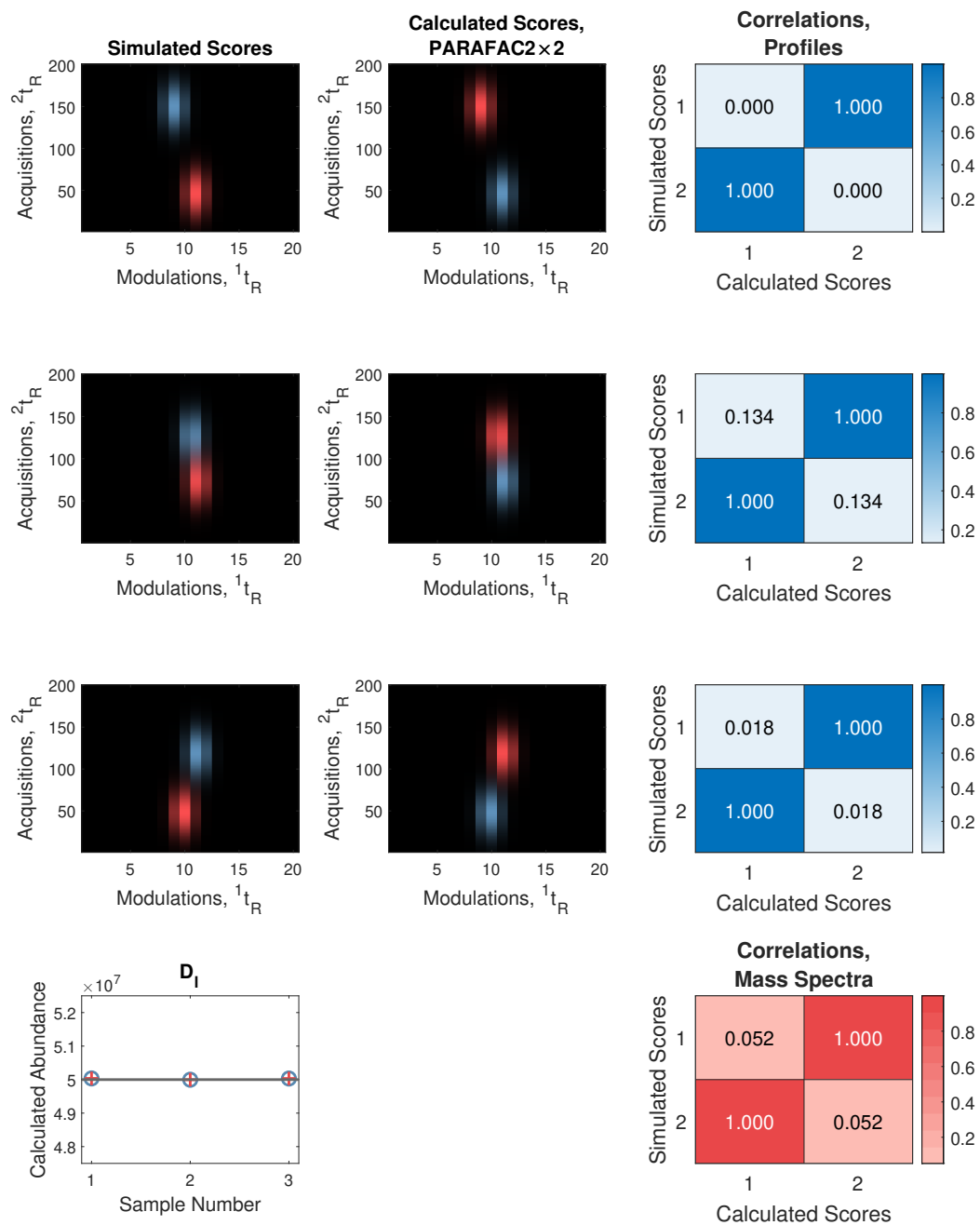


Figure 5.1: A simulated two-component model with a nominal SNR of 500. The percent variance explained using this model was 99.9959%. The calculated model demonstrates almost perfect agreement with the synthetic data.

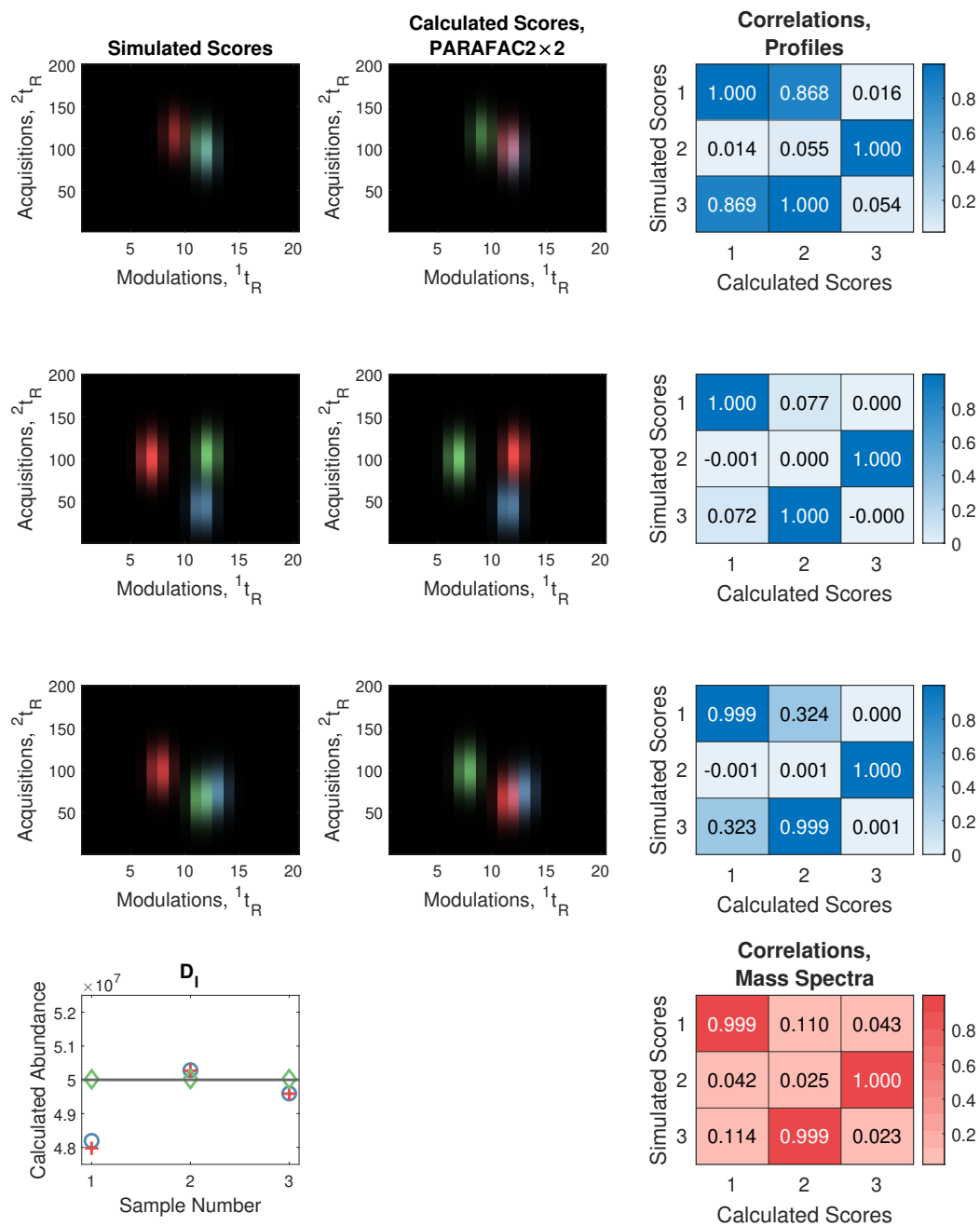


Figure 5.2: A simulated three-component model with a nominal SNR of 500. The percent variance explained using this model was 99.9565%. These results are good, despite almost complete overlap between two components in sample 1.

The components in the synthetic data are indexed in different positions relative to

the calculated features, but they can be related to each other using a permutation matrix, so this difference is inconsequential.

PARAFAC2 \times 2 is stable for analyses with seven components, based on the synthetic data that was tested. Beyond this level of complexity, the algorithm appears to have issues with reliable deconvolution of the signals. These issues may have arisen due to some limitation of the algorithm to deconvolve very complex data, or due to the fact that there is a limited amount of separation space for the synthetic data. However, for GC \times GC-TOFMS data from a well-optimized separation and with a properly-sized region of interest, this limitation should be of little practical import. Further experiments will be needed to explore the limitations of this algorithm on a variety of different datasets.

5.4.2 Analysis of Calibration Data using PARAFAC2 \times 2

Calibration data from a metabolomics study were used to test the PARAFAC2 \times 2 algorithm. Presumably, calibration data follows a predictable trend that can be used to judge the utility of PARAFAC2 \times 2 for quantitative and targeted analyses.

Experimental Data Collection

A region of interest was excised from a calibration experiment, containing 67 different calibrants that were dissolved in an amenable organic solvent mixture (either 50% Acetonitrile - 50% Water, or 50% Isopropanol - 50% Toluene for polar and non-polar compounds respectively). Standard solutions were aliquotted at different volumes into 2-mL GC vials and blown down under nitrogen at 40°C. Residual moisture was removed by adding 100 μ L of toluene dried under anhydrous sodium sulfate, which was again blown down using a stream of nitrogen at the same temperature. The dry residual extract was derivatised following a standard two-step methoximation / silylation approach. Briefly, 50 μ L of 20 mg/mL methoxyamine HCl in pyridine at 60°C for two hours, followed by 100 μ L of MSTFA at 60° C for 1 hour. Based on

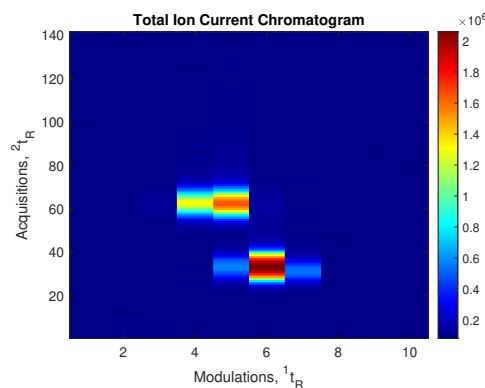
the expected concentration of each standard within the pyridine/MSTFA solvent, and the 1- μ L splitless injection volume, the relative quantification results from the PARAFAC2 \times 2 were plotted against the calculated pg of analyte that reach the head of the first GC \times GC-TOFMS column.

The two analytes present in this region of interest are the trimethylsilyl (TMS) derivatives of salicylic acid and adipic acid (in this case, both derivatised molecules contained two TMS groups). Their identities were confirmed by examining their retention indices, mass spectra, and analysing samples each containing a small fraction of analytes for confirmatory purposes.

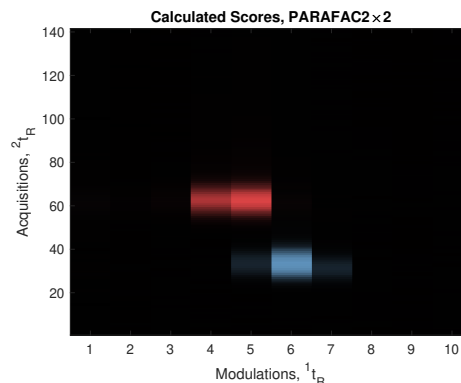
5.5 Extension to Multidimensional Separations Data

Higher-order separations present an exciting new avenue of research for the analysis of complex samples. However, while it is not impossible to model GC \times GC-TOFMS data by unfolding the retention times as a single retention mode, unfolding scales poorly for higher-order chromatographic separations (e.g. GC \times GC \times GC, LC \times LC \times LC, LC \times GC \times GC, etc). In addition to the excess degrees of freedom, there are practical issues for calculating excessively large matrices, related to the available memory on the computer system being used for the calculations. Utilising higher-order separations has found more favour in the relatively new field of multidimensional liquid chromatography, since the peak widths are generally much larger and there are fewer practical limitations with regard to the sampling rate of the mass spectrometer [136]. Some work has been done using comprehensive three-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry[137], but some issues persist with the published setup - since the instrumental sampling rate was limited to 200 Hz for third-dimension peaks eluting with a peak width of 50 ms, the sampling rate for a single peak is limited to about 3-4 acquisitions per peak.

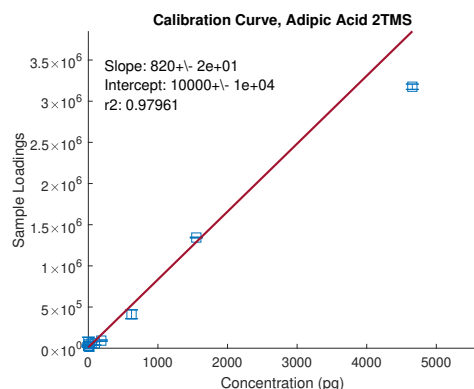
Consider a comprehensive three-dimensional separation, which can be described us-



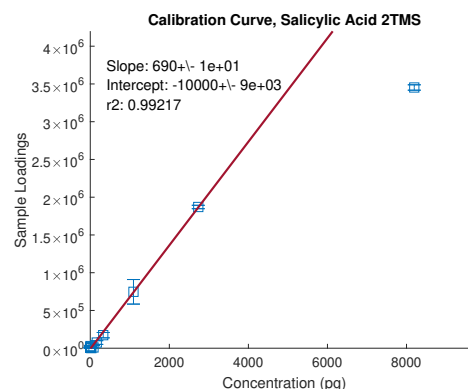
(a) Raw chromatographic data of a representative sample: Total Ion Current (TIC)



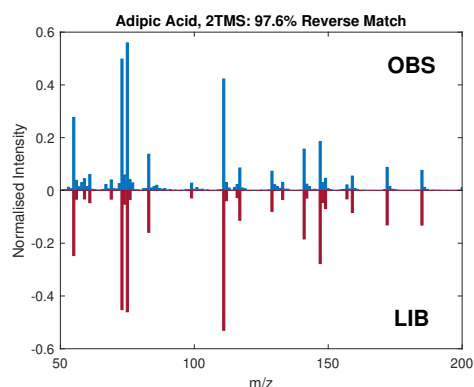
(b) Extracted scores for a representative sample. The upper left component is adipic acid (2TMS), and the lower right component is salicylic acid (2TMS)



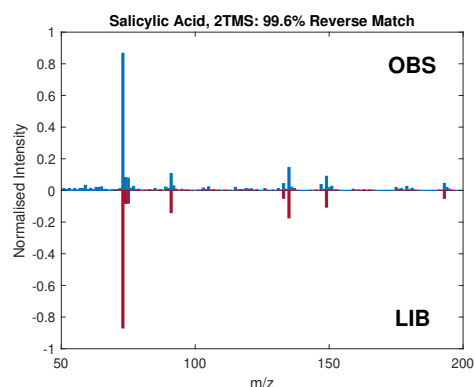
(c) Calibration curve for adipic acid (2TMS)



(d) Calibration curve for salicylic acid (2TMS)



(e) Library search results of the extracted mass spectrum for adipic acid (top), versus mass spectrum for salicylic acid (top), versus the library mass spectrum downloaded from the Golm Metabolome Database [134]



(f) Library search results of the extracted mass spectrum for salicylic acid (top), versus mass spectrum for adipic acid (top), versus the library mass spectrum downloaded from the Golm Metabolome Database [135]

Figure 5.3: Results of the analysis of calibration standards using PARAFAC2x2. The analysis utilised three-factors, but the noise component was not displayed for ease of visualisation. Plots for the scores and TIC chromatograms are displayed for the 8th sample in the calibration, where the calculated mass of the analyte on column was 2732.1 pg for salicylic acid (2TMS) and 1551.9 pg for adipic acid (2TMS)

ing an intuitive extension of Equation 5.2 for a 5th order tensor, $\mathcal{X}_{ijklm} \in \mathbb{R}^{I \times J \times K \times L \times M}$ structured so that it contains I acquisitions along the third retention mode, K modulations from the second to the third dimension, and L modulations from the first to the second separation dimension. D_M represents the quantitative loadings for the M^{th} sample, and A is the matrix of $J \times R$ mass spectra.

Assuming that the practical aspects of higher-order separations with hyphenated multivariate detection methods (such as mass-spectrometers or spectroscopic methods) are overcome, it is possible to model N -dimensional drift using the same principles that guided the expressions developed previously for GC \times GC-TOFMS data, where $\mathcal{X} \in \mathbb{R}^{I \times J \times K \times L \times M}$

$$\mathcal{X} = F_3(D_M \odot F_2 \odot F_1 \odot A)^T \quad (5.20)$$

Or using unfolded data, where $\mathcal{X} \in \mathbb{R}^{I \times K \times L \times J \times M}$:

$$\mathcal{X} = (F_3 \odot F_2 \odot F_1) D_M A^T \quad (5.21)$$

A model similar to the one proposed for GC \times GC-TOFMS data can be constructed for an X_{ijklm} tensor with drift in three modes.

$$\begin{aligned} \mathcal{X}_{ijklm} = \operatorname{argmin}(& ||X_{klm} - B_{klm} D_{klm} A_{klm}^T||_F^2 + \mu_{klm} ||B_{klm} - P_{klm} B_{klm}^{*T}||_F^2 \\ & + ||X_{ilm} - B_{ilm} D_{ilm} A_{ilm}^T||_F^2 + \mu_{ilm} ||B_{ilm} - P_{ilm} B_{ilm}^{*T}||_F^2 \\ & + ||X_{ikm} - B_{ikm} D_{ikm} A_{ikm}^T||_F^2 + \mu_{ikm} ||B_{ikm} - P_{ikm} B_{ikm}^{*T}||_F^2 \\ & + \mu_A ||A_{klm} - A_{ilm}||_F^2 + \mu_A ||A_{klm} - A_{ikm}||_F^2 + \mu_A ||A_{ilm} - A_{ikm}||_F^2) \end{aligned} \quad (5.22)$$

Where $X_{klm} \in \mathbb{R}^{I \times J \times K \times L \times M} = \mathbb{X}_{:, :, k, l, m}$, and other 3rd order X tensors follow similar convention, with the subscripts indexing what slices are being considered. Matrices associated with the non-negative PARAFAC2 decomposition are denoted by similar subscripts.

The number of terms that restrict the dissimilarity of the mass spectra with respect to the different methods of unfolding that data are related to the number of possible combinations of each of the terms, which is equal to the binomial coefficient, ${}_NC_2$. For even higher orders of separations further extensions are possible, but it is not convenient nor especially useful to come up with a generalised notation for these circumstances. The authors leave this exercise to the interested reader.

5.6 Conclusions

A general theory of modelling separations data with drift in multiple modes is proposed, and has been shown to work on experimental and synthetic data that are close to the worst possible scenario for independent chromatographic drift in two modes. The presents a parsimonious method for the deconvolution of signals and extraction of both qualitative and quantitative metrics from GC×GC-TOFMS data, and eliminates the need for unreliable dynamic programming routines that may contribute to peak splitting and/or peak drop-out commonly encountered in GC×GC-TOFMS peak tables.

For targeted analysis of GC×GC-TOFMS data, this algorithm is sufficient. Determining appropriate regions of interest and a value for the component number is a relatively simple task for a handful of components. Since the component number is specific to each region, the number of parameters required scales with the number of components being analysed in a series of chromatograms. While the number of data analysis parameters does not scale with the number of components being analysed using the currently available commercial offerings, there is still certainly a high degree of complexity inherent to analysing entire chromatograms using a single set of parameters, and there is no guarantee that a single set of parameters will be sufficient to analyse all of the desired targets with a high degree of accuracy. The approach proposed by the authors is more flexible, similar to the application of PARAFAC2 to GC-MS experiments, but requires skilled user intervention. This is a significant

first step towards a holistic chemometric method for pre-processing entire GC \times GC-TOFMS experiments, but additional work is required to automate the selection of regions of interest and choosing appropriate component numbers for each region.

It is also possible to model single samples using PARAFAC2 \times 2, since a high number of replicates are inherent even with a single GC \times GC-TOFMS sample. This makes modelling of single chromatograms extensible to larger numbers of chromatograms using the same model, and the results of the analysis for one sample may be extrapolated to several more.

Additional investigations are needed to evaluate this technique in relation to different methods for unfolding the data, which is not a trivial task, and is deserving of its own article. Considerations for the practicality of unfolding data in a way that generates more replicates at the expense of degrees of freedom, and computational efficiency must be considered. However, to the best of the authors' knowledge this algorithm is the first of its kind applied to multidimensional chromatographic data, and represents a significant leap forward for the field of GC \times GC-TOFMS data analysis.

Chapter 6

A priori prediction of chemical rank in gas chromatography - mass spectrometry data with projection pursuit analysis

6.1 Introduction

Matrix decomposition techniques are a family of ubiquitous mathematical, and/or chemometric methods often used to extract chemical information from overlapping multivariate signals, commonly observed in gas chromatography-mass spectrometry (GC-MS) data. Interpretation of coelutions presents a significant data analysis challenge for which matrix decomposition techniques are well-suited. The natural competitor of GC-MS, High-Performance Liquid Chromatography - Mass Spectrometry (LC-MS), largely evades the problem of overlapping signals through the use of tandem mass analyzers, high-resolution mass spectrometers, and soft ionization (hence little-to-no fragmentation of ions). Conversely, GC-MS is typified by hard electron impact ionization (abundant fragmentation, with potential for common fragments in coeluting compounds), and single-stage, unit-mass spectrometers. This necessitates the development of additional software tools for processing data.

Techniques based on Multivariate Curve Resolution (MCR) [20], Independent Component Analysis (ICA) [138], and PARAFAC/PARAFAC2 for N-way analyses [42]

[43], have all been applied to the analysis of closely co-eluting factors in GC-MS data to extract identifiable mass spectra and quantitative peak profiles. The advantages of these techniques over various commercial offerings, is that matrix decomposition techniques are generally more robust against integration artefacts which can arise from random fluctuations of background noise, sub-optimal user parameters, or outright failure of proprietary algorithms in commercial software [139]. Problems with current commercial implementations are numerous [140], and can be especially significant for commercial GC \times GC-TOFMS software [141] [142]. To counteract the relatively low industry standards for GC-MS data analysis, a platform implementing PARAFAC2 has been released: “PARAFAC2 based Deconvolution and Identification System” (PARADISE) [139]. PARADISE allows users to extract high quality data, but relies on experienced-user intervention for the selection of regions of interest (ROIs), and the selection of an appropriate number of significant signal components in each region. Determining an appropriate component number, k , has been a bottleneck for the application of matrix decomposition techniques for automated data processing, and is often considered to be a fundamental problem in the practice of chemometrics [143].

Establishing a threshold for % variance explained, or determining the “elbow” in the calculated eigenvalues [144] [145] [146] are somewhat arbitrary. The correct threshold can vary depending on a number of factors, such as each components’ signal-to-noise ratio, and the characteristics of the baseline noise. The elbow method calculates the distance from each eigenvalue to a diagonal line from the first- and last-calculated eigenvalues. This method is reliant on an appropriate number of calculated eigenvalues, which dictate the slope of the line, and tends to underestimate k when there is more than one elbow in the series of eigenvalues. Cross-validation techniques are another common way of determining the correct number of components [147], where the predictive ability of models with particular values for k are measured, and the value of k for the best performing model is selected as the optimum. Calculating several

models, which is necessary for cross-validation techniques, is computationally intensive, and can be somewhat deceptive when it comes to deconvolution-type problems, as there is not a clear metric by which to assess the number of useful chemical components for two-way data. Even so, calculating several models for cross-validation purposes is still time-consuming given widespread reliance on the alternating least squares algorithm and nonetheless would benefit immensely from an initial estimate for the number of components, so as to reduce the number of models to calculate.

Worth mentioning is the Core Consistency Diagnostic (CORCONDIA) [148] for tensor decomposition, which solves for the super-diagonal entries of the core tensor for a Tucker model using the calculated loading matrices of a PARAFAC model for a particular number of components. For a model with an appropriate number of components, the super-diagonal entries should be close to 1, the off-diagonal entries should be close to zero, and a model quality measure between 0 and 1 (typically displayed as a percentage) can be determined. CORCONDIA has been generalised for use with PARAFAC2, and has been deployed for semi-supervised manual interpretation of raw GC-MS data [149] [139]. Convolutional neural networks have also been applied within the framework of PARADISE to classify different components as baseline noise, shouldering peaks, or useful chemical features [150].

Methods introducing a degree of statistical rigour have been influential, but many theorems have too broad a scope for applications in matrix decomposition [151] [152] and are generally not used. Other techniques have been published, but are somewhat utilitarian in nature [153] [131].

Choosing the correct number of components for a principal component analysis (PCA) [154], or for k -type clustering methods [155] have received considerably more attention in the literature. Although it is worth noting that an appropriate number of principal components is much less critical for principal component analysis than it is for the decomposition techniques used in deconvolution. Overestimating the number of components in PCA for example, only risks over-fitting, and does not dramatically

change the results of the analysis as it would for MCR, ICA, or PARAFAC2. That is, principal component $N + 1$, does not affect how principal components 1 to N are calculated, even though the reverse is true. It is nonetheless worth mentioning there are a number of methods to estimate an appropriate number of principal components automatically and, at least in theory, some of these techniques could be applied to deconvolution. A major drawback of techniques based on observing the diminishing significance of subsequent principal components, is that it poorly accounts for chemical factors with highly similar mass spectra. In these cases, as many as one principal component could reasonably account for the variance explained by dozens of factors with highly similar mass spectra. This is a common problem for regions of interest that present a high number of coeluting branched alkanes for example.

In this submission, we present an approach that decomposes a chromatographic region via the automated pursuit of a projection index that minimises the kurtosis of the resultant score matrix. If the data is scaled in an appropriate way, the scores cluster in such a way that an appropriate prediction for the chemical rank of the matrix can be made using a density-based clustering approach.

6.2 Motivation

For an $m \times n$ matrix, X , of m observations (sequential spectral acquisitions for GC-MS-type data) and n variables (Mass-to-charge ratios (m/z) for GC-MS-type data) where $(m, n) \in \mathfrak{R}$, there will exist k significant chemical factors that can be extracted as linearly independent parallel vectors via a matrix decomposition:

$$X = CS^T \tag{6.1}$$

C is an $m \times k$ matrix containing the scores, or elution profiles, and S^T a $k \times n$ matrix of the loadings, or characteristic mass spectra, for the k^{th} factor of the matrix X . Typically, S is normalised to each columns' euclidean norm such that the matrix

C contains the quantitative information. There are several ways of solving for these two factors, either by selecting a number of components that maximise the statistical independence of the elution profiles, C , as is the case for ICA, or by approximating each C , and S matrix via their least-squares solution iteratively, as in MCR-ALS. The chemical rank of the matrix, is the integer value for k that corresponds to the number of parallel vectors in the matrices C and S . An appropriate value for k is generally understood as being the value that best represents the latent chemical phenomena present within the matrix. A human analyst is typically the best judge of this, but it is possible to imitate an analysts’ thought process by making some assumptions about what projections of the data are worth considering.

Projection Pursuit Analysis (PPA) is a technique first proposed by Friedman and Tukey [35] that seeks to find “interesting” projections based on the pursuit of an interesting projection index. This is a general enough description to encompass a other common linear decomposition techniques: ICA maximises a projection index of statistical independence, and PCA maximises an index of the explained variance of the data. All projection indices make assumptions about what characteristics of the data are most interesting for the analyst, save for those instances where the projection index is selected manually. Hou and Wentzell in 2011 [30] first described the minimisation of kurtosis as a projection index to reveal resultant clustering of the data. This was motivated by the observation that highly resolved score clusters present a very low kurtosis K , or tendency for the data to feature a relatively low number and extremity of outliers. Similar observations with a larger measure of kurtosis are generally unimodally distributed, with a higher number of more extreme outliers. Kurtosis can be described by the following equation:

$$K = \frac{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^4}{\frac{1}{n} (\sum_{i=1}^n (z_i - \bar{z})^2)^2} \quad (6.2)$$

Where z_i refer to the score of each sample, as projected along a vector, \mathbf{v} such that

$z_i = \mathbf{x}_i^T \mathbf{v}$. Calculation of the projection vector is tantamount to minimising:

$$K = \frac{n \sum_{i=1}^n (\mathbf{v}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v})^2}{(\mathbf{v}^T X^T X \mathbf{v})^2} \quad (6.3)$$

With respect to \mathbf{v} , and determination of subsequent projection vectors can be made on the deflated matrix. Equation 6.3 was minimised by Wentzell and Hou [30] using a quasi-power method, and their algorithm was used for this work.

In order to make GC-MS data cluster in a way that's directly observable using a density-based metric, an appropriate scaling and pre-treatment of the data must be applied. Normally, the change in intensity of a Gaussian peak increases until it reaches an inflection point at $x = \mu \pm \sigma$ for an idealised distribution. The change in intensity continues until the peak at $x = \mu$ is reached, and the observed change in intensity increases and decreases similarly in the opposite direction. This idealised Gaussian behavior is frequently observed when modelling the scores of any decomposition of GC-MS data, and the resultant scores are typically rather diffuse for mean-centred and/or autoscaled data (*i.e.* the resultant scores' rate of change tends towards a reduction in density, rather than an increase). To counteract this, each acquisition can be scaled such that the intensity of each mass channel is bounded between 0 and 1 - commonly referred to as row-wise *min-max* scaling. In effect, this encourages the resultant scores to cluster, by diminishing the effect of the changing intensity for each of the peaks being analysed. Since the PPA algorithm works best on low-dimensional data, the data was reduced to be approximately 5 times more variables than the number of observations. Savitsky-Golay filtering was also employed to smooth the data prior to analysis, with a polynomial order of 5, and a window size equal to the minimum sample number for clustering + 4 (in this case, an odd number was always selected).

Density-based spatial clustering of applications with noise (DBSCAN) was used to estimate the number of clusters within the projection pursuit scores. Two major parameters, the ϵ neighbourhood of each point, and the minimum number of neigh-

bouring points required to form a cluster, can be estimated based on acquisition rate of the instrument and expected peak-width (2σ) for each peak. These parameters are also relatively easy to optimise, depending on the needs of the user. A typical value for ϵ was 0.3, and a typical value for the minimum number of samples for inclusion into a cluster was 5. A drawback of using DBSCAN here is its dependence on assigning a single class membership to each observation. For poorly resolved components, despite obvious clustering in the PPA scores, it is not always possible to assign representative class membership. However the most significant benefit of using DBSCAN is its ability to return an estimate for the number of clusters.

6.3 Materials and Methods

Test data was acquired from different sources to best evaluate the effectiveness of the technique across different operating conditions. Gas Chromatography - Time-of-Flight Mass Spectrometry (GC-TOFMS) data with a relatively high sampling rate was previously published with the ICA-OSD package, and the operating conditions under which it was collected are highlighted in the original study from which it was obtained [156]. Gas Chromatography - quadrupole Mass Spectrometry (GC-qMS) data with a relatively low sampling rate was extracted from an anonymous study performed on an Agilent Technologies 5975C MSD (Agilent Technologies, Mississauga, Ontario, Canada) equipped with a nominal 30 m (5% phenyl / 95% methyl)-equivalent stationary phase, 0.25 mm internal diameter \times 0.25 μ m film thickness capillary column. Ultra-high purity helium (5.0 grade; Praxair, Edmonton, AB) was used as a carrier gas operating under speed-optimised flow (2 mL/min) and an optimal heating rate of 10 °C per minute of system dead time. The temperature program was initially held for two minutes at 40 °C and at the end of the run at 325 °C for 5 minutes. The single quadrupole collected three mass spectra per second from 25 to 500 m/z.

GC \times GC-TOFMS data was collected on a Leco Pegasus 4D GC \times GC-TOFMS with a quad-jet liquid nitrogen cooled thermal modulator. The first-dimension column was

a 60 m \times 0.25 mm internal diameter \times 0.25 μ m film thickness Rxi-5SilMS (5% phenyl) and the second-dimension column used was a 1.2 m \times 0.25 mm internal diameter \times 0.25 μ m Rtx-200MS (trifluoropropyl). Helium was used as the carrier gas at speed optimised flow (2 mL/min) and an optimal heating rate of 3.5 $^{\circ}$ C/min from an initial temperature of 80 $^{\circ}$ C held for 4 minutes, and a final temperature of 315 $^{\circ}$ C held for 10 minutes. Mass spectra were collected at 200 Hz between 50 and 660 m/z. The detector voltage offset was optimized according to the internal quality control parameters of the instrument with an electron impact energy of -70 eV. The ion source temperature was 225 with a transfer line temperature of 225 .

Synthetic GC-MS data was generated with an in-house MATLAB[®] function using a manually curated list of 100 different mass spectra obtained from the 2017 version of the National Institute of Standards and Technology Mass Spectral Database (NIST MS Database). Among the list were 40 common targets for ignitable liquid and pesticide analysis, and 60 randomly selected chemical components that were judged to be amenable to analysis by GC based on their experimental or estimated retention indices as reported in the NIST MS Database. Many of the common targets for ignitable liquid analysis, such as the alkylbenzenes and structurally similar alkanes, produce fragmentation mass spectra with many common ions and a high degree of similarity. Consequently, the benefits and limitations of this technique for coelutions of compounds with highly similar mass spectra are demonstrated in Appendix D.

Computations were performed on a Lenovo ThinkCentre M700 running Ubuntu 18.04 LTS “Bionic Beaver” with 8 GB RAM, and an Intel i3-6100T CPU @ 3.20 GHz. All routines were implemented using MATLAB[®] 2020b. Multivariate Curve Resolution (MCR), and Savitsky-Golay Smoothing were performed using PLSToolbox 8.91.

Projection pursuit analysis was performed using the quasi-power described by Hou and Wentzell [30]. 7 projection score vectors were used for analysis of subsequent clustering. 50 initialisations of the data were used to ensure convergence to a global

optimum. The DBSCAN algorithm used was part of the scikit-learn Python package, which appeared to offer better performance on these data [157].

The algorithms in this work have been published elsewhere, and the proposed method is trivial to implement. The experimental data used in this work can be found online at: <https://doi.org/10.7939/DVN/NTLBGY>.

6.4 Results and Discussion

6.4.1 GC-MS Results

MCR is a well-established technique for the deconvolution of 2-way GC-MS data, and non-negativity constraints were used to ensure the results corresponded to quantitative latent phenomena. k component numbers were calculated from the clustering of the PPA scores using DBSCAN. MCR loadings were normalized to their maximum m/z value and compared to library spectra from the NIST database via dot product calculations. The highest scoring library spectra are listed as a match, with the dot product converted to a percentage for the corresponding match factor.

The following dataset was collected on a quadrupole GC-MS, presenting a relatively low sampling rate. The results of the automated analysis are summarized in figure 6.1.

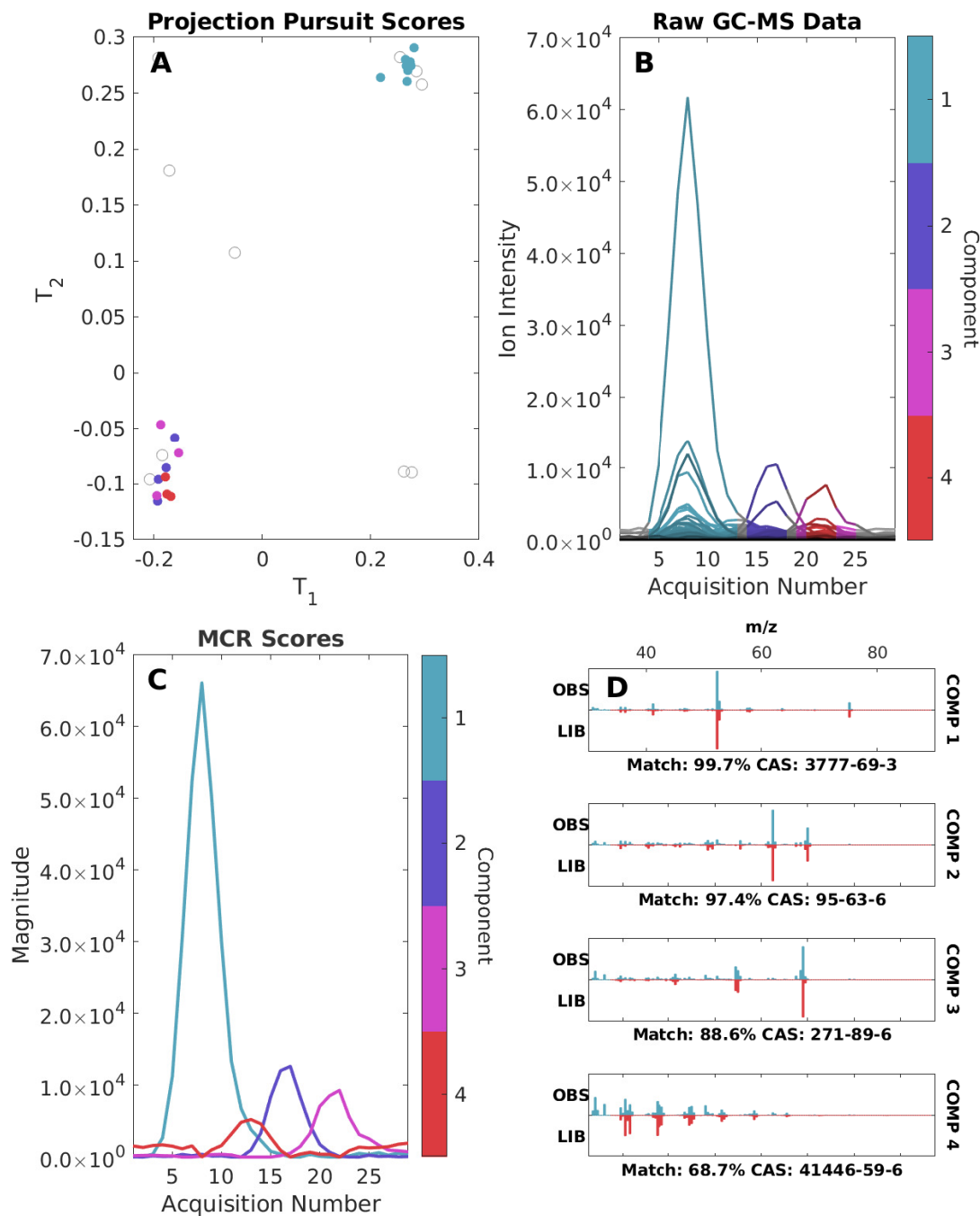


Figure 6.1: Summary analysis of a chromatographic region, with each of the components accounting for 92.7%, 96.5%, 98.6%, 99.7% percent variance explained, respectively. Projection pursuit scores with their predicted cluster membership are found in A, and the cluster memberships are visualised using the raw data in B. Using the predicted number of chemical factors, the resultant MCR scores are presented in C, and the results of the library searches are presented in D. ϵ value used for clustering was 0.3, and the minimum number of samples for clustering was 3, owing to the low sampling rate of the instrument.

The algorithm was able to identify several chromatographic regions that corresponded well with the latent chemical factors, as subsequently deconvolved by MCR. Although this appears to be a reasonable estimate for the chemical rank of the matrix, one of the factors as identified by PPA-DBSCAN (Component 4) was split across multiple acquisitions in a manner that suggests that this factor was spuriously identified. This behaviour, combined with the relatively low library match statistic for this component suggests the presence of a significant amount of noise being captured as (or in conjunction with real signal) in Component 4.

6.4.2 GC-TOFMS Results

Analysing one of the datasets from the ICA-OSD package, “gcms1”:

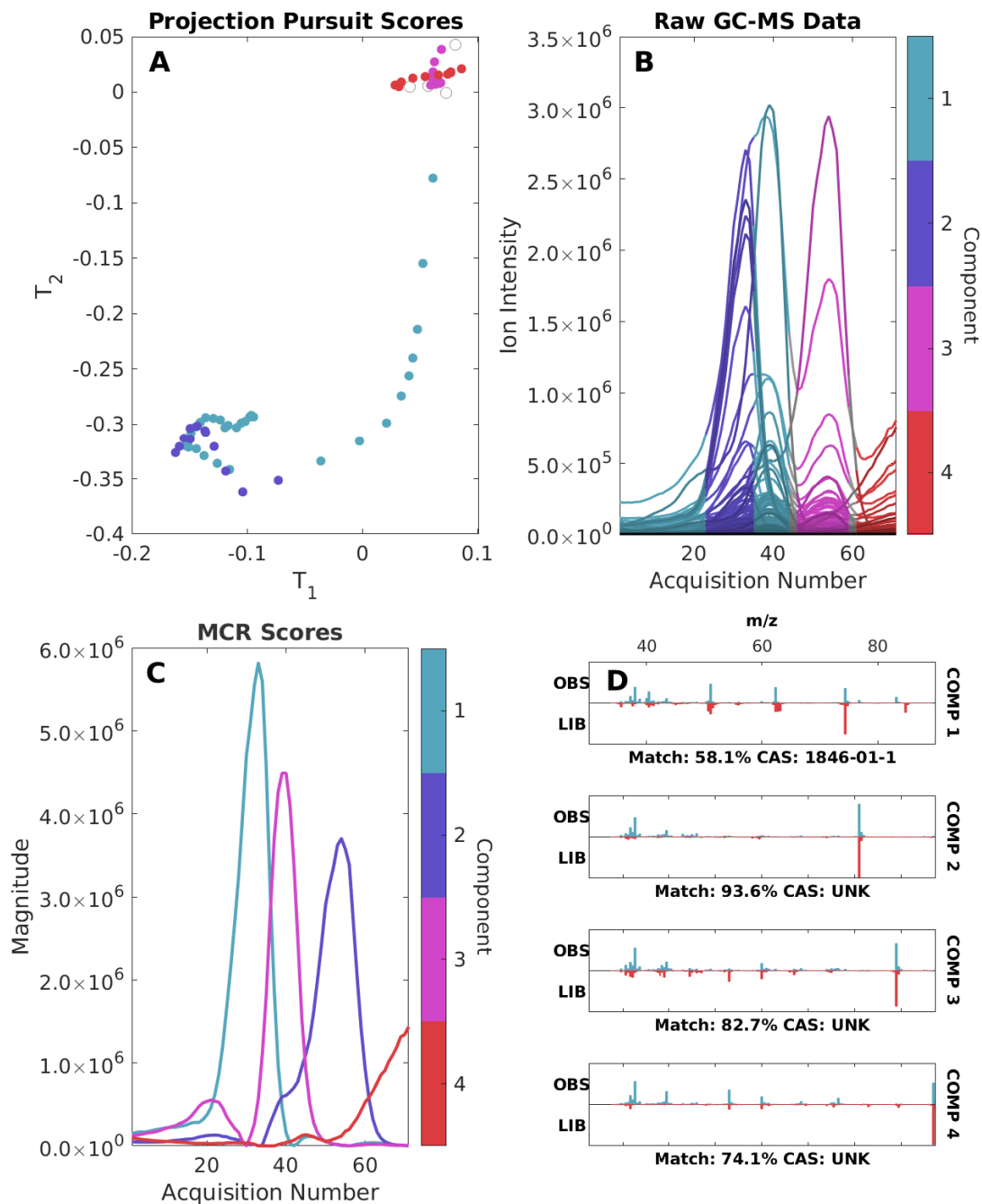


Figure 6.2: Summary analysis of a chromatographic region, shipped as part of the ICA-OSD package. Each component accounts for 46.8%, 70.8%, 97.1%, and 99.9% of the variance explained, respectively. ϵ value used for clustering was 0.3, and the minimum number of samples for clustering was 5, reflecting the higher sampling rate of the instrument used for the analysis.

The library hits for each of the components score relatively low, but each component profile appears to correspond to a real chemical feature that was identified by the PPA-DBSCAN algorithm. This could be due to the relatively low number of derivatized metabolite standards in the NIST database that was used for this work. The benefits of a higher sampling rate are apparent in this analysis, since each identified cluster appears to correlate to real chemical features in the data.

6.4.3 Considerations for GC \times GC-TOFMS Data

This algorithm can in theory be used for GC \times GC-TOFMS unfolded along a single retention axis; however, there are practical limitations to using PPA for large datasets, since the global optimum for even smaller datasets is not always easy to find. A workaround is to sum the modulations, or second-dimension acquisitions to reconstruct either the first- or second-dimension retention times. This presents some challenges for automation, since it is difficult to tell along which axis multiple components are best resolved, without first calculating the model. As well, reconstructed first-dimension chromatograms present much fewer observations for a typical region of interest, but reconstructed second-dimension chromatograms typically offer less resolution due to the very short length of the second-dimension column. An example of this can be found in Appendix D.

6.4.4 Limitations of the approach

This approach is primarily limited by the clustering step, which makes it difficult to assign class membership to components where there is significant overlap with other co-eluting chemical factors. These factors are typically excluded from the clusters as noise, or incorrectly assigned to a more prominent component.

For small regions of interest with relatively low sampling rates, there are few observations in the resultant matrix. Fewer observations make the clustering step more difficult, but also analysis by projection pursuit analysis since the kurtosis of the

resultant scores is not as obvious.

Examples of algorithm failure can be seen in Appendix D.

6.5 Conclusion

Kurtosis minimisation appears to be a useful projection index for predicting the number of significant components in GC-MS data. While DBSCAN has been used to make an automated estimate for the k component number, it presents a number of drawbacks related to the necessary input parameters, and the fact that one observation must be assigned a single class membership. Nonetheless, our preliminary results show that it is possible to discover useful projection vectors leading to clearly identifiable clusters, that correlate with a useful prediction of k .

GC-MS data is among the most challenging types of data to analyse, since chemical factors present in co-elutions often feature highly similar mass spectra. Projection pursuit analysis can be used to discover small differences in these mass spectra, and the resultant score clusters may resolve well enough to distinguish them. Correctly indicating the number of chemical components is only half of the required theory for an objective and parameter-free analysis of an entire GC-MS chromatogram. As it currently stands, user input is still required to identify regions of interest that are well-enough resolved from neighbouring regions of interest. A current standard for region of interest selection and component number determination is currently absent, and demands a great deal of innovation on both fronts. It is the authors' hope that this work presents an exciting new avenue for the automated determination of k component numbers, and will lead to parameter-free analysis of entire GC-MS experiments.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The primary aim of this thesis was to improve the methods by which users of GC \times GC-TOFMS can transform their raw data into useful information, with minimal impacts from artefacts and spurious signals.

The numerical solution for the cluster resolution metric has been proposed, and critically tested on a number of different datasets. Improvements to the computational efficiency of this algorithm may enable its deployment as an online tool sometime in the near future. As well, the speed with which the cluster resolution metric can be calculated enables more combinations of data, and has been shown to generate reliable results for classification problems that rely on peak table output from commercial software. Any feature selection routine is limited by the quality of the input data however, and external validation of the model remains the gold standard by which selected chemical factors can be judged as significant.

There are a number of challenges posed by commercial offerings to process GC \times GC-TOFMS data: subjective settings for integration parameters, and an inability to correctly process both trace analytes and overloaded analyte signals in the same chromatogram while accounting for drifting signals. As such, a new processing algorithm has been proposed. PARAFAC2 \times 2 allows users to analyse either single or multiple chromatograms within a particular retention window, without excessive parameters

and within a reasonable amount of time. Compared with more straightforward methods of analysis, such as unfolding the data along a single retention axis, this method preserves the high degree of redundancy for GC \times GC-TOFMS data that greatly benefits tensor decomposition methods. This has proven to be useful for targeted analyses, as demonstrated with calibration data, and removes much of the uncertainty to do with identifying features from the output of ChromaTOF[®], since the mass spectra are deconvolved in tandem with the sample-wise loadings. The results appear to be highly interpretable, and seem to correspond well to the latent chemical phenomena being studied.

Sometime in the near future the ultimate goal of this type of research will be automated application of chemometric tools for parameter-free pre-processing of GC \times GC-TOFMS data. To enable this, a novel technique for the estimation of the number of latent factors (i.e. the chemical rank) has been proposed, utilising PPA that minimises the kurtosis of the manifest scores. The clustering of these scores on appropriately pre-processed data appear to cluster in a way that correlates to the number of chemical factors in the matrix being studied. Although there are still some parameters that need to be tuned by the analysis, with respect to the clustering algorithm: DBSCAN, it appears as though this technique offers simple a theoretical framework for the estimation of the number of chemical factors within a dataset.

7.2 Future Work

7.2.1 Future work to do with FS-CR

There are two main categories of consideration for future work: the first is to do with the cluster resolution metric. The cluster resolution metric currently minimises a very complicated parametric equation using the Nedler-Mead Simplex algorithm. This is relatively easy for the computer to manage, since the complexity is largely due to arithmetic operations. However, scaling the cost function for determinations of

the cluster resolution metric in higher dimensions will become prohibitively difficult, and troubleshooting any errors even more so. The first step for the extension of the cluster resolution metric to higher dimensions will depend on the application of more standard matrix mathematics to find a minimum of a much simpler cost function. This should be easy to do using the ALS algorithm.

It is also suspected that the cluster resolution metric may be a superior metric for assessing model quality in classification methods, and by extension many wrapper-type routines for feature selection could stand to benefit from utilising the cluster resolution metric, as opposed to cross-validation metrics alone. As mentioned previously, the improvements to using cluster resolution have to do with the granularity of the model quality results. As a non-linear optimisation problem with a smooth surface close to the optimum, small differences in the sample positions in PCA space can have a measurable effect on the results of the analysis by cluster resolution, regardless of whether there are samples that have crossed the decision boundary based on the inclusion or exclusion of a variable being considered. Future work will involve building identical feature selection routines: one using cluster resolution, versus one using a cross-validation metric and comparing the results on various datasets with known solutions to the problem.

7.2.2 Future work to do with PARAFAC2 \times 2

Extension of the PARAFAC2 \times 2 algorithm to higher-order chromatographic, or ion mobility modes presents an exciting avenue for tensor decompositions, and more application-oriented analytical chemists. Doing so will be less straightforward, since it will require the derivation of new cost functions related to the expression postulated towards the end of Chapter 6. However, the rewards for reliable identification and integration of extremely complex chromatographic data may well be worth it.

Flexible coupling PARAFAC2, and PARAFAC2 \times 2 both rely on several initialisations to ensure convergence to a global optimum. This is a computationally intensive

process, and may slow down batch analyses of multiple chromatographic regions of interest considerably. Work is currently underway to perfect an algorithm that provides accurate initial guess for Flexible coupling PARAFAC2, and PARAFAC2 \times 2, using Independent Component Analysis (ICA). This approach appears to provide more informative initial estimates for the score profiles, and ensures convergence to the same solution every time. However, testing on further datasets is necessary to lend credence to the idea that the initial estimates as provided by ICA are close to a global optimum.

Examining the sample-wise loadings of PARAFAC2 \times 2 does not appear to be particularly sensitive for quantitative analysis, despite the relatively tight error bars for even extremely overloaded samples. To counteract this, it may be possible to use the calculated scores to solve for quantitative, multivariate information in the mass-spectral mode, and to perform a multivariate regression on those data, for each chemical factor within a region of interest. Doing so may improve the sensitivity, but a critical evaluation of the model performance will need to be done on external validation (in this case, spike recovery) samples for either case.

7.2.3 Future work on Applications

The second category of major avenues for future work revolve around application of the new tools presented within this thesis for the analysis of the existing datasets. Many tentative bio-markers have not been conclusively identified, and it is currently unknown if this is a problem with the ChromaTOF[®] software, the library search tool, or if the molecules themselves are unknown and lack any library standards. As mentioned previously, a damning critique of ChromaTOF[®] is an absolutely necessary, but not particularly interesting avenue of research, given the scale and difficulty of the task. Even so, research is currently underway to try and find a fair means of comparison.

7.2.4 Future work on automated k estimation

The current implementation of PPA is extremely slow, and may scale poorly to large-scale analysis of many regions of interest. The current implementation of the algorithm utilises a quasi-power method, and also requires a dimensionality reduction step prior to analysis. The quasi-power algorithm requires a large number of initialisations to confirm convergence to a global optimum, and this necessitates much redundant computation time.

Independent Component Analysis (ICA) is a different matrix decomposition technique that seeks to maximise the statistical independence of the score vectors relative to each other. A common implementation of the algorithm, Joint Orthogonalisation of Diagonal Eigenmatrices (JADE), exploits higher order cumulants (such as kurtosis) to calculate the maximally statistically independent components. It may be worth considering if a similar algorithm may be possible for PPA.

7.2.5 Region of Interest Selection

A fully automated routine for the analysis of GC \times GC-TOFMS data as yet requires the automated selection of regions of interest. While there have been a number of pragmatic, utilitarian solutions to this issue by a number of different researchers, the vast majority of these solutions operate on single chromatograms. In order for the PARAFAC2 \times 2 algorithm to work, a similar region of interest must be selected that is consistent across multiple samples, although similar to PARAFAC2, with some alterations it is possible for the retention windows to be of different sizes. Selecting a similar retention window for all samples is inflexible, and selecting different retention windows for each sample begs the question of how to properly associate them for analysis by PARAFAC2 \times 2. Especially for experiments with significant drifts in retention time, the former possibility is less than appealing. Currently there exists no mathematically satisfying way of selecting regions of interest, and the type of mathematics that may illuminate potential avenues of research typically fall within combinatorics,

which is far outside of the scope of this thesis. It is possible however, for a number of regions identified per sample, the inner product of their primary principal component vectors could be used to gauge a comparison of the spectral information encoded within each region of interest, such that along with a pair of retention times, an optimum could be achieved that minimises the total sum of residual squares for all such vectors. However, this is purely conjecture.

7.3 Outlook

While it may be possible to fully automate the analysis of GC \times GC-TOFMS data in the near future, care must always be given to ensure that proper model validation is used. That is, while new developments in chemometrics may enable the analyst to save time and obtain a more objective summary of the data, the added convenience may just as easily be abused by companies and research groups with less than honourable intentions regarding the data. Powerful instrumentation has increased the number of features we are able to observe in our samples, but the number of samples we have been able to acquire has not nearly kept pace. Routines such as feature selection, and rank-deficient solutions to supervised learning problems just as easily enable their users to tell lies about the data, as they do enable them to tell the truth. This has been a major problem in metabolomics research as of late, and will likely plague the next discipline to suddenly become encumbered with an overabundance of data.

The problem has grown despite the existing software solutions available to the analytical chemist, and it's unlikely that peer-reviewed tools for the analysis of GC \times GC-TOFMS data will make the problem any worse. However, it is the author's opinion that relatively few high-quality features are always preferred over an abundance of low-quality data. Despite the improvements made to the feature selection routine that is popular in our research group, if the chemical information in the validation set is significantly different than the chemical information found in the training set, no feature selection is possible that will allow the model to correctly indicate the

external samples.

In developing algorithms for GC \times GC-TOFMS, or any other hyphenated chromatographic separation, it is always important to resist the temptation to rely too much on the programmatic aspects of data analysis; too often, a number of logical statements can account for difficulties with the data that the algorithm is being tested on, but these algorithms break down when data is fed to it that the author of the algorithm did not expect to see. Early on in this research, a number of algorithms were tested that could be categorised as suffering from such short-sightedness. Indeed, as the demand for better and better tools for analysis are echoed in the chromatography community, so does the demand for papers on data analysis. The best analysis is always the analysis with the best mathematical theory that is applicable to the data. Arbitrary thresholds for significance and conditional statements cannot be proven to be unique, or even useful solutions, and their drawbacks are not clear when comparing them with other analyses. Much of the academic work in this field leads to a dead-end, even if the information is thoroughly peer-reviewed.

Nonetheless, it is an exciting time to be an analytical chemist. Many more applications are being discovered for the advanced instrumentation that is enjoying more widespread use; owing to the incredibly detailed chemical characterisation, new insights for disease are being found regularly, and new tests are being developed for complex multivariate problems. Especially with the use of PARAFAC2 \times 2, it may be possible to quantify features reliably using 2-dimensional chromatography, which may make such instrumentation more practical for routine applications.

Bibliography

- [1] M. A. (mykhaylo) (<https://math.stackexchange.com/users/692546/mykhaylo>), *Least-squares solution that minimizes two expressions*, <https://math.stackexchange.com/questions/3993537/least-squares-solution-that-minimizes-two-expressions/> (version 2021-01-22), 2021.
- [2] A. P. de la Mata, R. H. McQueen, S. L. Nam, and J. J. Harynuk, “Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics,” *Analytical and bioanalytical chemistry*, vol. 409, no. 7, pp. 1905–1913, 2017.
- [3] O. Fiehn, “Metabolomics—the link between genotypes and phenotypes,” *Functional genomics*, pp. 155–171, 2002.
- [4] A De Juan, E Casassas, and R Tauler, “Soft modeling of analytical data,” *Encyclopedia of analytical chemistry Applications, theory and instrumentation*, 2006.
- [5] R. G. Brereton and G. R. Lloyd, “Partial least squares discriminant analysis taking the magic away,” *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, 2014.
- [6] L. E. Frank and J. H. Friedman, “A statistical view of some chemometrics regression tools,” *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.
- [7] E. Fridman and E. Pichersky, “Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products,” *Current opinion in plant biology*, vol. 8, no. 3, pp. 242–248, 2005.
- [8] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.
- [9] C. B. Clish, “Metabolomics an emerging but powerful tool for precision medicine,” *Molecular Case Studies*, vol. 1, no. 1, a000588, 2015.
- [10] L. R. Snyder, J. J. Kirkland, and J. W. Dolan, *Introduction to modern liquid chromatography*. John Wiley & Sons, 2011.
- [11] R. L. Grob and E. F. Barry, *Modern practice of gas chromatography*. John Wiley & Sons, 2004.
- [12] M. S. Klee and L. M. Blumberg, “Theoretical and practical aspects of fast gas chromatography and method translation,” *Journal of chromatographic science*, vol. 40, no. 5, pp. 234–247, 2002.

- [13] L. Blumberg and M. Klee, "Optimal heating rate in gas chromatography," *Journal of Microcolumn Separations*, vol. 12, no. 9, pp. 508–514, 2000.
- [14] L. M. Blumberg, "Theory of fast capillary gas chromatography—part 3 column performance vs. gas flow rate," *Journal of High Resolution Chromatography*, vol. 22, no. 7, pp. 403–413, 1999.
- [15] C. R. Mallet, Z. Lu, and J. R. Mazzeo, "A study of ion suppression effects in electrospray ionization from mobile phase additives and solid-phase extracts," *Rapid Communications in Mass Spectrometry*, vol. 18, no. 1, pp. 49–58, 2004.
- [16] R. I. Reed, *Ion Production by Electron Impact*. Academic Press, New York, 1962.
- [17] R. E. March, "An introduction to quadrupole ion trap mass spectrometry," *Journal of mass spectrometry*, vol. 32, no. 4, pp. 351–369, 1997.
- [18] A. Samokhin, "Spectral skewing in gas chromatography–mass spectrometry: Misconceptions and realities," *Journal of Chromatography A*, vol. 1576, pp. 113–119, 2018.
- [19] K. Mavstovska and S. J. Lehotay, "Practical approaches to fast gas chromatography–mass spectrometry," *Journal of Chromatography A*, vol. 1000, no. 1-2, pp. 153–180, 2003.
- [20] F. Azimi and M. Fatemi, "Multivariate curve resolution-assisted gc-ms analysis of the volatile chemical constituents in iranian citrus aurantium l. peel," *RSC advances*, vol. 6, no. 112, pp. 111 197–111 209, 2016.
- [21] X. Domingo-Almenara, A. Perera, N. Ramirez, N. Canellas, X. Correig, and J. Brezmes, "Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation," *Journal of Chromatography A*, vol. 1409, pp. 226–233, 2015.
- [22] F. P. Abramson, "Automated identification of mass spectra by the reverse search," *Analytical Chemistry*, vol. 47, no. 1, pp. 45–49, 1975.
- [23] J. Hummel, N. Strehmel, C. Bolling, S. Schmidt, D. Walther, and J. Kopka, "Mass spectral search and analysis using the golm metabolome database," *The handbook of plant metabolomics*, pp. 321–343, 2013.
- [24] X. Wei, I. Koo, S. Kim, and X. Zhang, "Compound identification in gc-ms by simultaneously evaluating the mass spectrum and retention index," *Analyst*, vol. 139, no. 10, pp. 2507–2514, 2014.
- [25] J.-M. Dimandja, "Introduction and historical background: The "inside" story of comprehensive two-dimensional gas chromatography," in *Separation Science and Technology*, vol. 12, Elsevier, 2020, pp. 1–40.
- [26] J. H. Winnike, X. Wei, K. J. Knagge, S. D. Colman, S. G. Gregory, and X. Zhang, "Comparison of gc-ms and GC×GC-MS in the analysis of human serum samples for biomarker discovery," *Journal of proteome research*, vol. 14, no. 4, pp. 1810–1817, 2015.

- [27] A. Mostafa and T. Gorecki, "Sensitivity of comprehensive two-dimensional gas chromatography (gcxgc) versus one-dimensional gas chromatography (1d gc)," *LC GC Europe*, vol. 26, no. 12, pp. 672–79, 2013.
- [28] X. Wei, X. Shi, I. Koo, S. Kim, R. H. Schmidt, G. E. Arteel, W. H. Watson, C. McClain, and X. Zhang, "Metpp: A computational platform for comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics," *Bioinformatics*, vol. 29, no. 14, pp. 1786–1792, 2013.
- [29] S. Wold, M. Sjostrom, and L. Eriksson, "Pls-regression a basic tool of chemometrics," *Chemometrics and intelligent laboratory systems*, vol. 58, no. 2, pp. 109–130, 2001.
- [30] S. Hou and P. Wentzell, "Fast and simple methods for the optimization of kurtosis used as a projection pursuit index," *Analytica chimica acta*, vol. 704, no. 1-2, pp. 1–15, 2011.
- [31] R. Tauler, "Multivariate curve resolution applied to second order data," *Chemometrics and intelligent laboratory systems*, vol. 30, no. 1, pp. 133–146, 1995.
- [32] H. Wold, "Path models with latent variables the nipals approach," in *Quantitative sociology*, Elsevier, 1975, pp. 307–357.
- [33] M. Otto, *Chemometrics: statistics and computer application in analytical chemistry*. John Wiley & Sons, 2016.
- [34] M. Andreut, "Parallel gpu implementation of iterative pca algorithms," *Journal of Computational Biology*, vol. 16, no. 11, pp. 1593–1599, 2009.
- [35] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on computers*, vol. 100, no. 9, pp. 881–890, 1974.
- [36] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.
- [37] R. A. Harshman *et al.*, "Foundations of the parafac procedure models and conditions for an explanatory multimodal factor analysis," 1970.
- [38] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.
- [39] R. Bro, "Multi-way analysis in the food industry-models, algorithms, and applications," in *MRI, EPG and EMA*, "Proc ICSLP 2000, Citeseer, 1998.
- [40] R. Harshman, "Parafac2 extensions of a procedure for "explanatory" factor-analysis and multidimensional scaling," *The Journal of the Acoustical Society of America*, vol. 51, no. 1A, pp. 111–111, 1972.
- [41] H. A. Kiers, "An alternating least squares algorithm for parafac2 and three-way dedicom," *Computational Statistics & Data Analysis*, vol. 16, no. 1, pp. 103–118, 1993.

- [42] H. A. Kiers, J. M. Ten Berge, and R. Bro, "Parafac2—part i. a direct fitting algorithm for the parafac2 model," *Journal of Chemometrics A Journal of the Chemometrics Society*, vol. 13, no. 3-4, pp. 275–294, 1999.
- [43] R. Bro, C. A. Andersson, and H. A. Kiers, "Parafac2—part ii. modeling chromatographic data with retention time shifts," *Journal of Chemometrics A Journal of the Chemometrics Society*, vol. 13, no. 3-4, pp. 295–309, 1999.
- [44] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
- [45] J. E. Cohen, *Non-negative parafac2 a flexible coupling approach*, <https://bit.ly/3znViSq>, 2018.
- [46] J. E. Cohen and R. Bro, "Nonnegative parafac2: A flexible coupling approach," in *International Conference on Latent Variable Analysis and Signal Separation*, Springer, 2018, pp. 89–98.
- [47] S. De Jong, "Simpls: An alternative approach to partial least squares regression," *Chemometrics and intelligent laboratory systems*, vol. 18, no. 3, pp. 251–263, 1993.
- [48] R. Bro and S. De Jong, "A fast non-negativity-constrained least squares algorithm," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 11, no. 5, pp. 393–401, 1997.
- [49] M. H. Van Benthem and M. R. Keenan, "Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems," *Journal of Chemometrics A Journal of the Chemometrics Society*, vol. 18, no. 10, pp. 441–450, 2004.
- [50] R. Bro and N. D. Sidiropoulos, "Least squares algorithms under unimodality and non-negativity constraints," *Journal of Chemometrics*, vol. 12, no. 4, pp. 223–247, 1998.
- [51] T. Mehmood, K. H. Liland, L. Snipen, and S. Saebo, "A review of variable selection methods in partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012.
- [52] T. Mehmood, S. Saebo, and K. H. Liland, "Comparison of variable selection methods in partial least squares regression," *Journal of Chemometrics*, 2020.
- [53] R. E. Mohler, K. M. Dombek, J. C. Hoggard, K. M. Pierce, E. T. Young, and R. E. Synovec, "Comprehensive analysis of yeast metabolite GC×GC–TOFMS data combining discovery-mode and deconvolution chemometric software," *Analyst*, vol. 132, no. 8, pp. 756–767, 2007.
- [54] K. M. Pierce, J. L. Hope, K. J. Johnson, B. W. Wright, and R. E. Synovec, "Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis," *Journal of Chromatography A*, vol. 1096, no. 1-2, pp. 101–110, 2005.

- [55] M. Farr'es, S. Platikanov, S. Tsakovski, and R. Tauler, "Comparison of the variable importance in projection (vip) and of the selectivity ratio (sr) methods for variable selection and interpretation," *Journal of Chemometrics*, vol. 29, no. 10, pp. 528–536, 2015.
- [56] A. Rinnan, M. Andersson, C. Ridder, and S. B. Engelsen, "Recursive weighted partial least squares (rpls) an efficient variable selection method using pls," English, *Journal of Chemometrics*, vol. 28, no. 5, pp. 439–447, May 2014. DOI: 10.1002cem.2582.
- [57] H. Sereshti, S. Ataolahi, G. Aliakbarzadeh, S. Zarre, and Z. Poursorkh, "Evaluation of storage time effect on saffron chemical profile using gas chromatography and spectrophotometry techniques coupled with chemometrics," *Journal of food science and technology*, vol. 55, no. 4, pp. 1350–1359, 2018.
- [58] E. Correa and R. Goodacre, "A genetic algorithm-bayesian network approach for the analysis of metabolomics and spectroscopic data application to the rapid identification of bacillus spores and classification of bacillus species," *BMC bioinformatics*, vol. 12, no. 1, p. 33, 2011.
- [59] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134, pp. 93–101, 2019.
- [60] K. H. Liland and U. G. Indahl, "Powered partial least squares discriminant analysis," *Journal of Chemometrics A Journal of the Chemometrics Society*, vol. 23, no. 1, pp. 7–18, 2009.
- [61] S. Maldonado and J. Lopez, "Dealing with high-dimensional class-imbalanced datasets embedded feature selection for svm classification," *Applied Soft Computing*, vol. 67, pp. 94–105, 2018.
- [62] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, and J. A. Suykens, "A tutorial on support vector machine-based methods for classification problems in chemometrics," *Analytica Chimica Acta*, vol. 665, no. 2, pp. 129–145, 2010.
- [63] K. Wongravee, N. Heinrich, M. Holmboe, M. L. Schaefer, R. R. Reed, J. Trevejo, and R. G. Brereton, "Variable selection using iterative reformulation of training set models for discrimination of samples application to gas chromatographymass spectrometry of mouse urinary metabolites," *Analytical chemistry*, vol. 81, no. 13, pp. 5204–5217, 2009.
- [64] J. M. Cadenas, M. C. Garrido, and R. Martinez, "Feature subset selection filter–wrapper based on low quality data," *Expert systems with applications*, vol. 40, no. 16, pp. 6241–6252, 2013.
- [65] B. M. Lukasiak, S. Zomer, R. G. Brereton, R. Faria, and J. C. Duncan, "Pattern recognition and feature selection for the discrimination between grades of commercial plastics," *Chemometrics and intelligent laboratory systems*, vol. 87, no. 1, pp. 18–25, 2007.

- [66] L. Adutwum, A. de la Mata, H. Bean, J. Hill, and J. Harynuk, “Estimation of start and stop numbers for cluster resolution feature selection algorithm an empirical approach using null distribution analysis of fisher ratios,” English, *Analytical and Bioanalytical Chemistry*, vol. 409, no. 28, pp. 6699–6708, Nov. 2017. DOI: 10.1007/s00216-017-0628-8. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/28963623>.
- [67] G. E. Box, “Non-normality and tests on variances,” *Biometrika*, vol. 40, no. 34, pp. 318–335, 1953.
- [68] T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik, and O. M. Kvalheim, “Biomarker discovery in mass spectral profiles by means of selectivity ratio plot,” *Chemometrics and Intelligent Laboratory Systems*, vol. 95, no. 1, pp. 35–48, 2009.
- [69] N. A. Sinkov and J. J. Harynuk, “Three-dimensional cluster resolution for guiding automatic chemometric model optimization,” *Talanta*, vol. 103, pp. 252–259, 2013, Cited By 3. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84878424748&partnerID=40&md5=60b3ef68e7cca29aeeb42e826df9df95>.
- [70] L. A. Adutwum, “Data reduction and feature selection for chemometric analysis,” English, PhD thesis, University of Alberta, 2017. DOI: 10.7939/R3X92213T. [Online]. Available: <https://search.datacite.org/works/10.7939/R3X92213T>.
- [71] N. A. Sinkov, B. M. Johnston, P. M. L. Sandercock, and J. J. Harynuk, “Automated optimization and construction of chemometric models based on highly variable raw chromatographic data,” *Analytica Chimica Acta*, vol. 697, no. 1-2, pp. 8–15, 2011. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-79957969441&partnerID=40&md5=06b07e8e95750c2eaeda6a6f2abca9f7>.
- [72] R. J. Abel and J. J. Harynuk, *abelrobin/LoadPEG: First public release of LoadPEG.m*, version 1.9, Sep. 2020. DOI: 10.5281/zenodo.4035154. [Online]. Available: <https://doi.org/10.5281/zenodo.4035154>.
- [73] B. A. Weggler, L. M. Dubois, N. Gawlitta, T. Groger, J. Moncur, L. Mondello, S. Reichenbach, P. Tranchida, Z. Zhao, R. Zimmermann, *et al.*, “A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software,” *Journal of Chromatography A*, vol. 1635, p. 461 721, 2021.
- [74] A Takagi, E Fujimura, and S Suehiro, “A new method of statokinesigram area measurement application of a statistically calculated ellipse,” in *Vestibular and visual control on posture and locomotor equilibrium*, Karger Publishers, 1985, pp. 74–79.
- [75] MathWorks, *Chi-square cumulative distribution function*, 2018. [Online]. Available: <https://www.mathworks.com/help/stats/chi2cdf.html>.

- [76] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder–mead simplex method in low dimensions," English, *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, Jan. 1998. DOI: 10.1137/S1052623496303470. [Online]. Available: <https://search.proquest.comdocview920036749>.
- [77] R. H. McQueen, J. J. Harynuk, W. V. Wismer, M. Keelan, Y. Xu, and A. P. de la Mata, "Axillary odour build-up in knit fabrics following multiple use cycles," *International Journal of Clothing Science and Technology*, 2014.
- [78] A. P. de la Mata, R. H. McQueen, S. L. Nam, and J. J. Harynuk, "Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics," *Analytical and bioanalytical chemistry*, vol. 409, no. 7, pp. 1905–1913, 2017.
- [79] N. F. Perez, J. Ferre, and R. Boque, "Calculation of the reliability of classification in discriminant partial least-squares binary classification," *Chemometrics and Intelligent Laboratory Systems*, vol. 95, no. 2, pp. 122–128, 2009.
- [80] V. Spruyt, *How to draw a covariance error ellipse*, 2014. [Online]. Available: <https://www.visiondummy.com/2014/04/draw-error-ellipse-representing-covariance-matrix>.
- [81] J. C. Lansey, *Beautiful and distinguishable line colors + colormap*, 2015. [Online]. Available: <https://www.mathworks.com/matlabcentral/fileexchange/42673-beautiful-and-distinguishable-line-colors-colormap>.
- [82] V. Martinez-Cagigal, *Shaded area error bar plot*, 2015. [Online]. Available: <https://www.mathworks.com/matlabcentral/fileexchange/58262-shaded-area-error-bar-plot>.
- [83] H. Tang, N. Abunasser, M. Garcia, M. Chen, K. S. Ng, and S. O. Salley, "Potential of microalgae oil from *dunaliella tertiolecta* as a feedstock for biodiesel," *Applied Energy*, vol. 88, no. 10, pp. 3324–3330, 2011.
- [84] J. K. Bwapwa, A. Anandraj, and C. Trois, "Possibilities for conversion of microalgae oil into aviation fuel a review," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 1345–1354, 2017.
- [85] L. Zhu, "Microalgal culture strategies for biofuel production a review," *Biofuels, Bioproducts and Biorefining*, vol. 9, no. 6, pp. 801–814, 2015.
- [86] J.-L. Mouget, A. Dakhama, M. C. Lavoie, and J. de la Noue, "Algal growth enhancement by bacteria is consumption of photosynthetic oxygen involved," *FEMS Microbiology Ecology*, vol. 18, no. 1, pp. 35–43, 1995.
- [87] R. Ramanan, B.-H. Kim, D.-H. Cho, H.-M. Oh, and H.-S. Kim, "Algae–bacteria interactions evolution, ecology and emerging applications," *Biotechnology advances*, vol. 34, no. 1, pp. 14–29, 2016.

- [88] R. Willamme, Z. Alsafra, R. Arumugam, G. Eppe, F. Remacle, R. D. Levine, and C. Remacle, "Metabolomic analysis of the green microalga *Chlamydomonas reinhardtii* cultivated under daynight conditions," *Journal of biotechnology*, vol. 215, pp. 20–26, 2015.
- [89] M. Sun, Z. Yang, and B. Wawrik, "Metabolomic fingerprints of individual algal cells using the single-probe mass spectrometry technique," *Frontiers in plant science*, vol. 9, p. 571, 2018.
- [90] M. T. Croft, A. D. Lawrence, E. Raux-Deery, M. J. Warren, and A. G. Smith, "Algae acquire vitamin b₁₂ through a symbiotic relationship with bacteria," *Nature*, vol. 438, no. 7064, pp. 90–93, 2005.
- [91] S. Amin, L. Hmelo, H. Van Tol, B. Durham, L. Carlson, K. Heal, R. Morales, C. Berthiaume, M. Parker, B Djunaedi, *et al.*, "Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria," *Nature*, vol. 522, no. 7554, pp. 98–101, 2015.
- [92] M. R. Seyedsayamdost, R. J. Case, R. Kolter, and J. Clardy, "The jekyll-and-hyde chemistry of *Phaeobacter gallaeciensis*," *Nature chemistry*, vol. 3, no. 4, pp. 331–335, 2011.
- [93] R. J. Case, S. R. Longford, A. H. Campbell, A. Low, N. Tujula, P. D. Steinberg, and S. Kjelleberg, "Temperature induced bacterial virulence and bleaching disease in a chemically defended marine macroalga," *Environmental Microbiology*, vol. 13, no. 2, pp. 529–537, 2011.
- [94] J. S. Wirth and W. B. Whitman, "Phylogenomic analyses of a clade within the roseobacter group suggest taxonomic reassignments of species of the genera *aestuariivita*, *citricella*, *loktanella*, *nautella*, *pelagibaca*, *ruegeria*, *thalassobius*, *thiobacimonas* and *tropicibacter*, and the proposal of six novel genera," *International journal of systematic and evolutionary microbiology*, vol. 68, no. 7, pp. 2393–2411, 2018.
- [95] A. R. Bramucci, L. Labeeuw, T. J. Mayers, J. A. Saby, and R. J. Case, "A small volume bioassay to assess bacterial-phytoplankton co-culture using water-pulse-amplitude-modulated (water-pam) fluorometry," *Journal of visualized experiments JoVE*, no. 97, 2015.
- [96] L. R. Cavonius, N.-G. Carlsson, and I. Undeland, "Quantification of total fatty acids in microalgae: comparison of extraction and transesterification methods," *Analytical and bioanalytical chemistry*, vol. 406, no. 28, pp. 7313–7322, 2014.
- [97] E. G. Bligh and W. J. Dyer, "A rapid method of total lipid extraction and purification," *Canadian journal of biochemistry and physiology*, vol. 37, no. 8, pp. 911–917, 1959.
- [98] E. C. Y. Chan, K. K. Pasikanti, and J. K. Nicholson, "Global urinary metabolic profiling procedures using gas chromatography–mass spectrometry," *Nature protocols*, vol. 6, no. 10, pp. 1483–1499, 2011.

- [99] S. L. Nam, A. Mata, R. P. Dias, and J. J. Harynuk, "Towards standardization of data normalization strategies to improve urinary metabolomics studies by GC \times GC-TOFMS," *Metabolites*, vol. 10, no. 9, p. 376, 2020.
- [100] M. S. Armstrong, A. P. de la Mata, and J. J. Harynuk, "An efficient and accurate numerical determination of the cluster resolution metric in two dimensions," *Journal of Chemometrics*, e3346,
- [101] J. Hummel, N. Strehmel, J. Selbig, D. Walther, and J. Kopka, "Decision tree supported substructure prediction of metabolites from gc-ms profiles," *Metabolomics*, vol. 6, no. 2, pp. 322–333, 2010.
- [102] P. Draper and D. MacLean, "Mass spectra of alkylquinolines," *Canadian Journal of Chemistry*, vol. 46, no. 9, pp. 1487–1497, 1968.
- [103] C. Wagner, M. Sefkow, and J. Kopka, "Construction and application of a mass spectral and retention time index database generated from plant gcei-tof-ms metabolite profiles," *Phytochemistry*, vol. 62, no. 6, pp. 887–900, 2003.
- [104] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, *et al.*, "Proposed minimum reporting standards for chemical analysis," *Metabolomics*, vol. 3, no. 3, pp. 211–221, 2007.
- [105] R. J. Case, M. Labbate, and S. Kjelleberg, "Ahl-driven quorum-sensing circuits their frequency and function among the proteobacteria," *The ISME journal*, vol. 2, no. 4, pp. 345–349, 2008.
- [106] A. L. Schaefer, E. Greenberg, C. M. Oliver, Y. Oda, J. J. Huang, G. Bittan-Banin, C. M. Peres, S. Schmidt, K. Juhaszova, J. R. Sufrin, *et al.*, "A new class of homoserine lactone quorum-sensing signals," *Nature*, vol. 454, no. 7204, pp. 595–599, 2008.
- [107] M. Gardiner, N. D. Fernandes, D. Nowakowski, M. Raftery, S. Kjelleberg, L. Zhong, T. Thomas, and S. Egan, "Varr controls colonization and virulence in the marine macroalgal pathogen textitNautella italica R11," *Frontiers in microbiology*, vol. 6, p. 1130, 2015.
- [108] J. Hudson, M. Gardiner, N. Deshpande, and S. Egan, "Transcriptional response of textitNautella italica R11 towards its macroalgal host uncovers new mechanisms of host–pathogen interaction," *Molecular ecology*, vol. 27, no. 8, pp. 1820–1832, 2018.
- [109] A. M. Wolf, E. T. Fontham, T. R. Church, C. R. Flowers, C. E. Guerra, S. J. LaMonte, R. Etzioni, M. T. McKenna, K. C. Oeffinger, Y.-C. T. Shih, *et al.*, "Colorectal cancer screening for average-risk adults 2018 guideline update from the american cancer society," *CA a cancer journal for clinicians*, vol. 68, no. 4, pp. 250–281, 2018.

- [110] L. A. Siminoff, H. L. Rogers, M. D. Thomson, L. Dumenci, and S. Harris-Haywood, "Doctor, what's wrong with me factors that delay the diagnosis of colorectal cancer," *Patient education and counseling*, vol. 84, no. 3, pp. 352–358, 2011.
- [111] D. J. Ahnen, S. W. Wade, W. F. Jones, R. Sifri, J. M. Silveiras, J. Greenamyre, S. Guiffre, J. Axilbund, A. Spiegel, and Y. N. You, "The increasing incidence of young-onset colorectal cancer a call to action," in *Mayo Clinic Proceedings*, Elsevier, vol. 89, 2014, pp. 216–224.
- [112] C. T. F. on Preventive Health Care *et al.*, "Recommendations on screening for colorectal cancer in primary care," *Cmaj*, vol. 188, no. 5, pp. 340–348, 2016.
- [113] J. F. Collins, D. A. Lieberman, T. E. Durbin, and D. G. Weiss, "Accuracy of screening for fecal occult blood on a single stool sample obtained by digital rectal examination a comparison with recommended sampling practice," *Annals of internal medicine*, vol. 142, no. 2, pp. 81–85, 2005.
- [114] C. K. Wong, R. N. Fedorak, C. I. Prosser, M. E. Stewart, S. V. van Zanten, and D. C. Sadowski, "The sensitivity and specificity of guaiac and immunochemical fecal occult blood tests for the detection of advanced colonic adenomas and cancer," *International journal of colorectal disease*, vol. 27, no. 12, pp. 1657–1664, 2012.
- [115] R. L. Siegel, S. A. Fedewa, W. F. Anderson, K. D. Miller, J. Ma, P. S. Rosenberg, and A. Jemal, "Colorectal cancer incidence patterns in the united states, 1974–2013," *JNCI Journal of the National Cancer Institute*, vol. 109, no. 8, 2017.
- [116] S. Bouatra, F. Aziat, R. Mandal, A. C. Guo, M. R. Wilson, C. Knox, T. C. Bjorndahl, R. Krishnamurthy, F. Saleem, P. Liu, *et al.*, "The human urine metabolome," *PloS one*, vol. 8, no. 9, e73076, 2013.
- [117] Y. Qiu, G. Cai, M. Su, T. Chen, Y. Liu, Y. Xu, Y. Ni, A. Zhao, S. Cai, L. X. Xu, *et al.*, "Urinary metabonomic study on colorectal cancer," *Journal of proteome research*, vol. 9, no. 3, pp. 1627–1634, 2010.
- [118] Y. Cheng, G. Xie, T. Chen, Y. Qiu, X. Zou, M. Zheng, B. Tan, B. Feng, T. Dong, P. He, *et al.*, "Distinct urinary metabolic profile of human colorectal cancer," *Journal of proteome research*, vol. 11, no. 2, pp. 1354–1363, 2012.
- [119] A. Hirayama, K. Kami, M. Sugimoto, M. Sugawara, N. Toki, H. Onozuka, T. Kinoshita, N. Saito, A. Ochiai, M. Tomita, *et al.*, "Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry," *Cancer research*, vol. 69, no. 11, pp. 4918–4925, 2009.
- [120] L. Deng, D. Chang, R. R. Foshaug, R. Eisner, V. K. Tso, D. S. Wishart, and R. N. Fedorak, "Development and validation of a high-throughput mass spectrometry based urine metabolomic test for the detection of colonic adenomatous polyps," *Metabolites*, vol. 7, no. 3, p. 32, 2017.

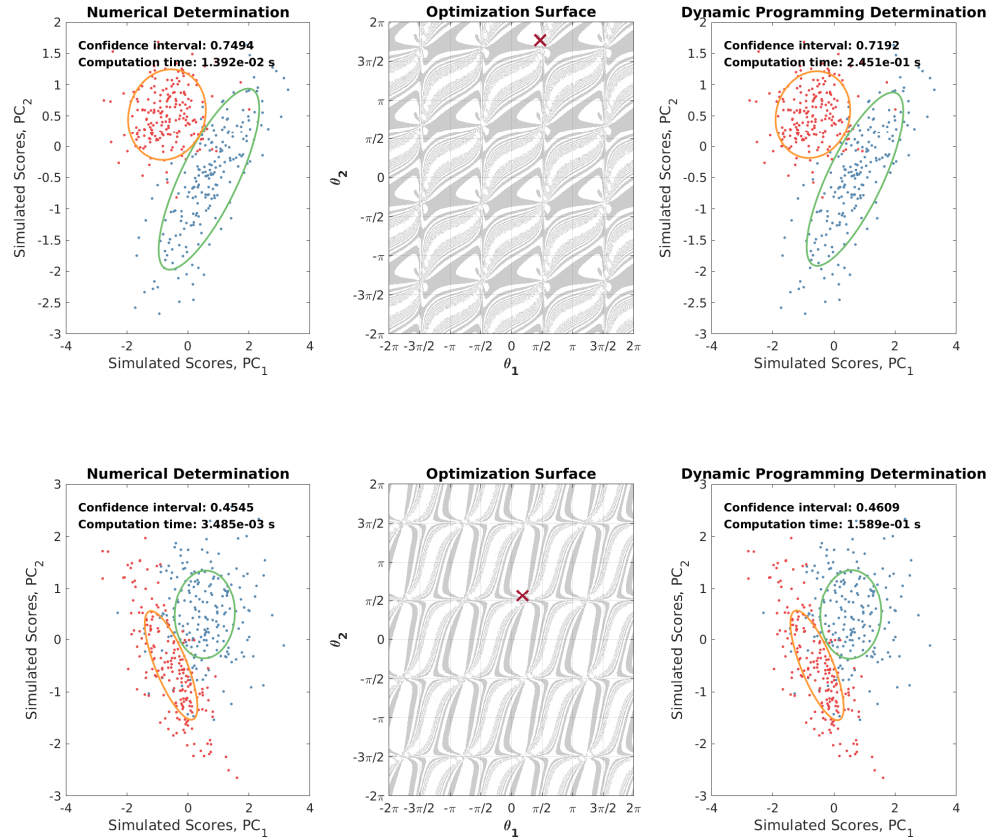
- [121] N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, *et al.*, “The human serum metabolome,” *PloS one*, vol. 6, no. 2, e16957, 2011.
- [122] C. Ibanez, C. Simo, M. Palazoglu, and A. Cifuentes, “Gc-ms based metabolomics of colon cancer cells using different extraction solvents,” *Analytica chimica acta*, vol. 986, pp. 48–56, 2017.
- [123] K. A. Perrault, K. D. Nizio, and S. L. Forbes, “A comparison of one-dimensional and comprehensive two-dimensional gas chromatography for decomposition odour profiling using inter-year replicate field trials,” *Chromatographia*, vol. 78, no. 15, pp. 1057–1070, 2015.
- [124] Z. Yu, H. Huang, A. Reim, P. D. Charles, A. Northage, D. Jackson, I. Parry, and B. M. Kessler, “Optimizing 2d gas chromatography mass spectrometry for robust tissue, serum and urine metabolite profiling,” *Talanta*, vol. 165, pp. 685–691, 2017.
- [125] J. Harynuk and T. Gorecki, “New liquid nitrogen cryogenic modulator for comprehensive two-dimensional gas chromatography,” *Journal of Chromatography A*, vol. 1019, no. 1-2, pp. 53–63, 2003.
- [126] L. Deng, K. Ismond, Z. Liu, J. Constable, H. Wang, O. I. Alatis, M. R. Weiser, T. P. Kingham, and D. Chang, “Urinary metabolomics to identify a unique biomarker panel for detecting colorectal cancer a multicenter study,” *Cancer Epidemiology and Prevention Biomarkers*, vol. 28, no. 8, pp. 1283–1291, 2019.
- [127] R. Eisner, R. Greiner, V. Tso, H. Wang, and R. N. Fedorak, “A machine-learned predictor of colonic polyps based on urinary metabolomics,” *BioMed research international*, vol. 2013, 2013.
- [128] N. A. Sinkov and J. J. Harynuk, “Cluster resolution a metric for automated, objective and optimized feature selection in chemometric modeling,” *Talanta*, vol. 83, no. 4, pp. 1079–1087, 2011. [Online]. Available: <http://www.scopus.com/inwardrecord.url?eid=2-s2.0-79251594688&partnerID=40&md5=dd9e381524bd757d0a1b>
- [129] M. R. Meyer and H. H. Maurer, “Current status of hyphenated mass spectrometry in studies of the metabolism of drugs of abuse, including doping agents,” *Analytical and bioanalytical chemistry*, vol. 402, no. 1, pp. 195–208, 2012.
- [130] H. Parastar, J. R. Radovic, J. M. Bayona, and R. Tauler, “Solving chromatographic challenges in comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry using multivariate curve resolution–alternating least squares,” *Analytical and bioanalytical chemistry*, vol. 405, no. 19, pp. 6235–6249, 2013.
- [131] J. C. Hoggard and R. E. Synovec, “Parallel factor analysis (parafac) of target analytes in GC×GC- TOFMS data automated selection of a model with an appropriate number of factors,” *Analytical chemistry*, vol. 79, no. 4, pp. 1611–1619, 2007.

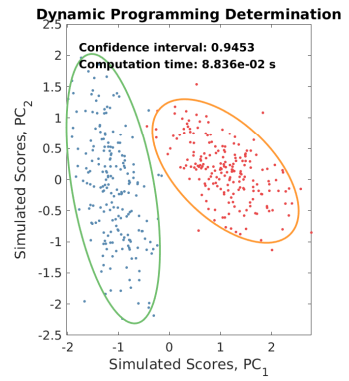
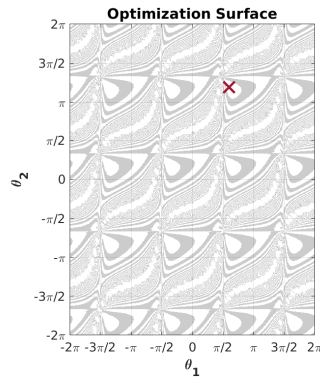
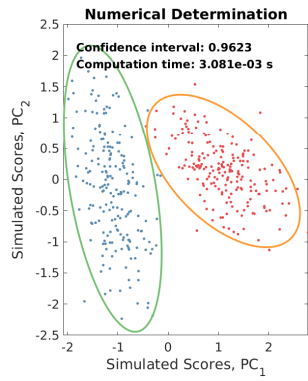
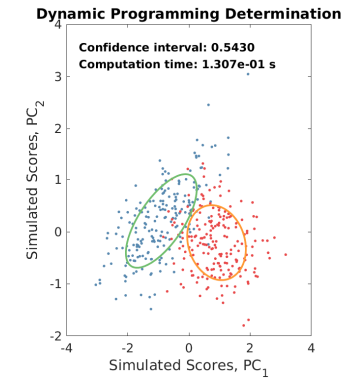
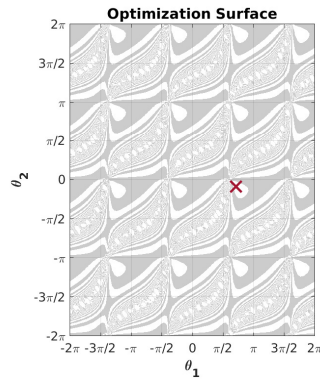
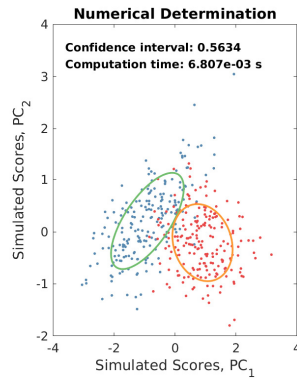
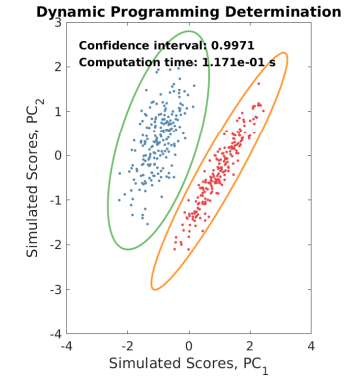
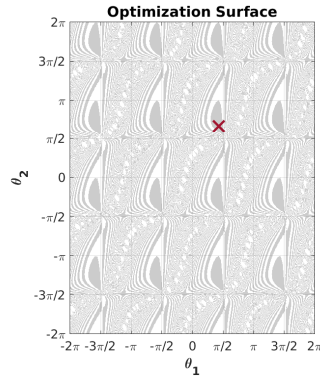
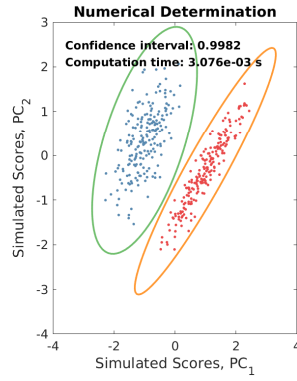
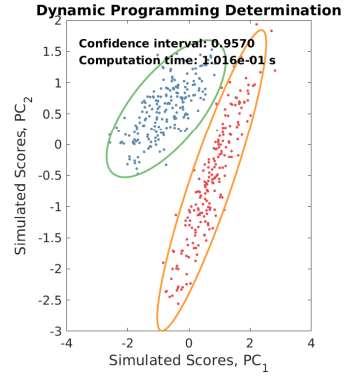
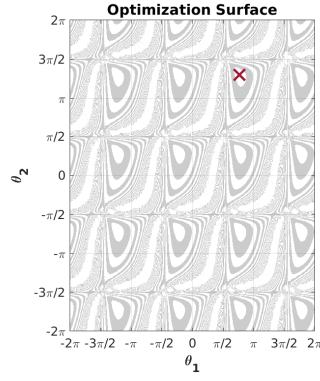
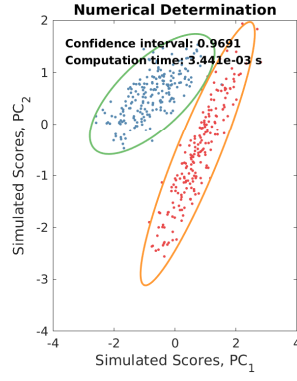
- [132] T. Skov, J. C. Hoggard, R. Bro, and R. E. Synovec, "Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling," *Journal of Chromatography A*, vol. 1216, no. 18, pp. 4020–4029, 2009.
- [133] B. J. Asher, L. A. D'Agostino, J. D. Way, C. S. Wong, and J. J. Harynuk, "Comparison of peak integration methods for the determination of enantiomeric fraction in environmental samples," *Chemosphere*, vol. 75, no. 8, pp. 1042–1048, 2009.
- [134] B. C, L. F, E. A, and K. J, *Mass spectrum of adipic acid, 2tms*, <http://gmd.mpimp-golm.mpg.deSpectrumsbaf089ca-eebf-4c05-ba7e-7aa412da17d3.aspx>, Accessed 2021-07-16, 2007.
- [135] L. F and E. A, *Mass spectrum of salicylic acid, 2tms*, <http://gmd.mpimp-golm.mpg.deSpectrumse68bea7d-defb-4d13-b64e-50c2148c9930.aspx>, Accessed 2021-07-16, 2005.
- [136] P. Venter, M. Muller, J. Vestner, M. A. Stander, A. G. Tredoux, H. Pasch, and A. de Villiers, "Comprehensive three-dimensional LC_{times} LC_{times} ion mobility spectrometry separation combined with high-resolution ms for the analysis of complex samples," *Analytical chemistry*, vol. 90, no. 19, pp. 11 643–11 650, 2018.
- [137] N. E. Watson, H. D. Bahaghighat, K. Cui, and R. E. Synovec, "Comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry," *Analytical chemistry*, vol. 89, no. 3, pp. 1793–1800, 2017.
- [138] X. Domingo-Almenara, A. Perera, N. Ramirez, and J. Brezmes, "Automated resolution of chromatographic signals by independent component analysis–orthogonal signal deconvolution in comprehensive gas chromatographymass spectrometry-based metabolomics," *Computer methods and programs in biomedicine*, vol. 130, pp. 135–141, 2016.
- [139] L. G. Johnsen, P. B. Skou, B. Khakimov, and R. Bro, "Gas chromatography–mass spectrometry data processing made easy," *Journal of Chromatography a*, vol. 1503, pp. 57–64, 2017.
- [140] H. Lu, Y. Liang, W. B. Dunn, H. Shen, and D. B. Kell, "Comparative evaluation of software for deconvolution of metabolomics data based on gc-tof-ms," *TrAC Trends in Analytical Chemistry*, vol. 27, no. 3, pp. 215–227, 2008.
- [141] L. Van Stee and U. T. Brinkman, "Peak detection methods for GC×GC an overview," *TrAC Trends in Analytical Chemistry*, vol. 83, pp. 1–13, 2016.
- [142] S. E. Reichenbach, X. Tian, C. Cordero, and Q. Tao, "Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography," *Journal of Chromatography A*, vol. 1226, pp. 140–148, 2012.
- [143] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components models," *Technometrics*, vol. 20, no. 4, pp. 397–405, 1978.

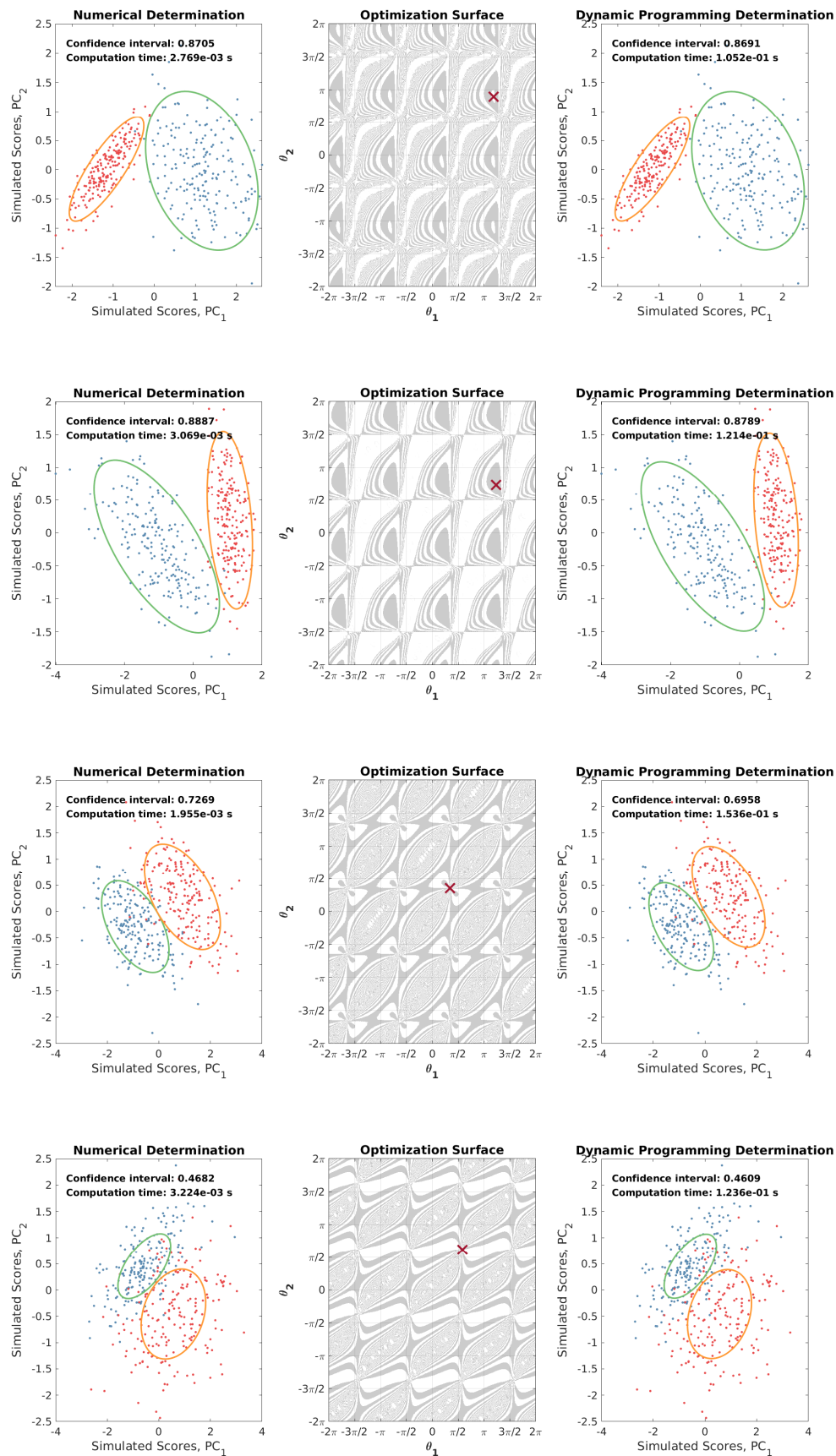
- [144] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J—Multidisciplinary Scientific Journal*, vol. 2, no. 2, pp. 226–235, 2019.
- [145] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," *Computational Statistics & Data Analysis*, vol. 51, no. 2, pp. 918–930, 2006.
- [146] R. L. Thorndike, "Who belongs in the family," *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.
- [147] S. Peters, H.-G. Janssen, and G. Vivo-Truyols, "A new method for the automated selection of the number of components for deconvolving overlapping chromatographic peaks," *Analytica chimica acta*, vol. 799, pp. 29–35, 2013.
- [148] R. Bro and H. A. Kiers, "A new efficient method for determining the number of components in parafac models," *Journal of Chemometrics A Journal of the Chemometrics Society*, vol. 17, no. 5, pp. 274–286, 2003.
- [149] M. H. Kamstrup-Nielsen, L. G. Johnsen, and R. Bro, "Core consistency diagnostic in parafac2," *Journal of Chemometrics*, vol. 27, no. 5, pp. 99–105, 2013.
- [150] A. B. Risum and R. Bro, "Using deep learning to evaluate peaks in chromatographic data," *Talanta*, vol. 204, pp. 255–260, 2019.
- [151] J. M. Davis and J. C. Giddings, "Statistical method for estimation of number of components from single complex chromatograms theory, computer-based testing, and analysis of errors," *Analytical chemistry*, vol. 57, no. 12, pp. 2168–2177, 1985.
- [152] S. Coppi, A. Betti, and F. Dondi, "Analysis of complex mixtures by capillary gas chromatography with statistical estimation of the number of components," *Analytica chimica acta*, vol. 212, pp. 165–170, 1988.
- [153] H. Motegi, Y. Tsuboi, A. Saga, T. Kagami, M. Inoue, H. Toki, O. Minowa, T. Noda, and J. Kikuchi, "Identification of reliable components in multivariate curve resolution-alternating least squares (mcr-als) a data-driven approach across metabolic processes," *Scientific reports*, vol. 5, p. 15710, 2015.
- [154] Y. Choi, J. Taylor, R. Tibshirani, *et al.*, "Selecting the number of principal components estimation of the true rank of a noisy matrix," *The Annals of Statistics*, vol. 45, no. 6, pp. 2590–2617, 2017.
- [155] P. J. Rousseeuw, "Silhouettes a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [156] S. Dhakshinamoorthy, N.-T. Dinh, J. Skolnick, and M. P. Styczynski, "Metabolomics identifies the intersection of phosphoethanolamine with menaquinone-triggered apoptosis in an in vitro model of leukemia," *Molecular bioSystems*, vol. 11, no. 9, pp. 2406–2416, 2015.

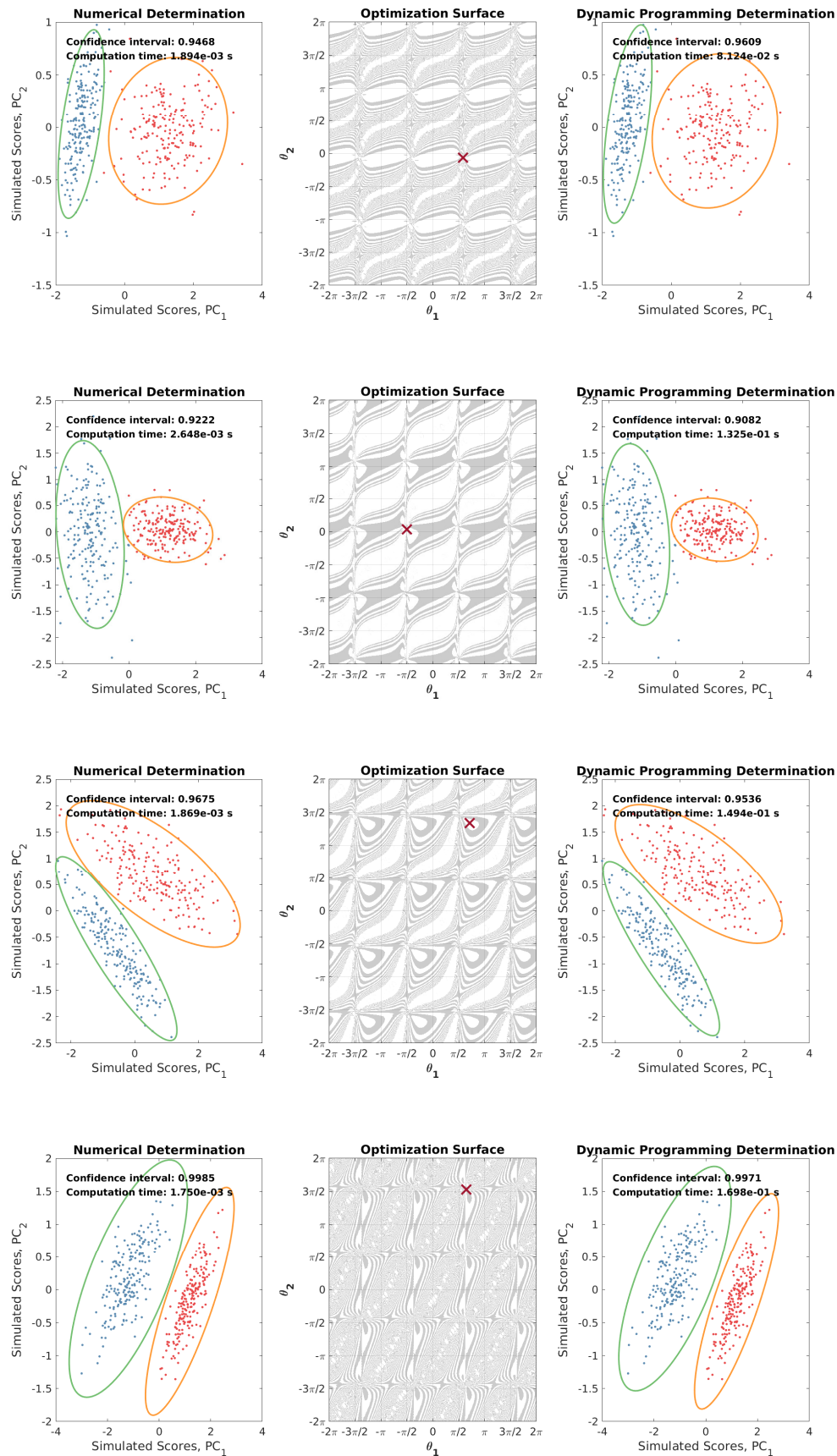
- [157] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

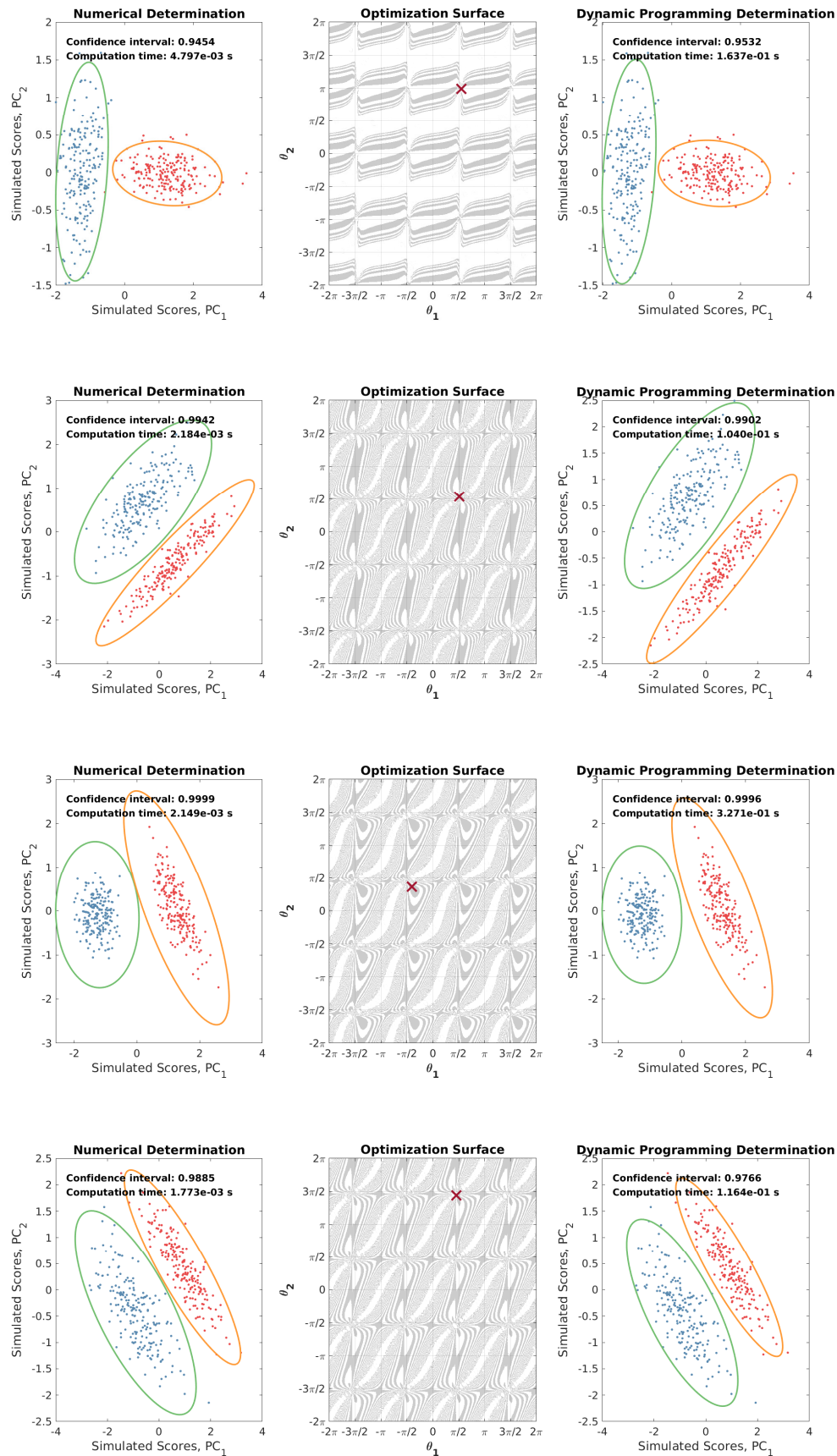
Appendix A: Sample Calculations for “An Efficient and Accurate Numerical Determination of the Cluster Resolution Metric in 2 Dimensions”

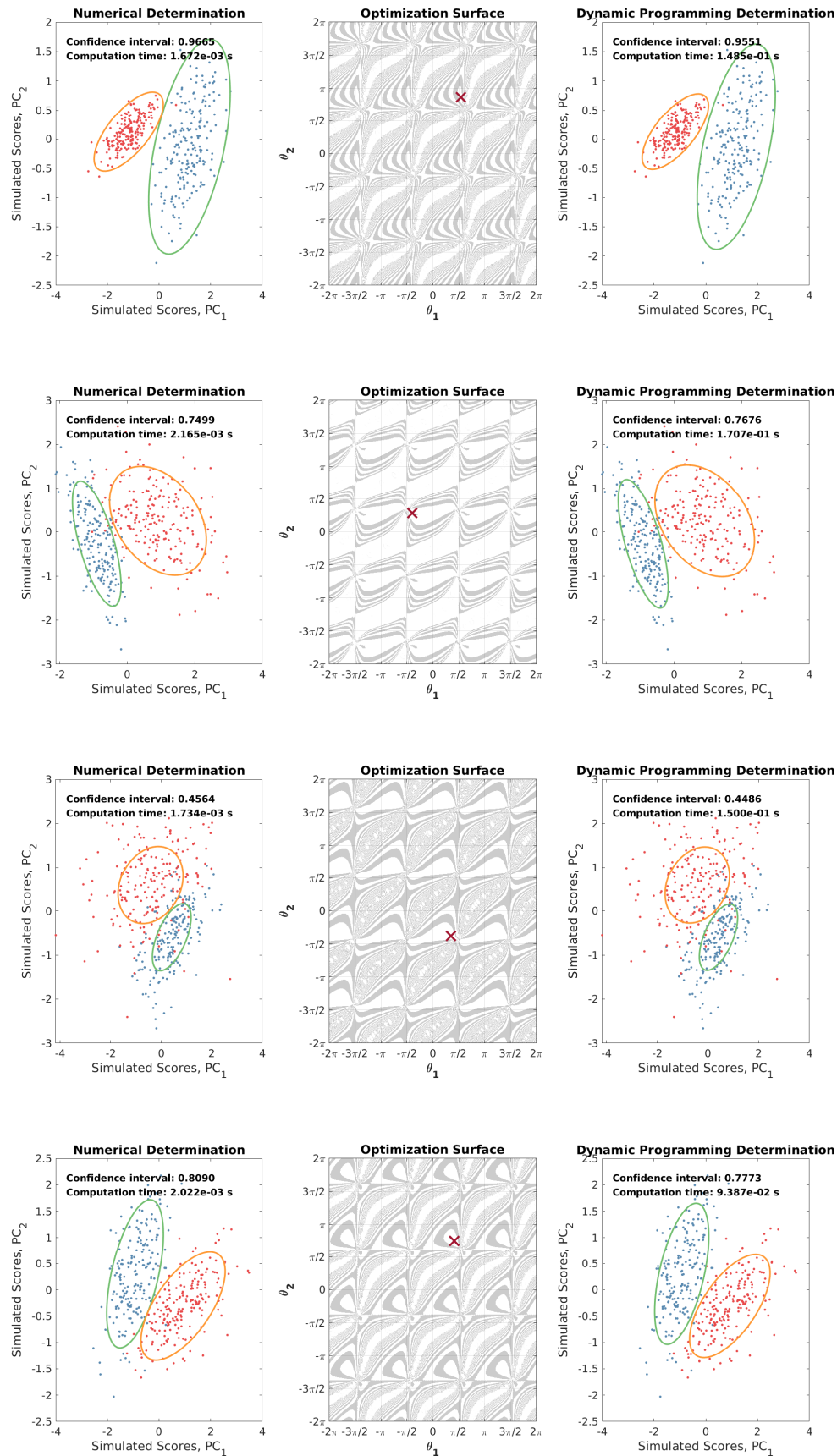


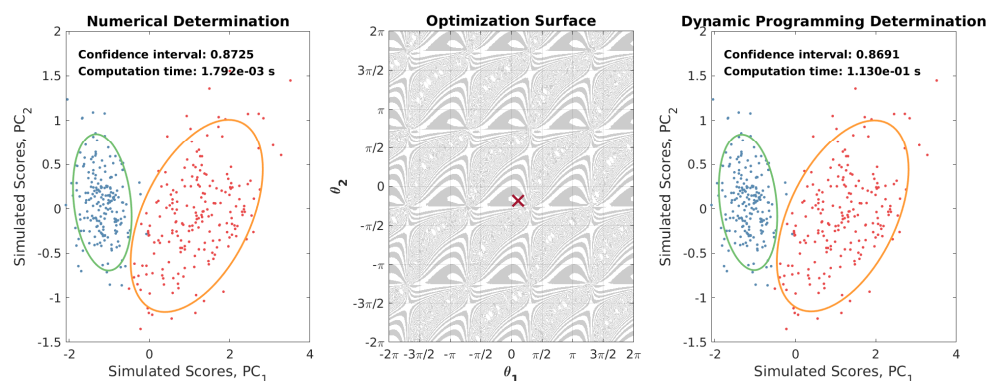
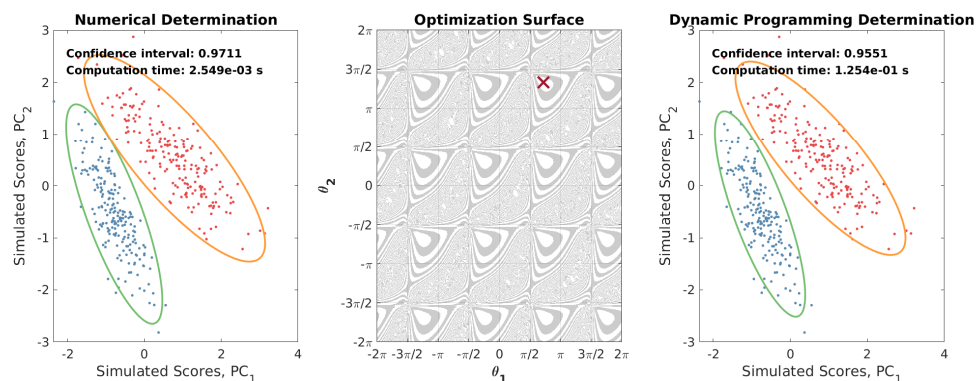
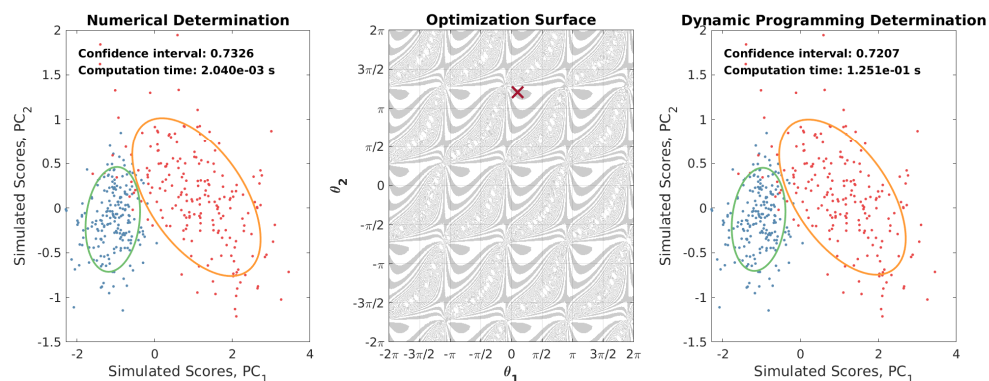
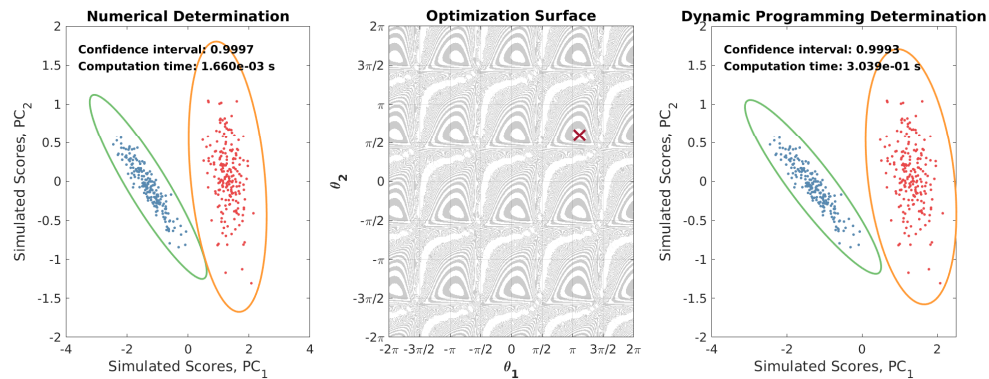


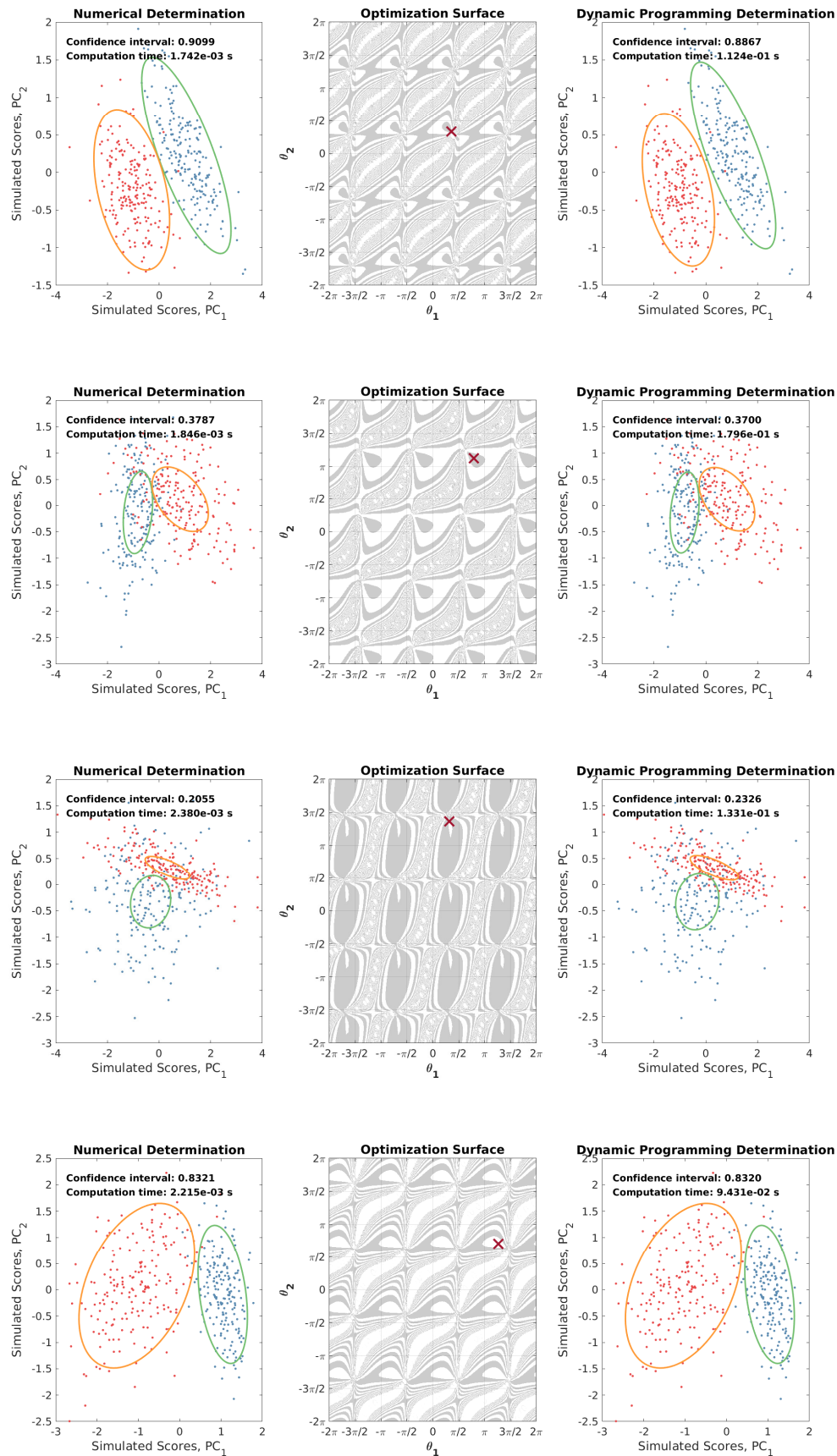


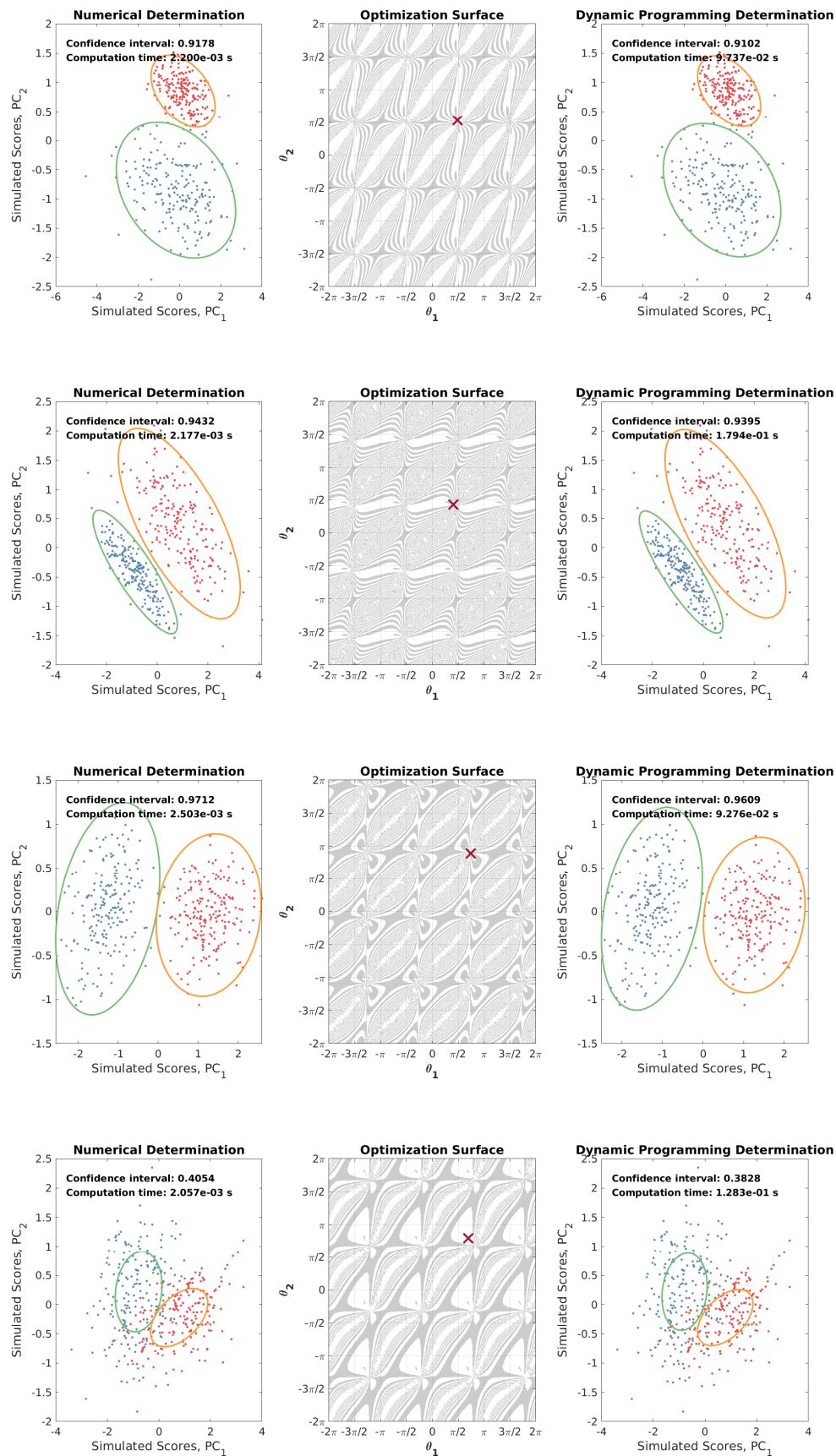


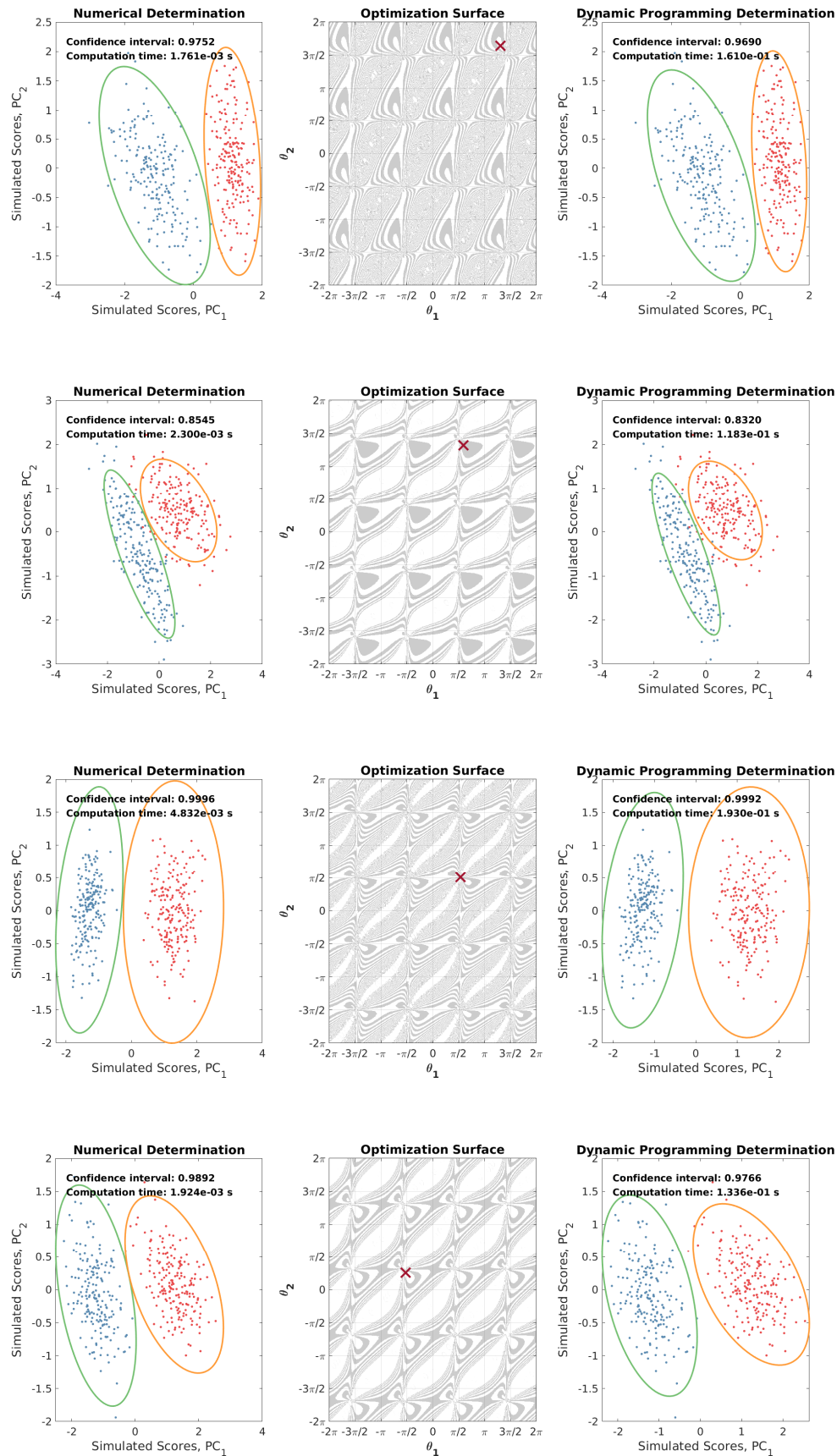


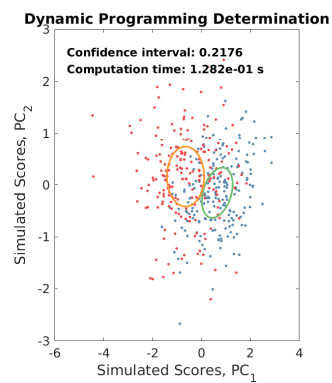
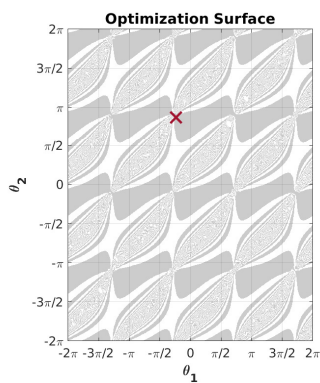
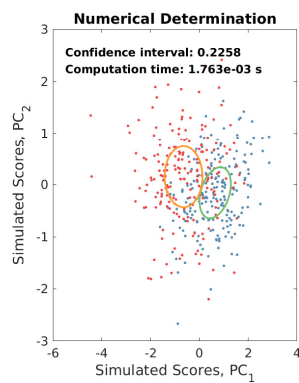
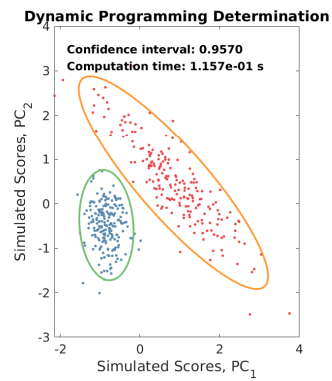
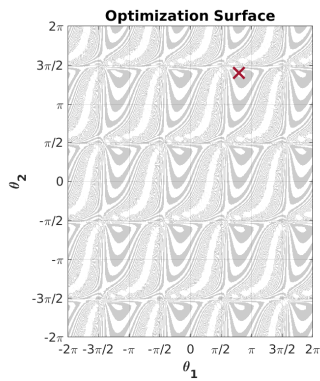
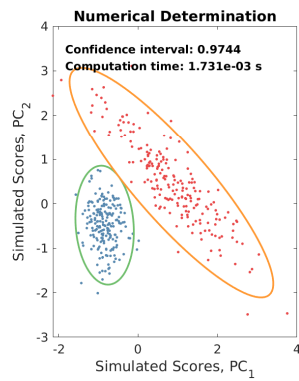












Appendix B: Supporting Information
for “Global Metabolome Analysis of
Dunaliella tertiolecta, *Rueger
iaitalica* Co-cultures using Thermal
Desorption - Comprehensive
2-Dimensional Gas Chromatography -
Time-of-Flight Mass Spectrometry
(TD-GC×GC-TOFMS)”

B.1 Example Instrument Blank, Reagent Blank

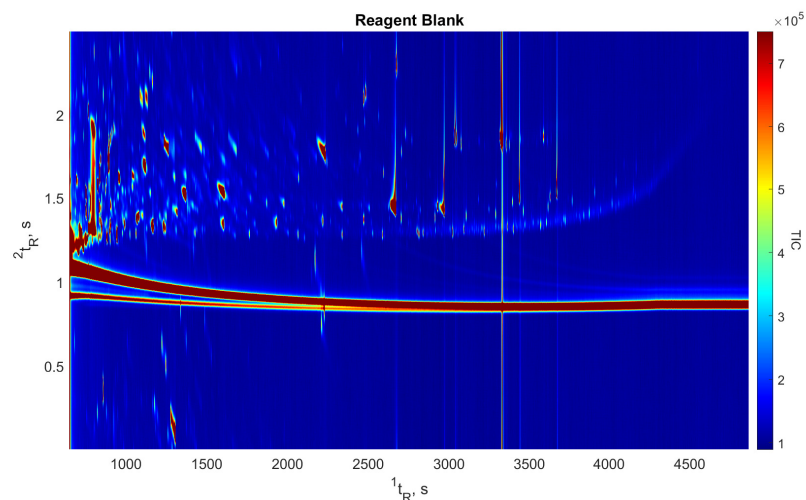


Figure B.1: Example reagent blank, demonstrating the low purity of derivatisation reagents. Subsequent analysis will show that despite the high number of interfering chemical components, the extracted features were generally not found in the reagent blanks.

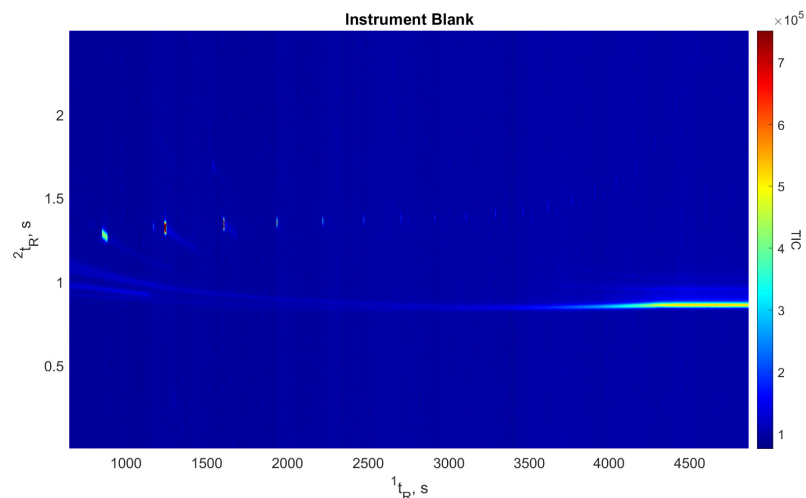


Figure B.2: Example instrument blank, demonstrating that the instrument was largely free of contamination between runs, despite a heavy sample load. The observed peaks are cyclic siloxanes, a byproduct of column degradation. These peaks were not identified as significant in the analysis of the data.

B.2 Overview of Extracted Features

Below are the summary analyses of the features determined to be most significant according to the FS-CR routine. For each peak, the most relevant details about the features are presented. The chromatograms display the peak using the quantitative ion identified by ChromaTOF[®], with the nominal retention times indicated by the white cross. A representative sample for each class is displayed, including a reagent blank. A plot comparing the mass spectra, in addition to a box plot summarising the significance of each feature per class is also presented.

Linear alkanes between $n = 12$ and $n = 32$ were run at the start of the analysis, and retention indices were calculated for all analytes that eluted within this window - otherwise the retention index is listed as undefined.

Poor library matches within the Golm Metabolome Database do not reflect the quality of the selected metabolites. The library may not contain an appropriate standard, or the library mass spectra may have been collected on a different type of mass spectrometer. Definitive identification of these analytes remains an open research avenue for future work.

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 967.5	
$^2t_R(s)$: 2.070	
Quantification Ion (m/z): 143	1-dotprod (Spectral Dissimilarity): 0.1923
Analyte Name (ChromaTOF): Analyte5083	Analyte Name (Library): Alanine, beta- (1TMS)
Retention Index (Observed): Undefined	Retention Index Difference (Library): Undefined

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/784e1232-c517-423f-98f9-ae3eb5351dac.aspx>

Contributor: Jäger C, Schomburg D Department of Bioinformatics and Biochemistry, Technische Universität Carolo-Wilhelmina Braunschweig, Langer Kamp 19B, D-38106 Braunschweig, Germany

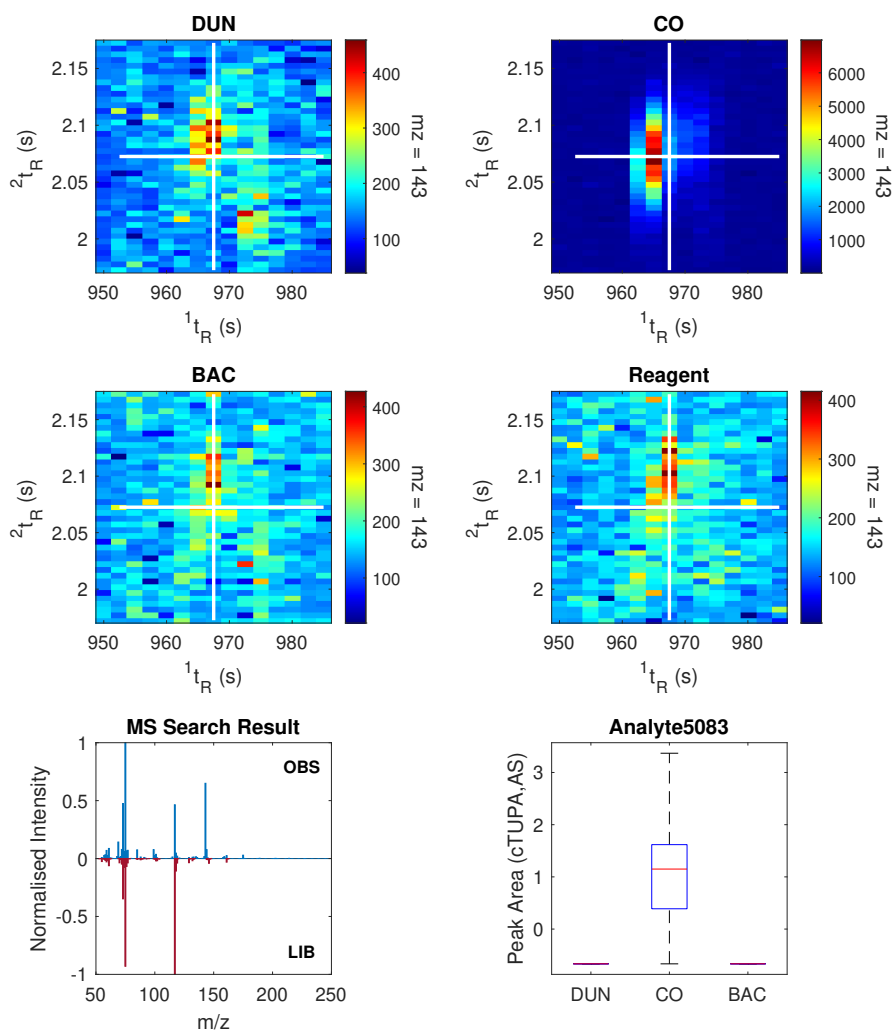


Figure B.3: Summary analyses, Analyte 5083

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 1212.5	
$^2t_R(s)$: 1.695	
Quantification Ion (m/z): 203	1-dotprod (Spectral Dissimilarity): Unknown
Analyte Name (ChromaTOF): Analyte8351	Analyte Name (Library): Unknown
Retention Index (Observed): 1284.99	Retention Index Difference (Library): Unknown

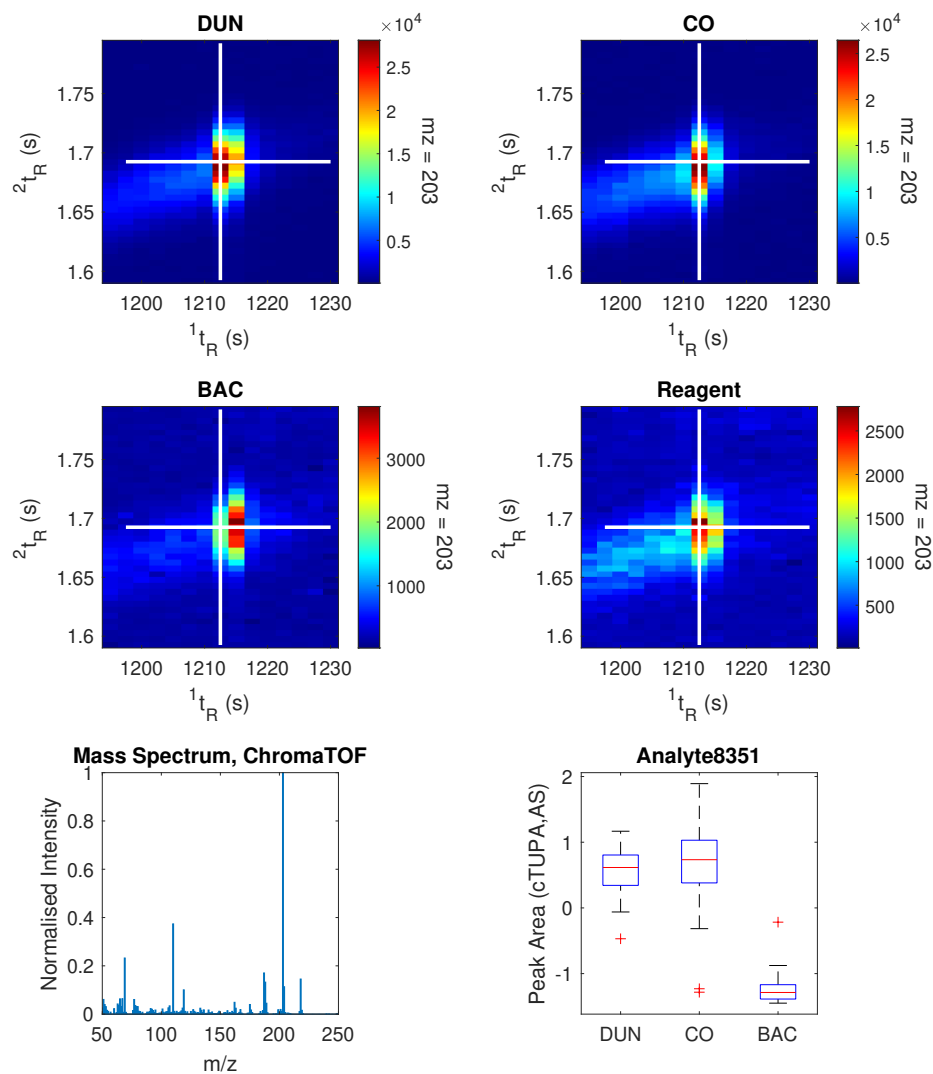


Figure B.4: Summary analyses, Analyte 8351

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 1642.5	
$^2t_R(s)$: 1.905	
Quantification Ion (m/z): 137	1-dotprod (Spectral Dissimilarity): Unknown
Analyte Name (ChromaTOF): Analyte13706	Analyte Name (Library): Unknown
Retention Index (Observed): 1480.80	Retention Index Difference (Library): Unknown

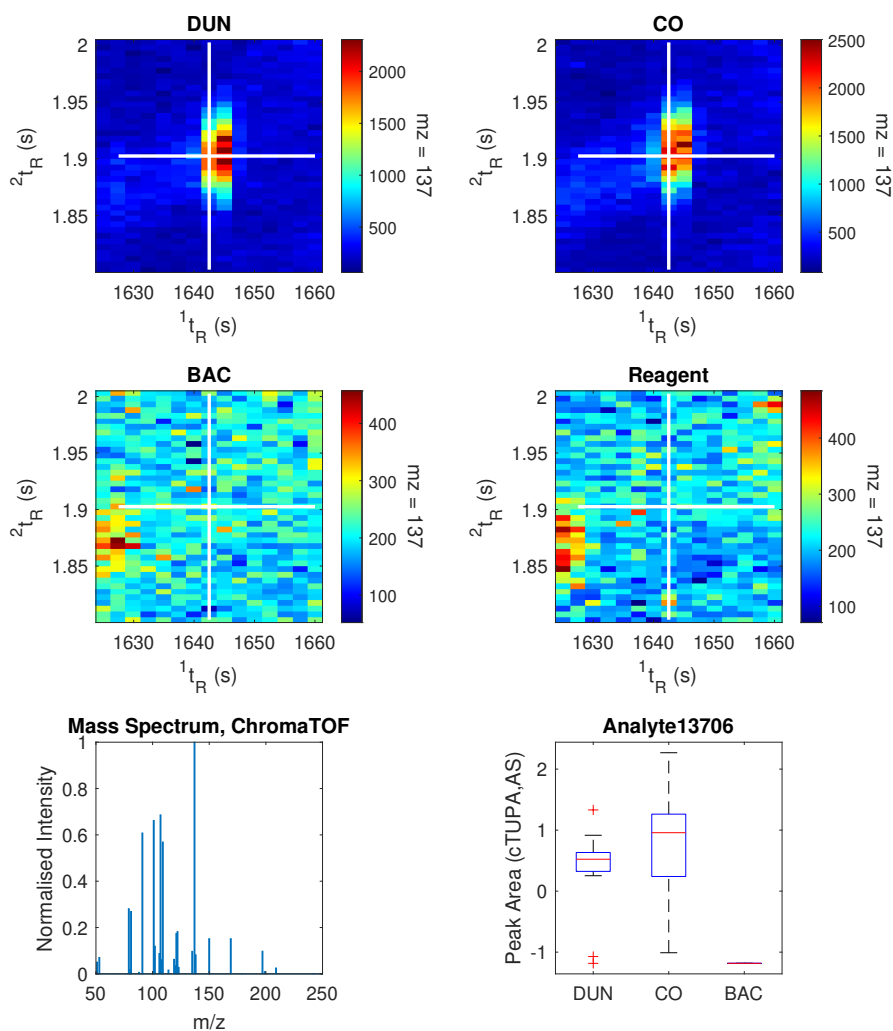


Figure B.5: Summary analyses, Analyte 13706

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 1677.5	
$^2t_R(s)$: 1.270	
Quantification Ion (m/z): 57	1-dotprod (Spectral Dissimilarity): 0.0846
Analyte Name (ChromaTOF): Hexadecane	Analyte Name (Library): Pentadecane, n-
Retention Index (Observed): 1497.08	Retention Index Difference (Library): 2.92

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/e6650dda-783d-483f-bb3d-1d5c083da4ec.aspx>

Contributor: Jäger C, Schomburg D Department of Bioinformatics and Biochemistry, Technische Universität Carolo-Wilhelmina Braunschweig, Langer Kamp 19B, D-38106 Braunschweig, Germany

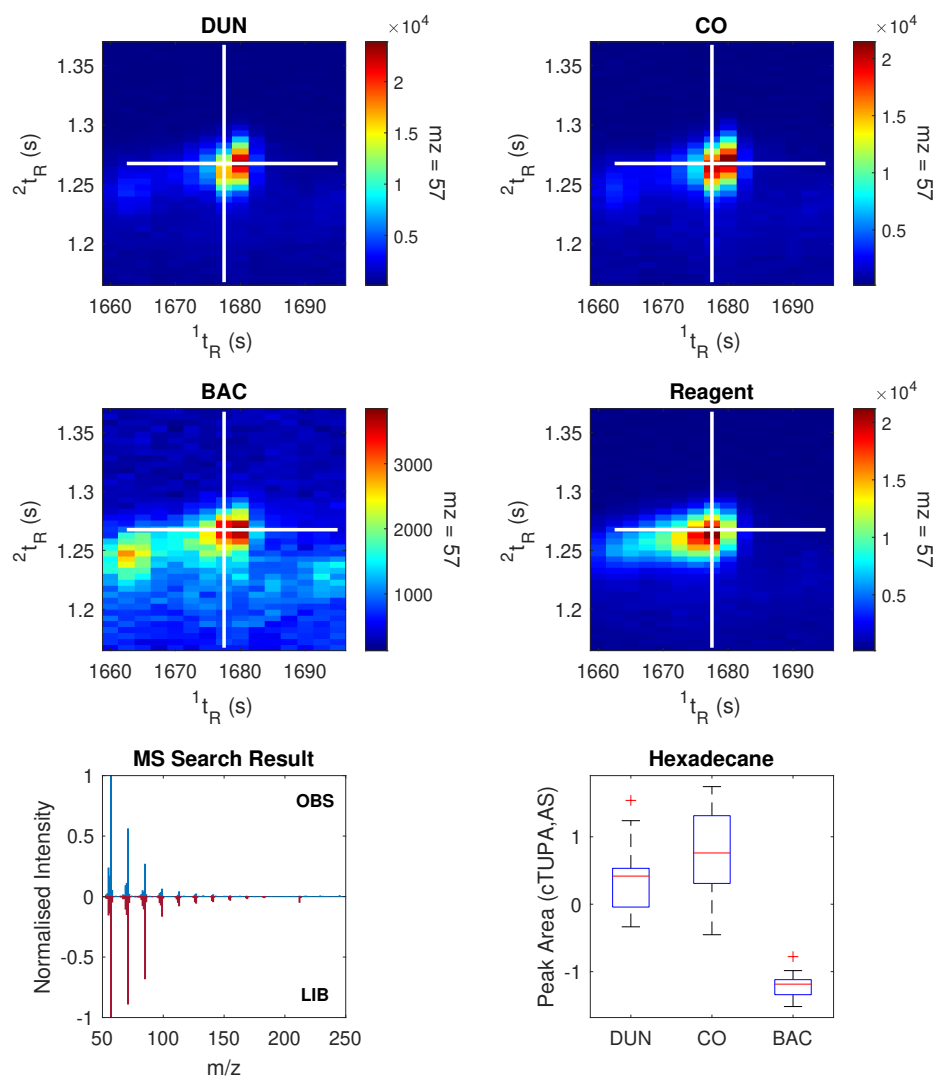


Figure B.6: Summary analyses, analyte: "Hexadecane"

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 1932.5	
$^2t_R(s)$: 1.480	
Quantification Ion (m/z): 117	1-dotprod (Spectral Dissimilarity): 0.2868
Analyte Name (ChromaTOF): Analyte16951	Analyte Name (Library): 2-Aminoadipic-acid (2TMS)
Retention Index (Observed): 1624.02	Retention Index Difference (Library): 3.79

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/bec90e9c-eafd-4bec-9ad6-d2cde2ecdbbe.aspx>

Contributor: Jäger C, Schomburg D Department of Bioinformatics and Biochemistry, Technische Universität Carolo-Wilhelmina Braunschweig, Langer Kamp 19B, D-38106 Braunschweig, Germany

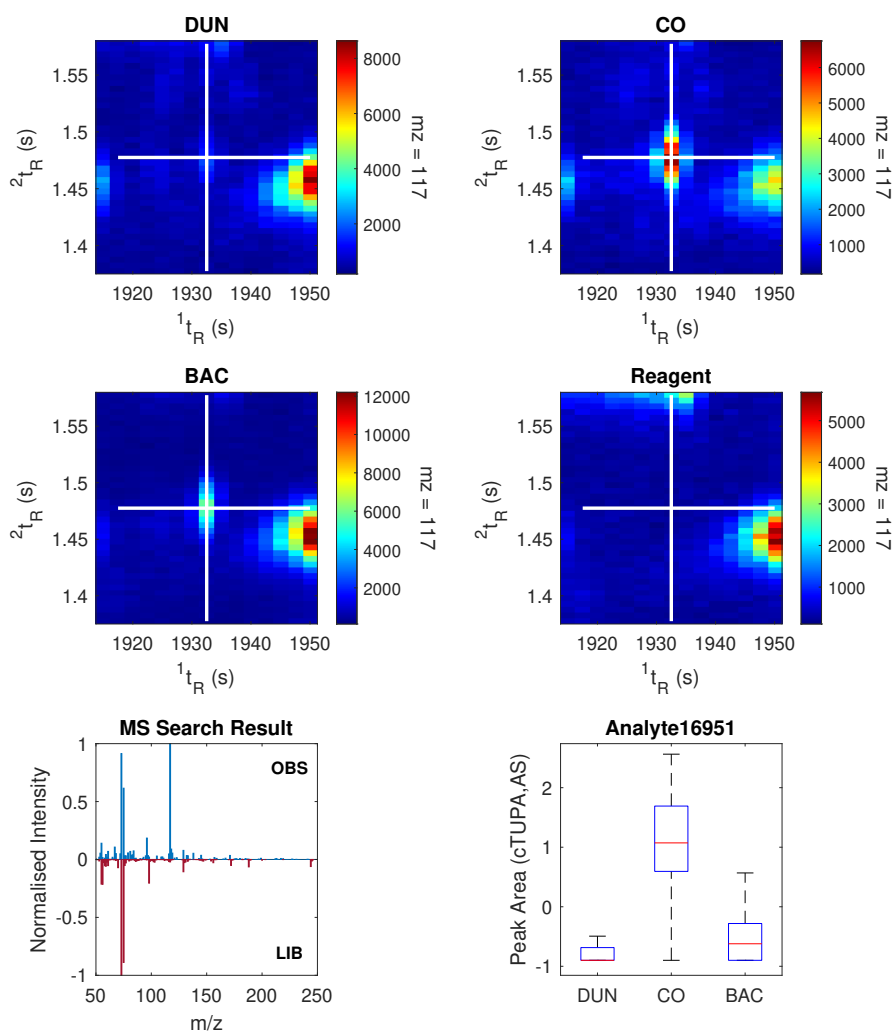


Figure B.7: Summary analyses, Analyte 16951

Peak Information	Top Golm Metabolome Database Search Result
1t_R (s): 2025.0	
2t_R (s): 1.990	
Quantification Ion (m/z): 228	1-dotprod (Spectral Dissimilarity): 0.2952
Analyte Name (ChromaTOF): Analyte17909	Analyte Name (Library): Oxaloacetate (1MEOX) (3TMS) MP
Retention Index (Observed): 1672.07	Retention Index Difference (Library): 0.64

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/c0839970-b4b6-4089-9639-50f2aa8fd7d4.aspx>

Contributor: Boelling C, Liebig F, Erban A, Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

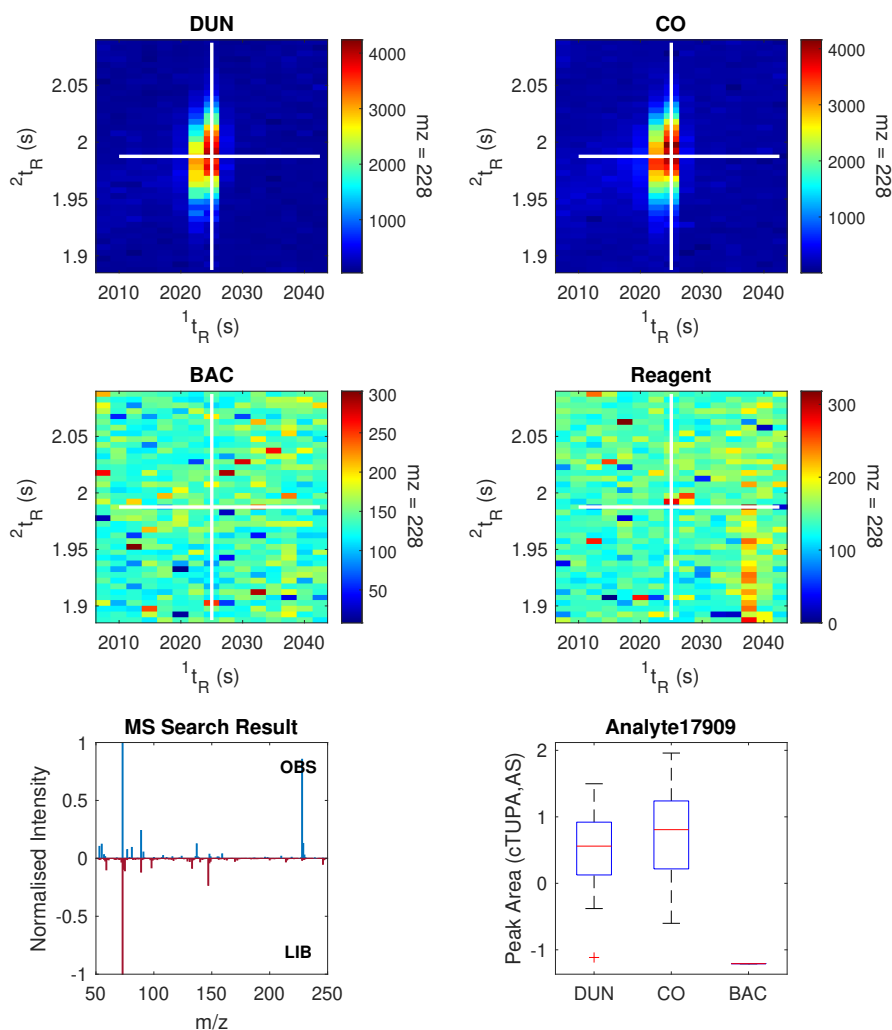


Figure B.8: Summary analyses, Analyte 17909

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 2180.0	
$^2t_R(s)$: 1.600	
Quantification Ion (m/z): 69	1-dotprod (Spectral Dissimilarity): Unknown
Analyte Name (ChromaTOF): Analyte19404	Analyte Name (Library): Unknown
Retention Index (Observed): 1754.72	Retention Index Difference (Library): Unknown

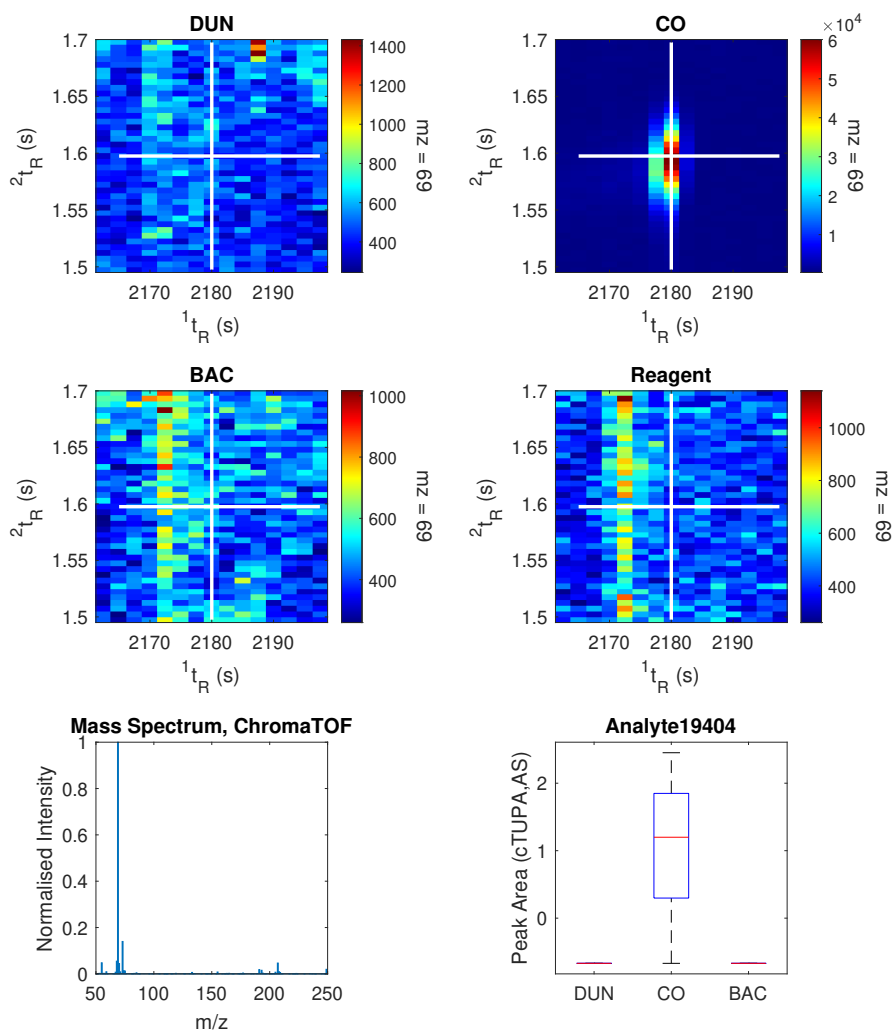


Figure B.9: Summary analyses, Analyte 19404

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 2370.0	
$^2t_R(s)$: 1.365	
Quantification Ion (m/z): 195	1-dotprod (Spectral Dissimilarity): 0.1310
Analyte Name (ChromaTOF): Analyte21239	Analyte Name (Library): Homoserine lactone, N-2-oxocaproyl- (1MEOX) (1TMS) BP
Retention Index (Observed): 1860.70	Retention Index Difference (Library): 0.51

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/7d96c6e2-42fc-46c8-85c2-0e11137c7210.aspx>
Contributor: Boelling C, Liebig F, Erban A, Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

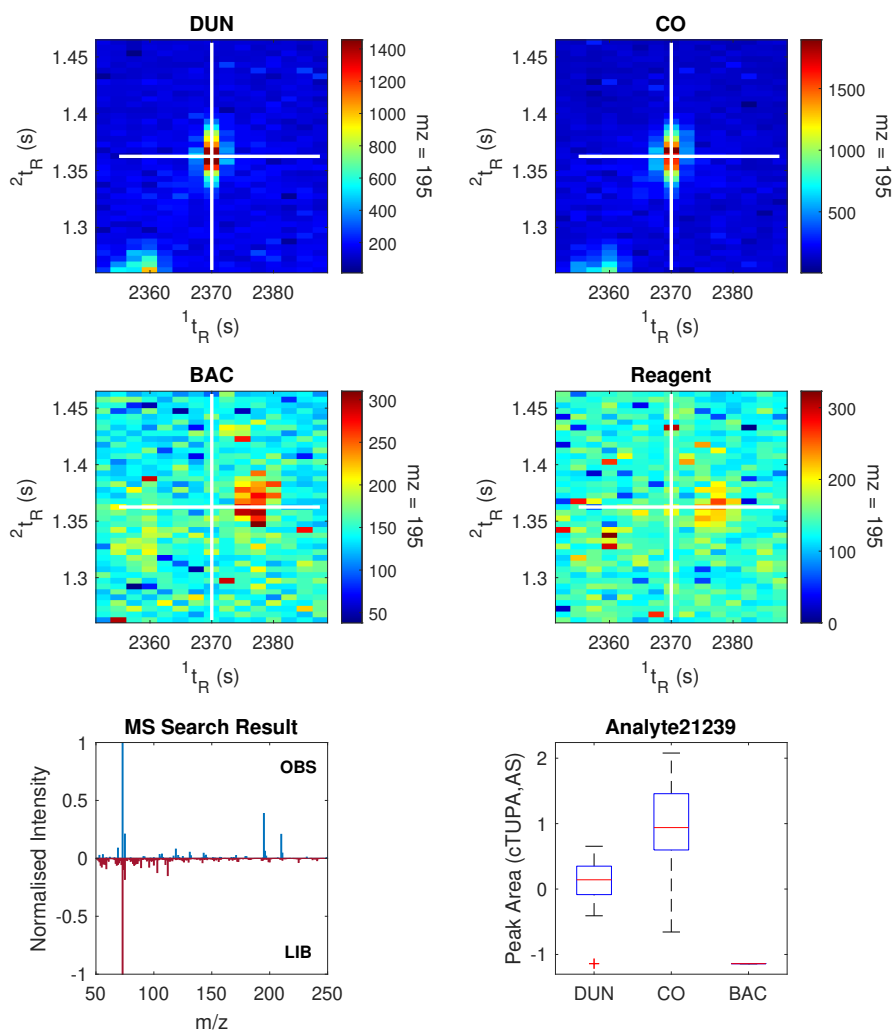


Figure B.10: Summary analyses, Analyte 21239

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 2380.0	
$^2t_R(s)$: 1.775	
Quantification Ion (m/z): 158	1-dotprod (Spectral Dissimilarity): Unknown
Analyte Name (ChromaTOF): Analyte21318	Analyte Name (Library): Unknown
Retention Index (Observed): 1866.42	Retention Index Difference (Library): Unknown

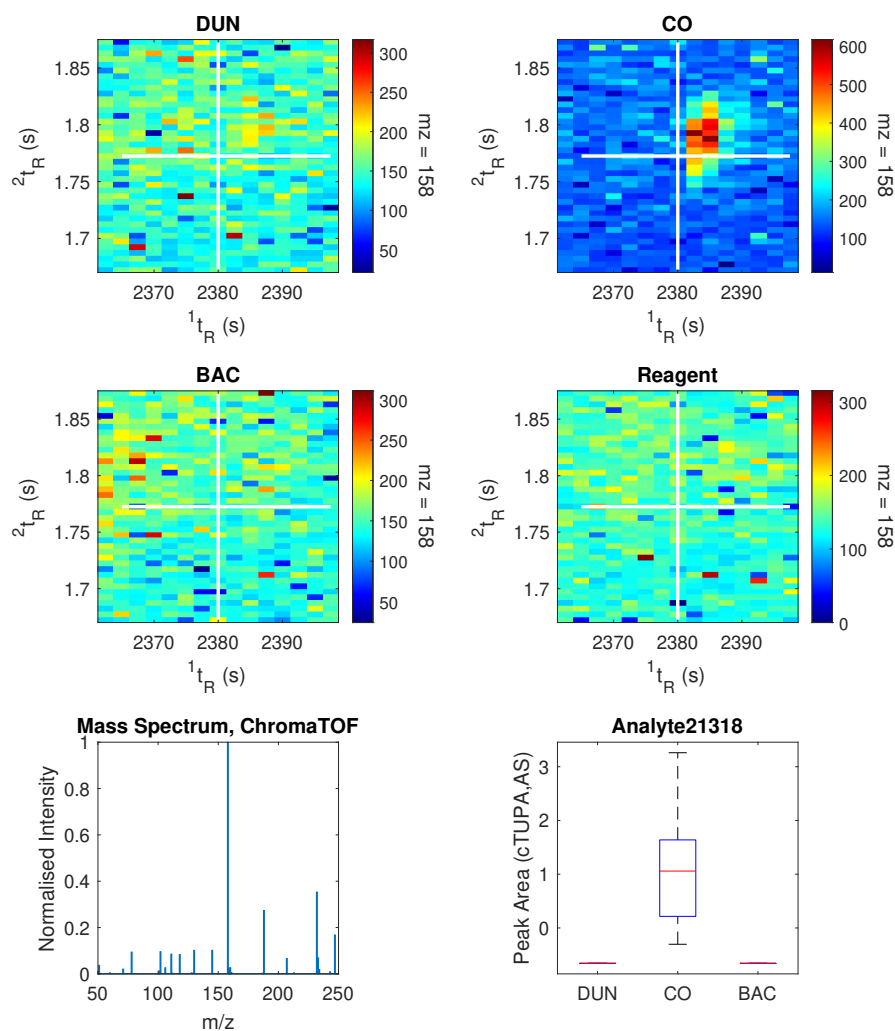


Figure B.11: Summary analyses, Analyte 21318

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 2497.5	
$^2t_R(s)$: 1.320	
Quantification Ion (m/z): 143	1-dotprod (Spectral Dissimilarity): 0.3436
Analyte Name (ChromaTOF): Analyte22226	Analyte Name (Library): Indole-3-acetic acid, 1H- (1TMS)
Retention Index (Observed): 1935.06	Retention Index Difference (Library): 0.35

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/35e55241-8b1a-42c3-a602-c1f05b31f818.aspx>

Contributor: Liebig F, Erban A, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

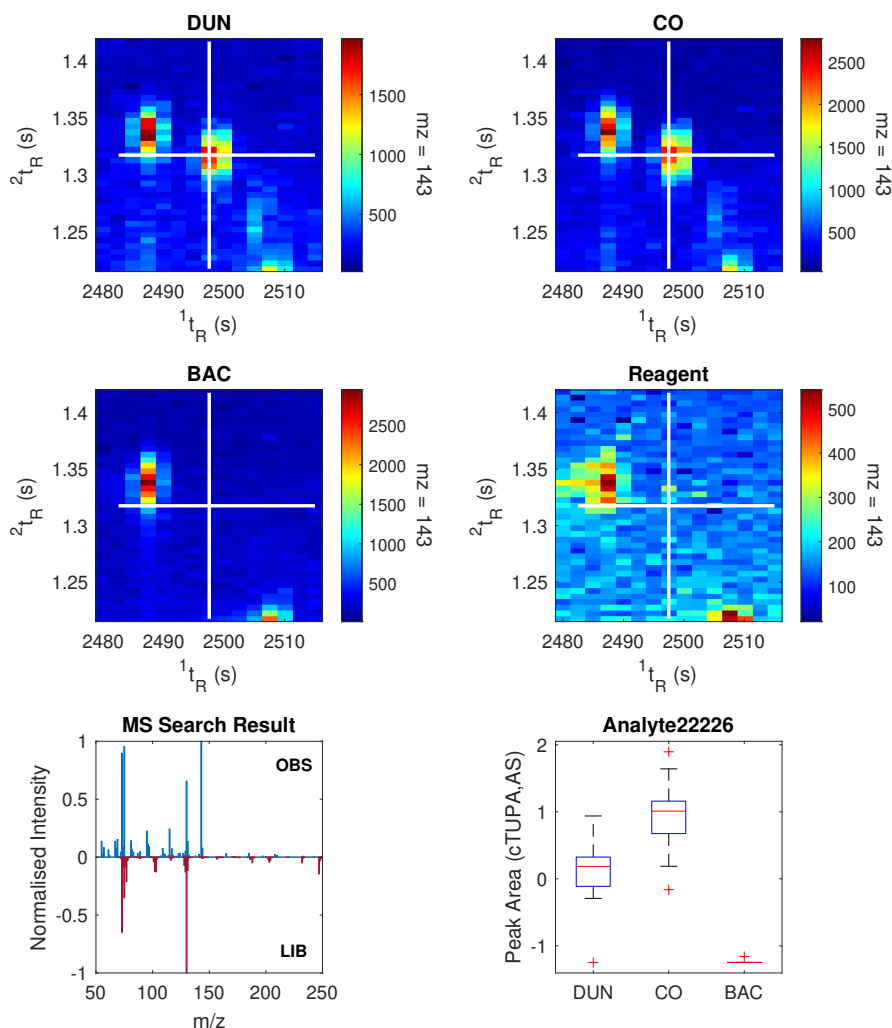


Figure B.12: Summary analyses, Analyte 22226

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 2597.5	
$^2t_R(s)$: 1.335	
Quantification Ion (m/z): 165	1-dotprod (Spectral Dissimilarity): 0.2324
Analyte Name (ChromaTOF): Analyte23041	Analyte Name (Library): Pyruvic acid, 4-hydroxyphenyl-(1MEOX) (3TMS) MP
Retention Index (Observed): 1994.76	Retention Index Difference (Library): 1.63

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/14c81dc4-1c9f-4731-a2cd-974fdd692d47.aspx>

Contributor: Boelling C, Liebig F, Erban A, Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

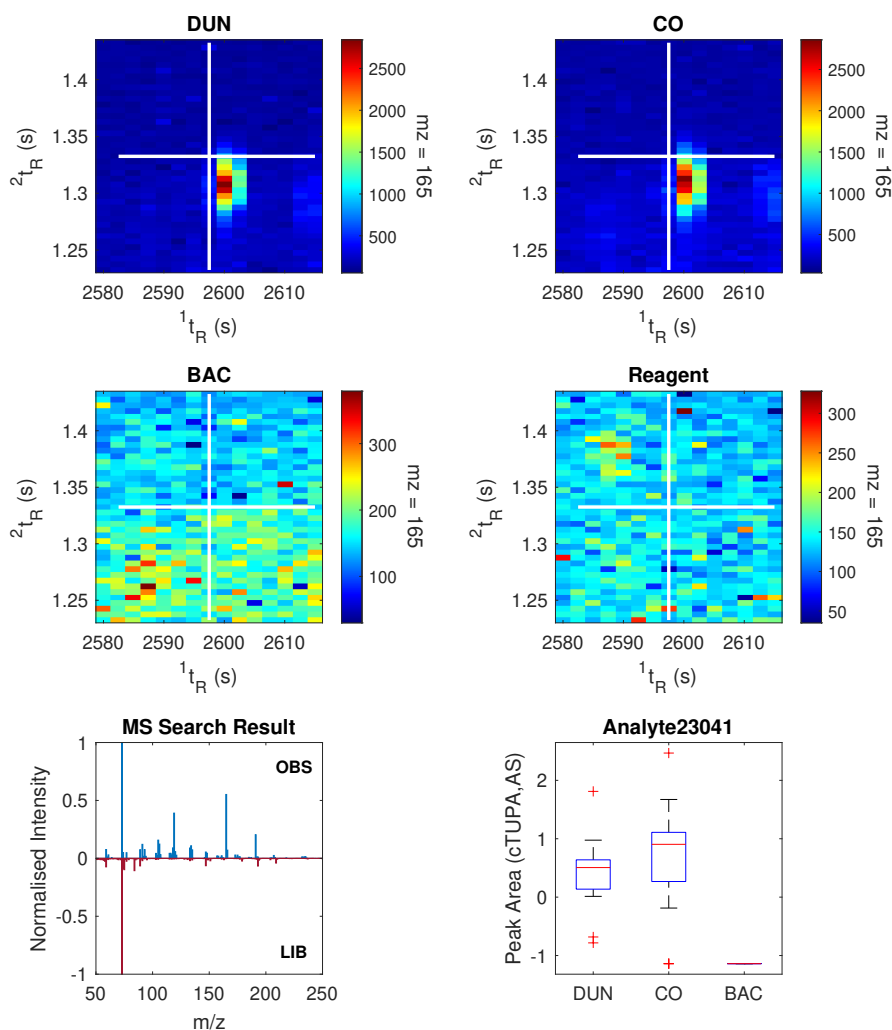


Figure B.13: Summary analyses, Analyte 23041

Peak Information	Top Golm Metabolome Database Search Result
1t_R (s): 2645.0	
2t_R (s): 2.045	
Quantification Ion (m/z): 376	1-dotprod (Spectral Dissimilarity): 0.0793
Analyte Name (ChromaTOF): Analyte23397	Analyte Name (Library): Gluconic acid, 2-amino-2-deoxy-(7TMS)
Retention Index (Observed): 2024.59	Retention Index Difference (Library): 1.14

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/bfddda38-5af0-4dc1-a200-152b9380b8d4.aspx>
Contributor: Boelling C, Liebig F, Erban A, Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

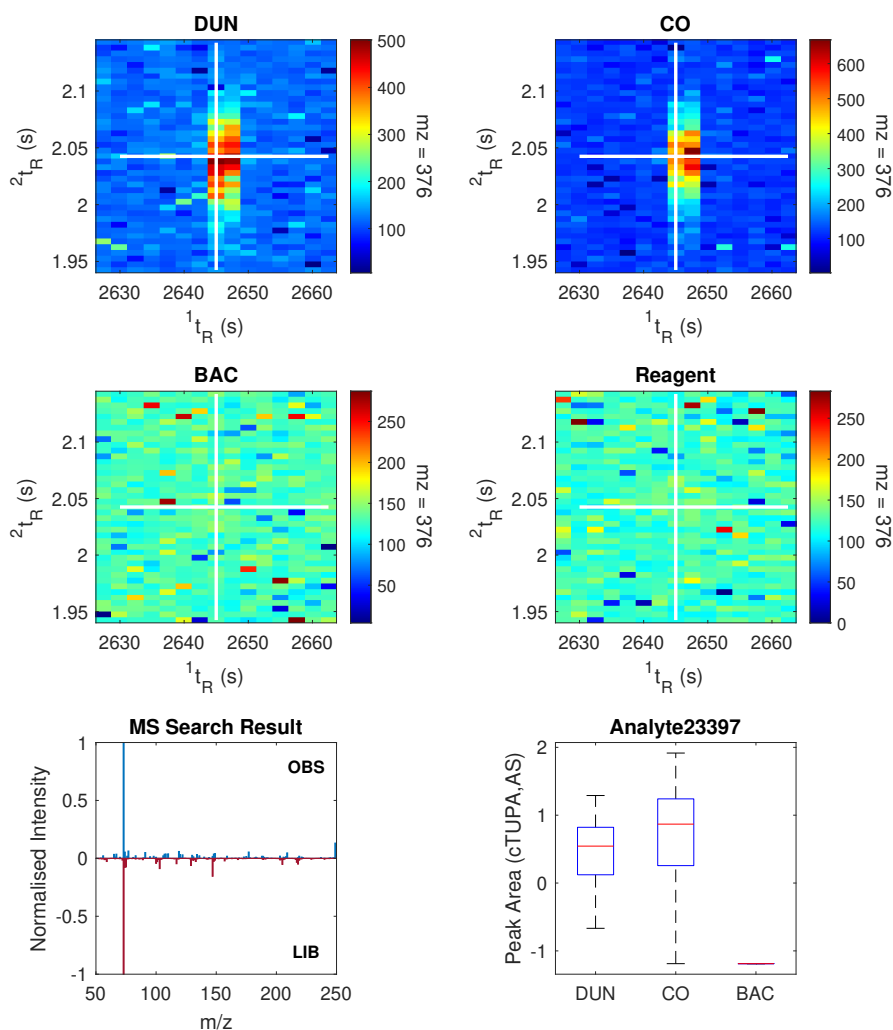


Figure B.14: Summary analyses, Analyte 23397

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 2832.5	
$^2t_R(s)$: 1.865	
Quantification Ion (m/z): 80	1-dotprod (Spectral Dissimilarity): Unknown
Analyte Name (ChromaTOF): Analyte24829	Analyte Name (Library): Unknown
Retention Index (Observed): 2145.06	Retention Index Difference (Library): Unknown

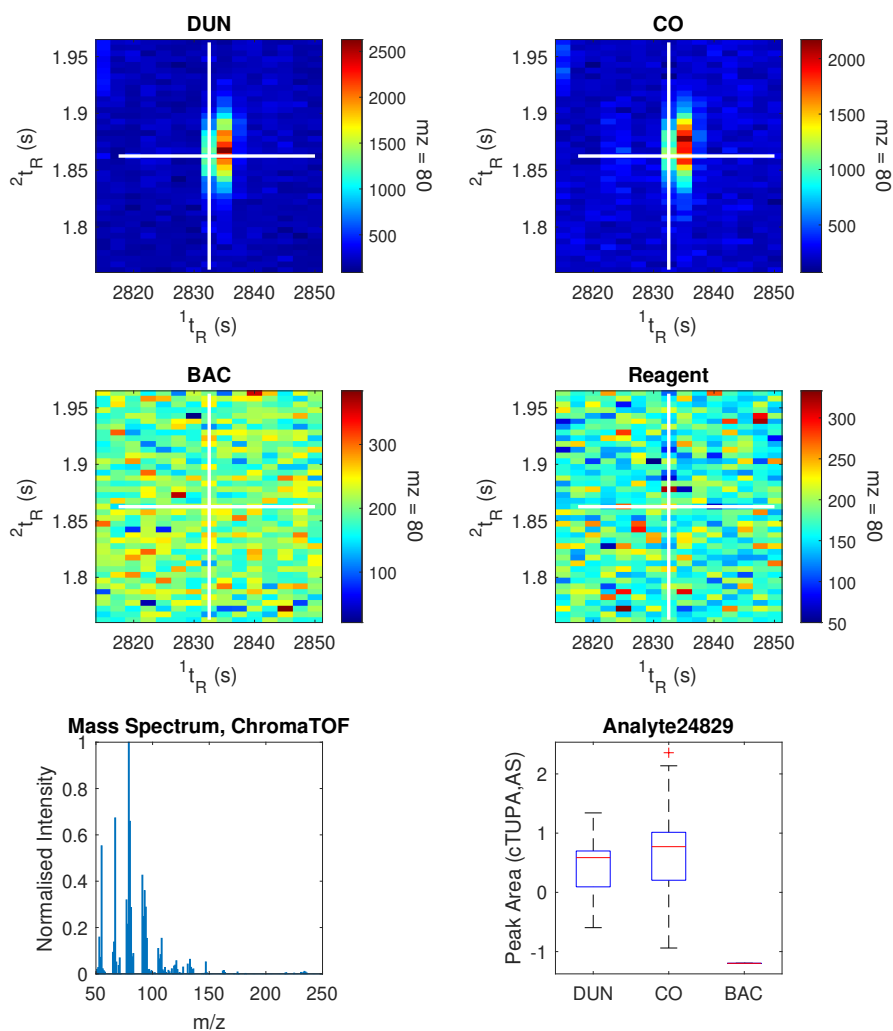


Figure B.15: Summary analyses, Analyte 24829

Peak Information	Top Golm Metabolome Database Search Result
1t_R (s): 3200.0	
2t_R (s): 1.415	
Quantification Ion (m/z): 233	1-dotprod (Spectral Dissimilarity): 0.0553
Analyte Name (ChromaTOF): Analyte27584	Analyte Name (Library): NA
Retention Index (Observed): 2334.57	Retention Index Difference (Library): 1.87

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/3ffe8941-e355-4f09-a740-ec684f396a2f.aspx>

Contributor: Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

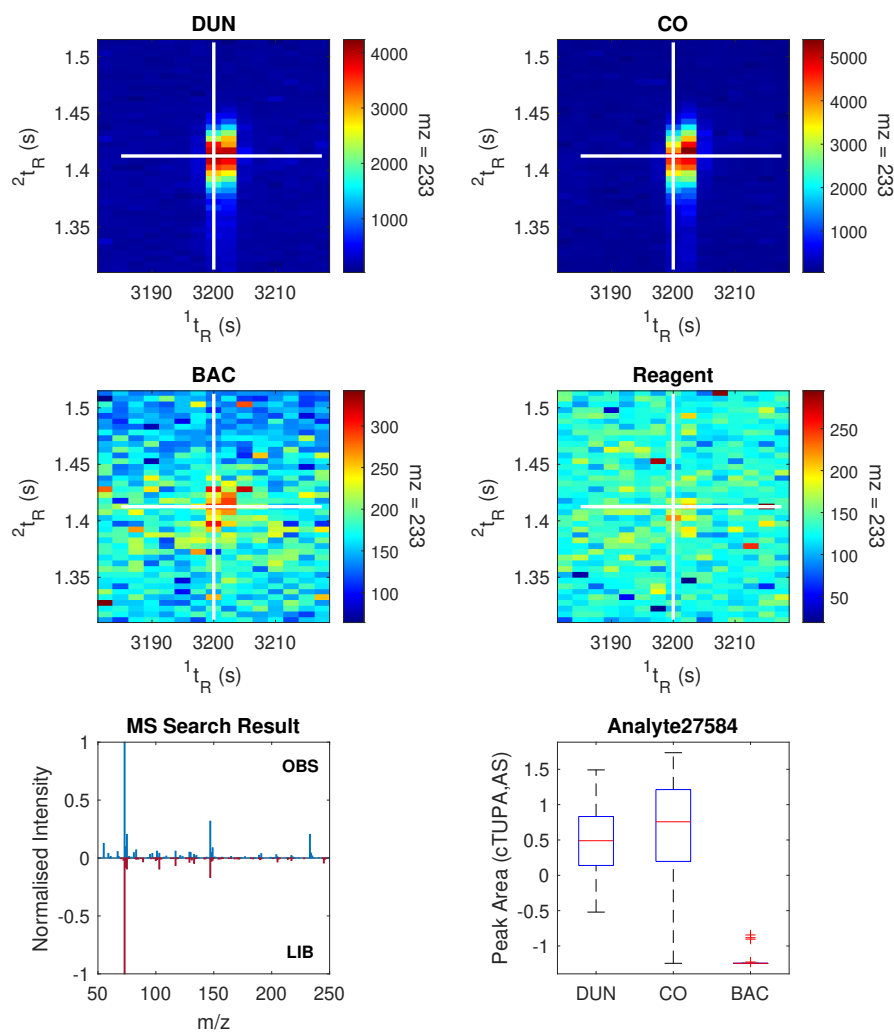


Figure B.16: Summary analyses, Analyte 27584

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 3242.5	
$^2t_R(s)$: 1.475	
Quantification Ion (m/z): 195	1-dotprod (Spectral Dissimilarity): 0.3819
Analyte Name (ChromaTOF): Analyte27833	Analyte Name (Library): Galactose-6-phosphate (1MEOX) (6TMS) BP
Retention Index (Observed): 2416.04	Retention Index Difference (Library): 0.42

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/84406ec0-9f61-42c0-858c-37079019ec05.aspx>

Contributor: Erban A, Strehmel N, Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany

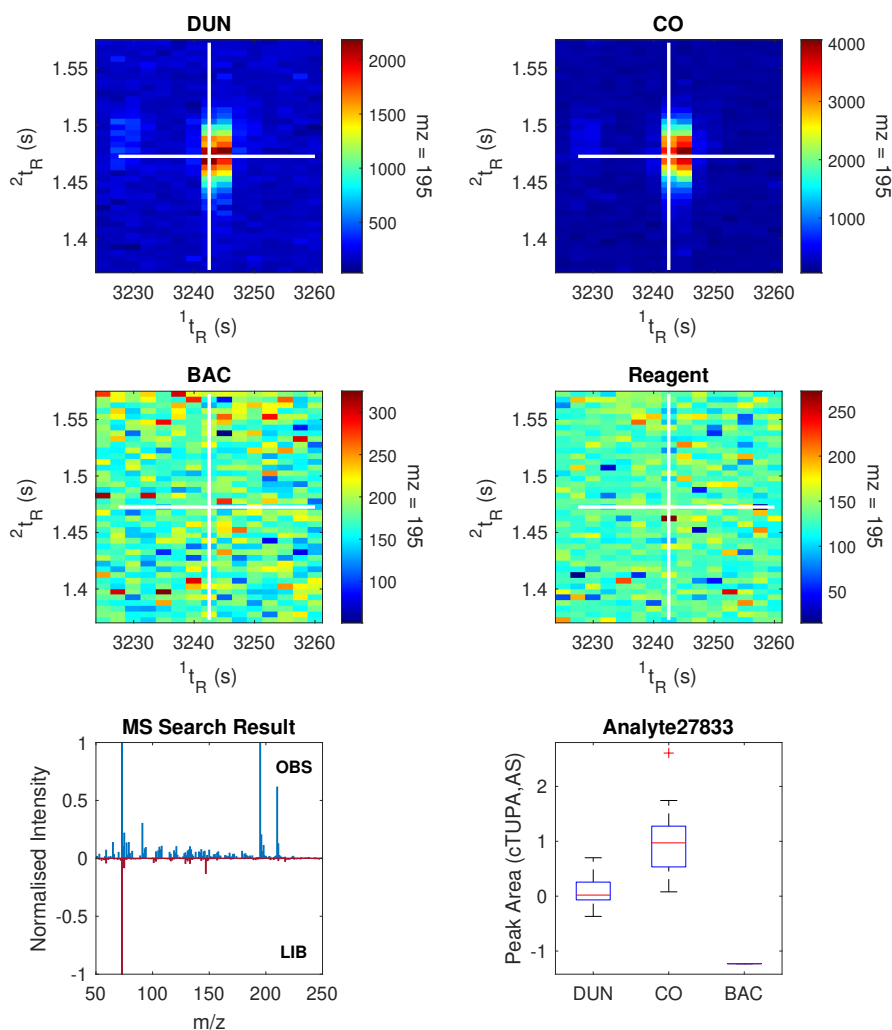


Figure B.17: Summary analyses, Analyte 27833

Peak Information	Top Golm Metabolome Database Search Result
$^1t_R(s)$: 4215.0	
$^2t_R(s)$: 1.840	
Quantification Ion (m/z): 324	1-dotprod (Spectral Dissimilarity): 0.3138
Analyte Name (ChromaTOF): Analyte33374	Analyte Name (Library): Cycloeucalenol (1TMS)
Retention Index (Observed): Undefined	Retention Index (Library): Undefined

Link: <http://gmd.mpimp-golm.mpg.de/Spectrums/e68eba0e-3d8c-44a8-8bbe-0001a2b2c23f.aspx>
Contributor: Moritz T, Umea Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 83 Umea, Sweden

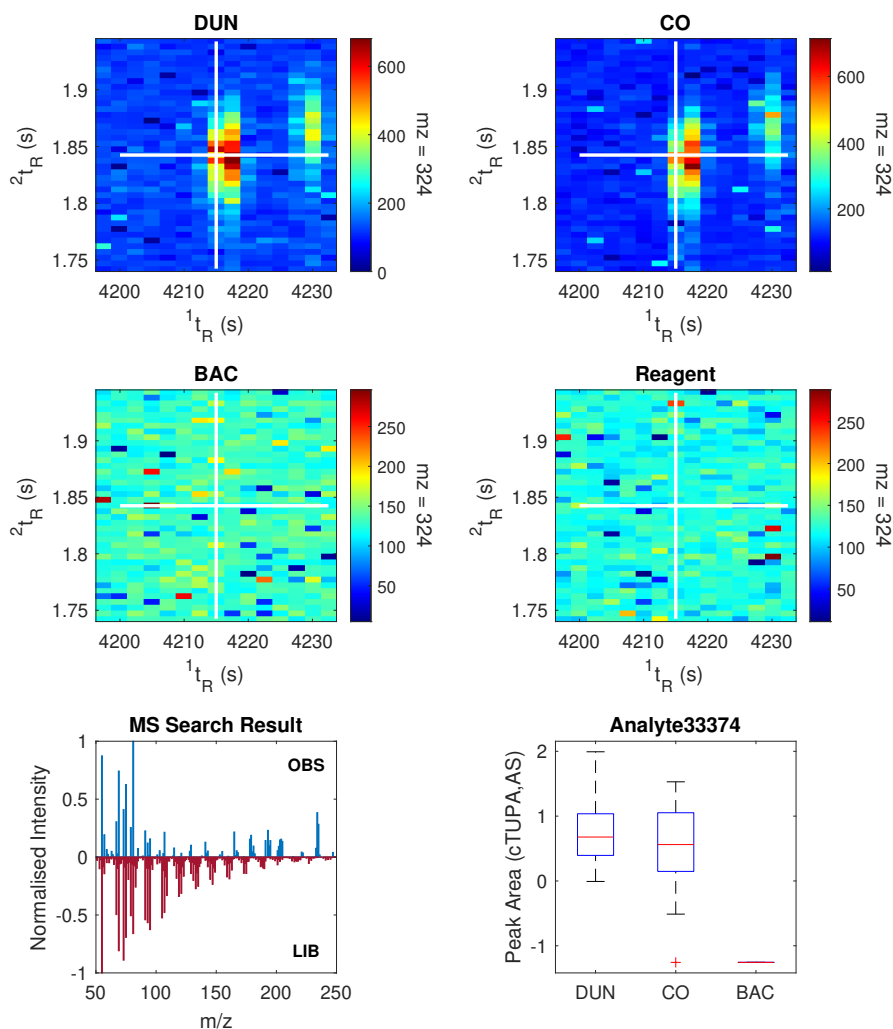


Figure B.18: Summary analyses, Analyte 33374

Appendix C: Derivation of Non-trivial Expressions used in Flexible Coupling PARAFAC2 - ALS, and Coupled PARAFAC2×2 - ALS

C.1 Derivation of an Expression to Solve for A_{kl} , and A_{il}

An expression that minimises the sum of squared residuals can also be described as the minimisation of the square of the Frobenius Norm:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\text{tr}(A^T A)} = \sqrt{\text{tr}(A A^T)}$$

An expression for an estimate of A_{kl} or A_{il} takes into account the coupling term that controls the difference between the two expressions relative to the mass spectral coupling constant, μ_A . Deriving an expression for A_{kl} begins with calculating the derivative of the following expression with respect to A_{kl} :

$$\frac{\partial}{\partial A_{kl}} (\|X_{kl} - B_{kl} D_{kl} A^T\|_F^2 + \mu_A \|A_{kl} - A_{il}\|_F^2) = 0$$

Which has been simplified from Equation 5.17, since the derivative with respect to A_{kl} of $\|X_{il} - B_{il} D_{il} A^T\|_F^2 + \mu_{il} \|B_{il} - P_{il} B^*\|_F^2$ and $\mu_{kl} \|B_{kl} - P_{kl} B^*\|_F^2$ are both zero. Expanding the terms in the previous expression and adding an arbitrary constant, $\frac{1}{2}$, to aid in simplification:

$$\frac{\partial}{\partial A_{kl}} \frac{1}{2} \text{tr} ((X_{kl} - B_{kl} D_{kl} A_{kl}^T)^T (X_{kl} - B_{kl} D_{kl} A_{kl}^T)) + \frac{\partial}{\partial A_{kl}} \text{tr} (\mu_A (A_{kl} - A_{il})^T (A_{kl} - A_{il})) = 0$$

$$\begin{aligned}
& \frac{\partial}{\partial A_{kl}} \text{tr}((X_{kl}^T X_{kl}) - \frac{\partial}{\partial A_{kl}} \text{tr}(X_{kl}^T B_{kl} D_{kl} A_{kl}^T) \\
& - \frac{\partial}{\partial A_{kl}} \text{tr}(A_{kl} D_{kl} B_{kl}^T X_{kl}) + \frac{\partial}{\partial A_{kl}} \text{tr}(A_{kl} D_{kl} B_{kl}^T B_{kl} D_{kl} A_{kl}^T) \\
& + \frac{\partial}{\partial A_{kl}} \text{tr}(\mu_A (A_{kl}^T A_{kl} - A_{kl}^T A_{il} - A_{il}^T A_{kl} + A_{il}^T A_{il})) = 0
\end{aligned}$$

Note that the term $\frac{\partial}{\partial A} (A_{kl}^T A_{kl} - A_{kl}^T A_{il}^T - A_{il}^T A_{kl} + A_{il}^T A_{il})$ with respect to A_{kl} is equivalent when the derivative is taken with respect to A_{il} . This reveals that the expression for the estimation of A_{kl} is the same expression as the expression that estimates A_{il} .

Using the following identities from the Matrix Cookbook:

$$\begin{aligned}
& \frac{\partial}{\partial A_{kl}} \text{tr}(X_{kl}^T X_{kl}) = 0 \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(X_{kl}^T B_{kl} D_{kl} A_{kl}^T) = X_{kl}^T B_{kl} D_{kl} \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(A_{kl} D_{kl} B_{kl}^T X_{kl}) = X_{kl}^T B_{kl} D_{kl} \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(A_{kl} D_{kl} B_{kl}^T B_{kl} D_{kl} A_{kl}^T) = 2A_{kl} D_{kl} B_{kl}^T B_{kl} D_{kl} \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(A_{kl}^T A_{kl}) = 2A_{kl} \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(A_{kl}^T A_{il}) = A_{il} \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(A_{il}^T A_{kl}) = A_{il} \\
& \frac{\partial}{\partial A_{kl}} \text{tr}(A_{il}^T A_{il}) = 0
\end{aligned}$$

Substituting into the previous expression yields:

$$-2X_{kl}^T B_{kl} D_{kl} + 2A_{kl} D_{kl} B_{kl}^T B_{kl} D_{kl} + 2\mu_A A_{kl} - 2\mu_A A_{il} = 0$$

$$2A_{kl} D_{kl} B_{kl}^T B_{kl} D_{kl} + 2\mu_A A_{kl} = 2X_{kl}^T B_{kl} D_{kl} + 2\mu_A A_{il}$$

Solving for A_{kl}

$$A_{kl} = \frac{\mu_A A_{il} + X_{kl}^T B_{kl} D_{kl}}{D_{kl} B_{kl}^T B_{kl} D_{kl} + \mu_A I_R}$$

C.2 Derivation of an Expression to Solve for B_k, B_{kl} , and B_{il}

An expression for B_{kl} and B_{il} can be solved for with respect to their respective coupled terms, and only differ with respect to the arrangement of the data, and is agnostic to the mass spectral coupling term, μ_A . As such, the following derivation can be described for the flexible coupling method of calculating non-negative PARAFAC2 with respect to B_k .

$$\frac{\partial}{\partial B_k} (\|X_k - B_k D_k A^T\|^2 + \mu_k \|B_k - P_k B^*\|^2) = 0$$

We can rearrange the original equation, adding an arbitrary constant as before, $\frac{1}{2}$, to aid with simplification later on.

$$\frac{\partial}{\partial B_k} \frac{1}{2} \text{tr}((X_k - B_k D_k A^T)(X_k - B_k D_k A^T)^T) + \frac{\partial}{\partial B_k} \frac{1}{2} \mu_k (\text{tr}((B_k - P_k B^*)(B_k - P_k B^*)^T)) = 0$$

Expanding the equations, where $(X_k - B_k D_k A^T)^T = X_k^T - A D_k B_k^T$ and $(B_k - P_k B^*)^T = B_k^T - B^{*T} P_k^T$:

$$\begin{aligned} \frac{\partial}{\partial B_k} \frac{1}{2} (\text{tr}(X_k X_k^T) - \text{tr}(X_k A D_k B_k^T) - \text{tr}(B_k D_k A^T X_k^T) + \text{tr}(B_k D_k A^T A D_k B_k^T)) \\ + \frac{\partial}{\partial B_k} \frac{1}{2} \mu_k (\text{tr}(B_k B_k^T) - \text{tr}(B_k B^{*T} P_k^T) - \text{tr}(P_k B^* B_k^T) + \text{tr}(P_k B^* B^{*T} P_k^T)) = 0 \end{aligned}$$

This equation can be simplified by using some convenient identities from [the Matrix Cookbook](<https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>):

$$\frac{\partial}{\partial B_k} \text{tr}(X_k X_k^T) = 0$$

$$\frac{\partial}{\partial B_k} \text{tr}(X_k A D_k B_k^T) = X_k A D_k$$

$$\frac{\partial}{\partial B_k} \text{tr}(B_k D_k A^T X_k^T) = X_k A D_k$$

$$\frac{\partial}{\partial B_k} \text{tr}(B_k D_k A^T A D_k B_k^T) = 2 B_k D_k A^T A D_k$$

$$\frac{\partial}{\partial B_k} \text{tr}(B_k D_k A^T A D_k B_k^T) = 2 B_k D_k A^T A D_k$$

$$\frac{\partial}{\partial B_k} \text{tr}(B_k B^{*T} P_k^T) = P_k B^*$$

$$\frac{\partial}{\partial B_k} \text{tr}(P_k B^* B_k^T) = P_k B^*$$

$$\frac{\partial}{\partial B_k} \text{tr}(P_k B^* B^{*T} P_k^T) = 0$$

$$\frac{\partial}{\partial B_k} \text{tr}(B_k B_k^T) = 2B_k$$

Substituting into the previous equation yields:

$$-X_k A D_k + B_k D_k A^T A D_k + \mu_k B_k - \mu_k P_k B^* = 0$$

Multiplying by the inverse of $D_k A^T A D_k$:

$$-X_k A D_k (D_k A^T A D_k)^{-1} + B_k + \mu_k B_k (D_k A^T A D_k)^{-1} - \mu_k P_k B^* (D_k A^T A D_k)^{-1} = 0$$

Solving for B_k :

$$-X_k A D_k + \mu_k B_k - \mu_k P_k B^* = -B_k (D_k A^T A D_k)$$

$$-X_k A D_k - \mu_k P_k B^* = -B_k (D_k A^T A D_k) - \mu_k B_k$$

Yields the final form of the equation:

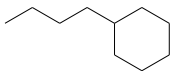
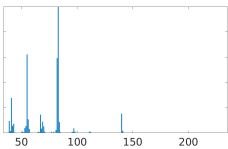
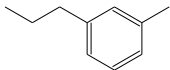
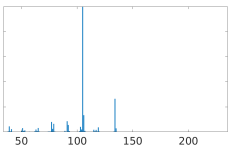
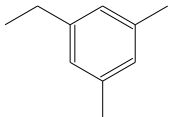
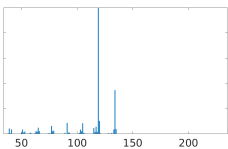
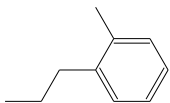
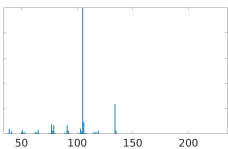
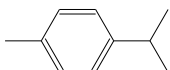
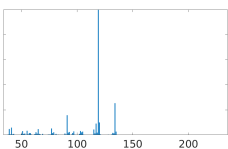
$$B_k = \frac{X_k A D_k + \mu_k P_k B^*}{D_k A^T A D_k + \mu_k I_R}$$

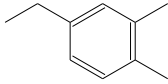
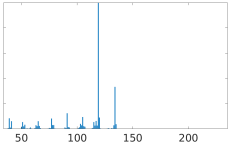

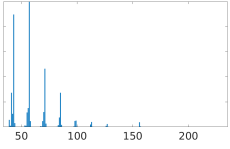
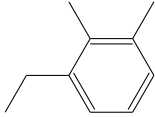
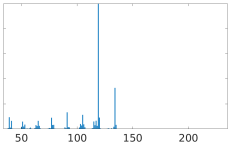
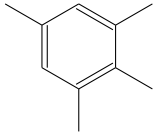
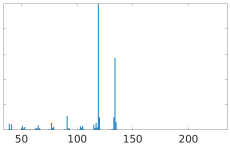
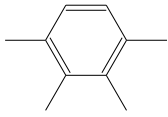
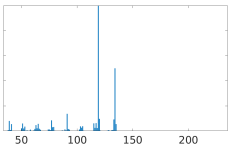
This solution is similar to the one released in Jeremy Cohen's software package [45], although the authors could not a previously published derivation.

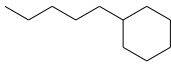
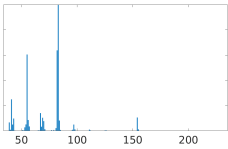
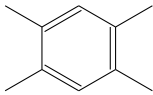
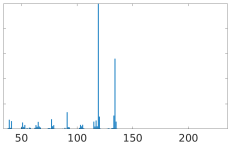
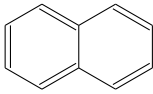
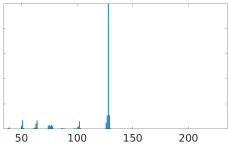
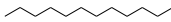
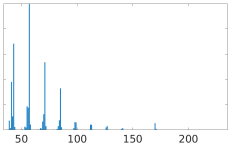
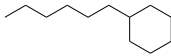
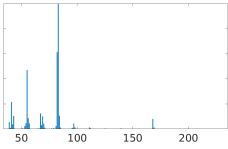
Appendix D: Supporting Information for: “*A-priori* prediction of chemical rank in gas-chromatography mass spectrometry data with projection pursuit analysis”

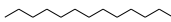
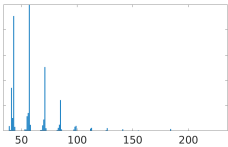
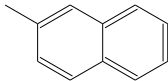
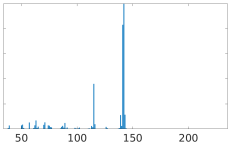
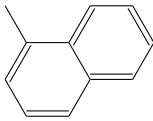
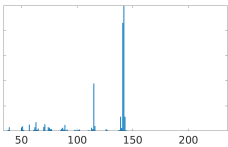
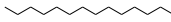
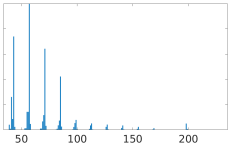

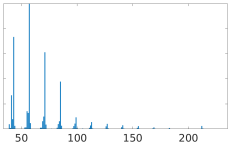
D.1 Library of Chemical Components used for Synthetic Data

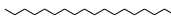
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
611143	C ₉ H ₁₂		
526738	C ₉ H ₁₂		
135988	C ₁₀ H ₁₄		
527844	C ₁₀ H ₁₄		
95636	C ₉ H ₁₂		


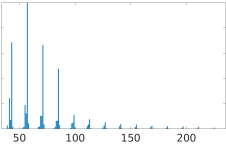

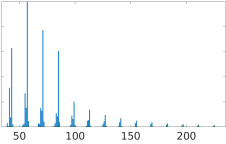
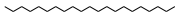
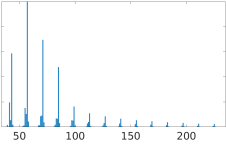
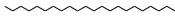
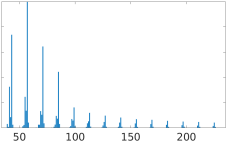
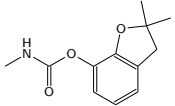
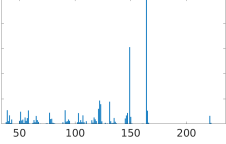
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
1678939	C ₁₀ H ₂₀		
1074437	C ₁₀ H ₁₄		
934747	C ₁₀ H ₁₄		
1074175	C ₁₀ H ₁₄		
99876	C ₁₀ H ₁₄		

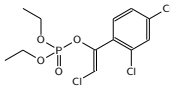
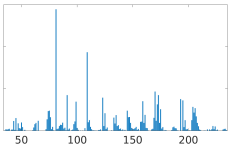
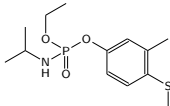
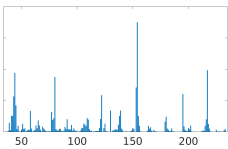
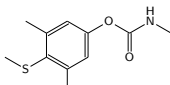
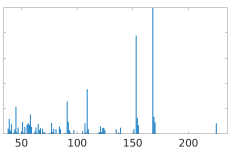
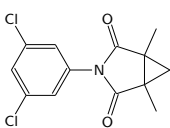
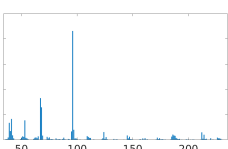
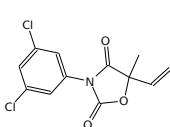
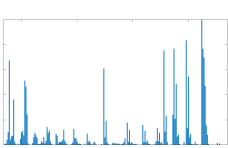
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
934805	C ₁₀ H ₁₄		
1120214	C ₁₁ H ₂₄		
933982	C ₁₀ H ₁₄		
527537	C ₁₀ H ₁₄		
488233	C ₁₀ H ₁₄		

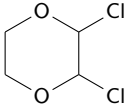
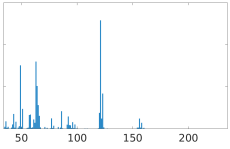
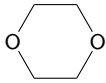
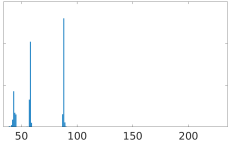
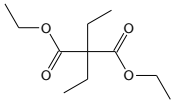
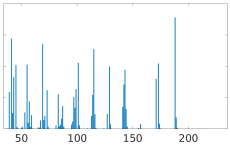
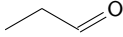
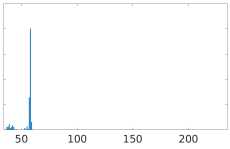
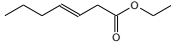
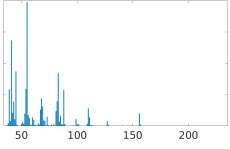
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
4292926	C ₁₁ H ₂₂		
95932	C ₁₀ H ₁₄		
91203	C ₁₀ H ₈		
112403	C ₁₂ H ₂₆		
4292755	C ₁₂ H ₂₄		

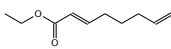
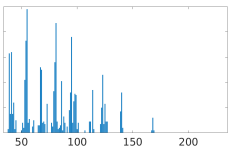
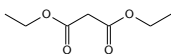
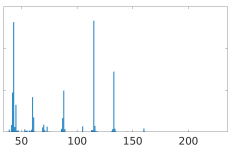
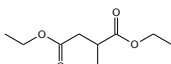
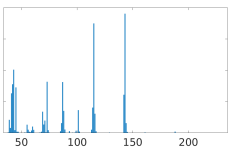
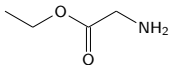
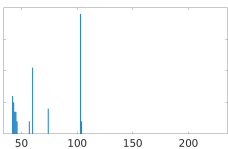
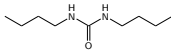
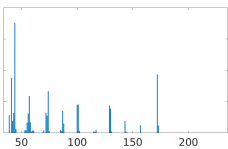
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
629505	C ₁₃ H ₂₈		
91576	C ₁₁ H ₁₀		
90120	C ₁₁ H ₁₀		
629594	C ₁₄ H ₃₀		
629629	C ₁₅ H ₃₂		

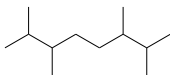
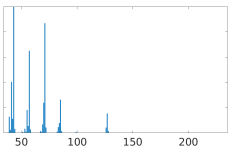
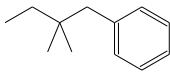
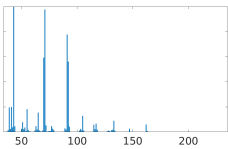
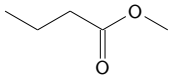
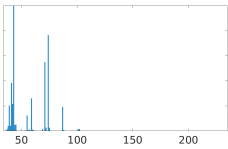
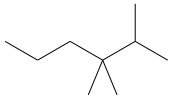
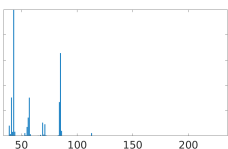
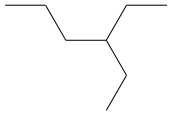
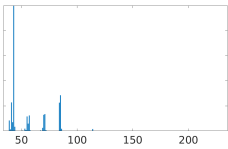
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
544763	C ₁₆ H ₃₄		
629787	C ₁₇ H ₃₆		
1921706	C ₁₉ H ₄₀		
593453	C ₁₈ H ₃₈		
638368	C ₂₀ H ₄₂		

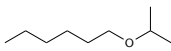
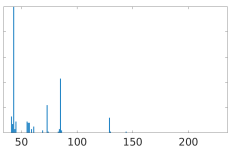
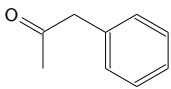
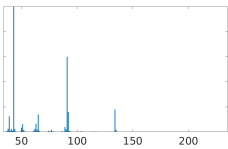
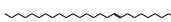
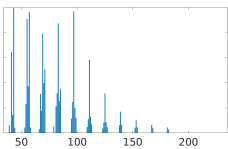
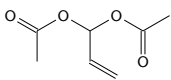
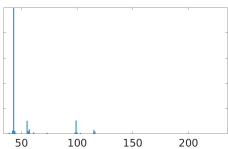
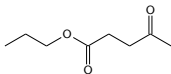
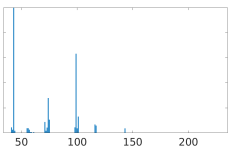
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
629925	C ₁₉ H ₄₀		
112958	C ₂₀ H ₄₂		
629947	C ₂₁ H ₄₄		
629970	C ₂₂ H ₄₆		
1563662	C ₁₂ H ₁₅ NO ₃		

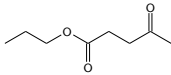
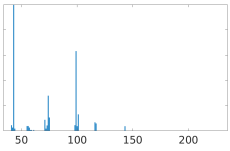
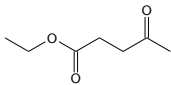
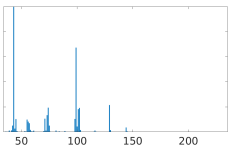

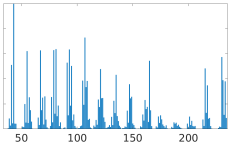
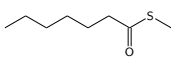
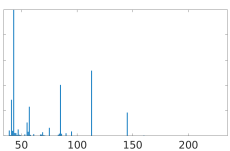
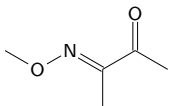
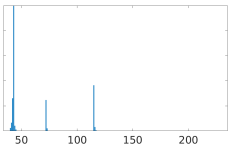
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
470906	C ₁₂ H ₁₄ Cl ₃ O ₄ P		
22224926	C ₁₃ H ₂₂ NO ₃ PS		
2032657	C ₁₁ H ₁₅ NO ₂ S		
32809168	C ₁₃ H ₁₁ Cl ₂ NO ₂		
50471448	C ₁₂ H ₉ Cl ₂ NO ₃		

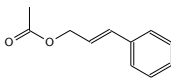
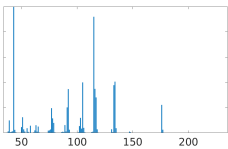
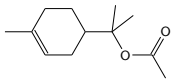
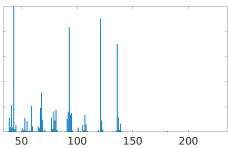
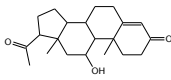
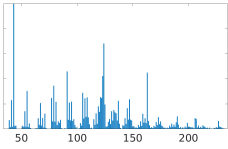
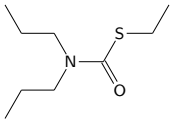
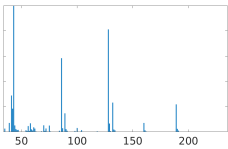
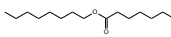
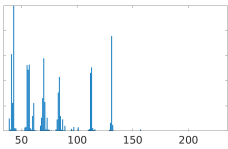
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
95590	C ₄ H ₆ Cl ₂ O ₂		
123911	C ₄ H ₈ O ₂		
77258	C ₁₁ H ₂₀ O ₄		
123386	C ₃ H ₆ O		
54340716	C ₉ H ₁₆ O ₂		

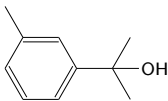
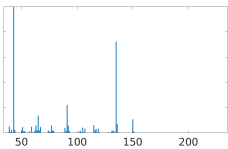
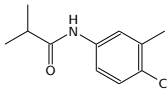
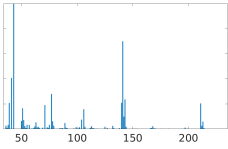
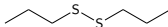
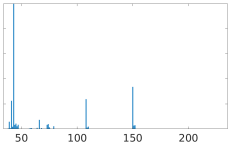
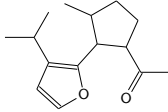
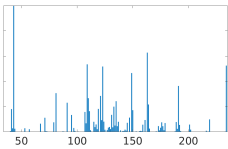
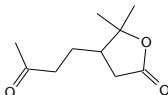
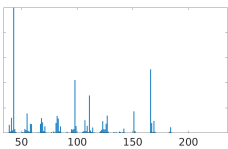
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
55282912	C ₁₀ H ₁₆ O ₂		
105533	C ₇ H ₁₂ O ₄		
4676511	C ₉ H ₁₆ O ₄		
459734	C ₄ H ₉ NO ₂		
1792172	C ₉ H ₂₀ N ₂ O		

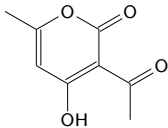
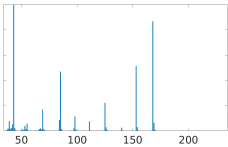
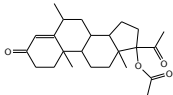
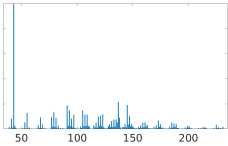
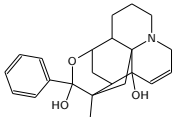
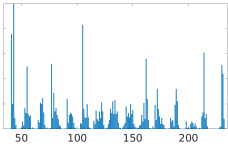
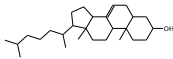
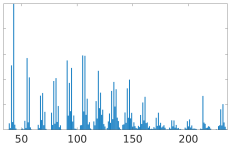
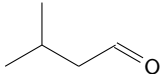
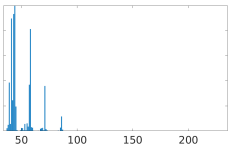
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
52670344	C ₁₂ H ₂₆		
28080866	C ₁₂ H ₁₈		
623427	C ₅ H ₁₀ O ₂		
16747287	C ₉ H ₂₀		
619998	C ₈ H ₁₈		

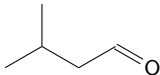
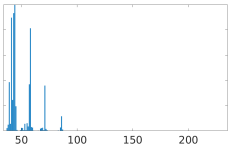
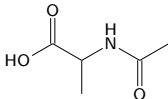
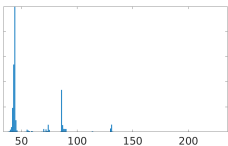
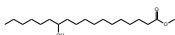
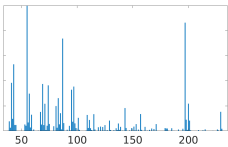
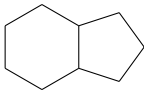
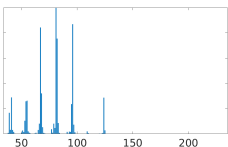
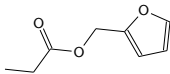
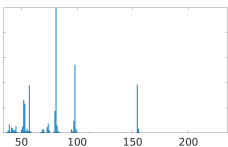
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
18636652	C ₉ H ₂₀ O		
103797	C ₉ H ₁₀ O		
71502224	C ₂₆ H ₅₂		
869294	C ₇ H ₁₀ O ₄		
645670	C ₈ H ₁₄ O ₃		

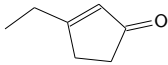
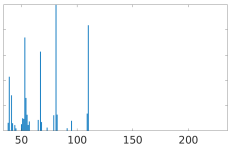
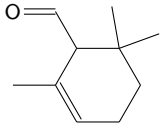
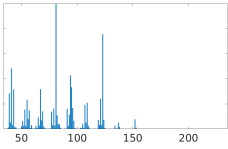
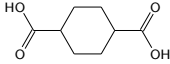
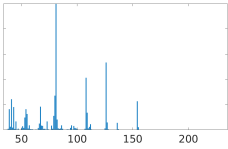
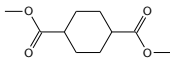
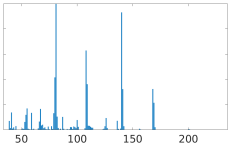
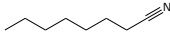
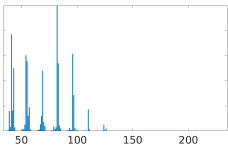
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
645670	C ₈ H ₁₄ O ₃		
539888	C ₇ H ₁₂ O ₃		
641827	C ₂₀ H ₃₄ O ₂		
2432828	C ₈ H ₁₆ OS		
617323	C ₅ H ₉ NO ₂		

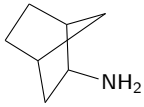
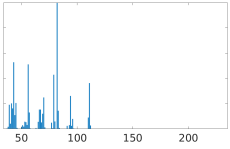
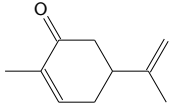
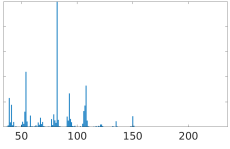

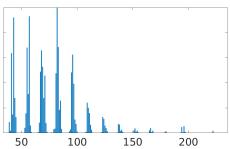
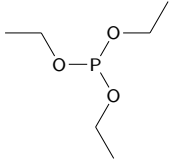
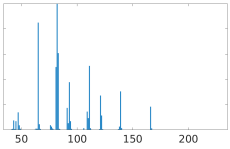
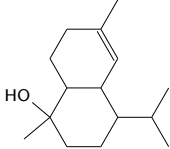
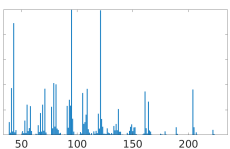
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
103548	C ₁₁ H ₁₂ O ₂		
80262	C ₁₂ H ₂₀ O ₂		
80751	C ₂₁ H ₃₀ O ₃		
759944	C ₉ H ₁₉ NOS		
5132752	C ₁₅ H ₃₀ O ₂		

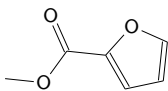
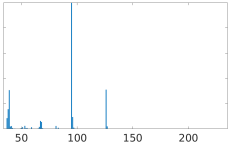
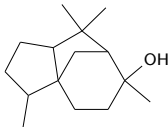
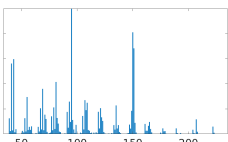
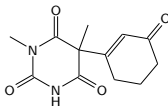
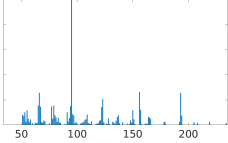
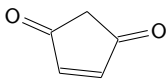
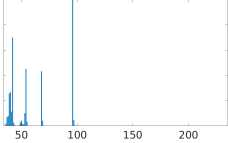
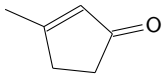
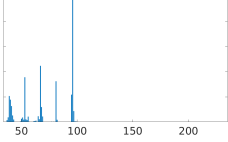
CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
5208377	C ₁₀ H ₁₄ O		
24051408	C ₁₁ H ₁₄ ClNO		
629196	C ₆ H ₁₄ S ₂		
1143460	C ₁₅ H ₂₂ O ₂		
4436811	C ₁₀ H ₁₆ O ₃		

CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
771039	C ₈ H ₈ O ₄		
71589	C ₂₄ H ₃₄ O ₄		
5096628	C ₂₂ H ₂₇ NO ₃		
80999	C ₂₇ H ₄₆ O		
590863	C ₅ H ₁₀ O		

CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
590863	C ₅ H ₁₀ O		
97698	C ₅ H ₉ NO ₃		
141231	C ₁₉ H ₃₈ O ₃		
4551513	C ₉ H ₁₆		
623198	C ₈ H ₁₀ O ₃		

CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
5682699	C ₇ H ₁₀ O		
432246	C ₁₀ H ₁₆ O		
619829	C ₈ H ₁₂ O ₄		
94600	C ₁₀ H ₁₆ O ₄		
124129	C ₈ H ₁₅ N		

CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
7242924	C ₇ H ₁₃ N		
2244168	C ₁₀ H ₁₄ O		
629801	C ₁₆ H ₃₂ O		
122521	C ₆ H ₁₅ O ₃ P		
481345	C ₁₅ H ₂₆ O		

CAS Number	Chemical Formula	Chemical Structure	Mass Spectrum
611132	C ₆ H ₆ O ₃		
77532	C ₁₅ H ₂₆ O		
427305	C ₁₂ H ₁₄ N ₂ O ₄		
930609	C ₅ H ₄ O ₂		
2758181	C ₆ H ₈ O		

D.2 Summary Analyses of GC-qMS Data

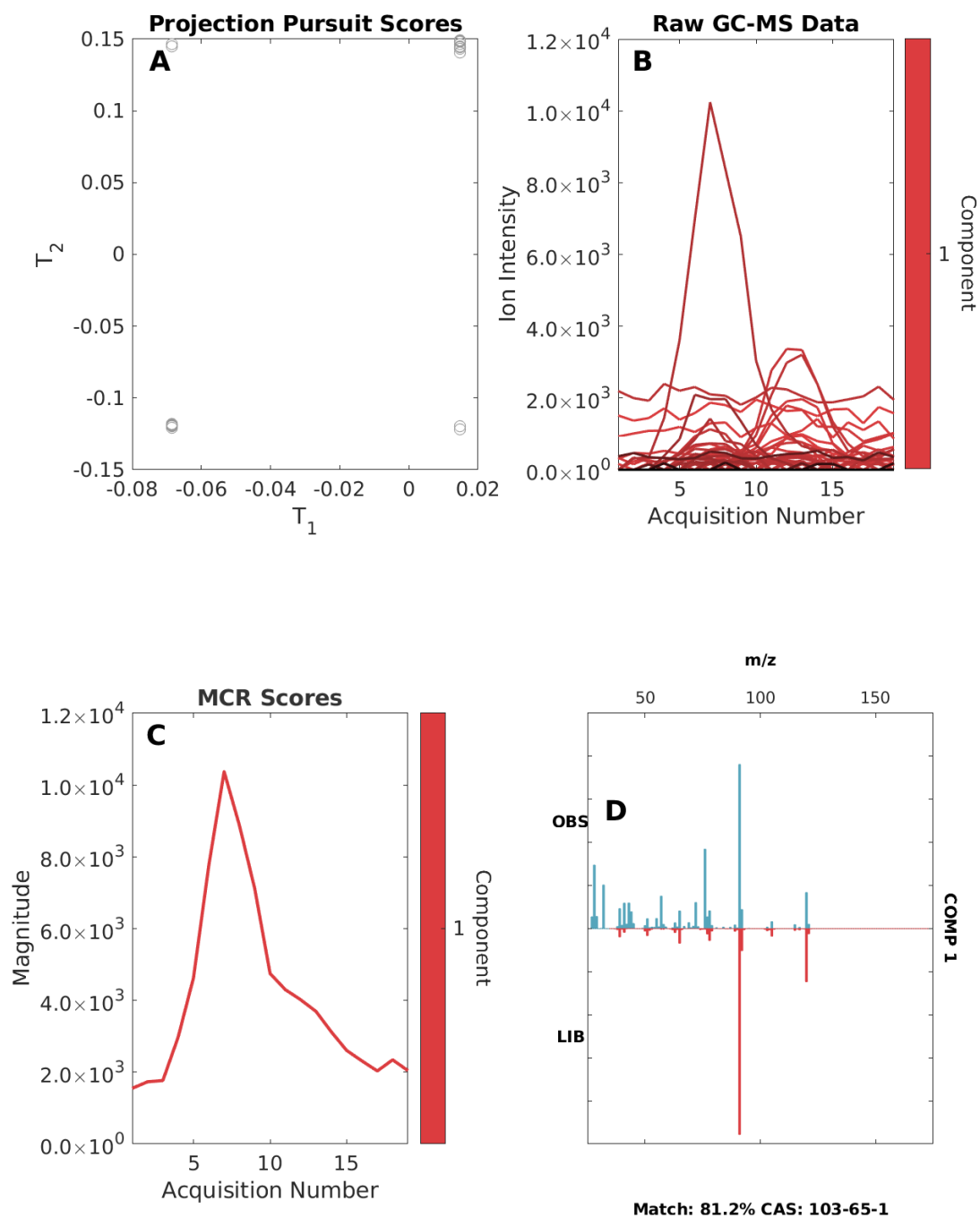


Figure D.1: Cumulative variances explained by component: 71.25

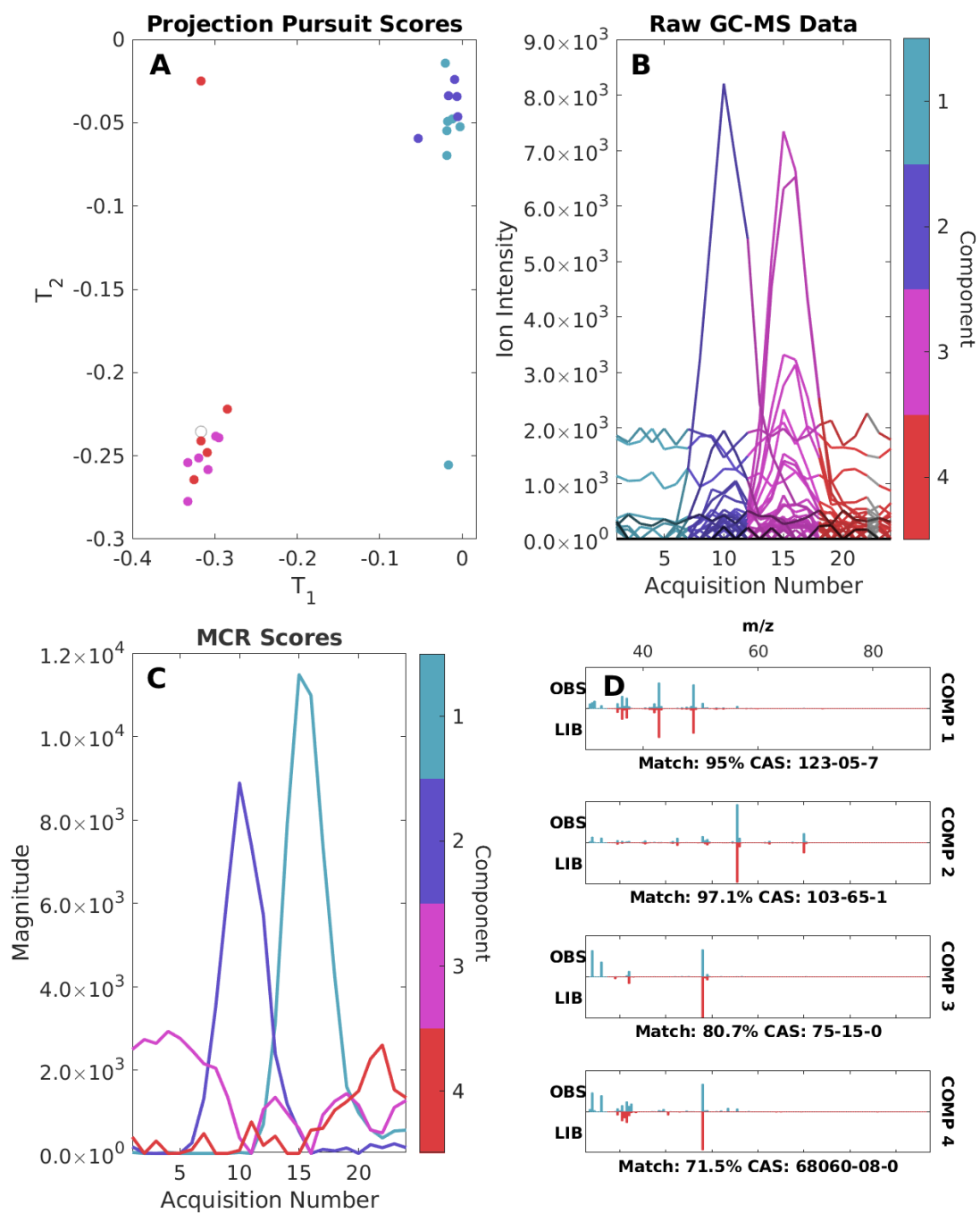


Figure D.2: Cumulative variances explained by component: 55.41 86.76 96.01 99.14

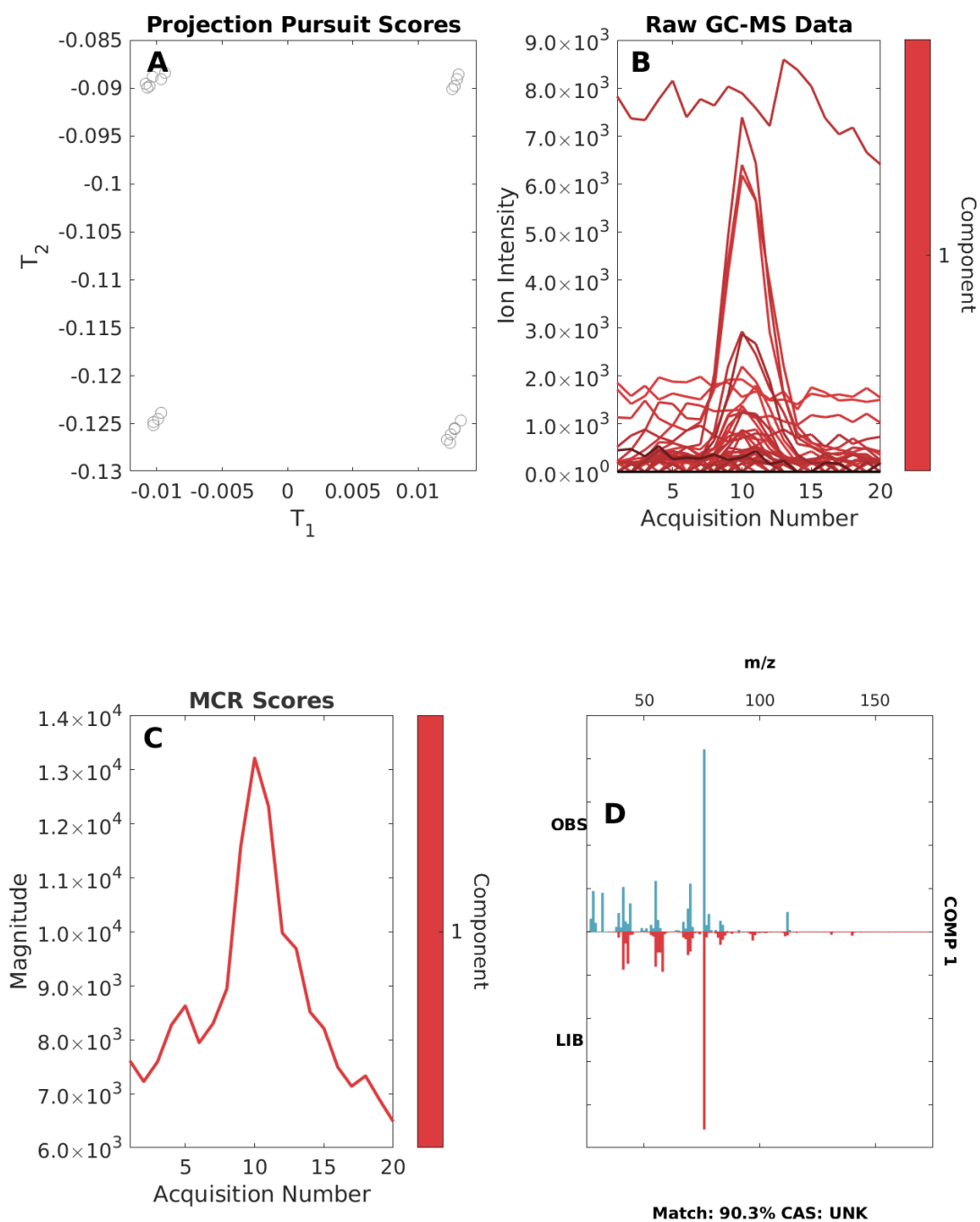


Figure D.3: Cumulative variances explained by component: 87.14

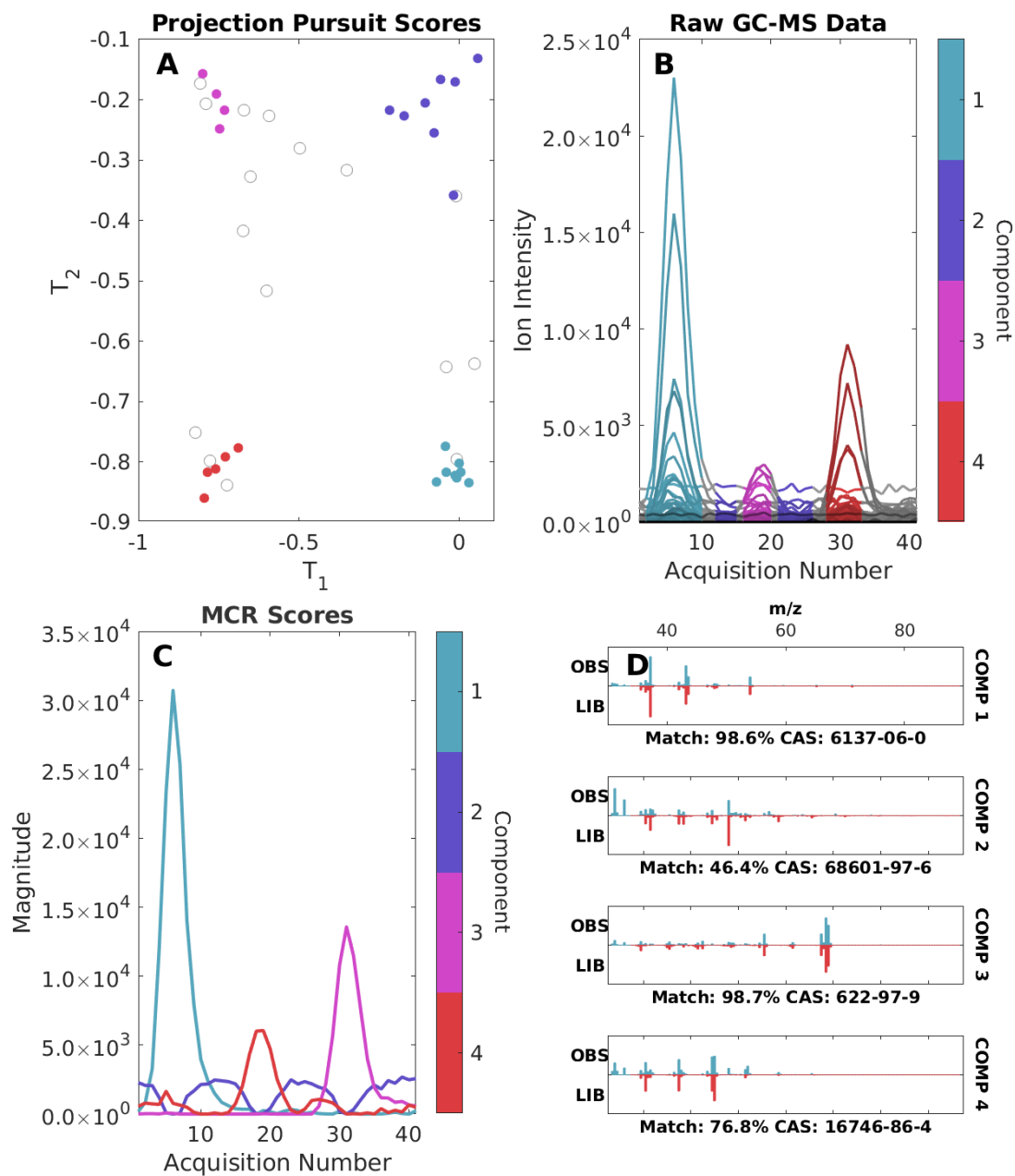


Figure D.4: Cumulative variances explained by component: 75.06 78.88 95.07 99.07

D.3 Summary Analysis of GC-TOFMS Data

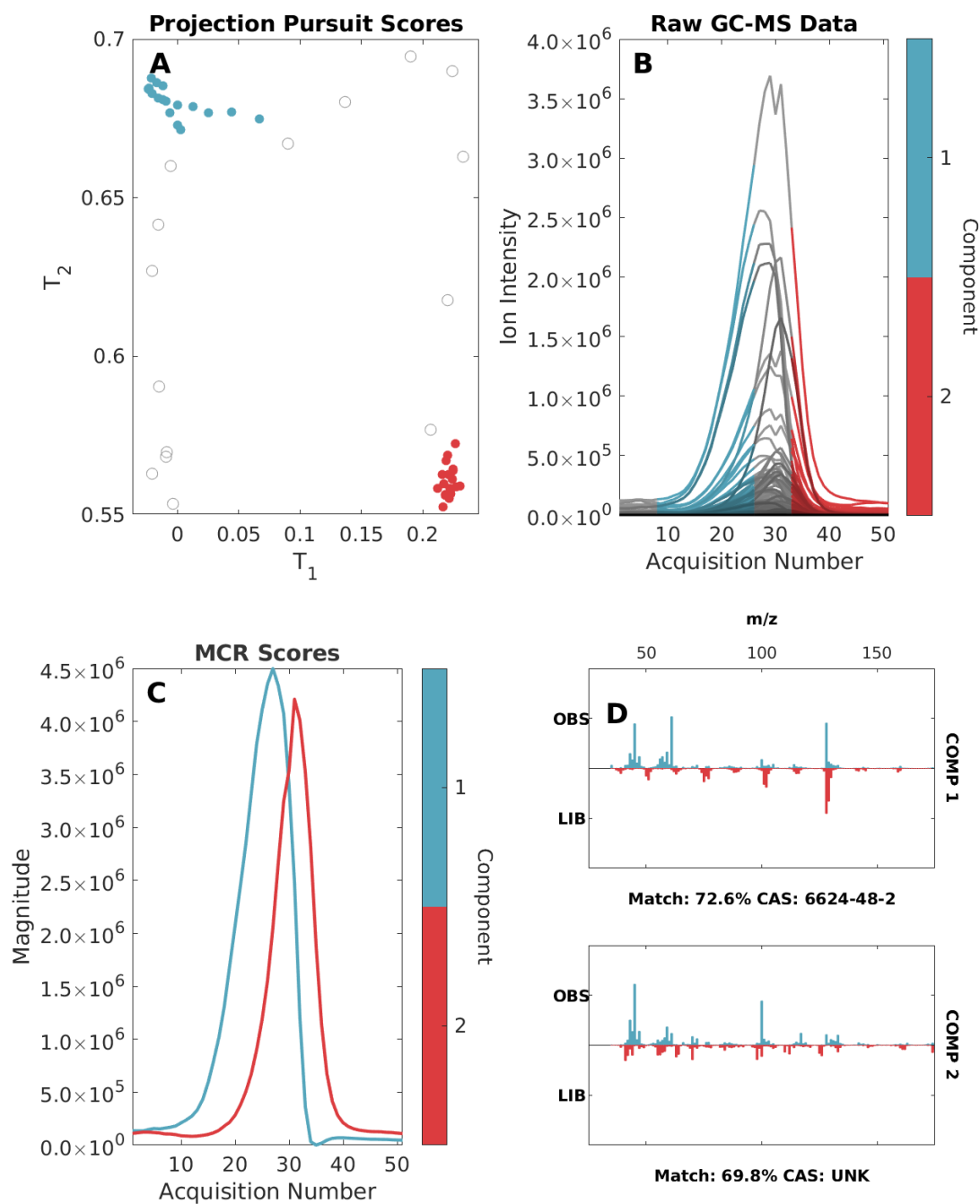


Figure D.5: Cumulative variances explained by component: 61.54 99.63

D.4 Summary Analysis of GC×GC-TOFMS Data

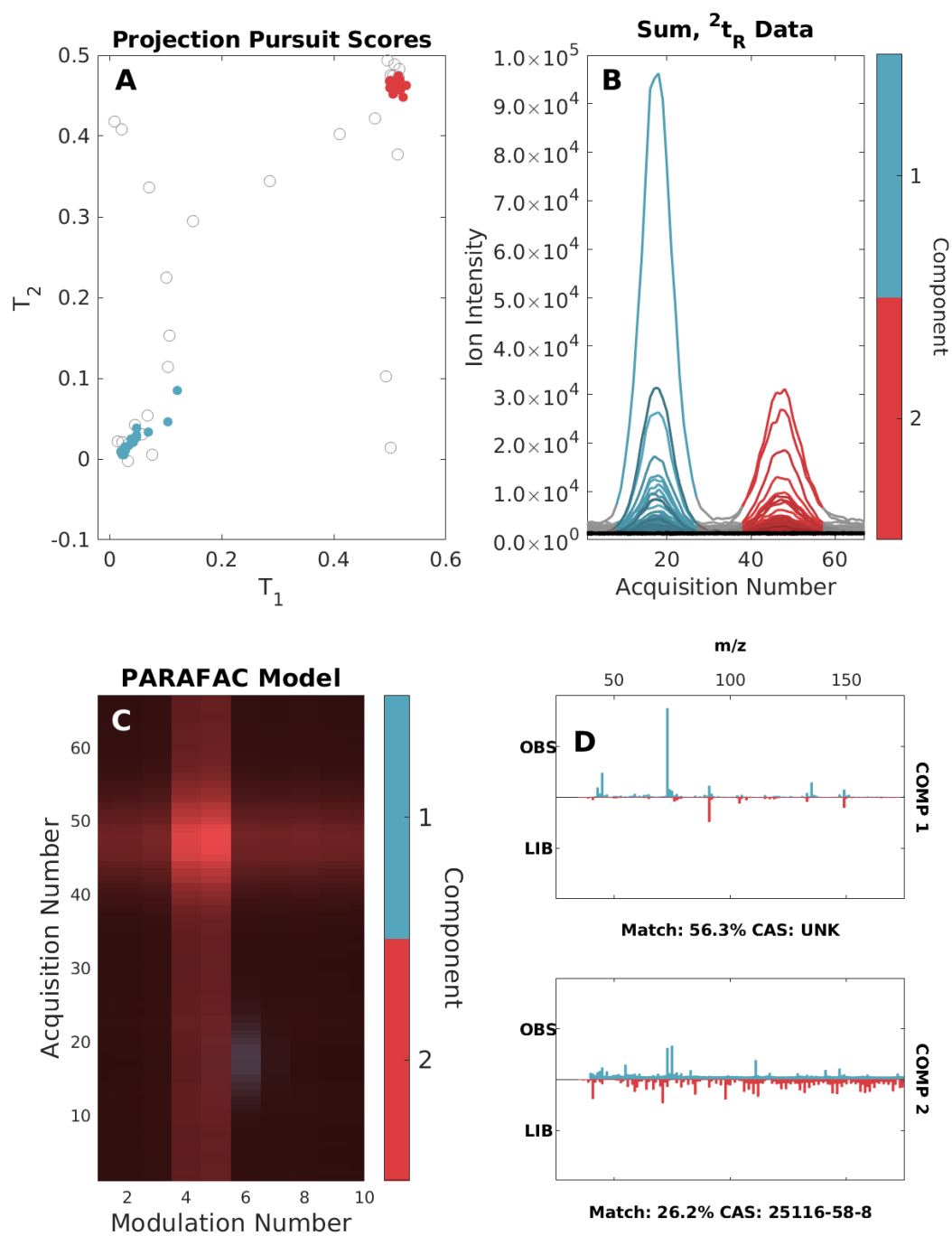


Figure D.6: Cumulative variances explained by component: 89.65

D.5 Summary Analyses of Synthetic Data

D.5.1 1 Factor Synthetic Data

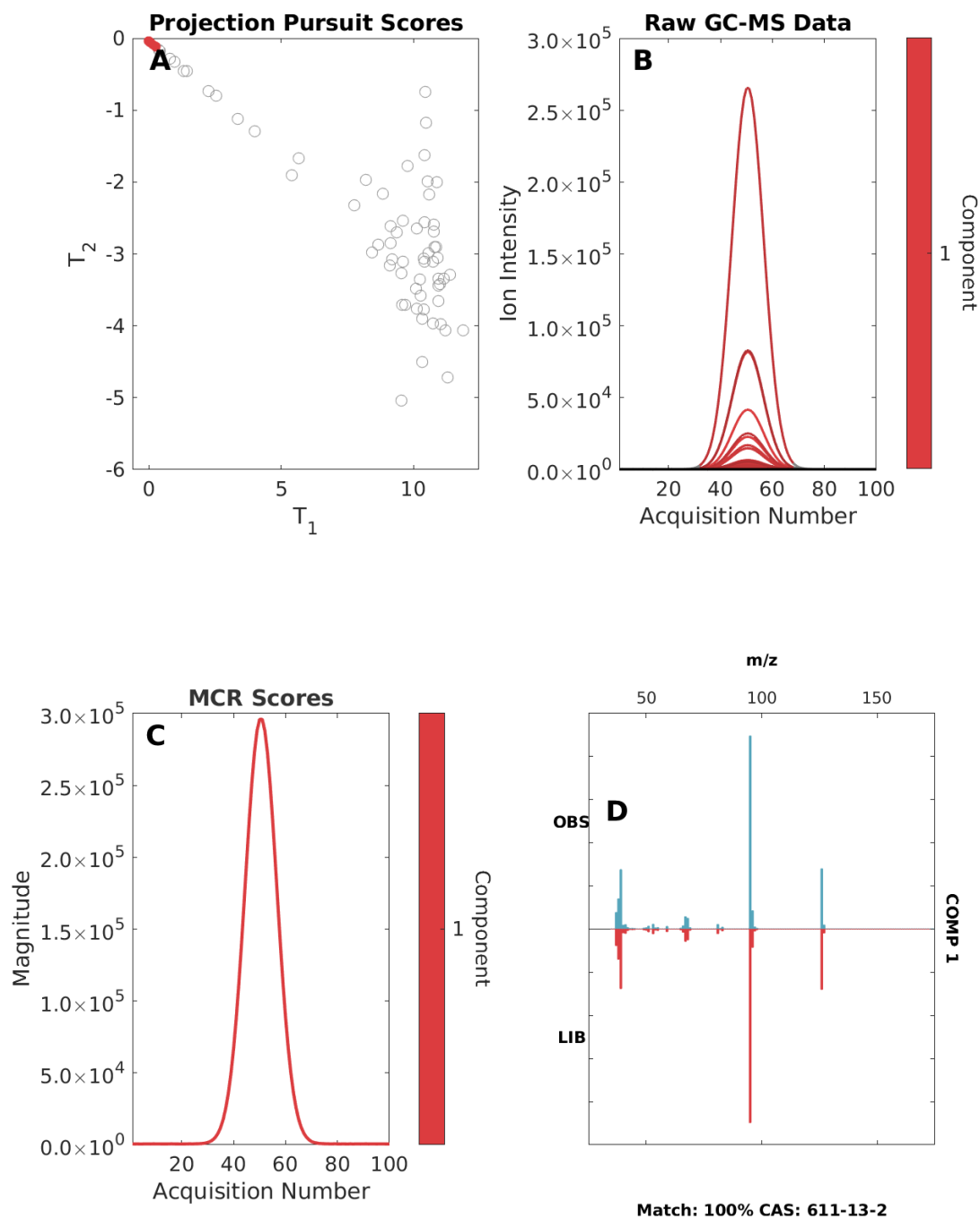


Figure D.7: Cumulative variances explained by component: 99.77

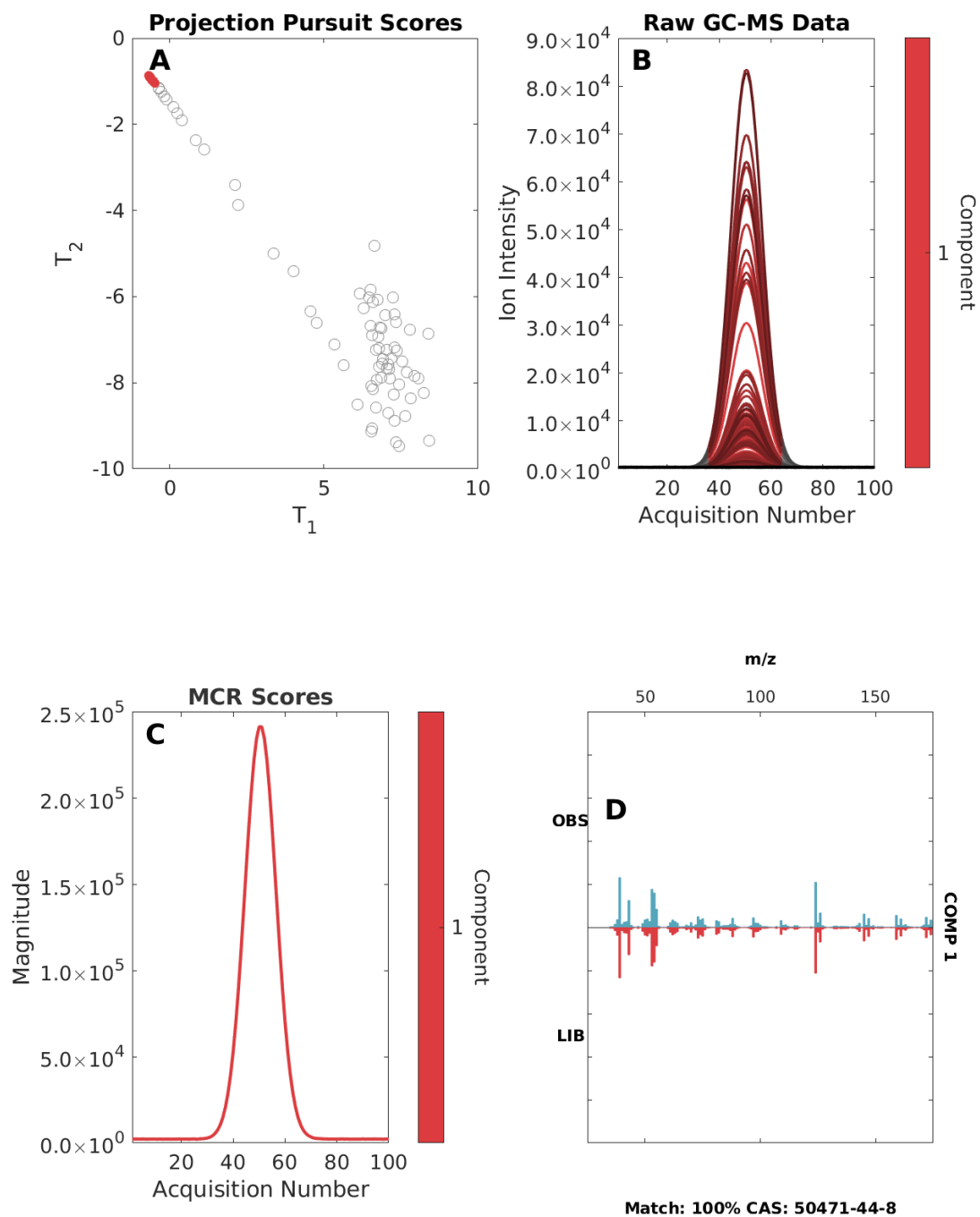


Figure D.8: Cumulative variances explained by component: 99.43

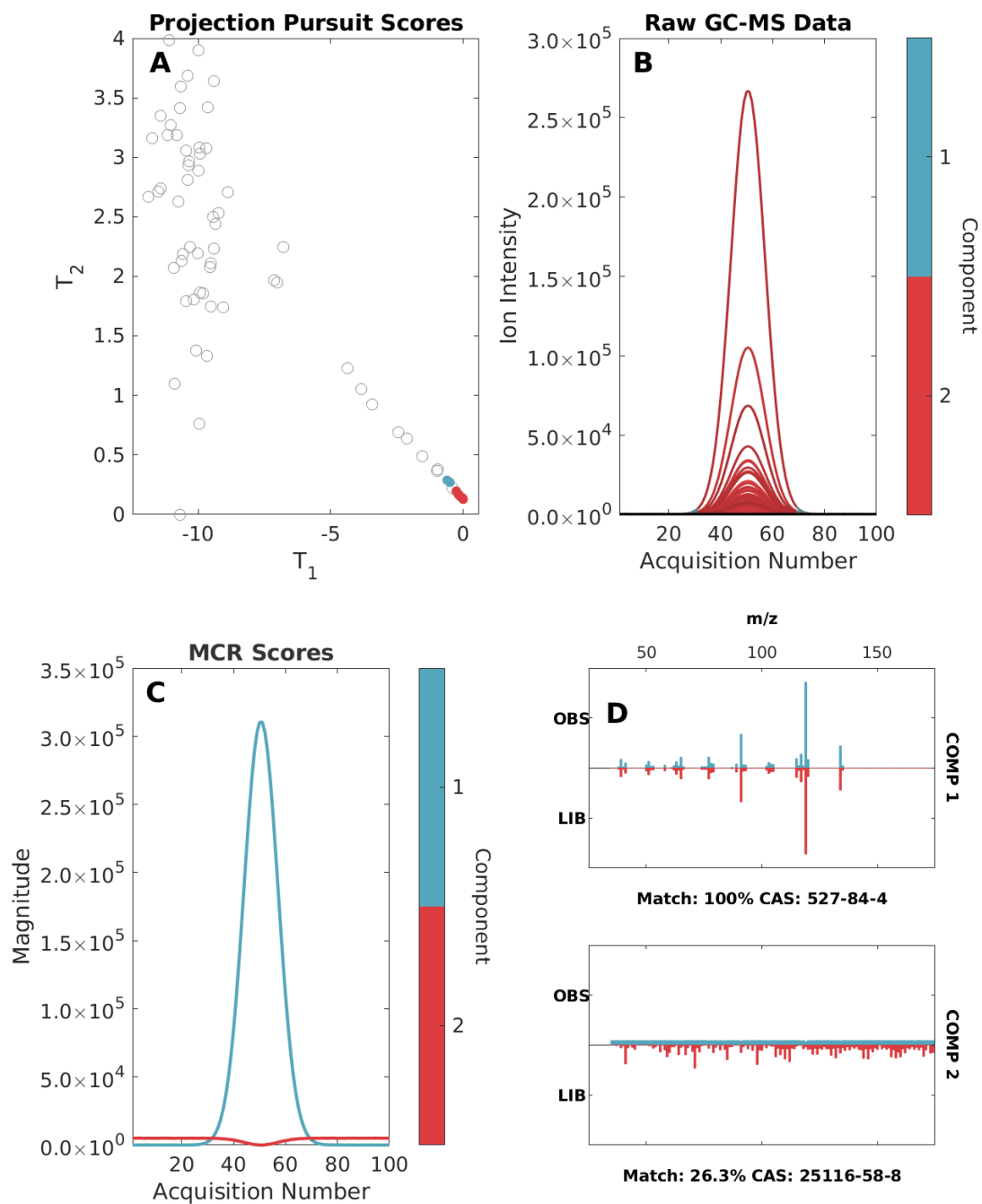


Figure D.9: Cumulative variances explained by component: 99.82 100.00

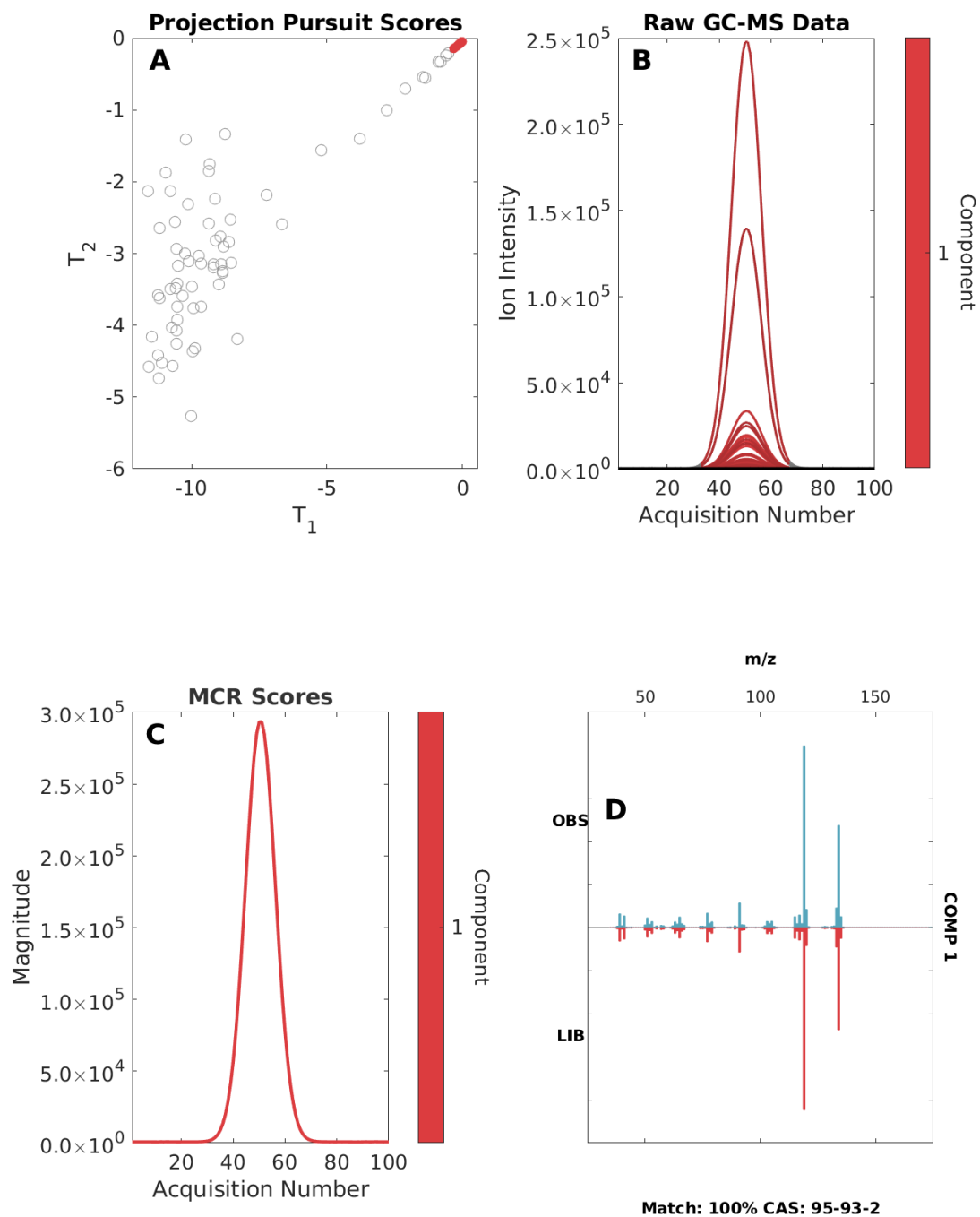


Figure D.10: Cumulative variances explained by component: 99.74

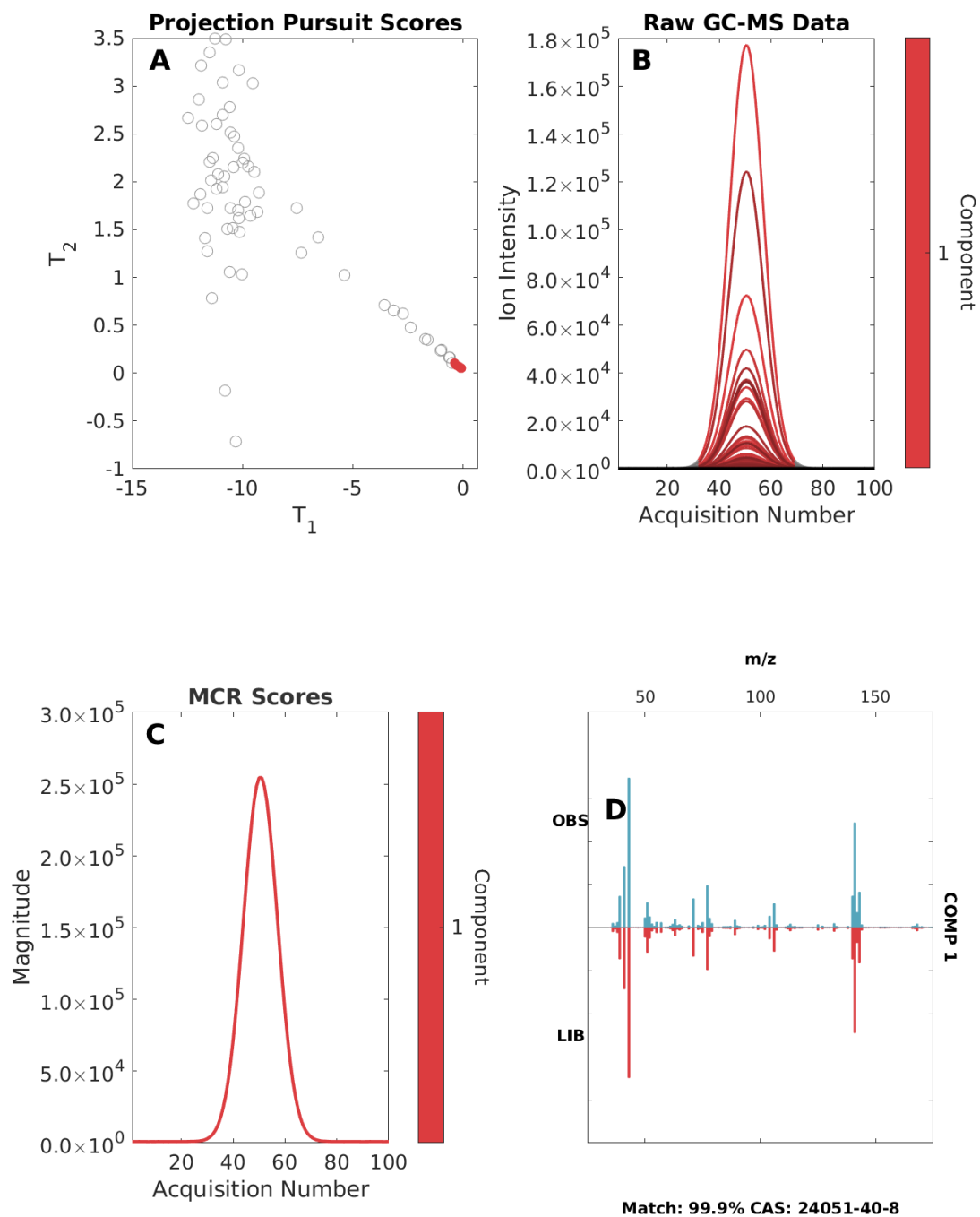


Figure D.11: Cumulative variances explained by component: 99.75

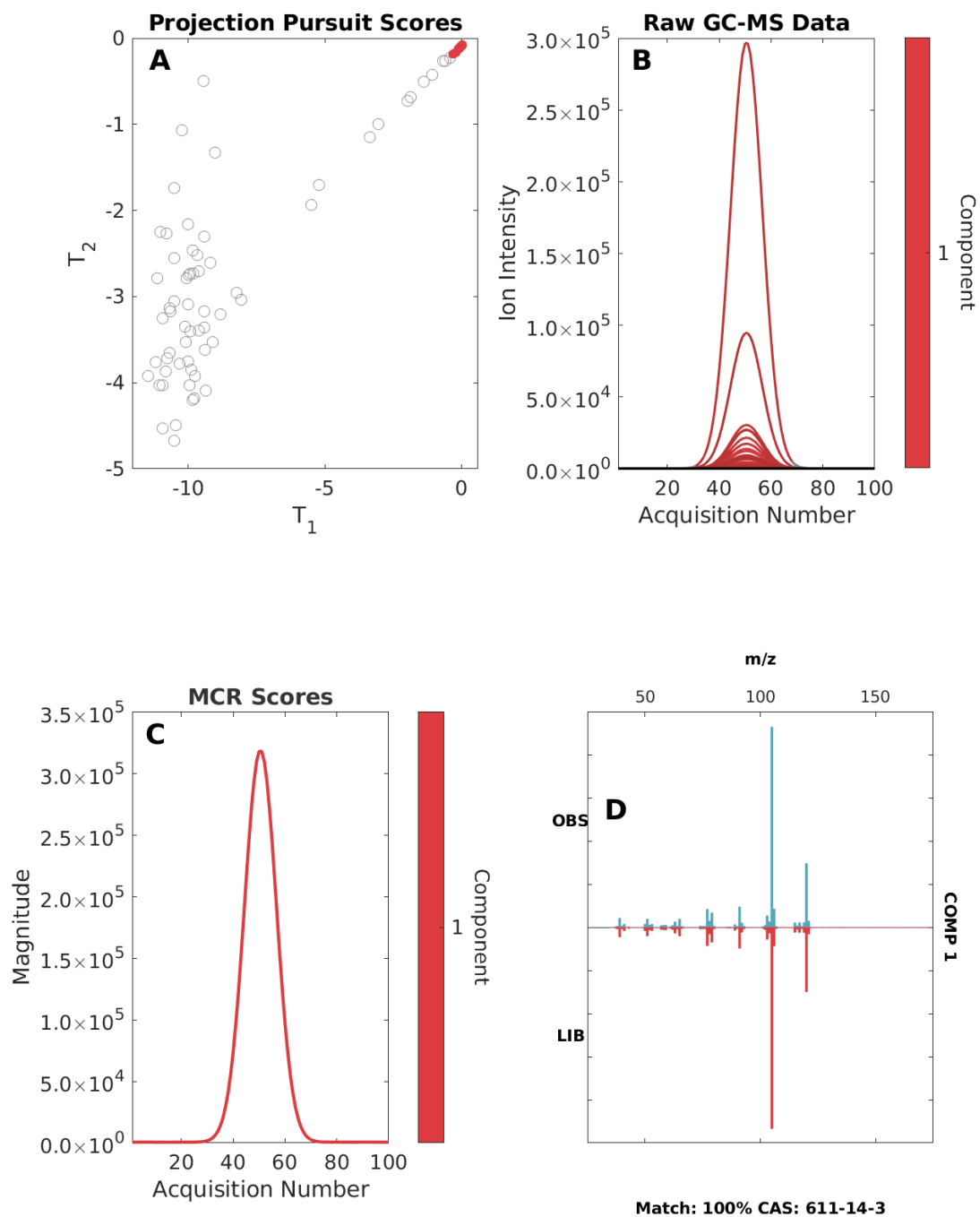


Figure D.12: Cumulative variances explained by component: 99.84 100.00

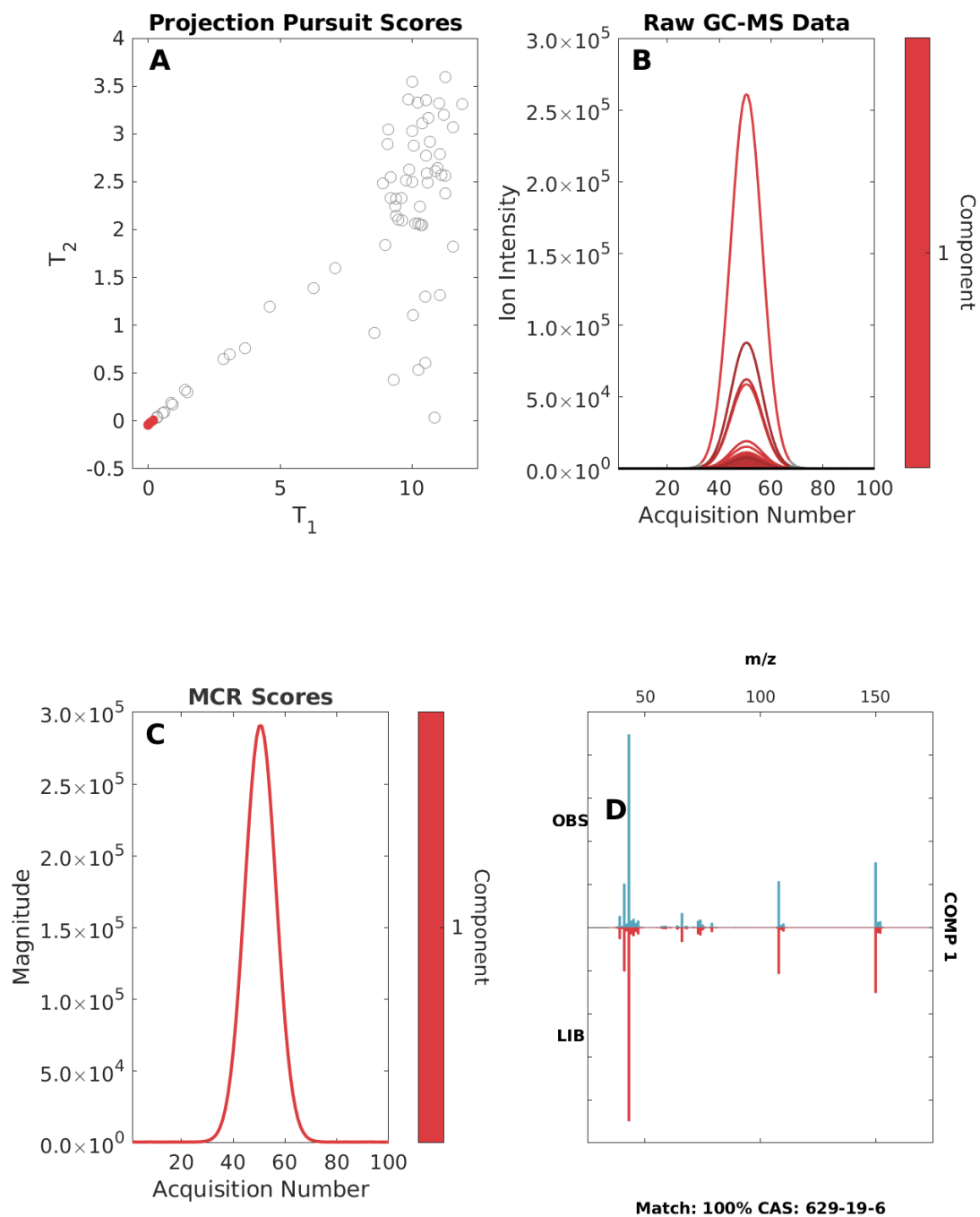


Figure D.13: Cumulative variances explained by component: 99.76

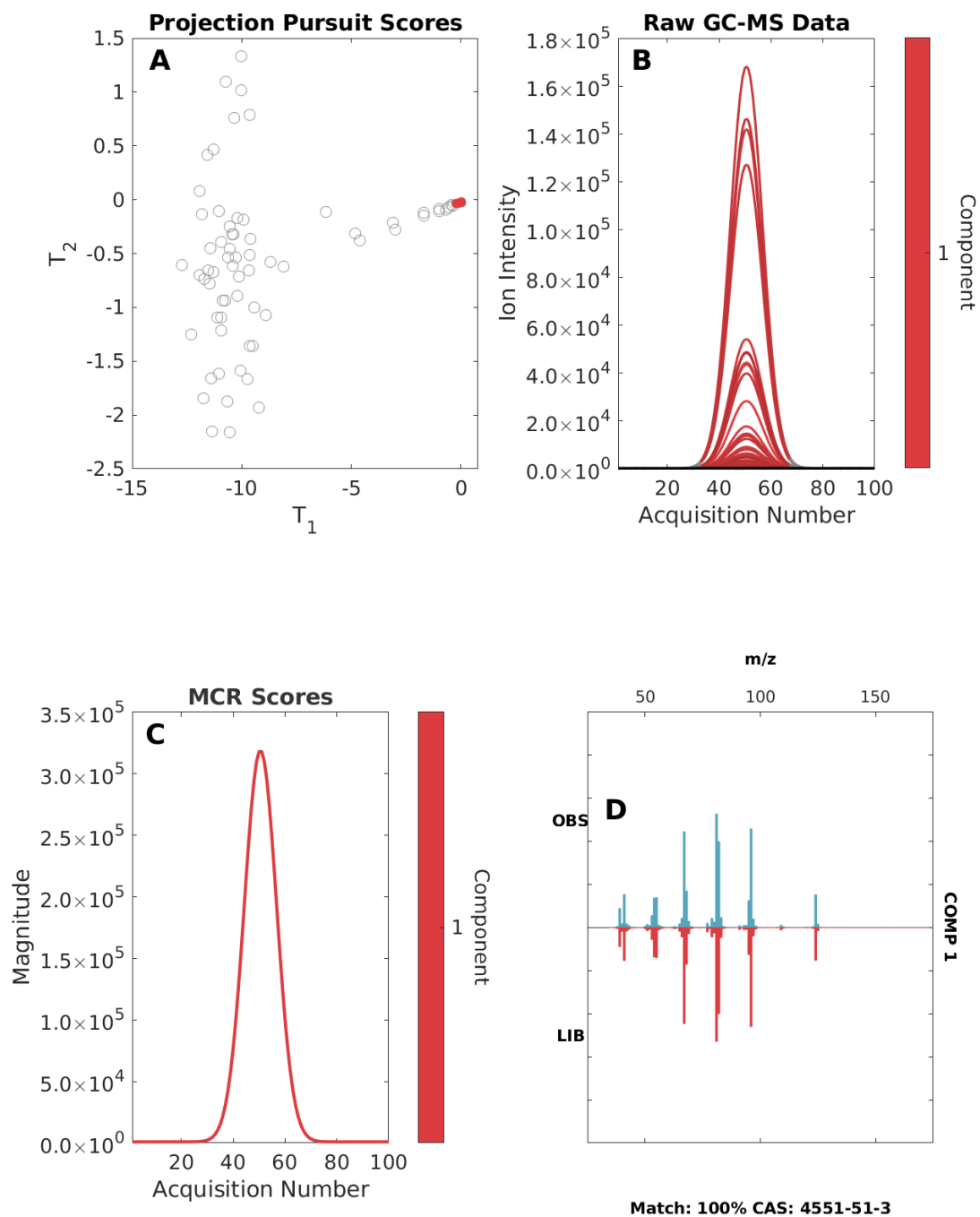


Figure D.14: Cumulative variances explained by component: 99.82

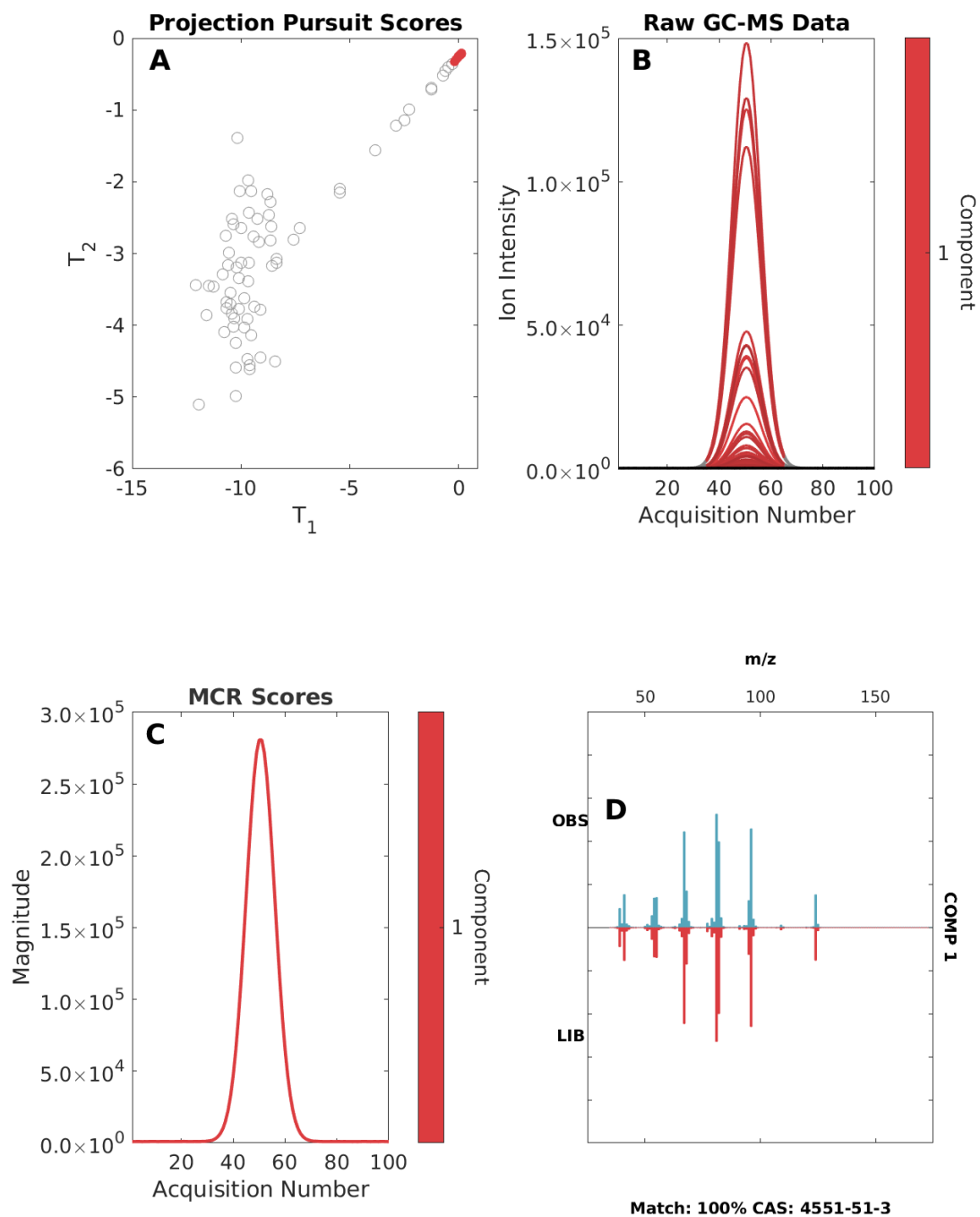


Figure D.15: Cumulative variances explained by component: 99.69

D.5.2 2 Factor Synthetic Data

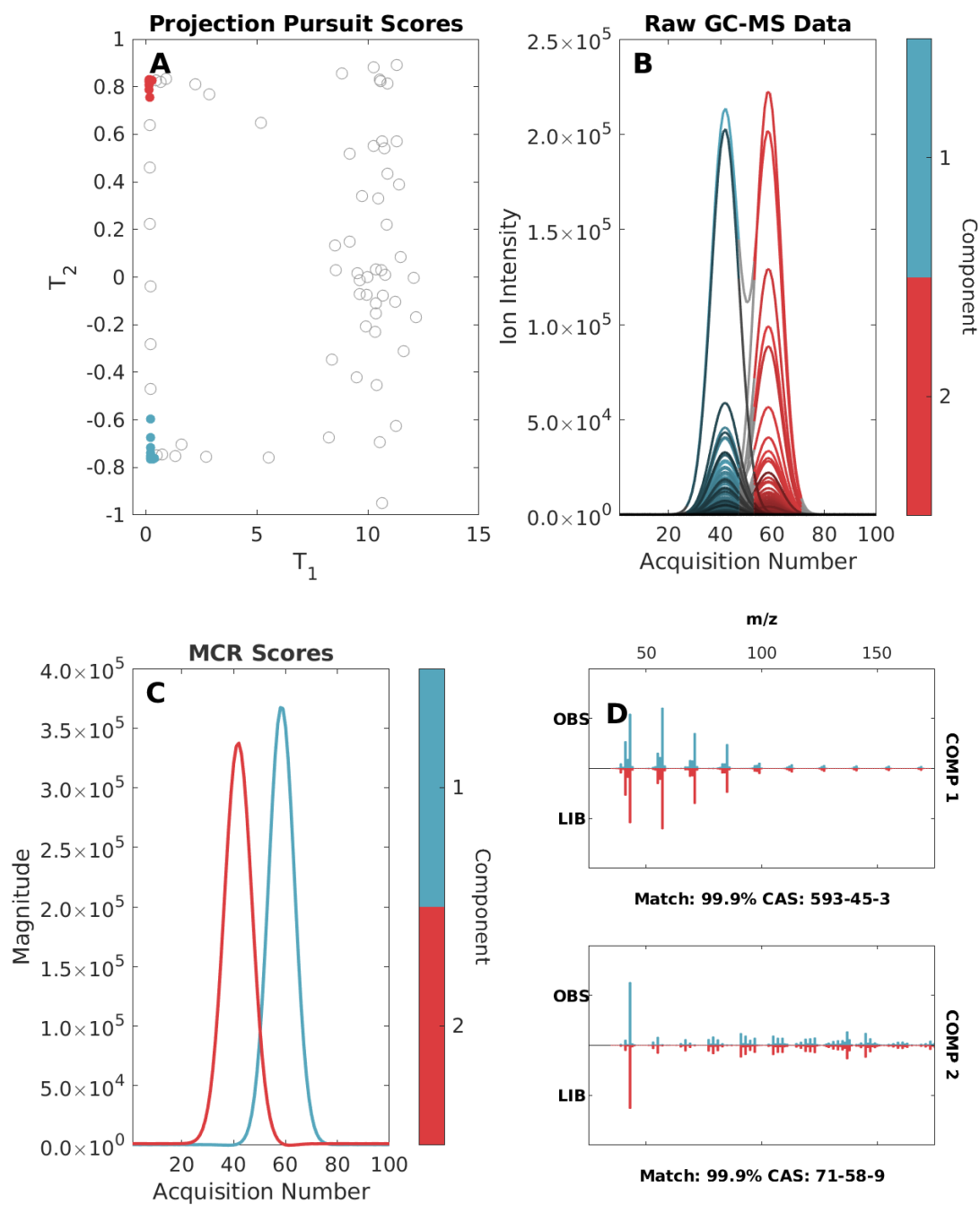


Figure D.16: Cumulative variances explained by component: 52.68 99.91

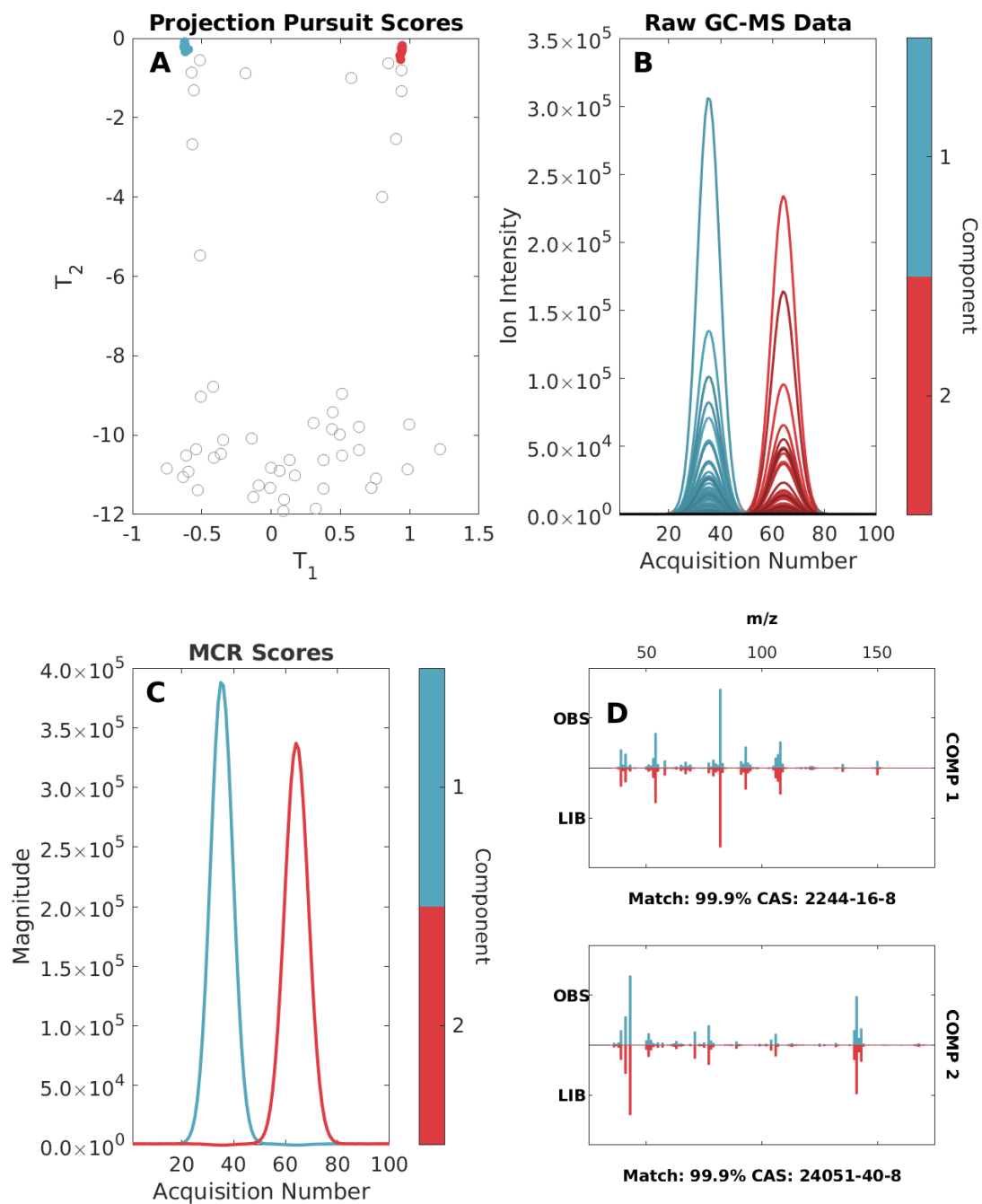


Figure D.17: Cumulative variances explained by component: 55.69 99.88

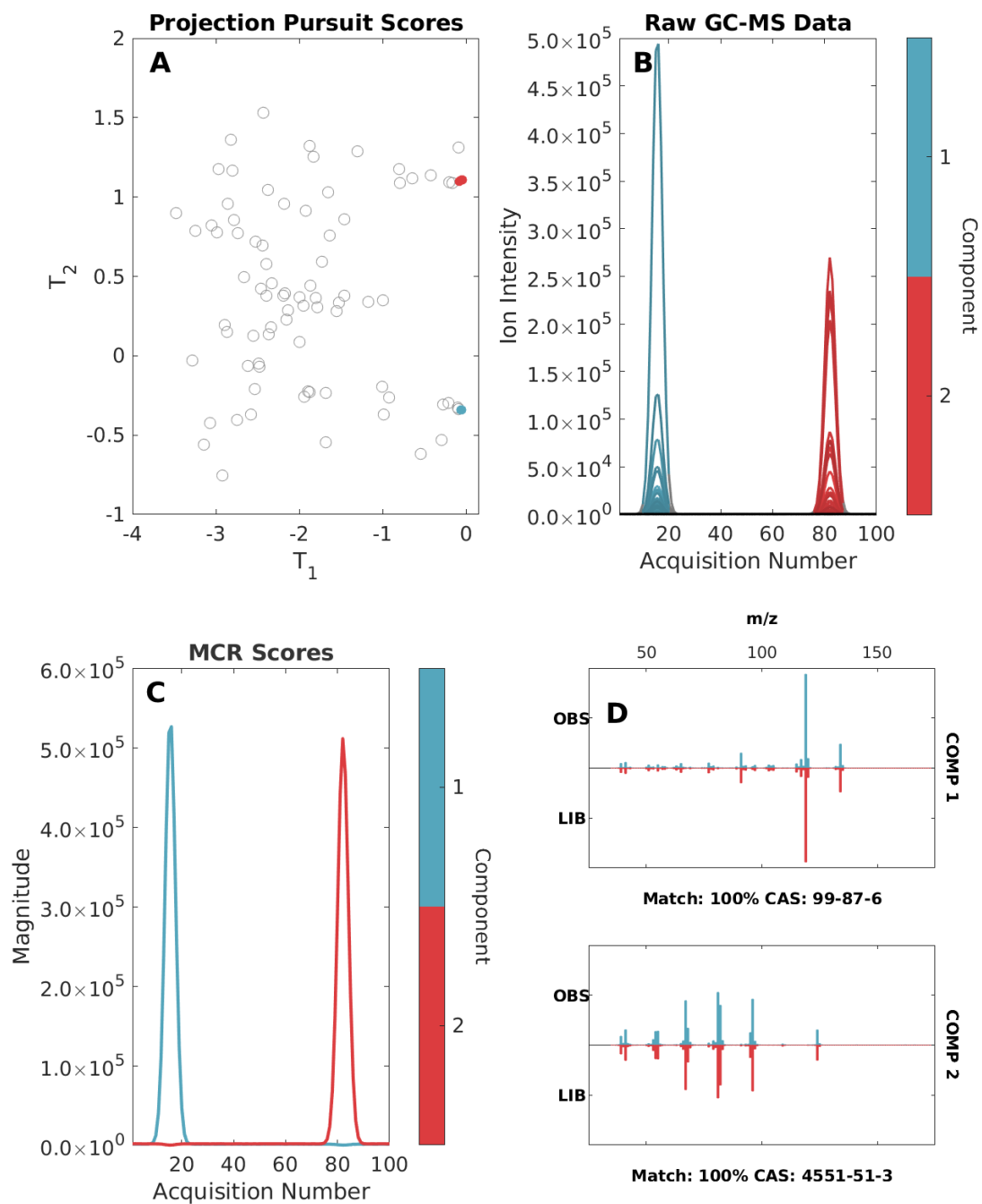


Figure D.18: Cumulative variances explained by component: 51.78 99.66

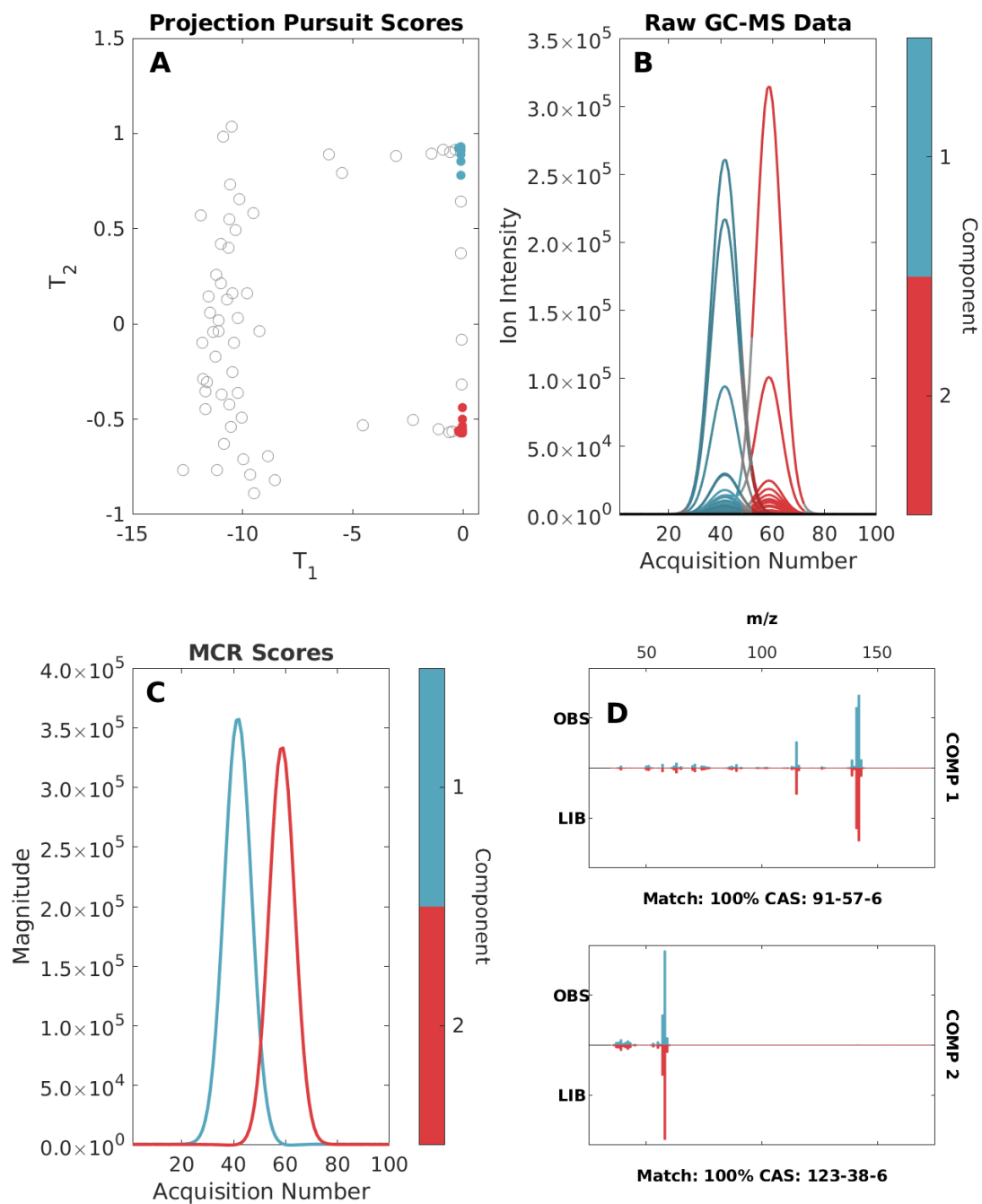


Figure D.19: Cumulative variances explained by component: 54.81 99.90

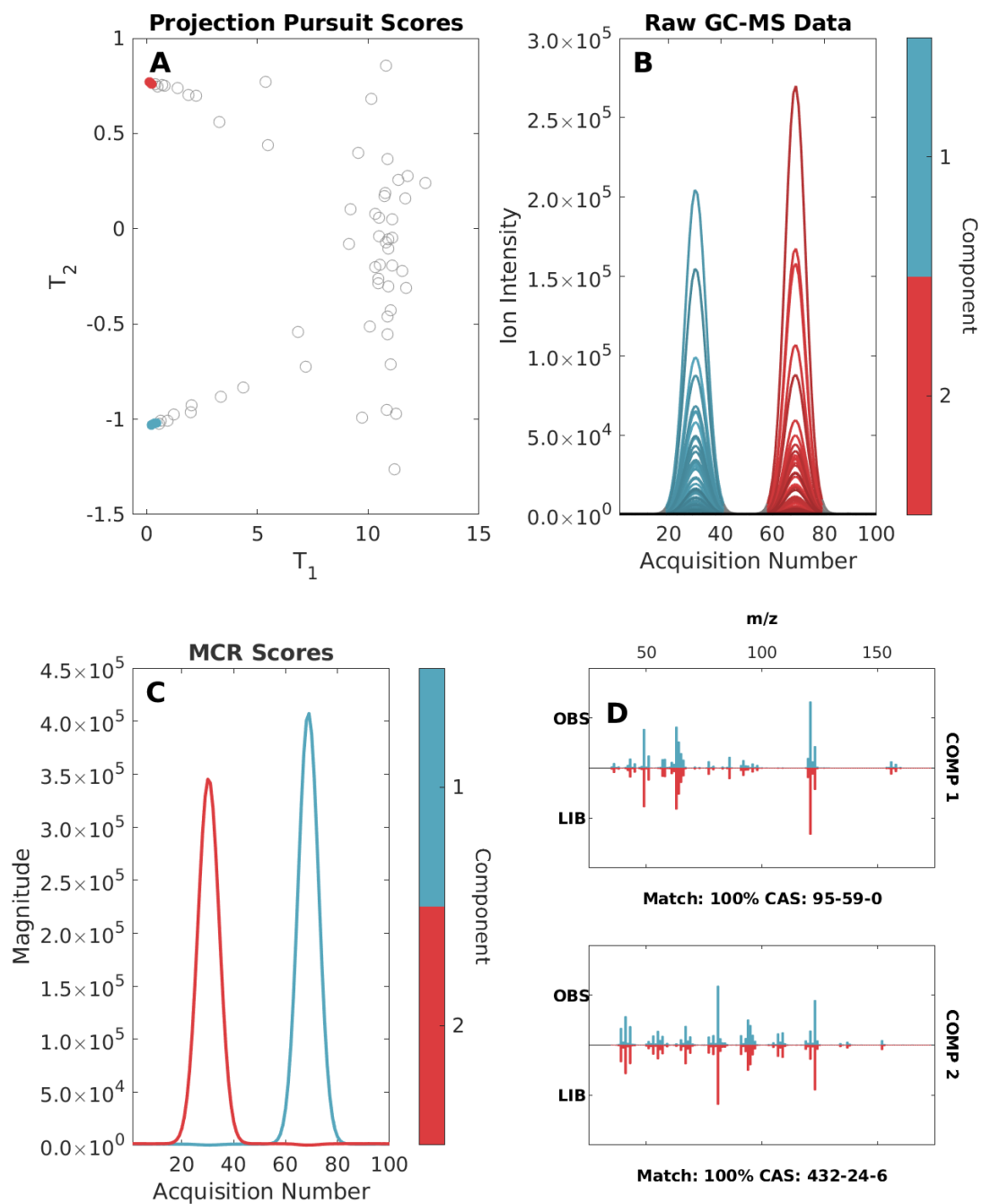


Figure D.20: Cumulative variances explained by component: 57.09 99.86

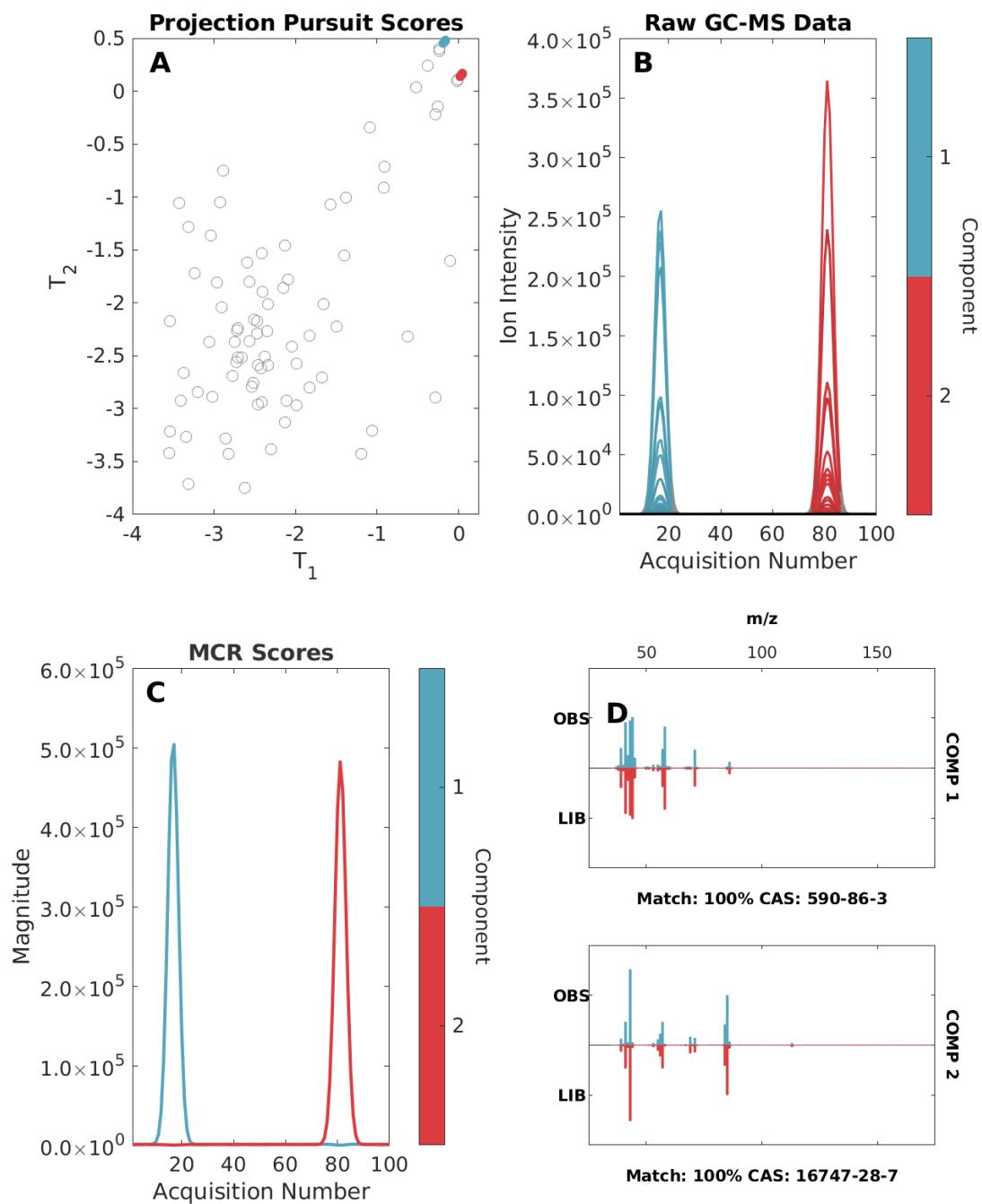


Figure D.21: Cumulative variances explained by component: 51.51 99.65

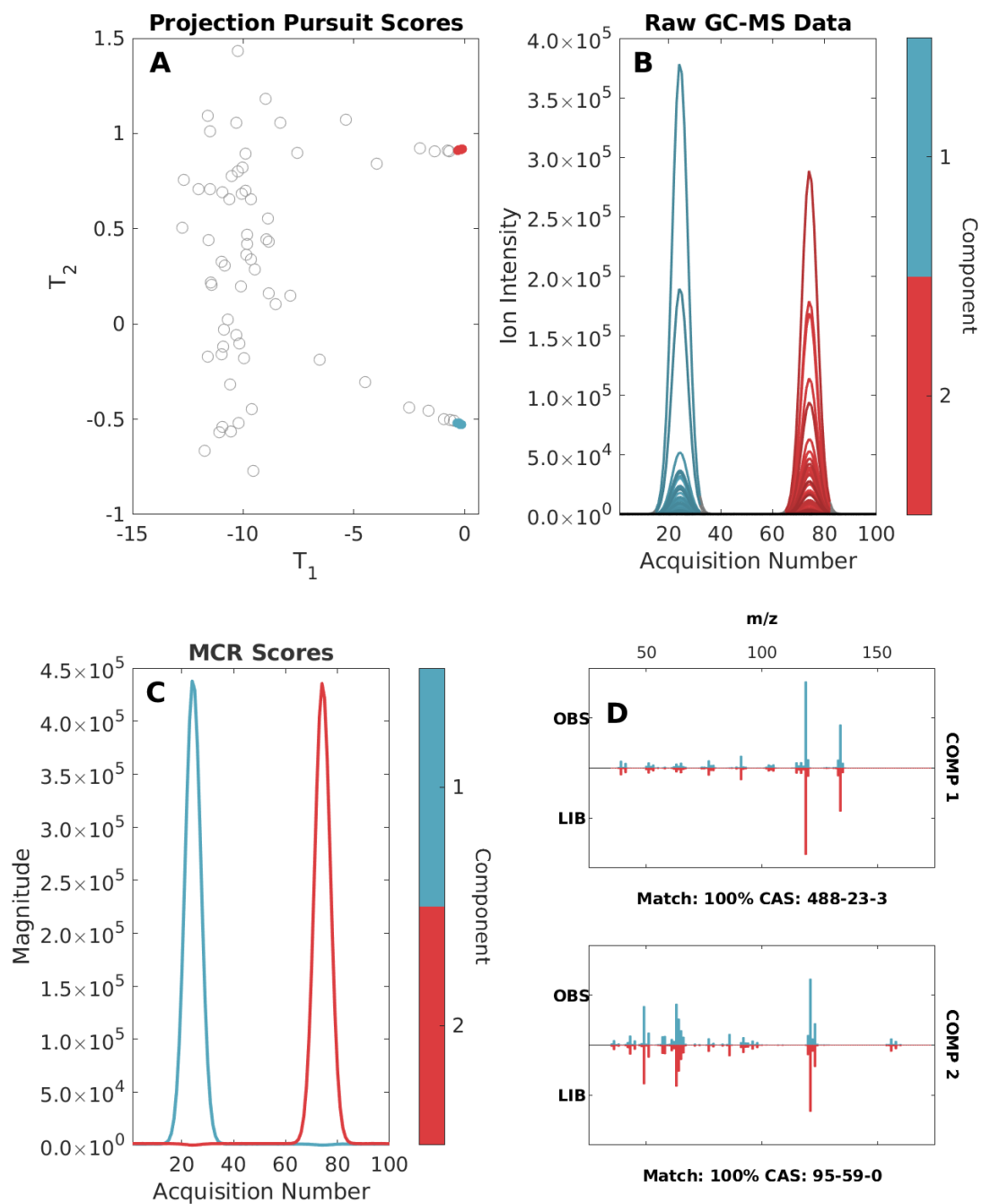


Figure D.22: Cumulative variances explained by component: 50.41 99.80

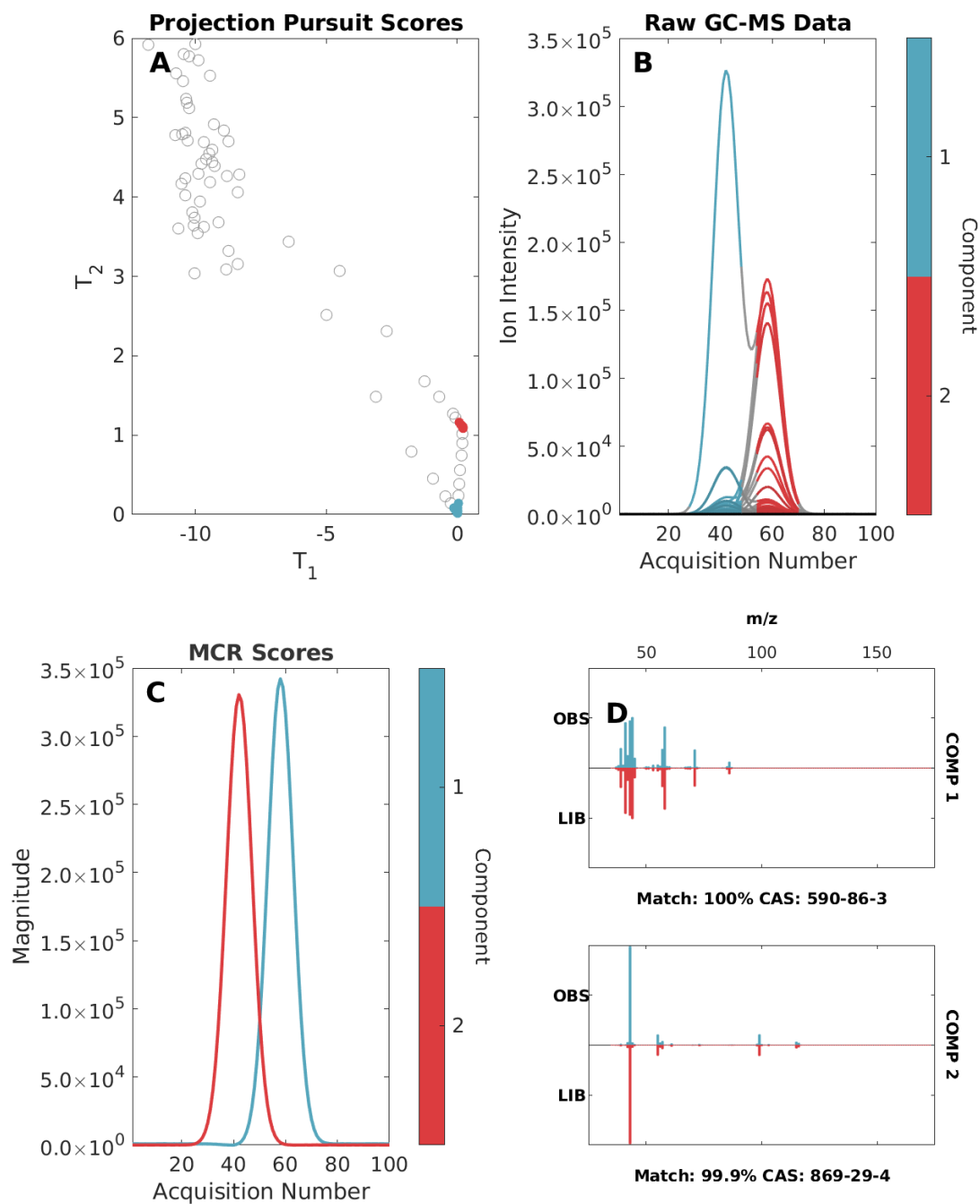


Figure D.23: Cumulative variances explained by component: 51.54 99.89

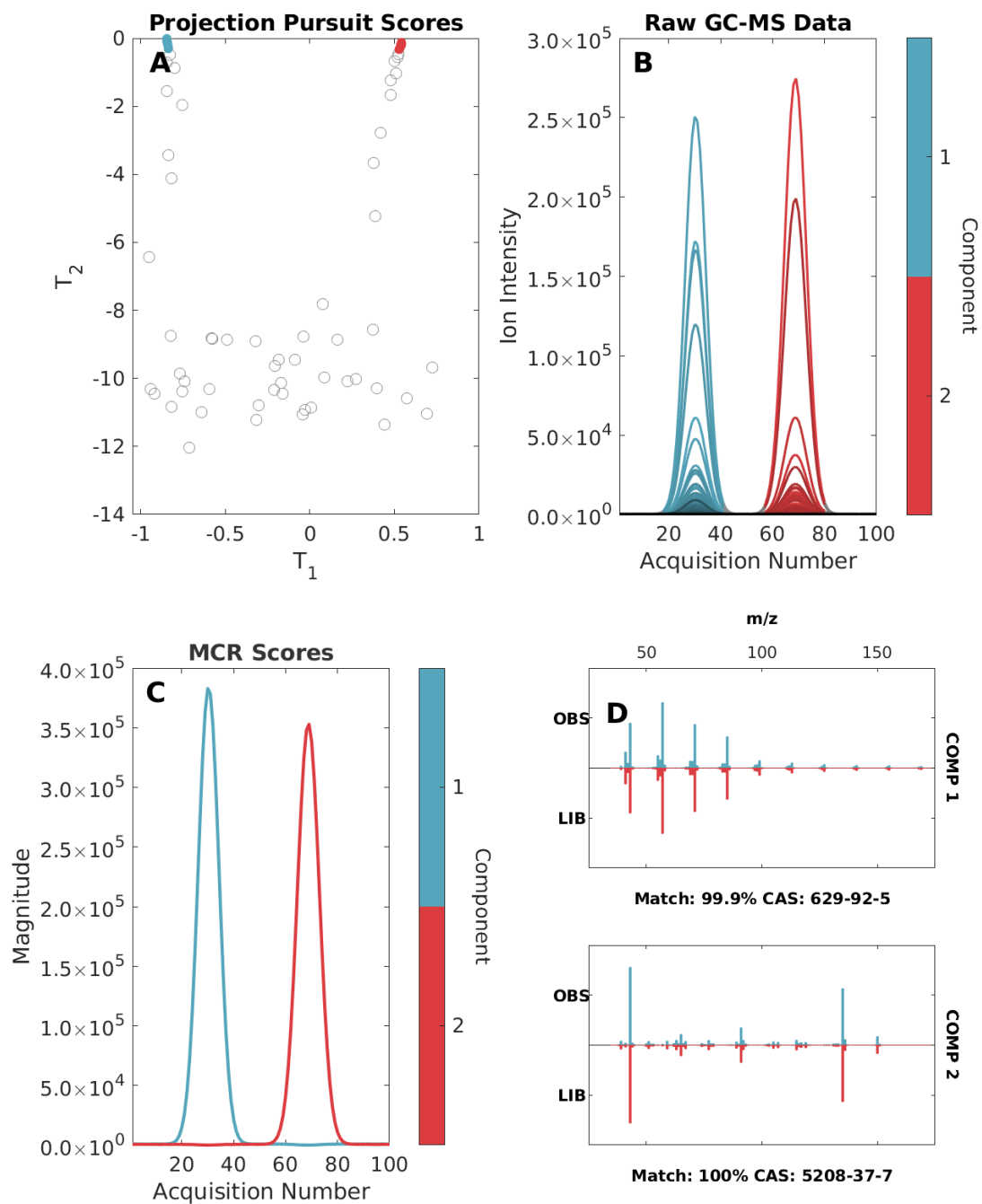


Figure D.24: Cumulative variances explained by component: 52.73 99.86

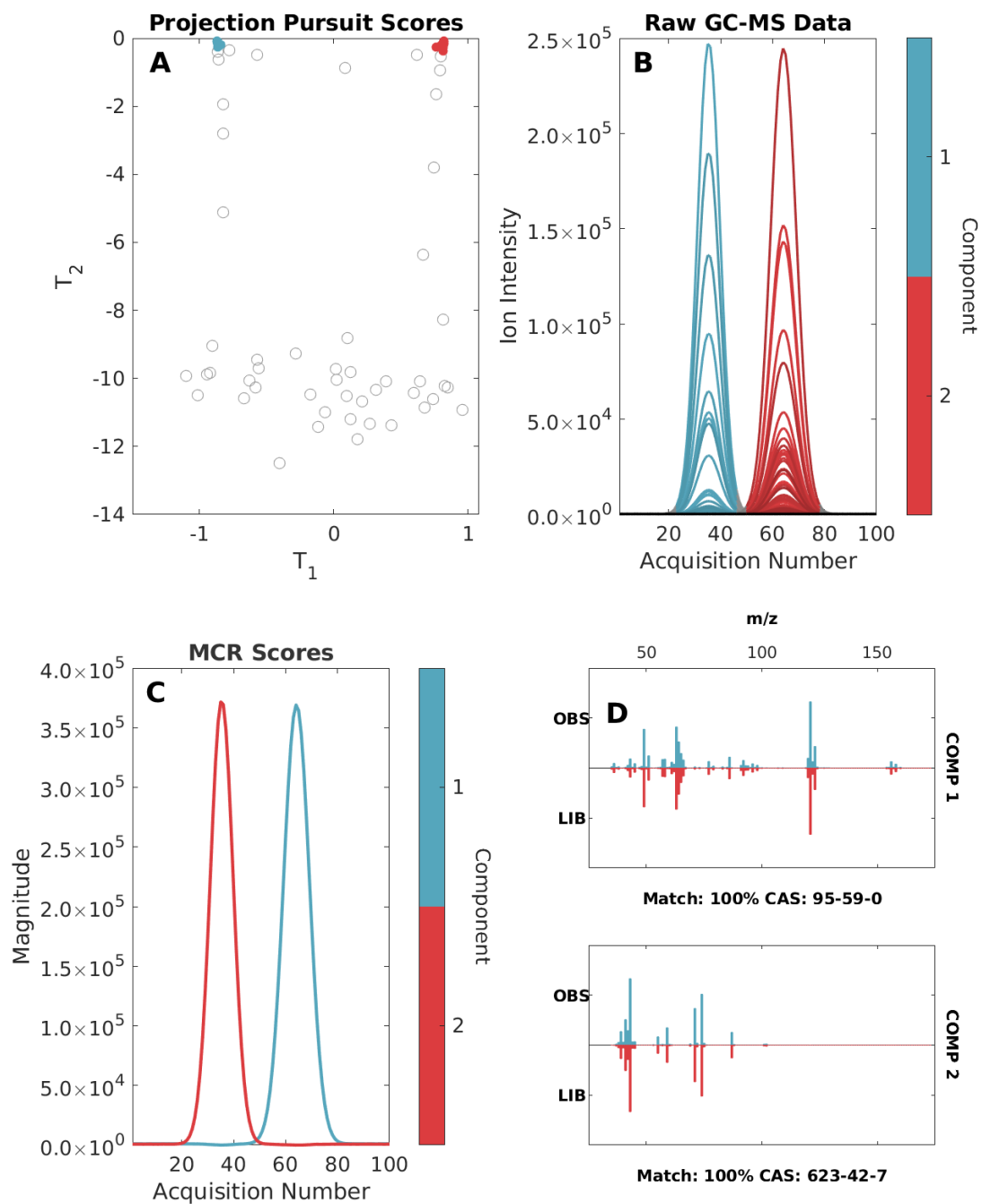


Figure D.25: Cumulative variances explained by component: 52.60 99.90

D.5.3 3 Factor Synthetic Data

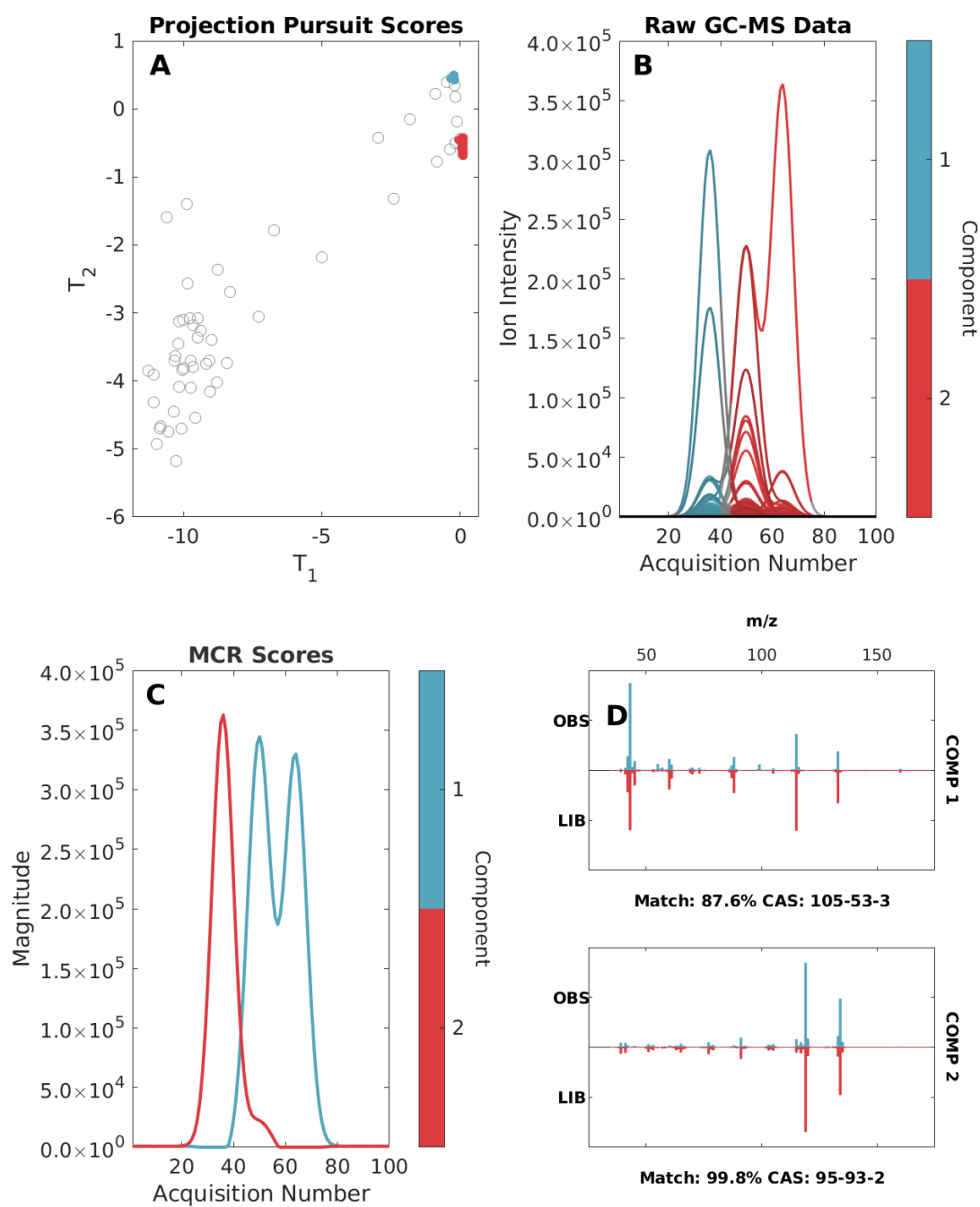


Figure D.26: Cumulative variances explained by component: 58.01 88.34

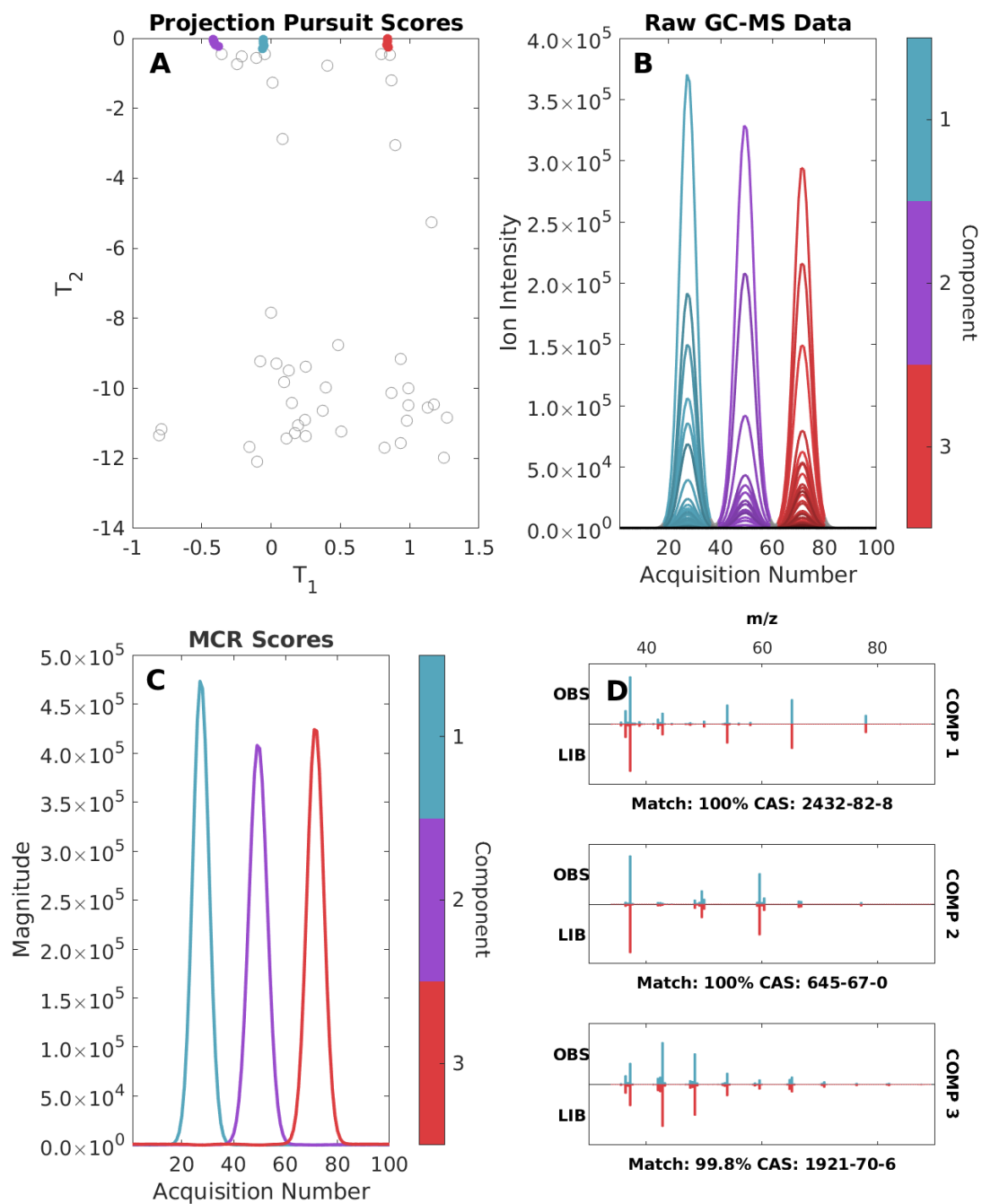


Figure D.27: Cumulative variances explained by component: 37.16 68.44 99.90

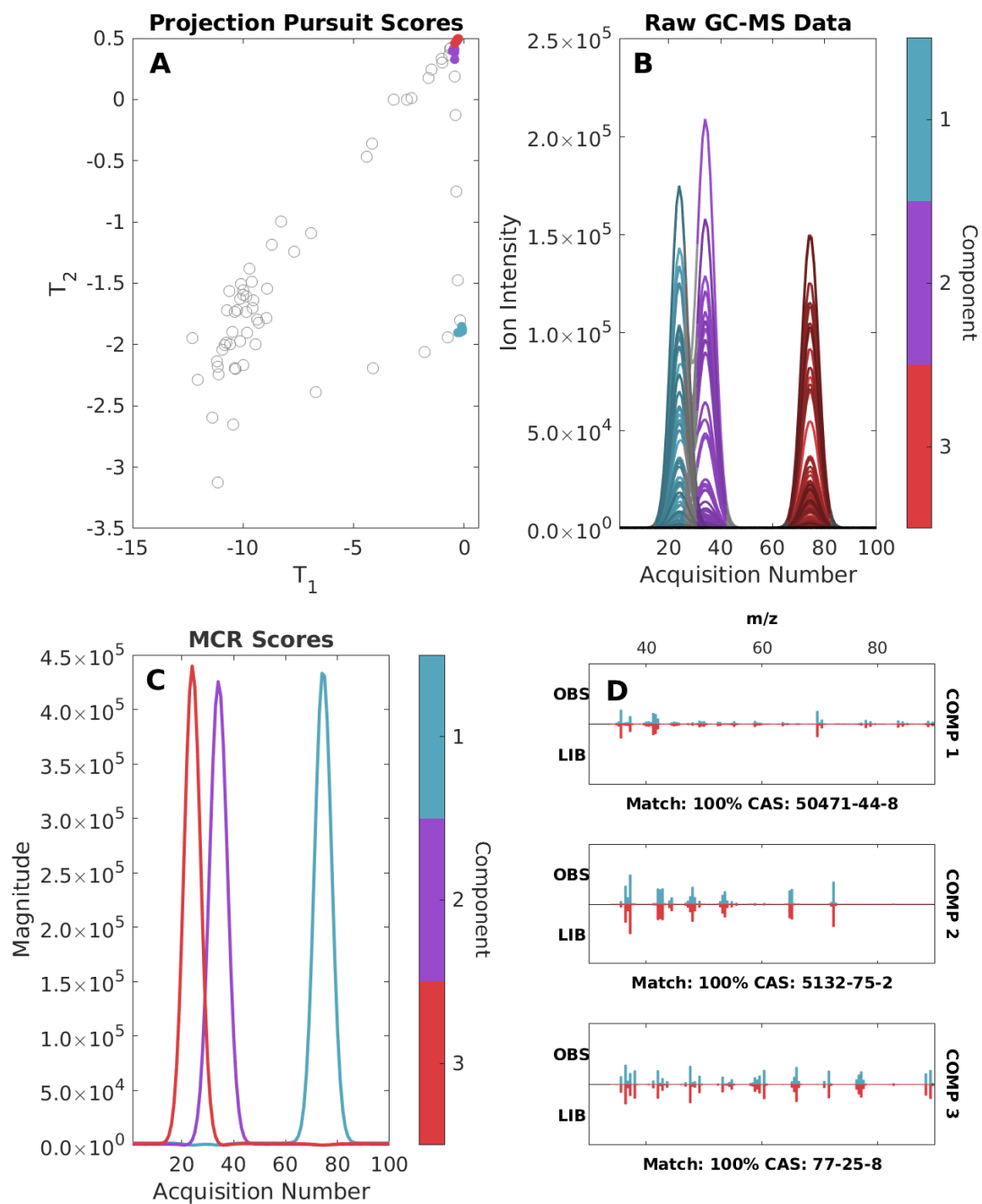


Figure D.28: Cumulative variances explained by component: 33.63 66.98 99.91

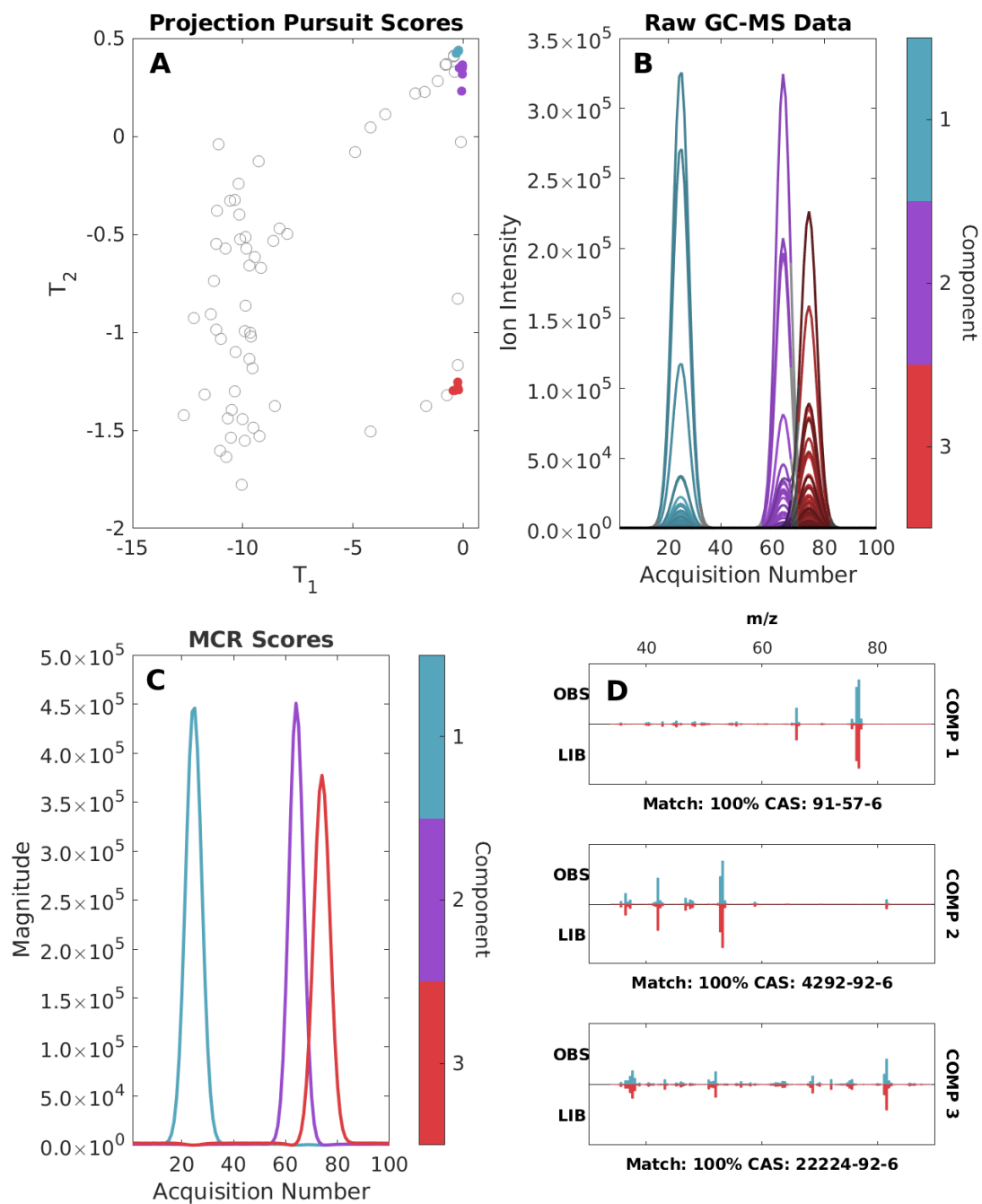


Figure D.29: Cumulative variances explained by component: 38.11 72.75 99.89

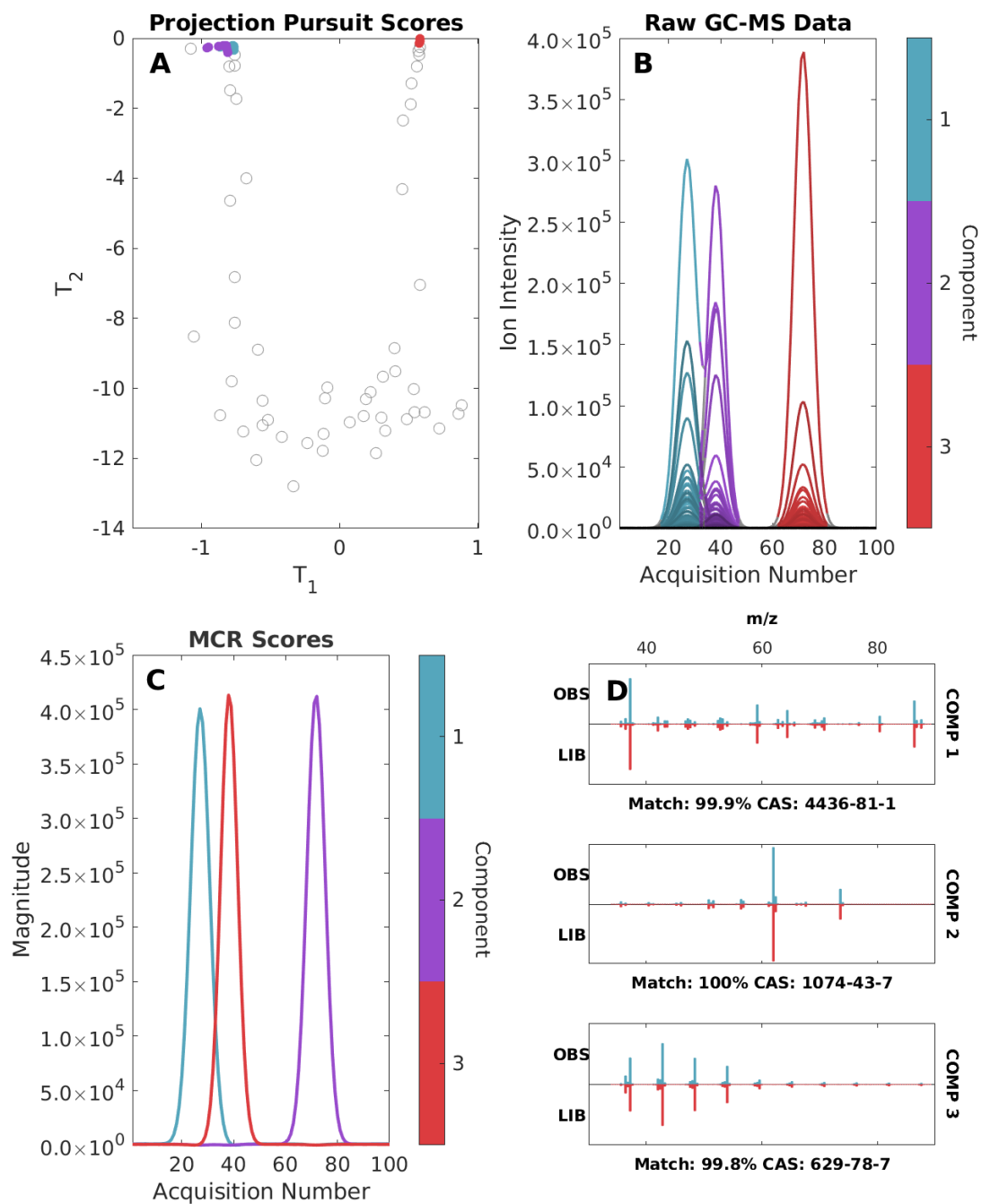


Figure D.30: Cumulative variances explained by component: 33.49 67.84 99.91

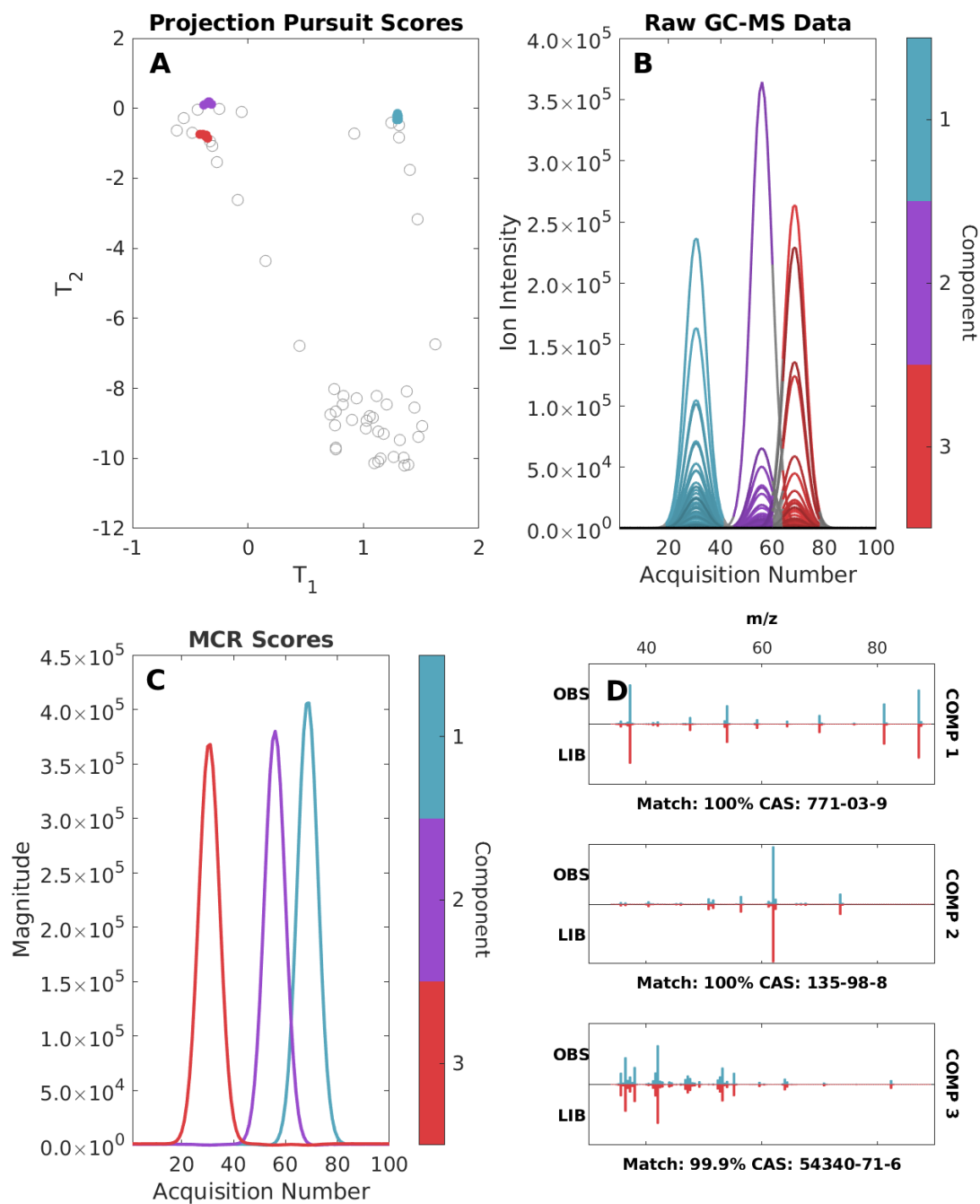


Figure D.31: Cumulative variances explained by component: 36.92 69.27 99.92

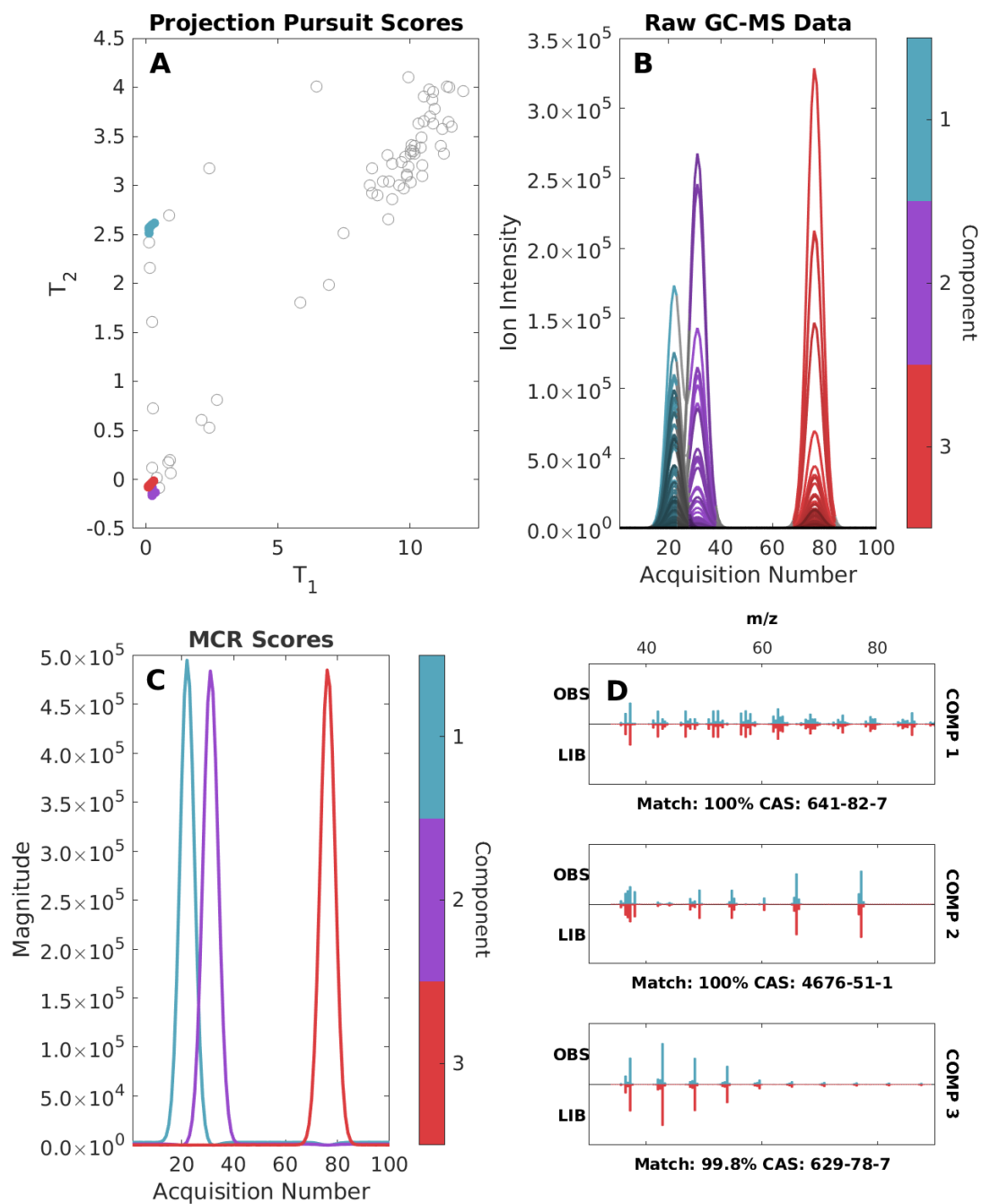


Figure D.32: Cumulative variances explained by component: 33.12 66.77 99.90

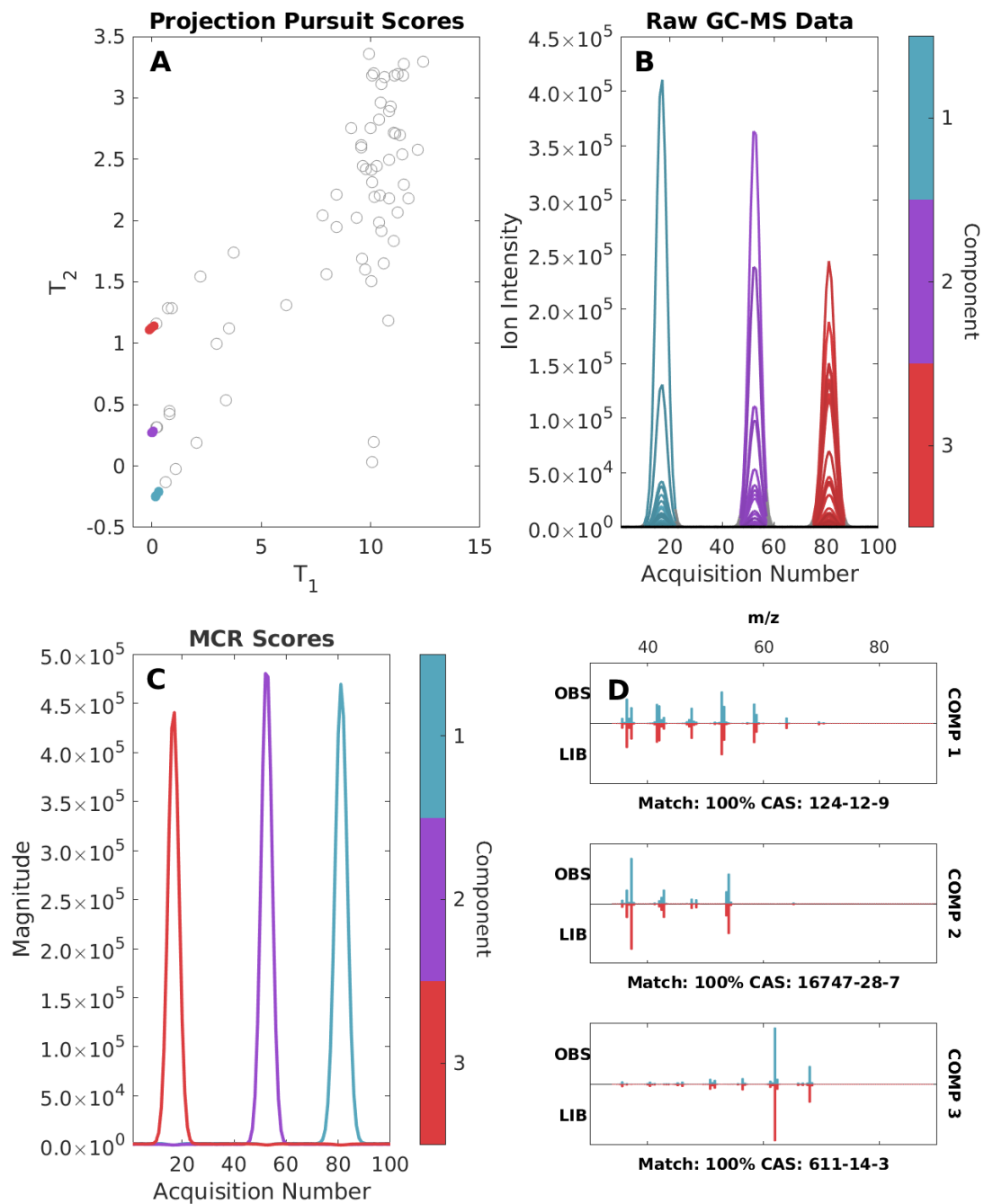


Figure D.33: Cumulative variances explained by component: 35.54 70.71 99.77

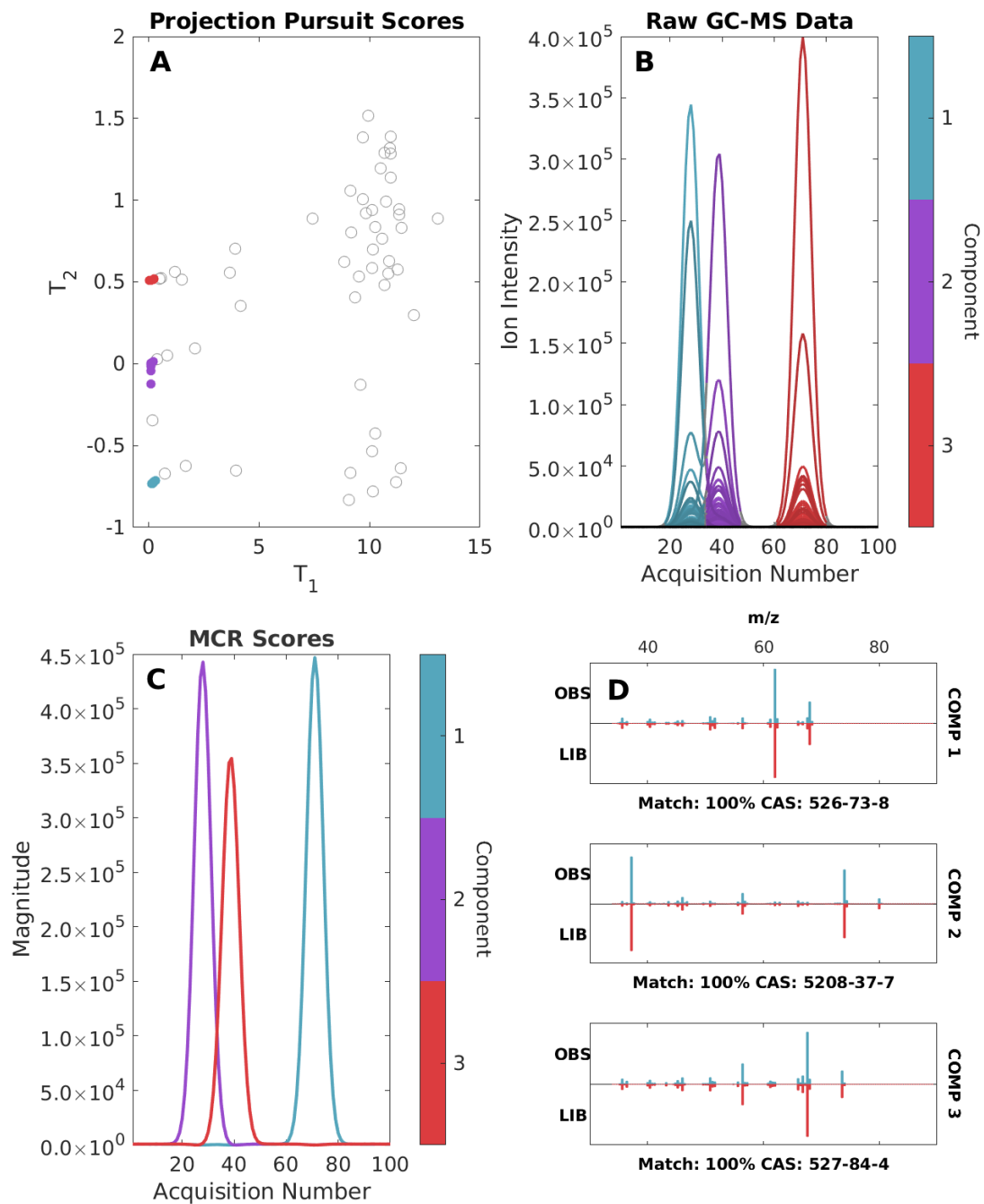


Figure D.34: Cumulative variances explained by component: 38.15 75.22 99.89

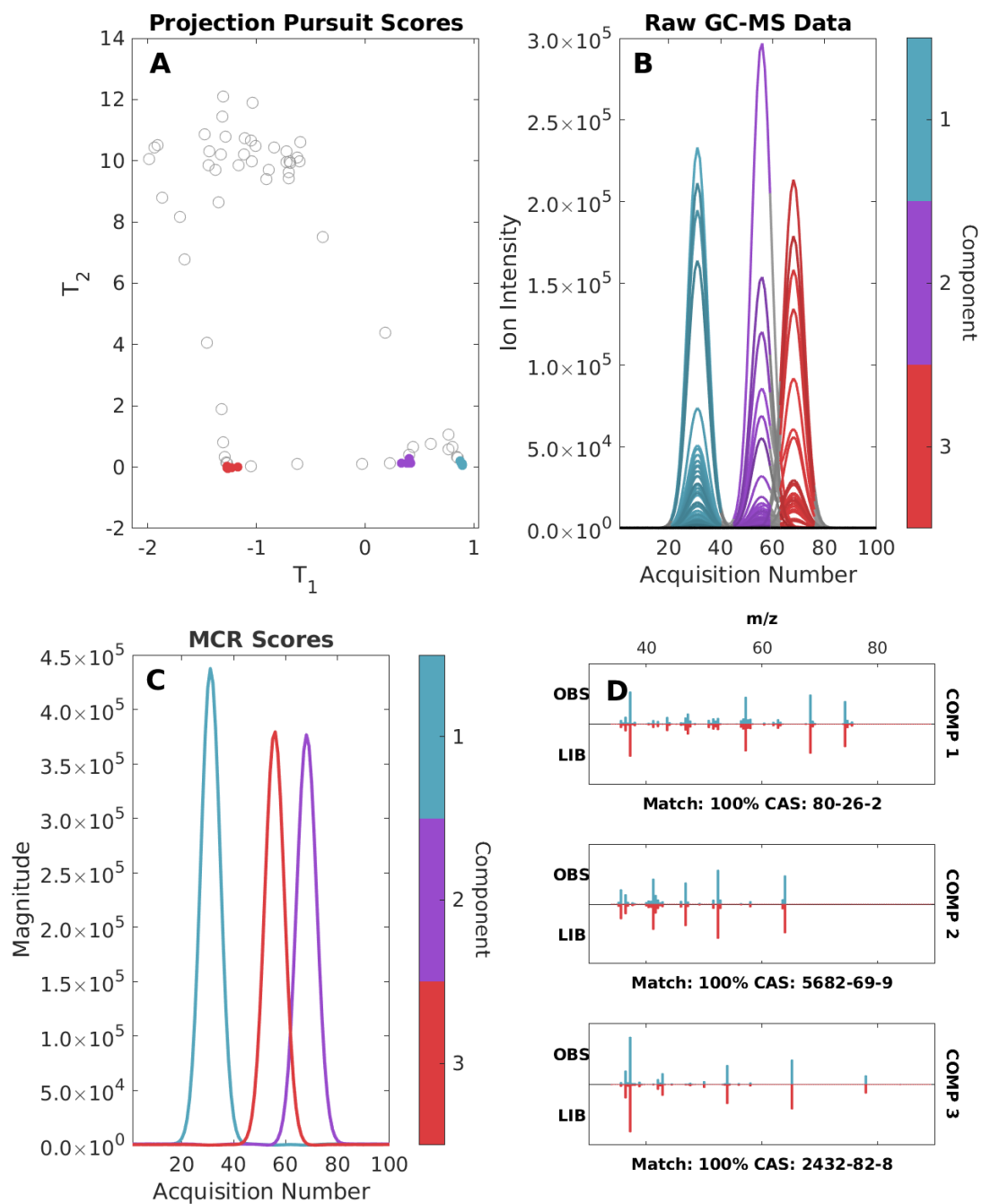


Figure D.35: Cumulative variances explained by component: 39.42 69.76 99.92

D.5.4 4 Factor Synthetic Data

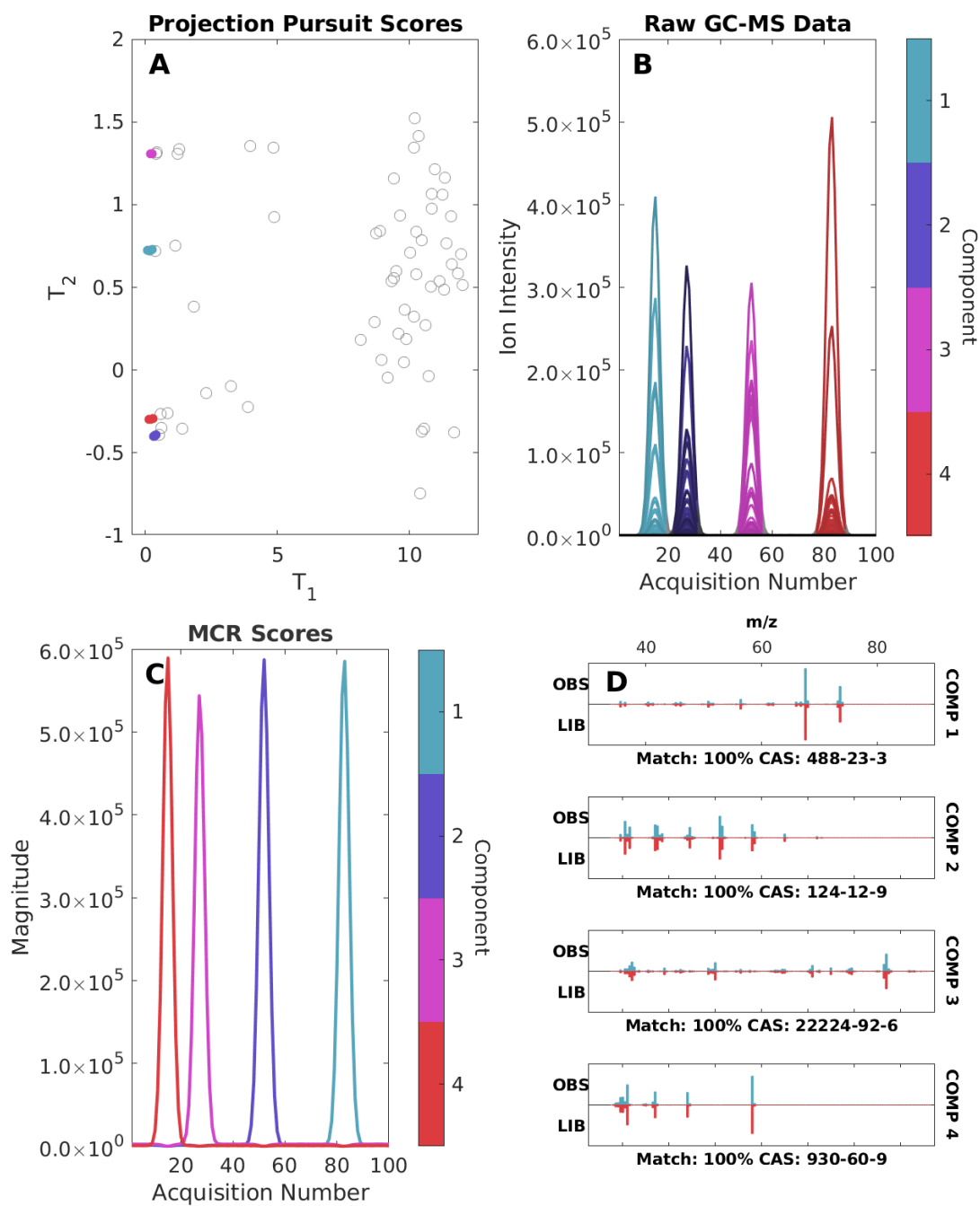


Figure D.36: Cumulative variances explained by component: 26.58 52.67 74.78 99.87

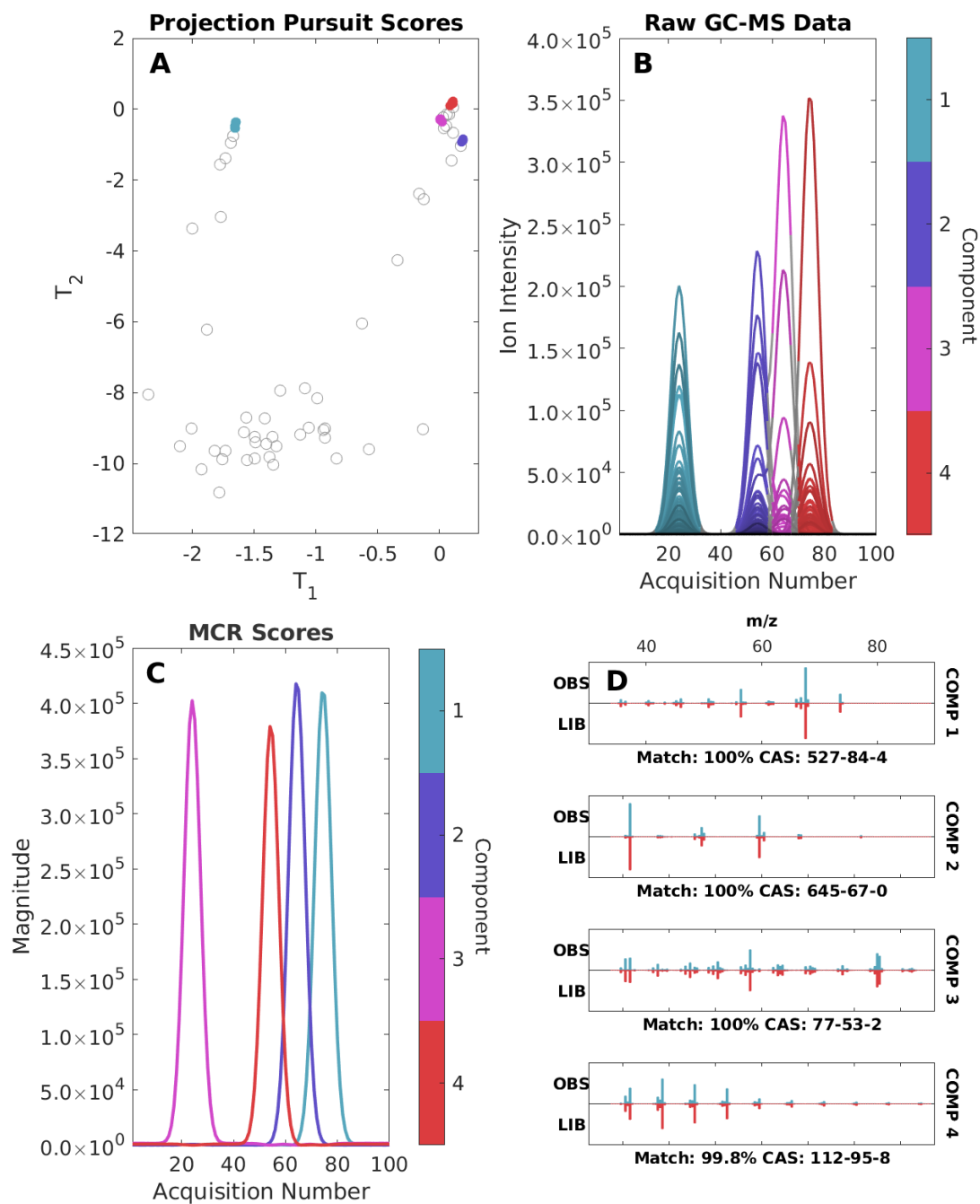


Figure D.37: Cumulative variances explained by component: 26.51 53.65 78.29 99.93

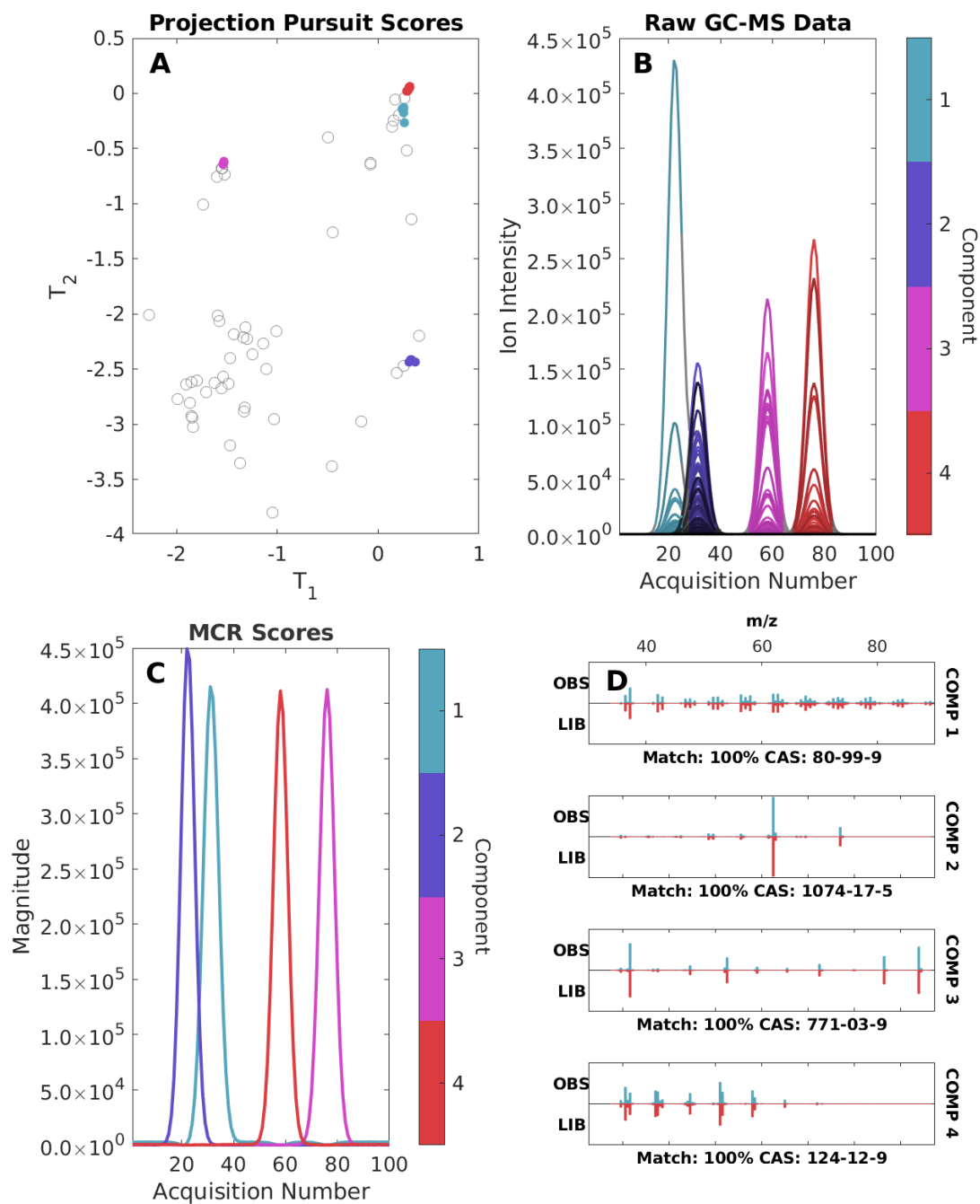


Figure D.38: Cumulative variances explained by component: 25.52 52.57 76.81 99.91

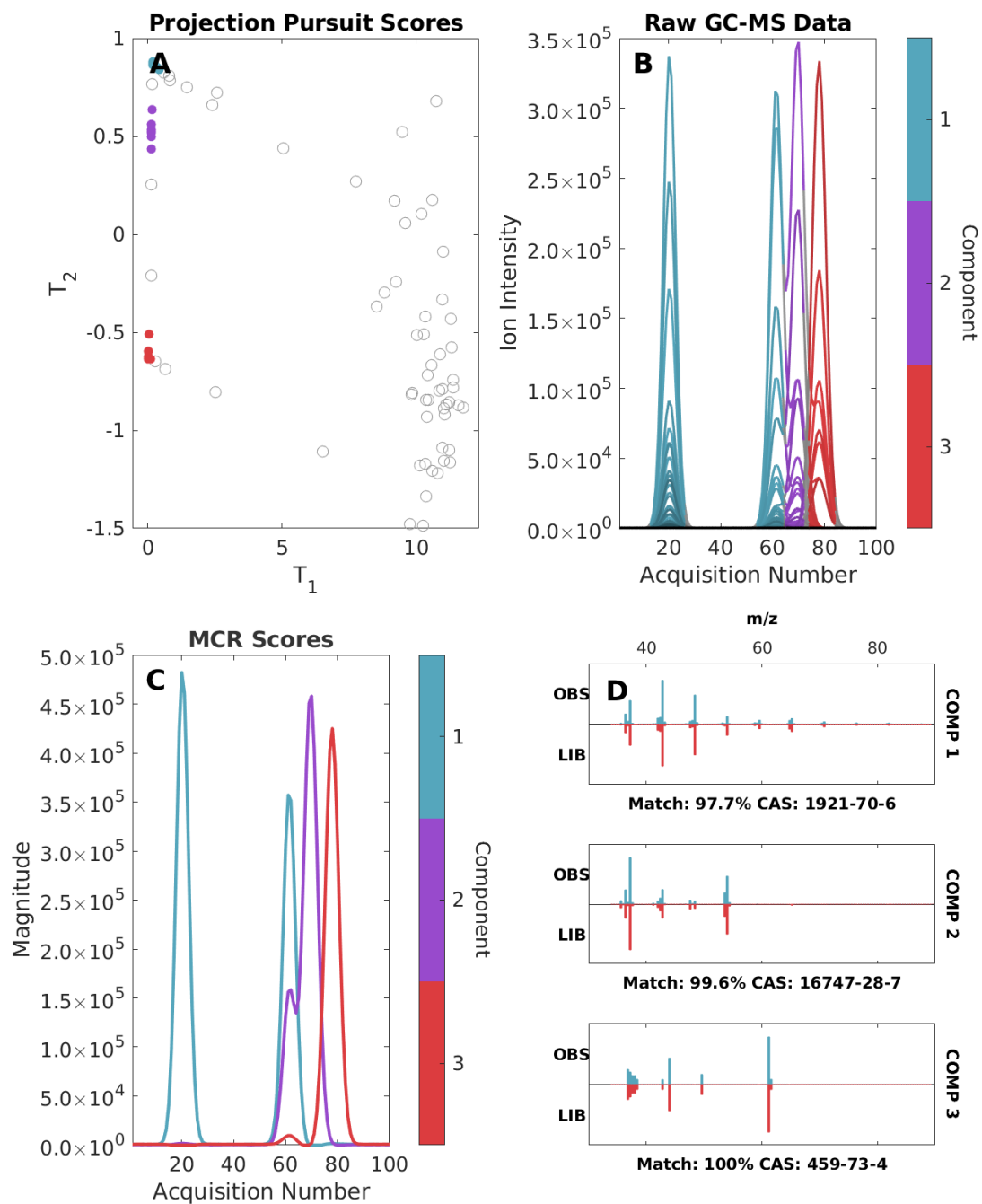


Figure D.39: Cumulative variances explained by component: 43.26 75.39 98.99

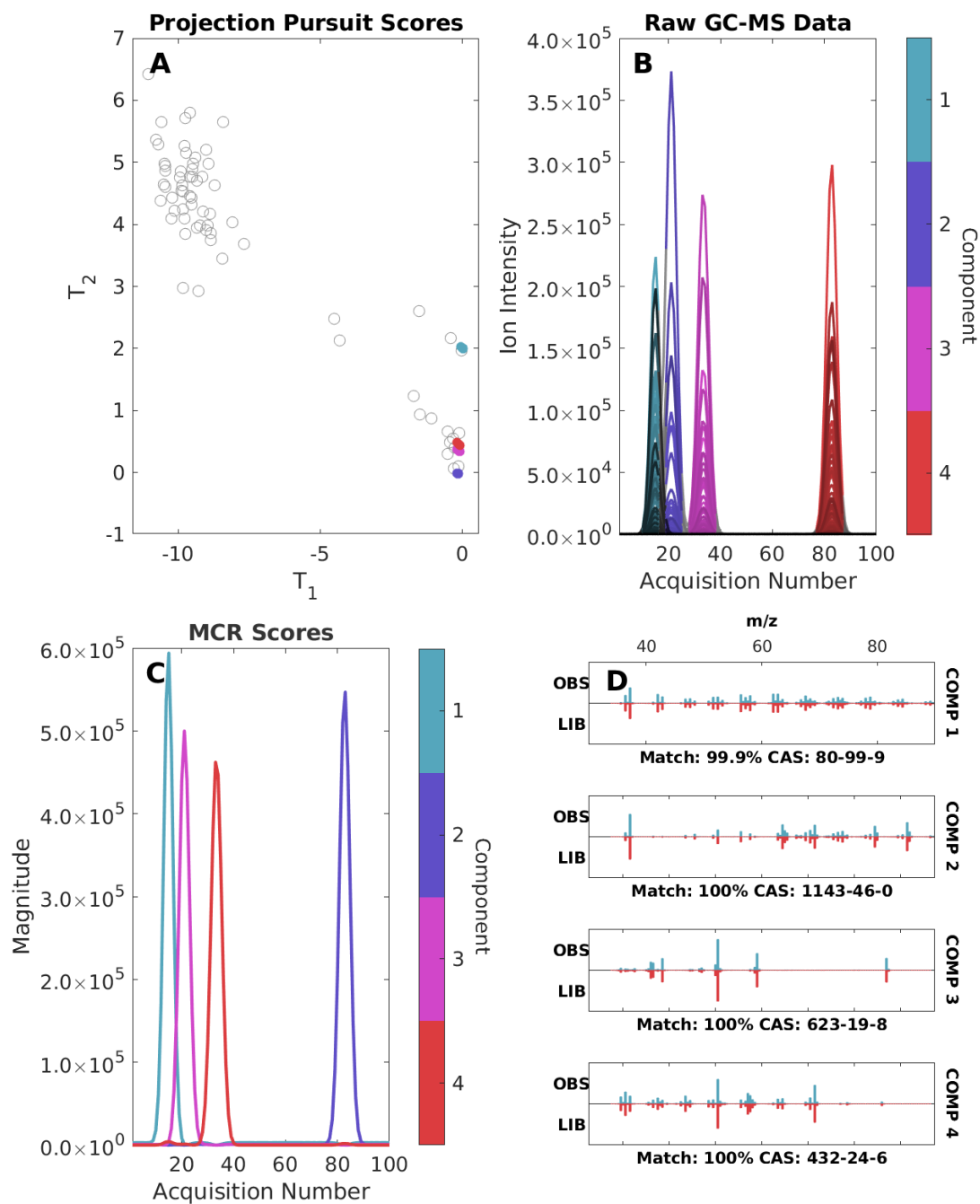


Figure D.40: Cumulative variances explained by component: 28.98 56.51 79.19 99.86

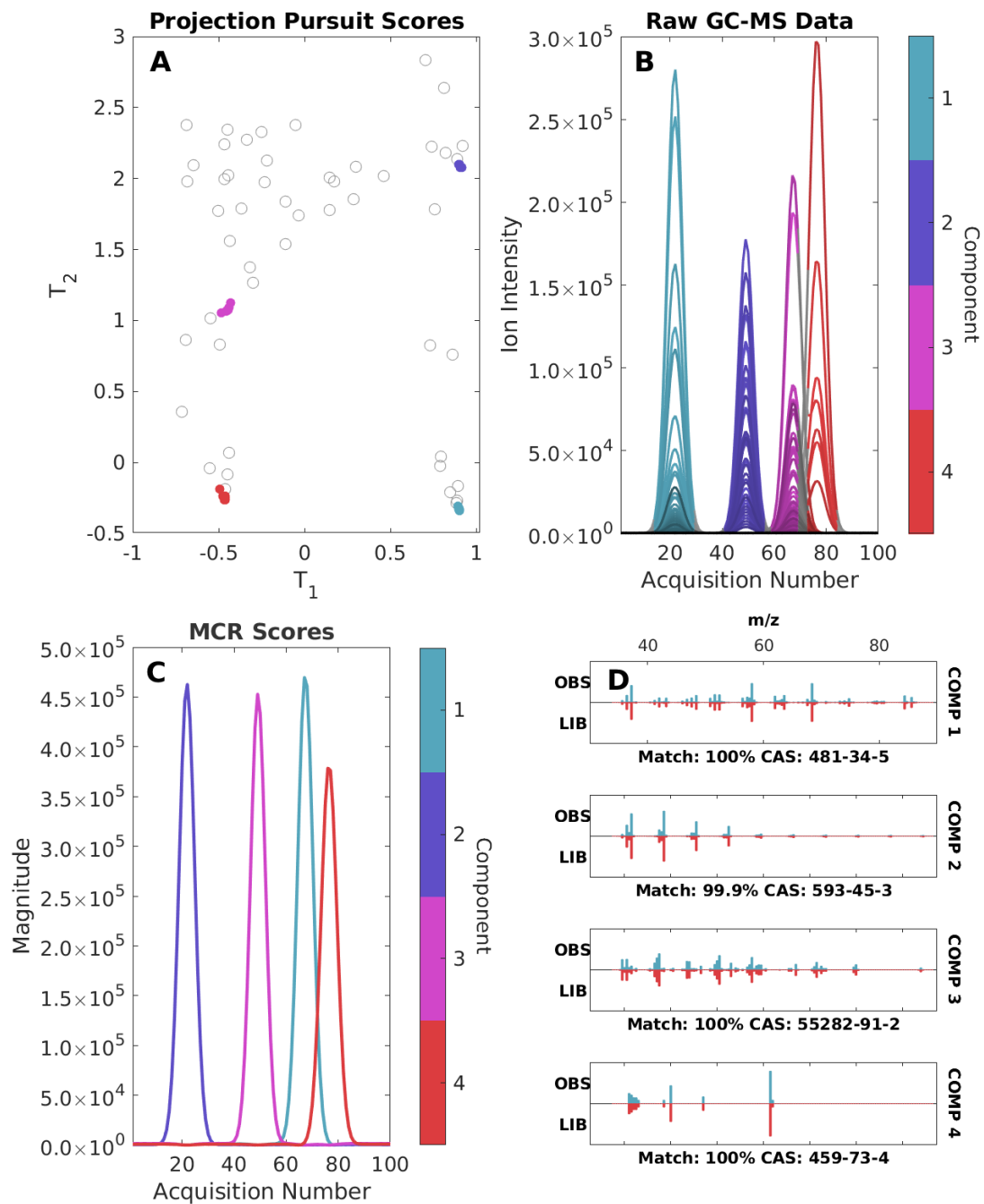


Figure D.41: Cumulative variances explained by component: 28.00 55.51 80.73 99.92

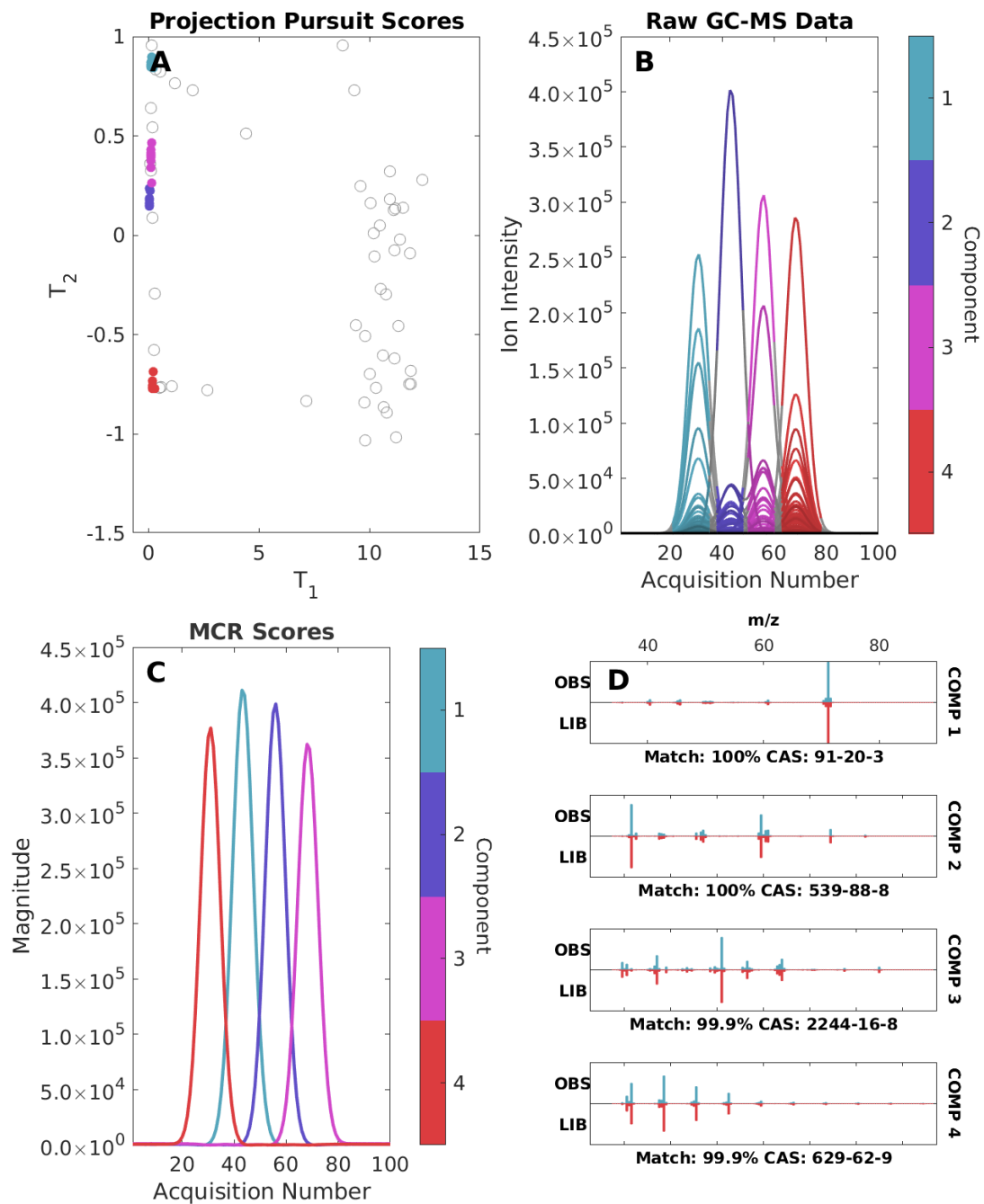


Figure D.42: Cumulative variances explained by component: 28.71 55.02 77.02 99.95

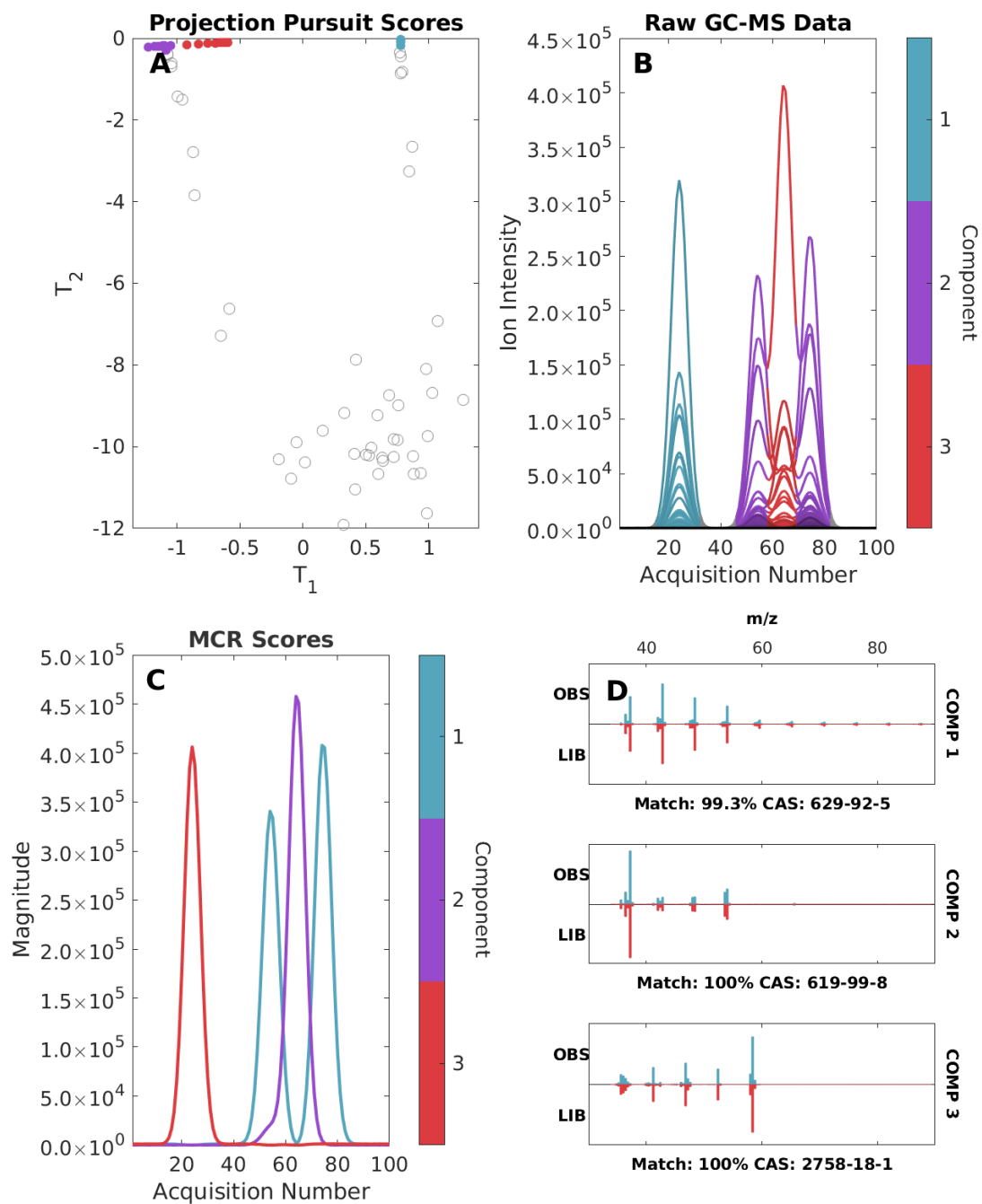


Figure D.43: Cumulative variances explained by component: 43.54 75.58 99.85

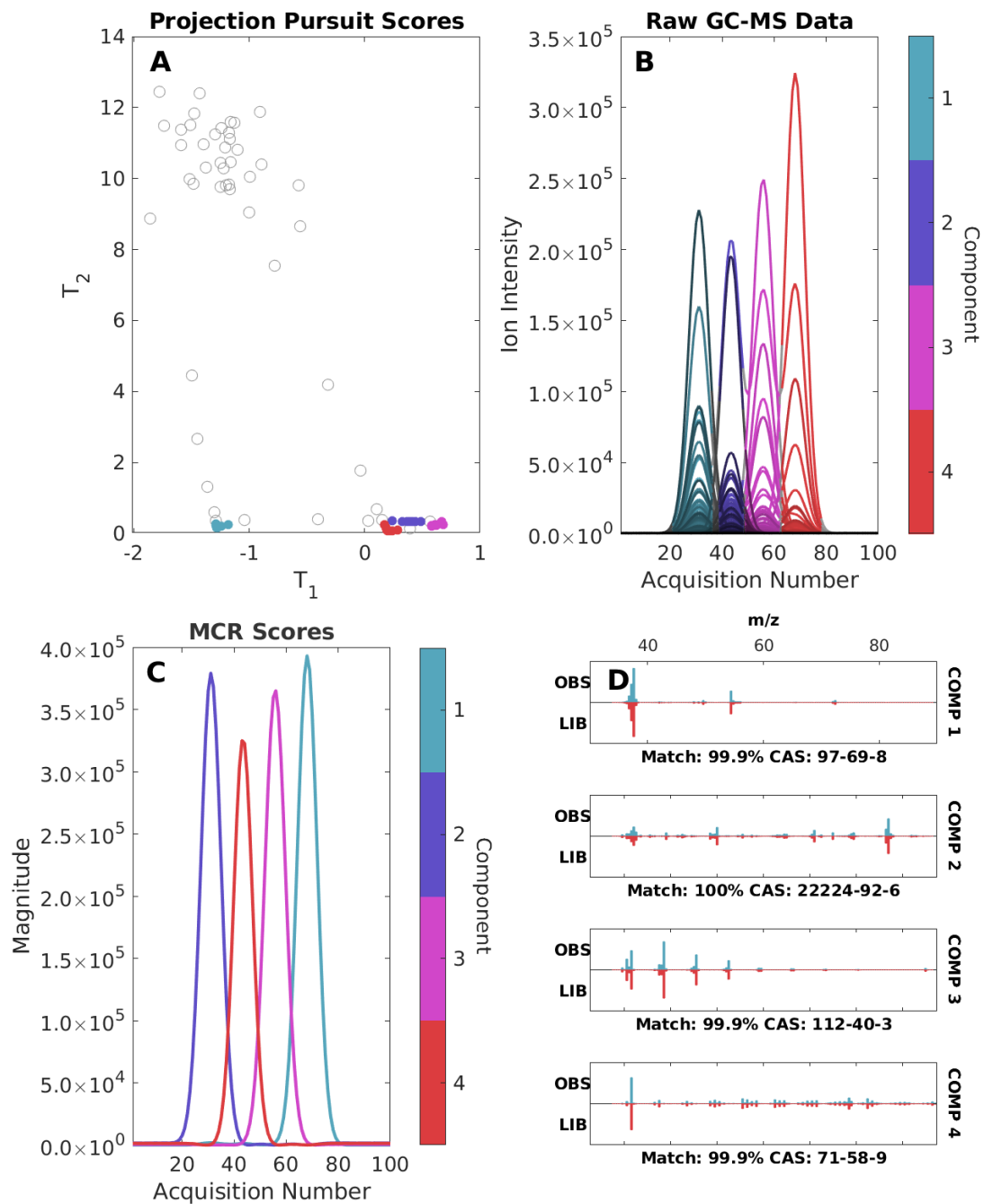


Figure D.44: Cumulative variances explained by component: 28.45 55.43 81.13 99.95

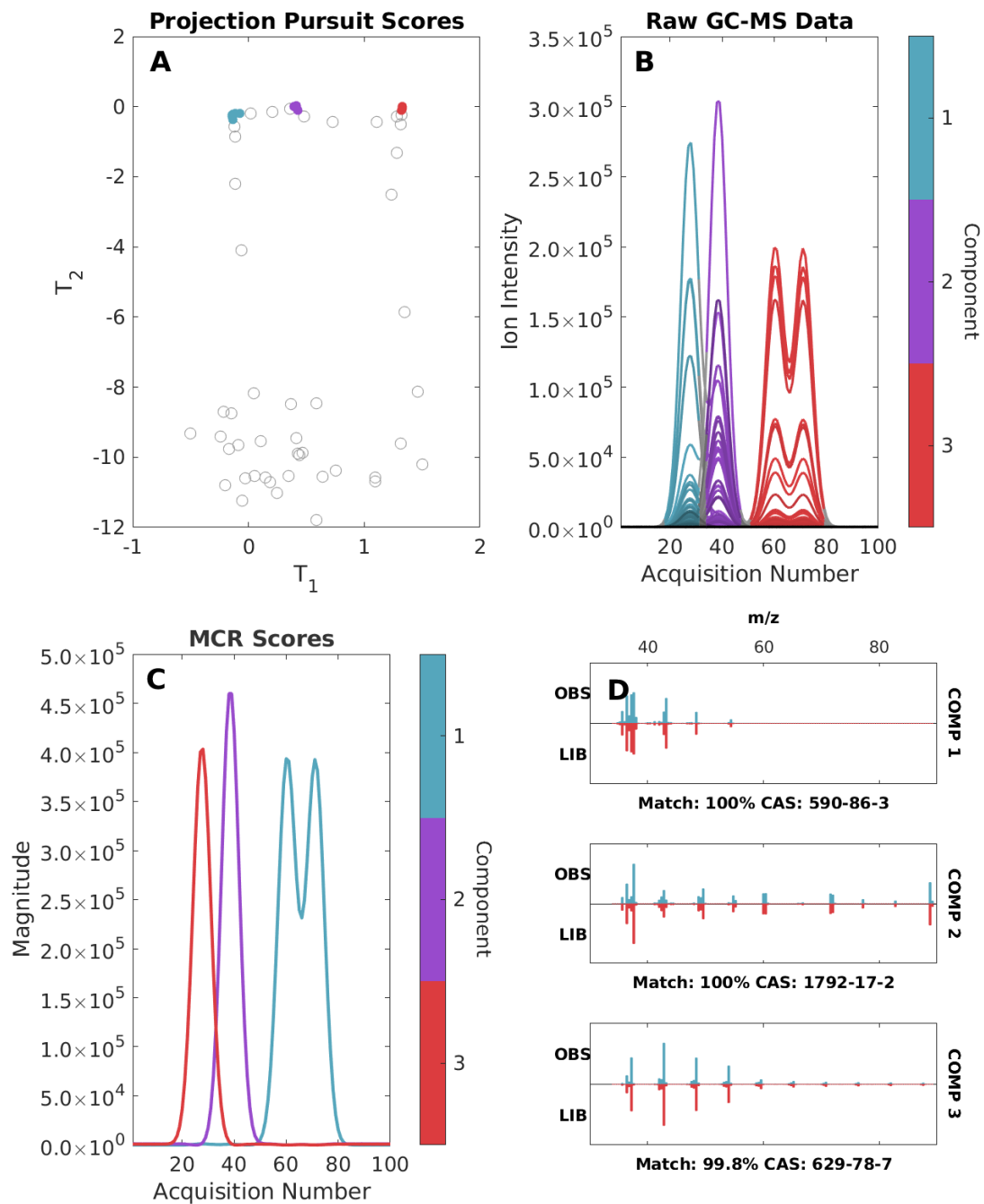


Figure D.45: Cumulative variances explained by component: 47.34 76.82 99.94

D.5.5 5 Factor Synthetic Data

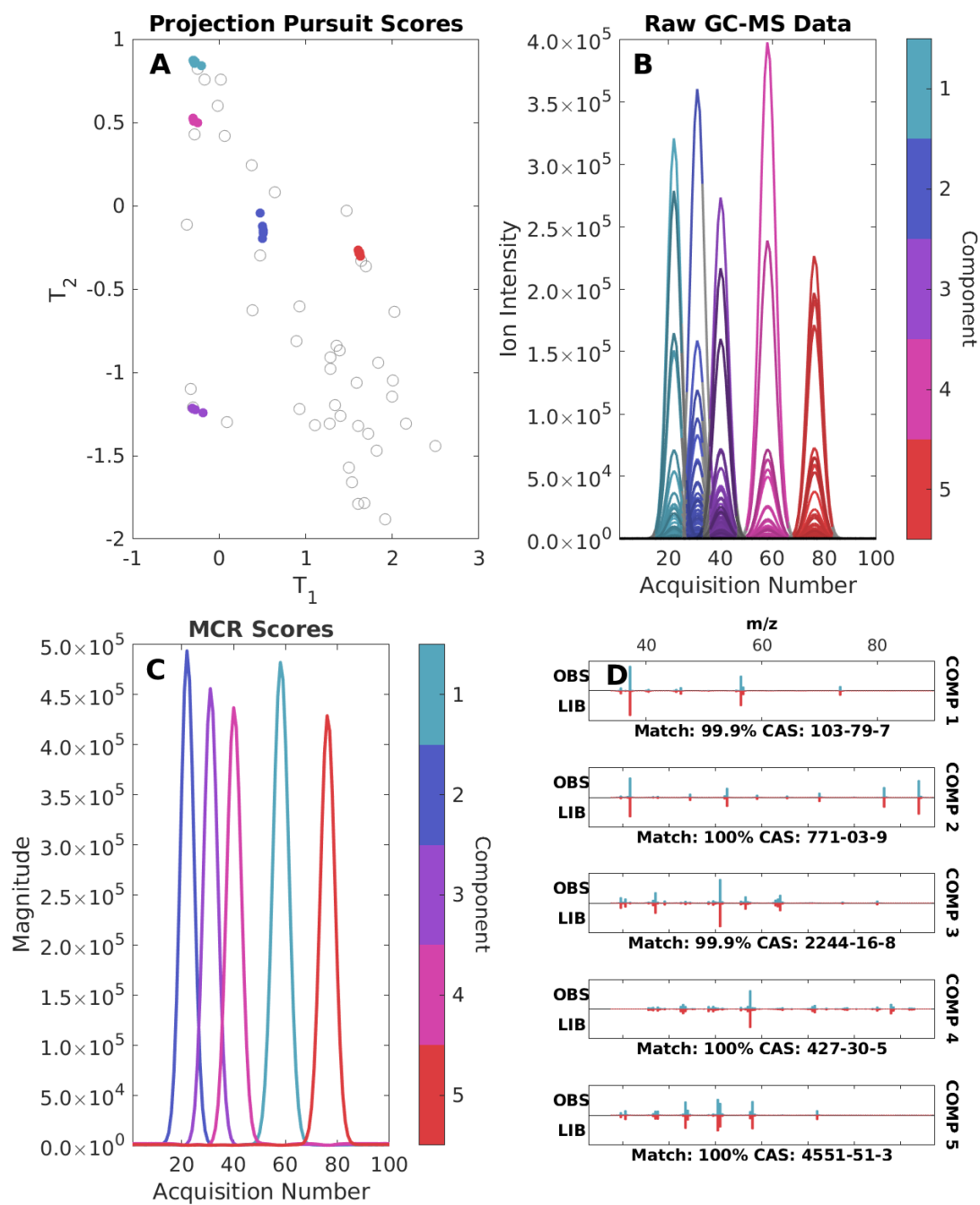


Figure D.46: Cumulative variances explained by component: 23.70 45.08 64.30 82.44 99.95

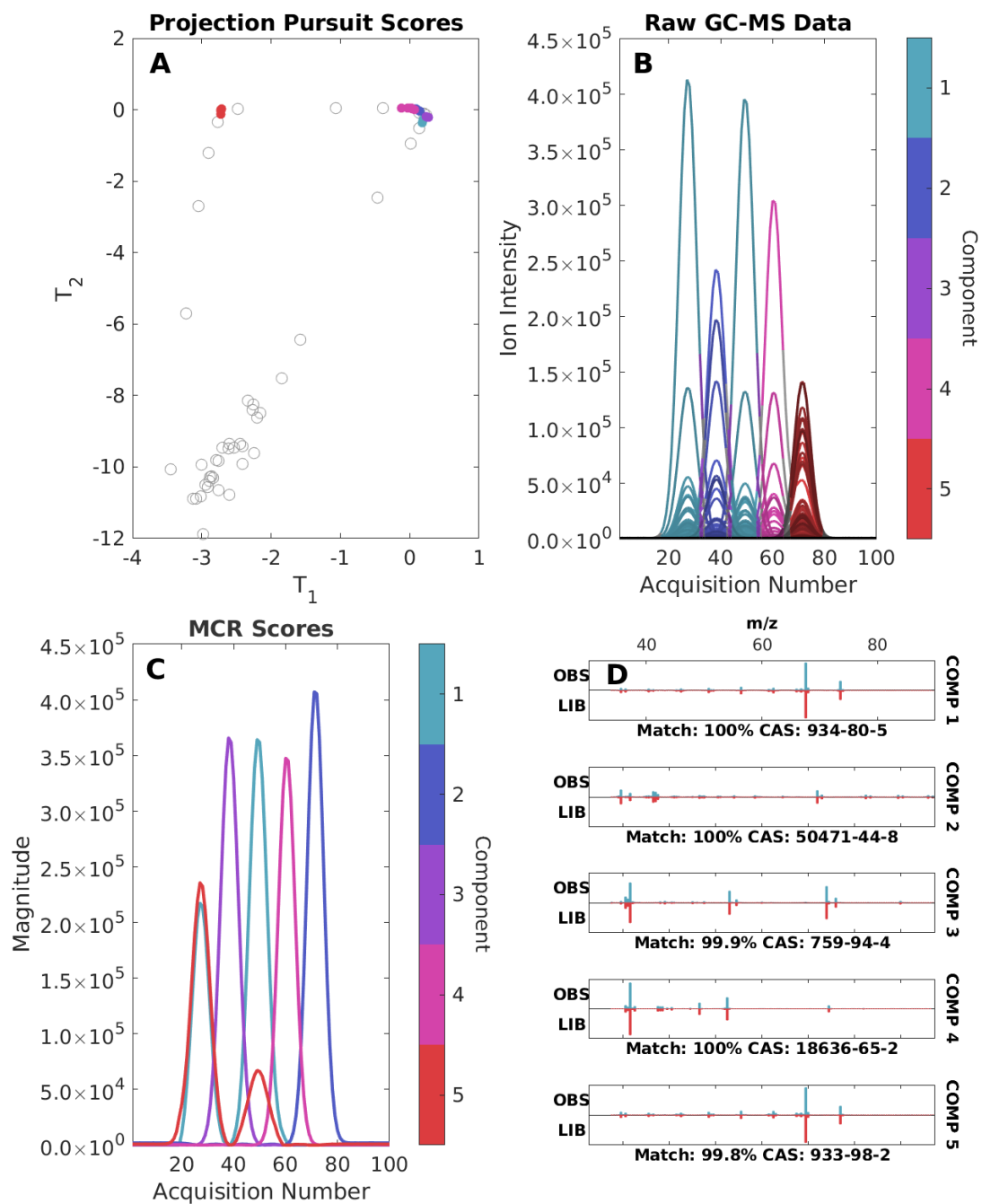


Figure D.47: Cumulative variances explained by component: 27.68 51.74 72.51 90.25 99.96

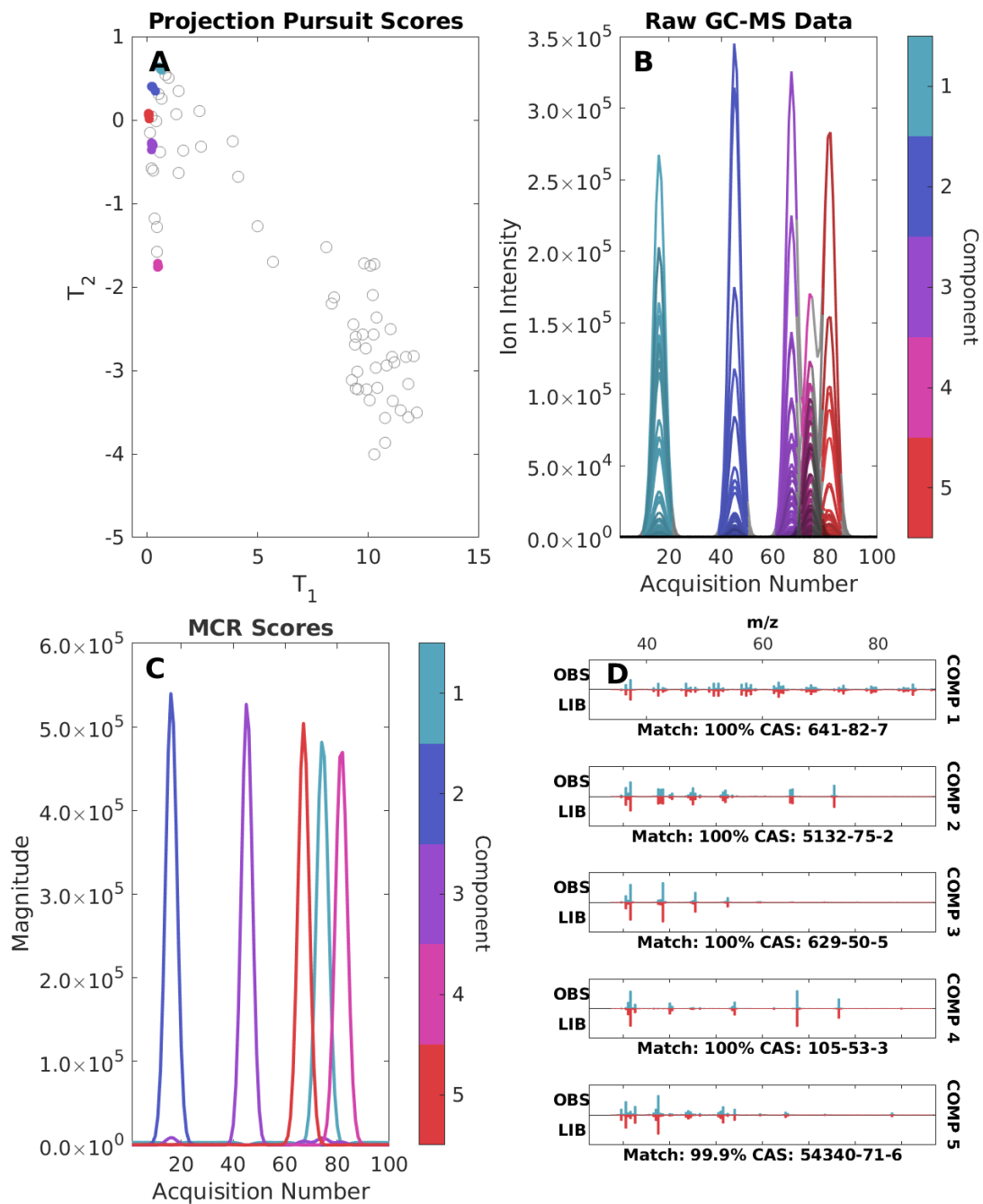


Figure D.48: Cumulative variances explained by component: 19.12 41.54 62.34 80.71 99.92

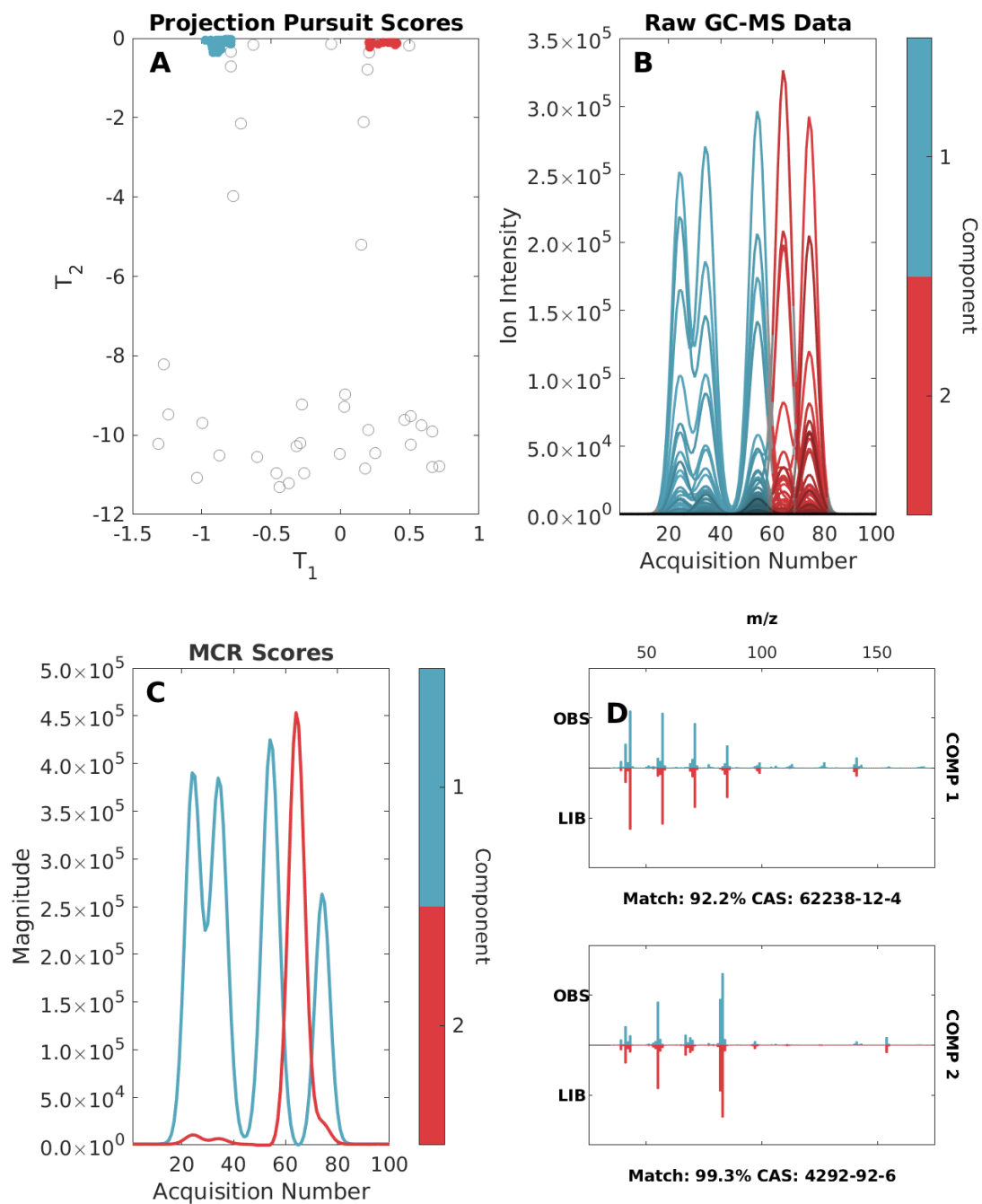


Figure D.49: Cumulative variances explained by component: 63.33 85.70

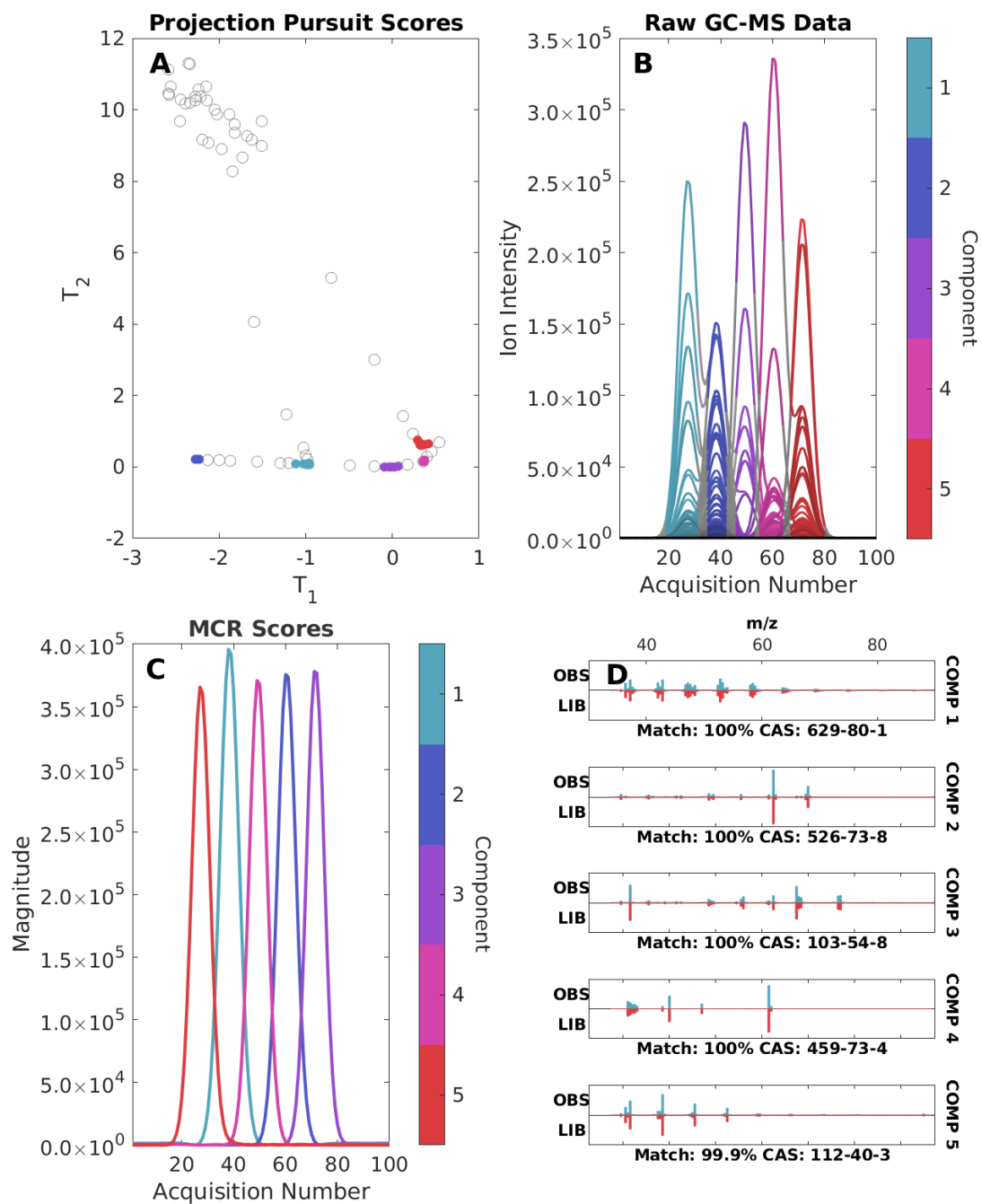


Figure D.50: Cumulative variances explained by component: 22.85 42.79 62.34 81.37 99.96

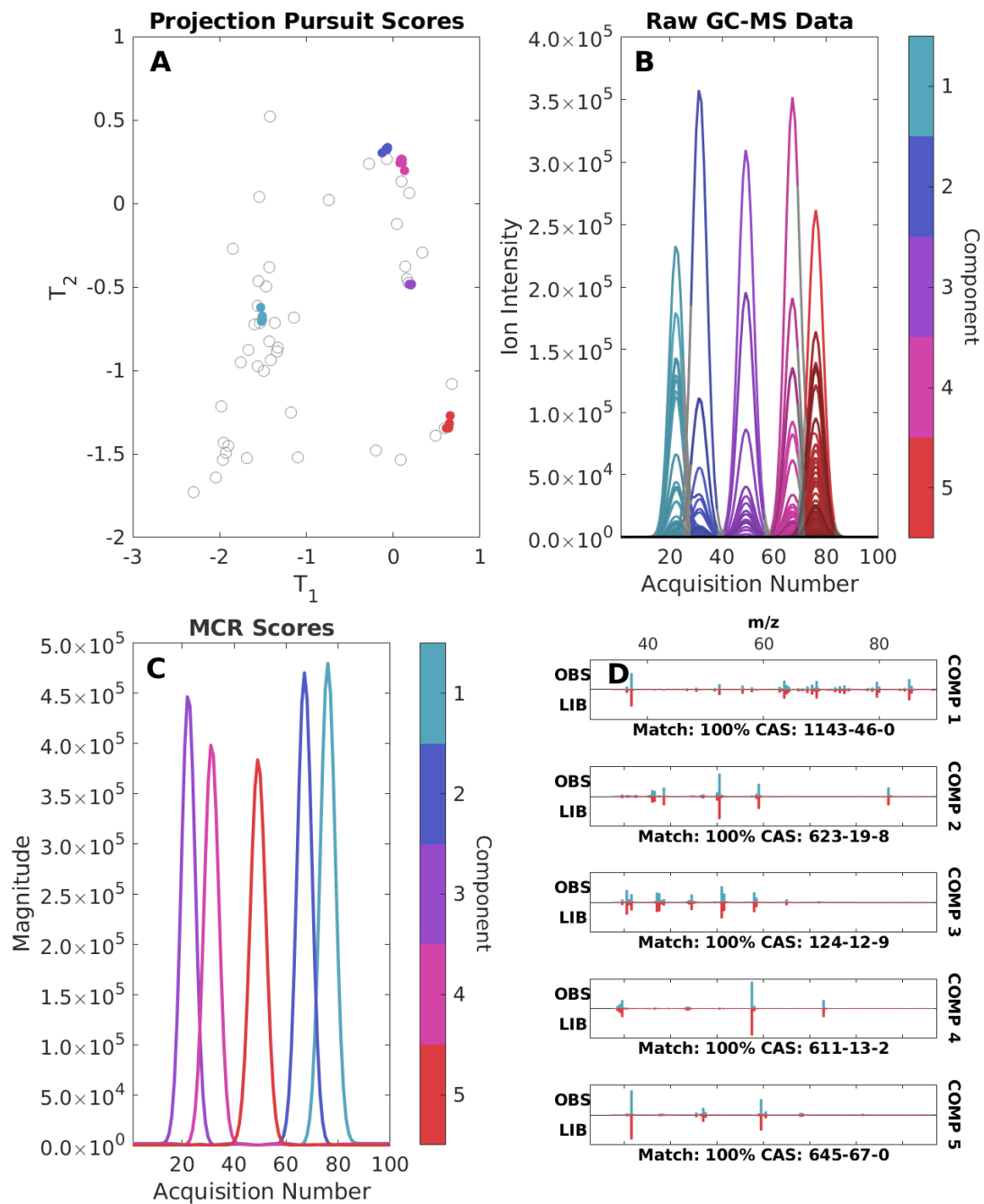


Figure D.51: Cumulative variances explained by component: 24.35 47.30 67.22 83.89 99.94

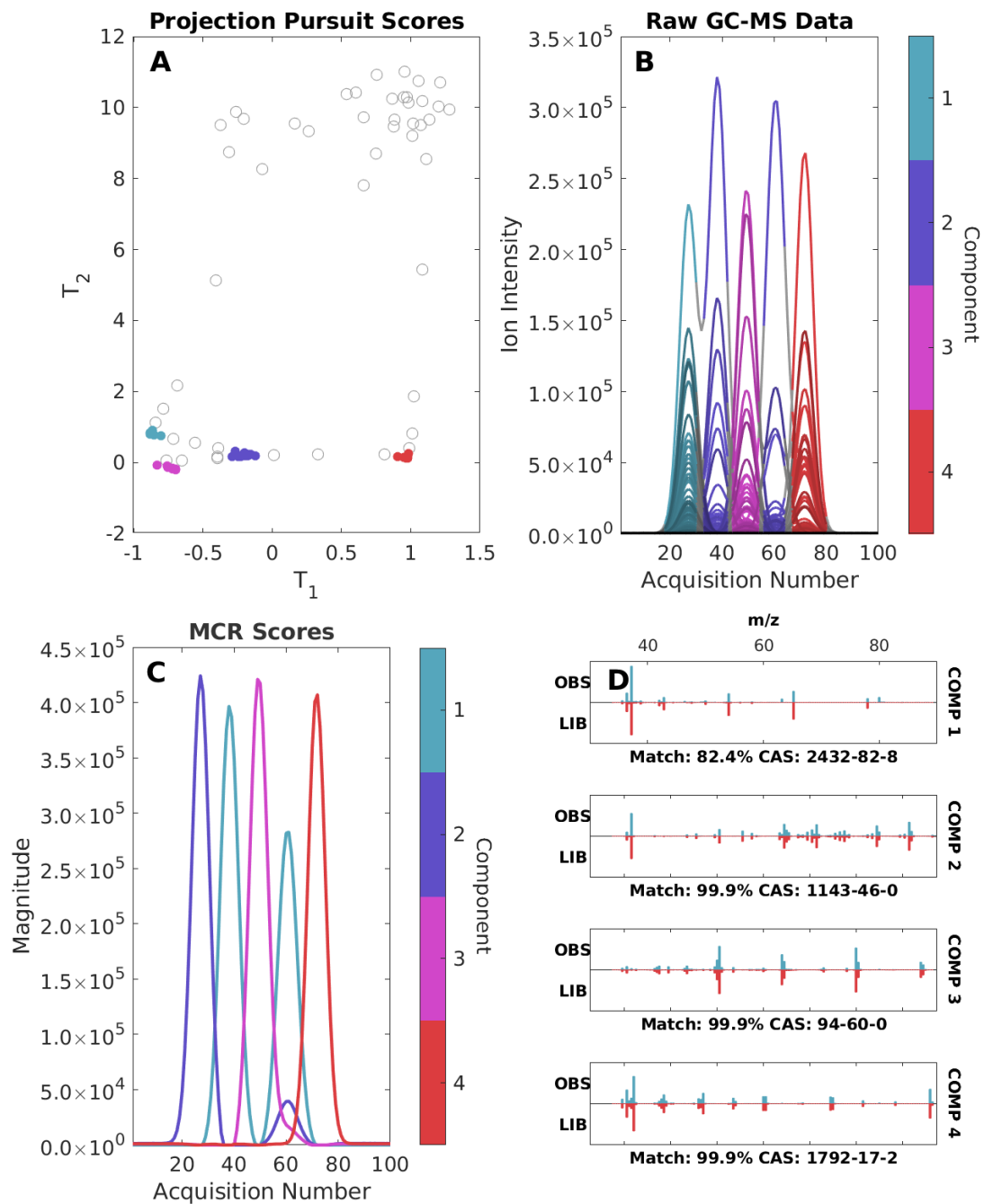


Figure D.52: Cumulative variances explained by component: 30.88 53.13 75.77 95.95

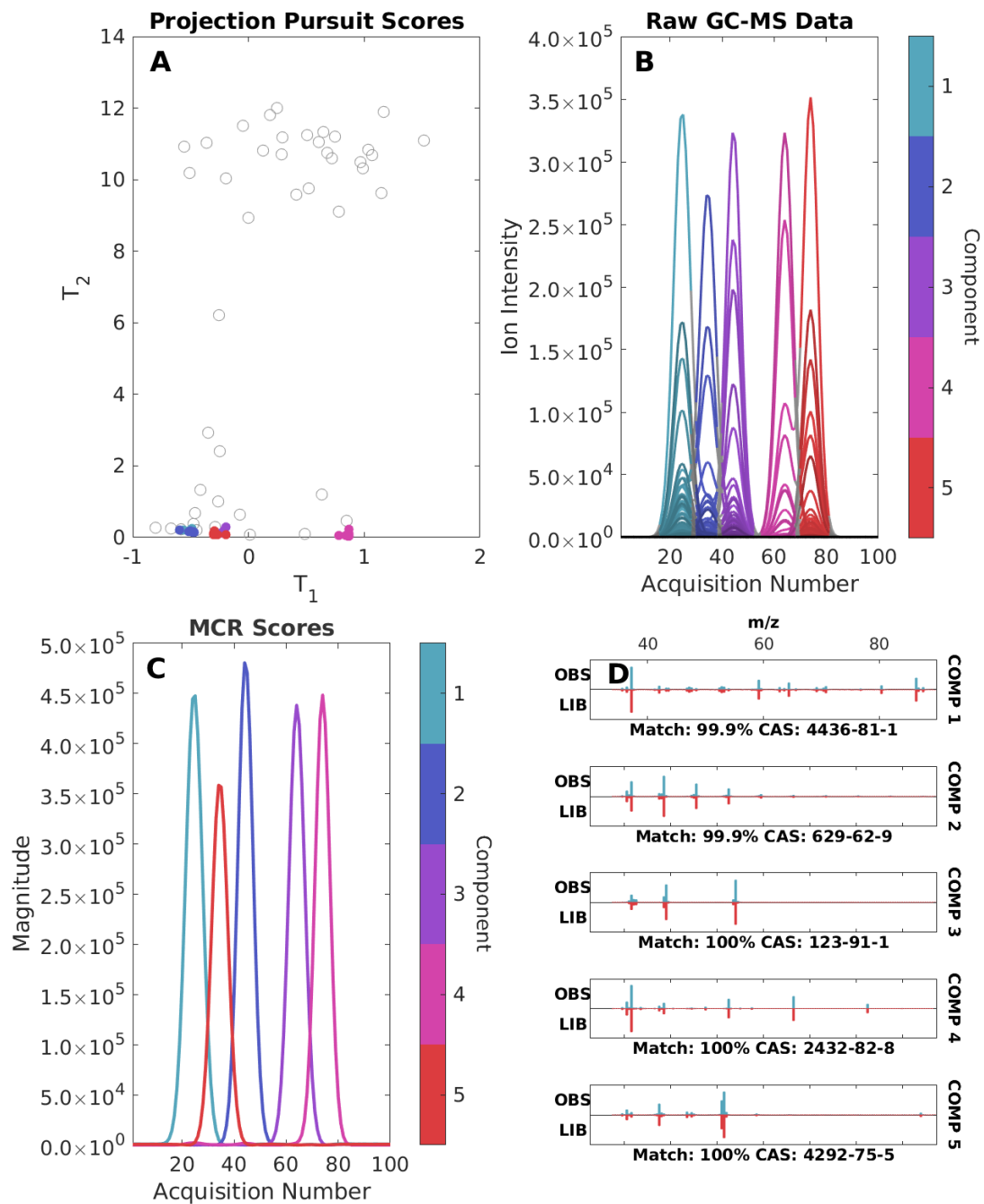


Figure D.53: Cumulative variances explained by component: 22.66 45.88 66.11 85.93 99.95

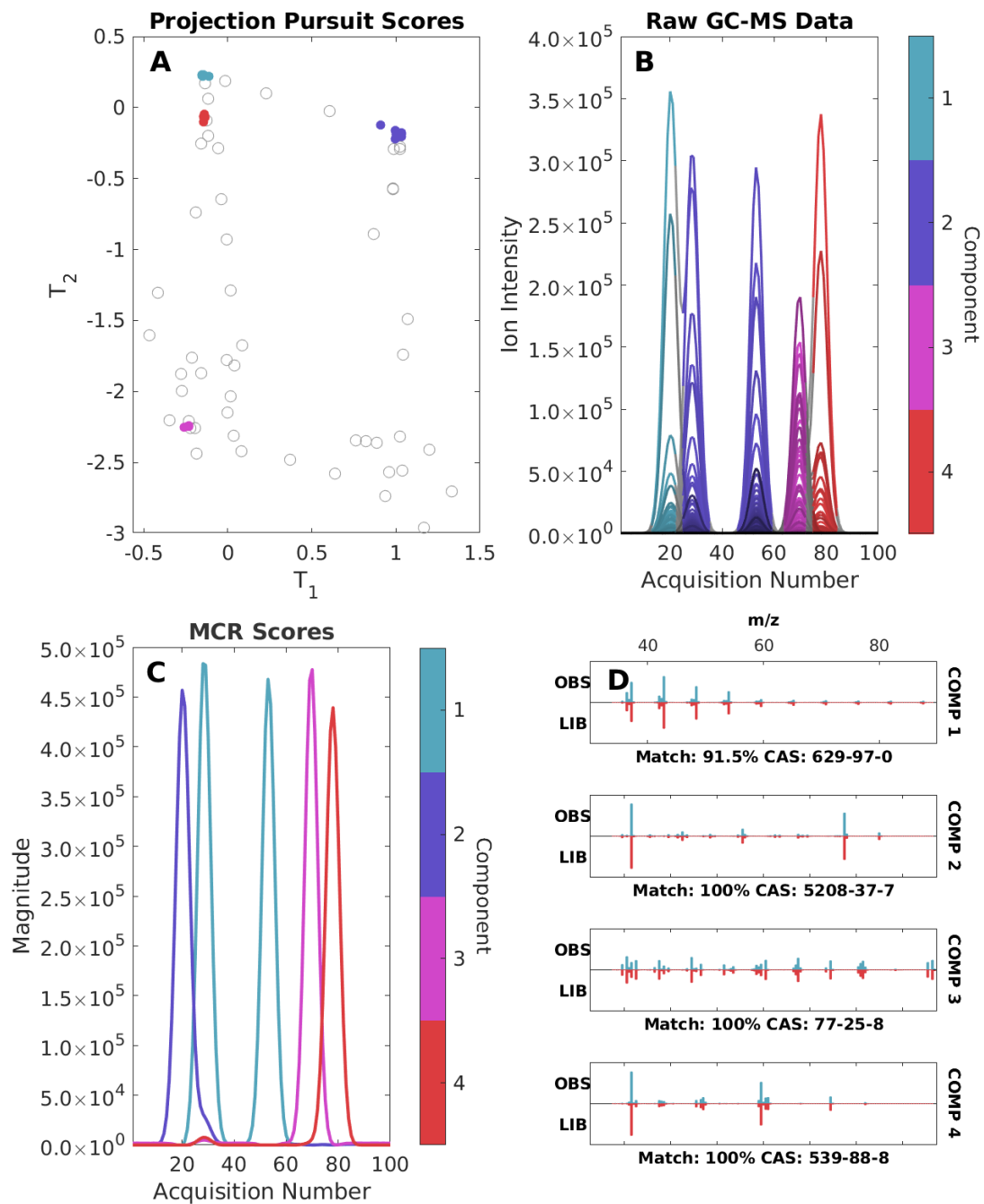


Figure D.54: Cumulative variances explained by component: 40.85 61.05 81.62 99.51

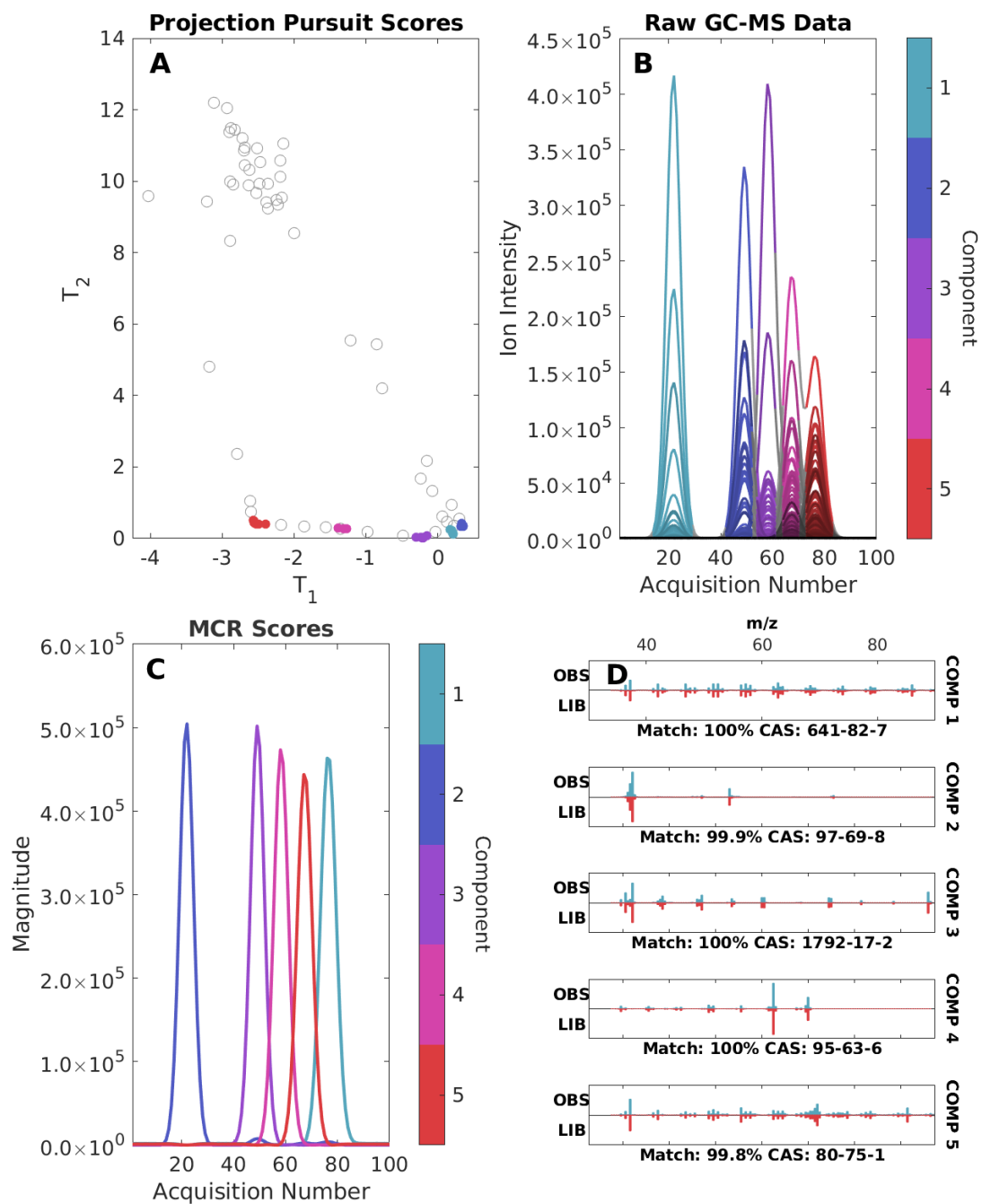


Figure D.55: Cumulative variances explained by component: 20.40 41.69 62.68 82.20 99.95