DEVELOPMENT OF DATA-DRIVEN AND MODEL-BASED TOOLS FOR SPECTROSCOPIC SLURRY MONITORING: CASE STUDIES ON LOW-ACTIVITY AND HIGH-LEVEL NUCLEAR WASTE

A Dissertation
Presented to
The Academic Faculty

By

Steven Howard Crouse

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemical and Biomolecular Engineering

Georgia Institute of Technology

Dec 2025

DEVELOPMENT OF DATA-DRIVEN AND MODEL-BASED TOOLS FOR SPECTROSCOPIC SLURRY MONITORING: CASE STUDIES ON LOW-ACTIVITY AND HIGH-LEVEL NUCLEAR WASTE

Thesis committee:

Dr. Martha A. Grover, Advisor
Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Ronald W. Rousseau, Advisor
Department of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Joseph K. Scott
Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Andrew J. Medford
Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Thomas M. Orlando
Chemistry and Physics
Georgia Institute of Technology

Date approved: August 28th, 2025

Those swirls in the cream mixing into the coffee? That's us. Ephemeral patterns of complexity, riding a wave of increasing entropy from simple beginnings to a simple end. We should enjoy the ride.

Sean Carroll

To those who motivate my creative pursuits

To my parents, family, friends, mentors, and colleagues

ACKNOWLEDGMENTS

I am immensely thankful for my two PhD advisors at Georgia Tech, Dr. Martha Grover and Dr. Ronald Rousseau. Dr. Martha Grover has always noticed and questioned the subtle details that many others (i.e. me, the author) miss or ignore. She has also shown me how to maintain composure, integrity, and a positive outlook when personalities, wants, and needs often clash. For this unwavering demonstration of character, I will always be grateful. During my discussions with Dr. Ronald Rousseau, he has always extracted the key concepts and held them to the scientific literature and engineering application. Dr. Ronald Rousseau regularly offers the simplest, and yet most exposing, questions in any discussion. He operates with immense dignity, curiosity, and humility. I am grateful for our discussions that discuss society's obsession with happiness or the importance of a strong conclusion that has plenty of "bang-bang-bang." I am fortunate to have worked with Dr. Martha Grover and Dr. Ronald Rousseau, two of humanity's best; it is quite humbling.

I would like to thank my thesis committee for working with me and offering their contributions to my growth as a researcher. Dr. Joe Scott has worked with me to better myself as a scientific communicator and instructor. Dr. A.J. Medford's class on data analytics has greatly influenced my academic interests, and his research suggestions have challenged me to connect my application to the state-of-the-art machine-learning literature. Dr. Thom Orlando's understanding of the physical and chemical systems that are the subject matter of this thesis have pushed me to better motivate my scientific pursuits in the context of nuclear-waste sites.

I would like to thank Dr. Stefani Kocevska and Dr. Rupanjali Prasad for their guidance, collaboration, and friendship. Much of our time together is recorded in meticulous detail thanks to Dr. Rupanjali Prasad's insistence of capturing the moment via photo. I would like to thank my many group members who I have spent a considerable amount of time working alongside; you have made the work environment incredibly rewarding: Dr. Aaron Liu, Dr. Patrick Harris, Dr. Rahul Venkatesh, Dr. Hossein Salami, Viviana Cardenas, Luciana Cardenas, Jessica Bonsu, Nischal Maharjan, Nuna Dumakor, Hannah Jin, and Yeonwoo Jeon.

I would like to thank my three closest friends and roommates — Daniel Aziz, Fernando Martinez, and Alex Heiler — who have necessitated research-level commitment from non-research activities.

I would like to thank my many teachers, with some listed here: Ms. Carol Knight, Mr. Brian Otter, Mr. Eric Williamson, Dr. Edwar Shamshoum, Mr. Jason Turner, Mr. Matt Boerner, Ms. Elizabeth Moose, Mr. Chris Gann, Dr. Marion Brisk, Dr. Cecile Tougas, Dr. Samuel Wheeler, Mr. Charlie Payne, and Mr. Mickey Alcorn.

My immediate and extended family have been my support structure before and during my PhD. I would like to thank my parents Martin Crouse and Sandra Dickerson Crouse for inspiring my curiosity and work ethic. I would like to thank my sister, Rachel Crouse, for keeping me grounded; her sharp wit and blunt words could humiliate the strongest ego. My grandparents — Howard Crouse, Colene Crouse, Steven Dickerson, and Sharon Dickerson — whose love and humor I greatly appreciate. To all of my extended family, thank you.

Finally, to my girlfriend, Ananya Thakur: our never-ending conversations are a great source of joy and relief for me. Your adventurous spirit is my driving force for exploring the unknown. I am so fortunate to have found a partner that will elevate our triumphs and diminish our failures. Thank you.

Financial Acknowledgments

I would like to thank the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) for their financial[1] and career support. I would also like to thank the Real-Time Inline Monitoring Team for their financial[2] support and collaboration.

[1]Support by the U.S. Department of Energy under Cooperative Agreement DE-FC01-06EW07053, entitled "The Consortium for Risk Evaluation with Stakeholder Participation III," is gratefully acknowledged.

[2]Support by the Office of Environmental Management, U.S. Department of Energy (DOE), through the DOE National Laboratory Program Announcement Number LAB 23-EM001, Hanford Tank Waste Cleanup Research & Development, is gratefully acknowledged.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AIC Akaike information criterion

ATR attenuated total reflectance

ATR-FTIR attenuated total reflectance - Fourier transform infrared

BSS blind source separation

CDAE convolutional denoising autoencoder

CDF cumulative distribution function

CI$_{95\%}$ 95% confidence interval

CLS classical least squares

DWPF Defense Waste Processing Facility

FBRM focused beam reflectance measurement

FTIR Fourier transform infrared

GFC glass-forming chemical

GFCs glass-forming chemicals

HLW high-level waste

i.i.d. independent and identically distributed

IC ion chromatography

ICA independent component analysis

ICLS indirect classical least squares

ICP inductively coupled plasma

ICP-AES inductively coupled plasma - atomic emission spectroscopy

ICP-MS inductively coupled plasma - mass spectrometry

IOT iterative optimization technology

IR infrared

LAW low-activity waste

LIBS laser induced breakdown spectroscopy

LOD limit of detection

MAE mean absolute error

MCR-ALS multivariate curve resolution - alternating least squares

MFPV Melter Feed Preparation Vessel

MPE mean percent error

MSPM multivariate statistical process monitoring

NCCLS nonnegatively constrained classical least squares

NIPALS nonlinear iterative partial least squares

NIST National Institute of Standards and Technology

PAT process analytical technology

PCA principal component analysis

PLSR partial least squares regression

PNL Batelle Pacific Northwest Laboratories

PSD particle size distribution

RC1 Mettler Toledo Reaction Calorimeter

ReLU rectified linear unit

RMSE root mean squared error

SME Slurry Mix Evaporator

SNR signal-to-noise ratio

SNV standard normal variate

SPE squared prediction error

SRACLS spectral residual augmented classical least squares

SRAT Sludge Receipt and Adjustment Tank

SRNL Savannah River National Laboratory

SWPF Salt Waste Processing Facility

US DOE United States Department of Energy

WTP Waste Treatment Plant

SUMMARY

Nuclear waste is challenging to store, handle, and process. The United States Department of Energy is currently (as of August 2025) building a waste treatment plant to vitrify waste at the Hanford site, and there has been increasing interest in the use of real-time and remote spectroscopic sensors for monitoring radioactive waste at this and other sites. However, there are significant technological hurdles preventing the immediate application of spectroscopic sensors to nuclear waste. The radioactive waste is a multicomponent, inhomogeneous, multiphase, and radioactive slurry with the potential for ongoing chemical/nuclear reactions and batch-varying particle morphology. Spectroscopic sensors have seen regular use in laboratory-scale reaction monitoring, not industrial nuclear-waste processing. There remain many unsolved questions and unconquered hurdles precluding the use of spectroscopic sensors for a safer, more efficient, and more robust nuclear-waste processing outlook.

This thesis introduces methods, shows data, and presents studies for monitoring multicomponent solutions, quantifying signal attenuation in slurries, and utilizing data for real-time fault detection in multicomponent slurries. Given the legacy of nuclear waste that has been left for current and future generations, my engineering objective is to enable safe and efficient nuclear-waste processing using the most appropriate methods, whether those methods exist or require development.

CHAPTER 1
INTRODUCTION

Nuclear waste is a pressing issue of our time. As nuclear technologies emerged in the 20th century, improved engineering methods enabled the identification, isolation, and utilization of radioactive isotopes [1]. However, a different set of challenges followed the nuclear exploration of the 20th century. Radioactive waste has been left behind that presents a sustained environmental and anthropological hazard; this material has been called legacy waste.

Legacy nuclear waste is stored at several sites in the United States. This thesis focuses specifically on two sites: The Hanford site in Washington State and the Savannah River Site in South Carolina. Since the radioactive waste will remain an anthropological and ecological hazard for thousands of years and given the waste has historically leaked from its storage vessels into the surrounding groundwater, the United States Department of Energy (US DOE) has decided that the waste at these sites will be immobilized for long-term storage. The Savannah River Site has been processing high-level waste (HLW)[1] nuclear waste since 1996. However, a new Salt Waste Processing Facility (SWPF) has been recently constructed and has been operating since 2020. The Hanford sister site has also (as of the completion of this thesis) completed construction of the low-activity waste (LAW) processing facility as part of the larger Waste Treatment Plant (WTP), with the first radioactive processing expected to begin in 2025. Construction is currently underway for an analogous processing facility for HLW at the Hanford WTP.

Nuclear waste presents a uniquely challenging process control environment due to radioactivity that introduces new chemistry, precludes the use of unshielded electronic equip-

[1]Nuclear-waste literature makes abundant use of acronyms that are not common elsewhere. Acronyms are used where appropriate to match available literature while also being defined at first use in each chapter to improve readability of the content.

ment, and prevents regular human interaction with processing equipment [1, 2, 3]. However, nuclear-waste vitrification has a significant history dating back to the 1950's in Canada and the UK, with the first ceramic melter for the vitrification of HLW commissioned at Batelle Pacific Northwest Laboratories (PNL) in 1984 [4]. However, there are significant technical challenges with current waste-processing strategies:

- Effective process models do not exist [5, 6]
  - Legacy nuclear waste is highly multicomponent, multiphase, and inhomogeneous between and within storage tanks — models cannot accurately predict behavior
  - Time-varying radioactivity makes chemical equilibrium a moving target
- Measurements are collected by extracting radioactive waste from the process and having humans perform analyses on the waste sample in an onsite laboratory [7, 8, 9]
  - Time consuming — slow process feedback and process decisions
  - Exposes humans to radioactivity
- Slurries are challenging to mix at large scales [10, 11, 12]
  - Poor mixing may affect quality of both *in situ* measurements and grab samples
  - Tanks may not be well-mixed when moving from one vessel to the next
- Large amounts of radioactivity [13, 14]
  - human and ecological safety hazard
  - incompatible with electronic equipment
  - unanticipated chemical behavior through radiolysis

Given that model-based methods are insufficiently accurate and mass-balance modeling introduces uncertainty beyond what the US DOE is willing to accept, there is a need for process measurement — *feedback* [7]. However, measuring multicomponent, rheologically complex, and radioactive slurries is a complex issue with nuance; no available instruments have been implemented under the conditions expected during nuclear-waste processing. To address the aforementioned challenges, this thesis investigates process analytical technology (PAT) tools for providing accurate, real-time, process-relevant information so that nuclear-waste processing can be achieved efficiently and safely. The remainder of the introduction will be devoted to discussing the technical state-of-the-art for nuclear-waste process monitoring, the Hanford site, the Savannah River Site, and vibrational spectroscopy before finally providing the content organization for this thesis.

1.1 Real-Time In-Line Monitoring

Presently, most nuclear-waste processing decisions are made at process hold points where waste samples (grab samples) are taken to an on-site laboratory where an array of analytical techniques are performed to ascertain: 1) which glass-forming chemicals (GFCs) are required to make effective glass for a given waste batch 2) and what is in a given waste batch for regulatory, decision-making, and accounting purposes.

"Grab-sampling" (analytical measurements using ion chromatography (IC), inductively coupled plasma - mass spectrometry (ICP-MS), among other analytical tools) can provide necessary measurements of stream composition [7, 8]. However, sampling in this manner provides a radioactive exposure risk for workers and can delay downstream decision making until analytical lab results are provided (which may be on the order of hours to several days). A mass-balance approach has been considered that would eliminate some grab-sampling locations [15]. However, mass-balance modeling is a feed-forward endeavor; process upsets, faults, and uncharted conditions or reactions may lead to inaccuracies in predicted behavior.

Much published work has been done recently to develop optical spectroscopy for the needs of nuclear-waste systems. For measuring the composition of slurries, work has been done to develop Raman spectroscopy in turbid environments [16, 17]. Attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy has been shown to be effective in measuring the solution phase of nuclear-waste simulants [18], including in the presence of solids [19]. Research has been done in the field of chemometrics, where pre-processing can decrease the experimental burden of creating large calibration datasets [20, 21]. Instrument hardware has also been investigated as it relates to quantifying nuclear waste with multiple Raman excitation wavelengths [22] or multiple probes [16, 23]. Partial least squares regression (PLSR) has been well-studied for quantifying the composition of nuclear-waste mixtures [21, 24], but alternative model structures have been proposed including a piecewise-linear PLSR model that leverages spectral (specifically UV-Vis) linearity over a narrow concentration or process variable range [25]. A number of researchers have used multiple sensors for measuring non-radioactive nuclear-waste simulants [16, 19, 20, 21, 26], while others have measured radioactive waste samples [24, 27].

Despite all of this work, there are scientific gaps on using these sensors: 1) in dense slurries of multicomponent solids [19, 28], 2) for analysis of multicomponent mixtures with training data that do not represent process conditions [29, 30, 31], and 3) for fault detection in slurry systems [32, 33].

1.2 The Hanford Site

The Hanford Site in Washington State contains 56 million gallons of radioactive chemical waste (196 MCi as of 2003 [13]) with approximately 1 million gallons having leaked from underground tanks [34, 35, 36, 37]. To address environmental issues, the United States Department of Energy is acting to clean up the released waste and to immobilize waste that is still in containment. The waste still in the underground tanks will be separated into LAW and HLW before being vitrified into borosilicate glass via the addition of GFCs

[38]. The waste at Hanford is inhomogeneous; it is located in 177 underground tanks that have different compositions due to being filled over the course of 45 years of plutonium production (1945 – 1990) in addition to subsequent tank mixing and processing [36, 5]. To create a stable and optimum glass form from a nuclear-waste slurry, waste loading must be considered as a function of the waste's composition and physical properties.

Currently, the process at Hanford is expected to combine mass-balance with measurement. The planned system will utilize a Monte-Carlo mass-balance model combined with laboratory measurements taken of the incoming waste stream (before processing) and prior to being melted [10, 39]. The ability to monitor waste remotely at Hanford is imperative because of the dangerous and time-consuming nature of grab-sampling, along with the immense remaining cost ($364.0 billion or greater) and processing time (waste treatment completion scheduled for 2079) associated with waste processing [40].

1.3 The Savannah River Site

The Savannah River Site near Aiken, South Carolina, contains 36 million gallons of legacy waste from nuclear weapons production that includes 208 MCi of radioactivity as of June 2024 [41, 42, 43, 44]. The nuclear reactions and subsequent chemical separation processes used at the Savannah River Site have generated diverse products with myriad physical and chemical properties. The resulting waste must be managed to reduce the risk to humans and the environment; historic challenges with aging and leaking storage tanks (similar to the Hanford site) have prompted immobilization of the waste at the Savannah River Site [45, 46]:

Legacy nuclear-waste remediation efforts at the Savannah River Site consist of three facilities: the Defense Waste Processing Facility (DWPF), the SWPF, and the Saltstone Production Facility. A schematic of the processes are shown in Figure 1.1. The DWPF vitrifies sludge waste (characterized by large amounts of insoluble solids) into a glass form for long-term storage; the DWPF has been vitrifying waste since March of 1996 [43]. The

SWPF is a chemical separation facility that takes liquid supernate composed primarily of a sodium salt solution; removes the most radioactive soluble components to send to the DWPF; and feeds the nearby Saltstone Production Facility, which incorporates the remaining waste into a cementitious matrix. The SWPF has been processing waste since October of 2020. Although the Savannah River Site has an established history of processing nuclear waste, there is renewed interest in state-of-the-art nuclear processing as the Hanford nuclear-waste site also begins remediation [5, 6].

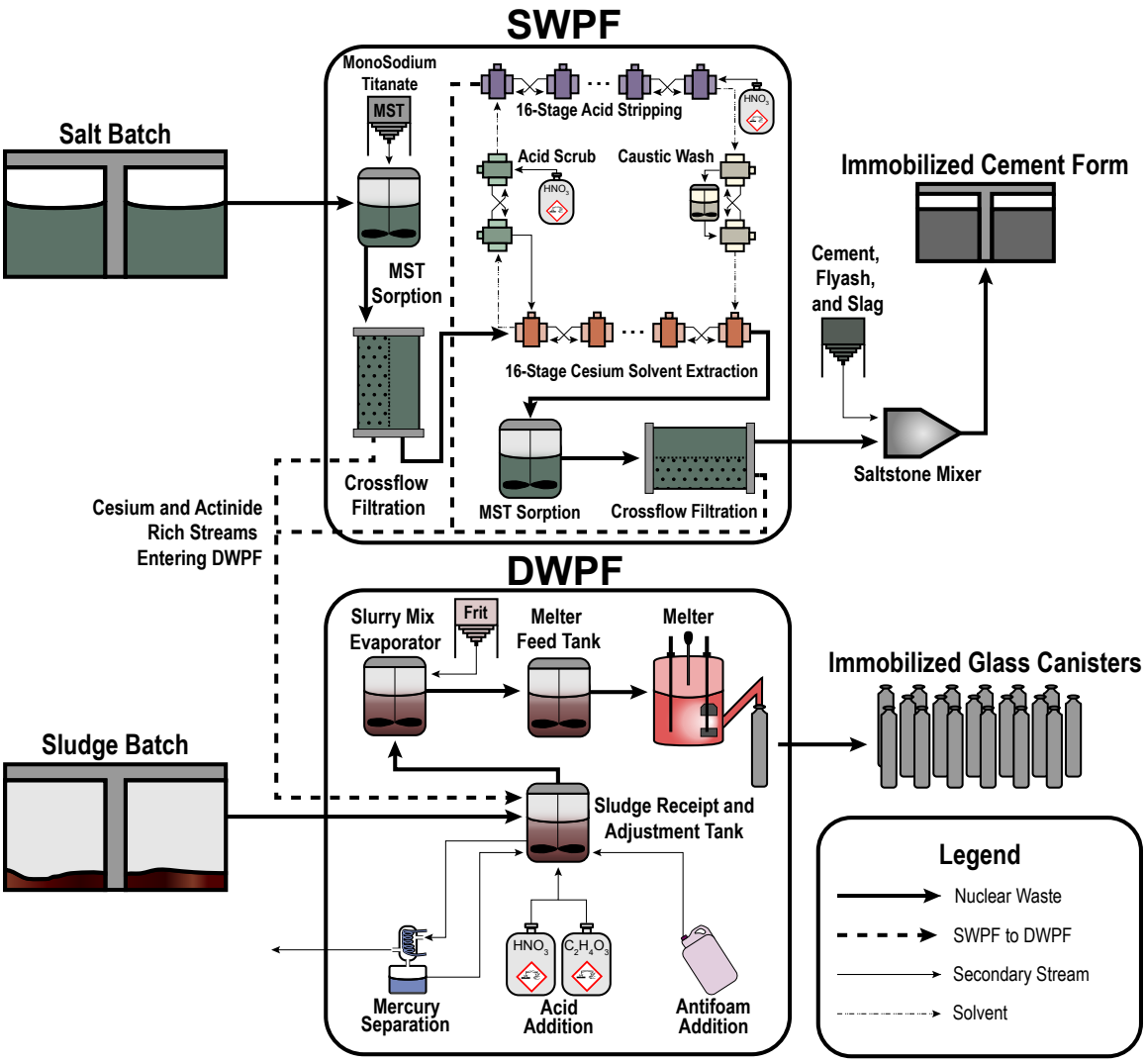


Figure 1.1: Overview of nuclear-waste processing at the Savannah River Site. Abbreviations: Defense Waste Processing Facility (DWPF), monosodium titanate (MST), and Salt Waste Processing Facility(SWPF). Figure used from [46].

1.4 Vibrational Spectroscopy

Vibrational spectroscopy, including Raman spectroscopy and ATR-FTIR spectroscopy, has been a useful tool for engineers to monitor industrial processes in real-time. This thesis investigates vibrational spectroscopy; specifically Raman spectroscopy and ATR-FTIR spectroscopy. These two complementary techniques provide molecular information about a solution or slurry of interest.

1.4.1 Classifying Target and Non-Target Species

When applying spectroscopy to quantify the composition of chemical components, unknown species (i.e. adulterants, interferents, etc.) can disrupt quantification models. The disruptive behavior of unknown species can lead to poor model robustness when analyzing new spectra, particularly when the peaks of target and non-target species overlap. Additionally, non-target species may be prohibitively resource-intensive to calibrate for, since common design of experiment schemes, such as full-factorial design, scale exponentially with the number of species of interest.

In this thesis, chemical constituents are classified by whether they are quantified using the spectra-to-concentration model. *Targets* are quantified species, while *non-targets* are not quantified despite possibly existing in the slurry. Species may be labeled non-targets because they are: at insignificant concentrations, not relevant for processing, or difficult to quantify because of model and data mismatch. To deal with overlapping non-target species, methods are introduced in Chapter 2 and Chapter 3.

1.5 Content Outline

This thesis explores PAT applied to nuclear-waste processing. Chapter 2 and Chapter 3 start by introducing the techniques of blind source separation (BSS) and nonnegatively constrained classical least squares (NCCLS) for monitoring multicomponent solutions where

unknown spectral components appear during chemical processing. Chapter 4 and Chapter 5 add complexity to the previous chapters by adding insoluble phases to the multicomponent solutions to yield slurries; these spectra are analyzed with linear methods in Chapter 4; the data have nonlinear transformations applied before analysis in Chapter 5. Lastly, the information from the previously-studied measurement models are put in the context of a chemical process of batch reactors in Chapter 6 and Chapter 7. Chapter 6 uses a dual-Kalman filter to combine model predictions with sensor measurements in a computational study, and Chapter 7 uses multivariate statistical process monitoring (MSPM) to detect faults in experimentally measured slurries. Lastly, a brief analysis of the content of thesis is made in Chapter 8.

CHAPTER 2

CONSTRAINED BLIND SOURCE SEPARATION: A CASE STUDY OF THE SAVANNAH RIVER SITE

This chapter answers a question: how can unanticipated chemical species be removed so that target species can be accurately quantified? A blind source separation (BSS) method was developed in the context of experimental data acquired through collaboration with Savannah River National Laboratory (SRNL). When applying existing BSS methods to the data from SRNL, poor non-target subtraction resulted; this motivated the development of an improved method. The method developed in this chapter makes use of prior information of known spectroscopic signals and constrains the identification of target species to match their known references. In applying this prior information and constraint, non-target removal was improved.[1]

2.1 Introduction

The Savannah River Site, as introduced in Section 1.3, has been immobilizing nuclear waste since 1996. The site operates the Defense Waste Processing Facility (DWPF), which processes and vitrifies high-level radioactive waste into borosilicate glass [47, 48, 49, 50, 51]. In the DWPF, the primary chemical preparation steps occur in the Sludge Receipt and Adjustment Tank (SRAT) and Slurry Mix Evaporator (SME). This chapter investigates data relevant to chemical processing in the DWPF and introduces a constrained BSS method to handle unanticipated chemistry in an abundant process chemical, glycolate. BSS is a class of techniques that separate a mixture signal into its source components without knowledge of the original signals nor their mixing matrix. The proposed method improves on prior methods by using the source separation algorithm to identify only unknown species, rather

[1] Much of the content and many of the figures in this chapter are reproduced from Crouse et al. [30]

than all species; this is hypothesized to improve the quality of non-target removal.

The first and second steps of the proposed BSS algorithm use classical least squares (CLS) followed by principal component analysis (PCA) to identify spectral estimates from mixture spectra. The third and final step of the algorithm uses multivariate curve resolution - alternating least squares (MCR-ALS) to iterate the estimates from the first two steps and any available reference spectra into a bilinear mixture model that matches available process spectra. The proposed BSS algorithm improves the estimation of target (quantified) species by removing non-target (unanticipated) contributions from mixture spectra (see Section 1.4.1 for motivation behind target and non-target species).

In the context of process monitoring, the proposed BSS algorithm may reduce the number and frequency of calibration experiments required for spectroscopic monitoring without compromising accuracy, since non-targets may be omitted from training data. In addition, BSS-preprocessing could facilitate real-time results in scenarios where decision-making is time-sensitive, but process spectra deviate from training data due to unforeseen process conditions. Ordinarily, deviations from training data require additional calibration experiments and corresponding process down-time.

In this chapter, attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy was used to quantify anions relevant to nuclear-waste processing at the Savannah River Site. An unknown spectral signature was observed in process data collected in collaboration with SRNL. To address the appearance of this spectral signature, a novel BSS algorithm was developed to preprocess spectral data by removing the influence of non-target species. Additionally, the utility of ATR-FTIR measurements for real-time monitoring was demonstrated in the context of nuclear-waste slurries at the Savannah River Site by quantifying static samples and a continuous 65-hour run of scaled-down SRAT and SME processes.

2.2 Materials and Methods

2.2.1 Overview of Savannah River Site Waste

A simplified process flowsheet for the DWPF is shown in Figure 1.1 in the introduction to this thesis. The physical and chemical composition of a representative sludge simulant (i.e. the SRAT feed) is shown in Table 2.1, while the composition of the final processing points (i.e. the SRAT and SME product slurries) are shown in Table 2.2. The SRAT feed undergoes chemical preparation (acid addition) in the SRAT before receiving glass frit additions and dewatering at the SME. Nitrate (NO_3^-) and nitrite (NO_2^-) were chosen as target species for estimation because of their process relevance (see Table 2.1 and Table 2.2) and potential for in-line monitoring. Other infrared-active species that may be present during waste processing and labeled as non-targets are: glycolate ($C_2H_3O_3^-$), carbonate (CO_3^{2-}), oxalate ($C_2O_4^{2-}$), formate (CHO_2^-), phosphate (PO_4^{3-}), and sulfate (SO_4^{2-}). As demonstrated in the following sections, glycolate spectra show deviations from available reference and training spectra, complicating glycolate quantification and motivating its label as a non-target component. Species other than glycolate are labeled as non-targets because of their small expected process concentrations. Known anion reactions that may occur during the SRAT and SME processes include: nitrite destruction, nitrite-to-nitrate conversion, glycolate destruction, glycolate-to-formate conversion, and glycolate-to-oxalate conversion. Detailed information on the expected anion conversions is provided by Woodham and coworkers [51]. Typically, anion concentrations at the Savannah River Site are quantified with ion chromatography (IC), which is associated with approximately 10% uncertainty and waiting times on the order of days [52]. In addition to ionic species, the simulant also contains the following solids as detected by inductively coupled plasma - atomic emission spectroscopy (ICP-AES): Ag, Al, Ba, Ca, Cr, Cu, Fe, Hg, K, Mg, Mn, Na, Ni, Pd, Rh, Ru, S, Si, Zn, and Zr [51].

Table 2.1: Simulant SRAT feed composition (i.e. expected process input), corresponding to nonradioactive simulated Tank 40-8 measured by [51]. The wt% total solids represents the solids that do not dissolve after heating to 110°C, while the wt% insoluble solids represents the difference between the measured total solids and dissolved solids.

Parameter	**Supernatant Target [mM]**
Na^+	1010
Cl^-	<2.82
NO_2^-	194
NO_3^-	119
SO_4^{2-}	13.7
$C_2O_4^{2-}$	7.00
PO_4^{3-}	<1.05
OH^-	219
CO_3^{2-}	36.1
Parameter	**Slurry Target [wt%]**
Insoluble Solids	8.39
Total Solids	13.90

Table 2.2: Simulant SRAT product and SME product concentrations (i.e. expected midpoint and process output, respectively) corresponding to nonradiaoctive simulated Tank 40-8 SRAT/SME product slurries reported by [51].

Parameter	**SRAT Product Slurry [mM]**	**SME Product Slurry [mM]**
HCO_2^-	17.1	21.4
Cl^-	3.70	<3.78
NO_2^-	<2.48	<2.92
NO_3^-	1070	953
PO_4^{3-}	<1.20	<1.41
SO_4^{2-}	17.5	16.8
$C_2O_4^{2-}$	25.3	30.6
$C_2H_3O_3^-$	906	863
CO_3^{2-}	<6.66	25.1
NH_4^+	<2.77	<2.77

2.2.2 Design of Experiments

Nonradioactive experiments mimicking the DWPF, SRAT, and SME cycles were performed in a 2-L Mettler Toledo Reaction Calorimeter (RC1) vessel equipped with temperature, pH, and ATR-FTIR probes by resesrchers at SRNL. Due to the high solids content, the sludge waste was probed with *in-situ* ATR-FTIR spectroscopy, which was able to measure infrared-active molecules in the solution phase without interference from solid particles because of a shallow laser penetration depth of 2–3 μm [53]. ATR-FTIR-measured concentrations were compared to IC measured concentrations collected by the SRNL Process Science Analytical Laboratory. Before IC measurement, samples were quenched by addition of 50% NaOH, which has been reported to increase the accuracy of ion measurements via IC [52]. Samples were taken from the SRAT/SME cycles at different processing points in two separate experiments, referred to as *Run 1* and *Run 2*. Run 1 process data correspond to five IC measurements from a SRAT/SME experiment (Table 2.3) with corresponding ATR-FTIR spectra. Run 2 process data correspond to three IC measurements from a SRAT/SME experiment (Table 2.4), also with corresponding ATR-FTIR data. Run 2 has an additional 3899 spectra (collected every minute over the course of 65 hours) that do not have associated IC measurements; the spectra will be used to show real-time changes in the SRAT and SME processes. The RC1 vessel was used to collect spectral data for model training[2] (eight experiments shown in Table 2.5), which were designed to match reported anion concentrations in the SRAT/SME sludge simulant (Table 2.1) and SRAT/SME product (Table 2.2) for the most abundant anions: nitrate, nitrite, and glycolate. Training data were collected using sodium salts without controlling pH, so training data pHs are basic but unmeasured.

[2]The eight training data were collected by former Grover Group member, Stefani Kocevska while participating in an internship at SRNL. The Run 1 and Run 2 data were collected by researchers at SRNL testing a sludge simulant [51]. All data preparation, data analysis, and most physical data retrieval was performed by the author of this thesis, partly while participating in an internship at SRNL.

Table 2.3: Process data from Run 1. Anion concentrations were measured by IC for a nonradioactive run of SRAT/SME in the RC1 experimental set-up (pH collected at 93°C and calibrated at 22°C).

Sample	Condition	pH	NO_3^- [M]	NO_2^- [M]	$C_2H_3O_3^-$ [M]
1	SRAT feed	10.12	0.16	0.18	0.00
2	Post nitric acid addition (SRAT)	5.81	0.79	0.15	0.00
3	SRAT product	2.94	0.79	0.03	0.59
4	Post dewater (SME)	3.08	1.28	0.00	0.82
5	SME product	3.14	1.43	0.00	0.86

Table 2.4: Process data from Run 2. Anion concentrations measured by IC for a nonradioactive run of SRAT/SME in the RC1 experimental set-up (pH collected at 93°C and calibrated at 22°C).

Sample	Condition	pH	NO_3^- [M]	NO_2^- [M]	$C_2H_3O_3^-$ [M]
1	SRAT feed	10.19	0.13	0.21	0.00
2	SRAT product	3.02	1.09	0.00	0.70
3	SME product	2.89	1.67	0.00	0.92

2.2.3 pH of Collected Data

The following tables (Table 2.3, Table 2.4, and Table 2.5) show concentration and relative pH data for the experimental data used in this chapter. The pH probe used for pH measurements was calibrated at room temperature (approximately 22°C). However, the training data and run data were collected at 93°C. The temperature difference will cause an under-approximation of pH at the elevated temperatures of the samples in Table 2.3, Table 2.4, and Table 2.5. The pH values are included here to provide relative pH information between measurements.

Table 2.5: Training data for the target anions and glycolate (pH collected at 93°C and calibrated at 22°C).

Sample	pH	NO_3^- [M]	NO_2^- [M]	$C_2H_3O_3^-$ [M]
1	5.82	0.00	0.00	0.00
2	6.53	0.95	0.00	0.00
3	7.49	0.00	0.28	0.00
4	7.87	0.00	0.00	0.79
5	7.58	0.14	0.04	0.81
6	7.50	0.88	0.26	0.75
7	7.36	0.95	0.05	0.00
8	7.16	0.16	0.30	0.00

2.2.4 Blind Source Separation Methodology (Using Principal Component Analysis)

BSS refers to techniques that infer original or source signals solely from measurements of signal mixtures [54]. In the context of spectroscopy and this thesis, BSS can be used to estimate and remove the pure chemical sources that make up a signal mixture, even when those sources are unknown. BSS has been shown by Maggioni et al. and Kocevska et al. to be effective in preprocessing spectra to remove non-target species from Raman and ATR-FTIR spectra in nuclear waste simulants [20, 21]. Kocevska's BSS algorithm incorporated available process information by augmenting mixture data to include known reference spectra, thereby guiding the identification of sources by independent component analysis (ICA) and MCR-ALS. The proposed algorithm of this chapter follows the general structure and justification of Maggioni et al. and Kocevska et al. However, the proposed method does not require algorithmic identification of known sources; known species are constrained to match user-supplied references, while unknown species are found in a "blind" manner similar to Maggioni's and Kocevska's BSS algorithms. The new preprocessing structure allows for available reference spectra to be directly incorporated into a BSS algorithm without estimating them algorithmically. In Kocevska's and Maggioni's works, sources identified with ICA and MCR-ALS, even well-known species, are limited in quality by the accuracy of the BSS algorithm; found sources may not always match user-supplied references. A

quantitative comparison of BSS methods can be found in this chapter in Section 2.3.6. The proposed algorithm gives the user control over which components are modified and subtracted by the BSS algorithm, while still identifying and subtracting unknown species that may arise in real-time due to changing process conditions. In the context of nuclear-waste monitoring, such an algorithm could facilitate real-time spectral preprocessing that utilizes available references for target species, while allowing for continual removal of non-target species that may appear or change throughout processing.

The proposed algorithm uses a combined CLS and PCA step, rather than ICA as reported previously [20, 21], to provide initial guesses for MCR-ALS with a nonnegativity constraint. CLS and PCA were chosen because they are not prone to finding local minima (i.e. introducing rotational ambiguity [55]) compared to ICA; this may improve algorithm consistency in an engineering environment. Deviation from a linear combination of known sources (the error matrix from CLS fitting) is used to identify sources in the mixture spectra that do not correspond to known species. Applying PCA to the CLS error matrix to identify unknown sources from mixture spectra has been reported by Haaland and Melgaard, who referred to this technique as spectral residual augmented classical least squares (SRACLS) [56, 57]. Their method improved the prediction accuracy and robustness of CLS by identifying sources of variation not present in their training data. In this chapter, the residuals of the CLS model fit are analyzed by PCA and then supplied to MCR-ALS to provide a bilinear model for source subtraction. The methodological pipeline is shown in Figure 2.1.

As shown in Figure 2.1a, the first step of the BSS algorithm is the well-established CLS relationship (Equation 2.1), which is used to model mixture spectra with all known reference spectra (targets and non-targets) so that any additional species present can be identified from CLS model error. Model error, in general, accounts for both measurement noise and model mis-match. In the application of spectroscopy, one cause of model mismatch is the presence of unknown species. Minimizing error, $\mathbf{E}_{\mathrm{CLS}}$, in the least-squares sense and solving for concentrations, $\mathbf{C}$, is shown in Equation 2.2 [58]. The least-squares

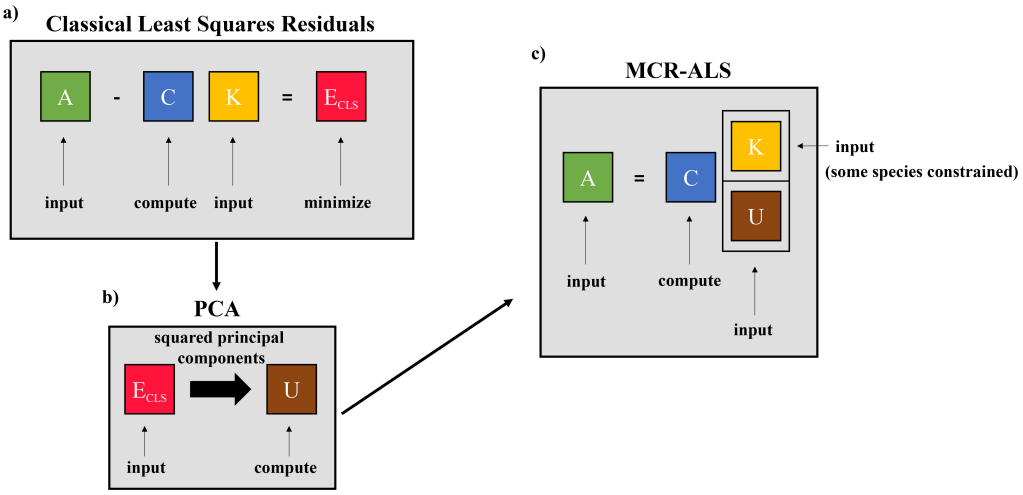


Figure 2.1: Methodological pipeline for the proposed BSS algorithm in three steps: **a)** calculate residuals of CLS where mixture spectra ($\mathbf{A}$) and reference spectra ($\mathbf{K}$) are used to find the concentration matrix ($\mathbf{C}$) that minimizes (in the least-squares sense) the residual matrix ($\mathbf{E}_{CLS}$), which provides an estimate of species that do not have known reference spectra, **b)** perform PCA on the residual matrix ($\mathbf{E}_{CLS}$) and extract the (elementwise) squared principal components ($\mathbf{U}$), and **c)** perform MCR-ALS using known reference spectra ($\mathbf{K}$) and estimates of unknown sources from PCA ($\mathbf{U}$) to identify a physically realistic mixture model that matches the mixture spectra. After the mixture model is identified, source subtraction can be done as is shown in Equation 2.9.

solution can be found by setting $\mathbf{E}_{\mathrm{CLS}} = 0$ and solving for the concentration matrix, ($\mathbf{C}$). Model error can then be solved for in terms of target references, $\mathbf{K}$, and process spectra, $\mathbf{A}$, by rearranging Equation 2.1 and inserting Equation 2.2, yielding Equation 2.3:

$$\mathbf{A} = \mathbf{CK} + \mathbf{E}_{\mathrm{CLS}} \tag{2.1}$$

$$\mathbf{C} = \mathbf{AK}^{\mathrm{T}}(\mathbf{KK}^{\mathrm{T}})^{-1} \tag{2.2}$$

$$\mathbf{E}_{\mathrm{CLS}} = \mathbf{CK} - \mathbf{A} = \mathbf{AK}^{\mathrm{T}}(\mathbf{KK}^{\mathrm{T}})^{-1}\mathbf{K} - \mathbf{A} \tag{2.3}$$

where $\mathbf{A}$ is an $n \times q$ matrix of mixture spectra, $\mathbf{C}$ is an $n \times p$ concentration matrix of pure components, $\mathbf{K}$ is an $p \times q$ matrix of pure component reference spectra of the target species,

and $\mathbf{E}_{\mathrm{CLS}}$ is an $n \times q$ matrix representing noise and error that is not modeled with linear combinations of the known reference spectra. In this chapter, there are n experimental observations, q is the dimension of the spectra (reported wavenumbers), and p is the number of sources for which references exist.

The second step of the proposed BSS algorithm, shown in Figure 2.1b, applies PCA to the residuals from the error matrix of CLS fitting, $\mathbf{E}_{\mathrm{CLS}}$. PCA reduces the dimensionality of data by projecting the data onto a space of reduced dimension while maximizing the variance in the projected data [59, 60]. PCA assumes that the data have been mean-centered, meaning that the data have been mean-centered by subtracting the mean row, $\bar{\mathbf{E}}_{\mathrm{CLS}}$, from each row of the original matrix. The principal components of a matrix can be found as the eigenvectors of the matrix's covariance matrix, $\mathbf{\Phi}$. Calculating the covariance matrix, shown in Equation 3.7, and finding the associated eigenvectors, shown in Equation 3.8, yield the principal components of PCA:

$$\mathbf{\Phi} = \frac{1}{n-1}(\mathbf{E}_{\mathrm{CLS}} - \bar{\mathbf{E}}_{\mathrm{CLS}})^{\mathrm{T}}(\mathbf{E}_{\mathrm{CLS}} - \bar{\mathbf{E}}_{\mathrm{CLS}}) \tag{2.4}$$

$$\mathbf{\Phi}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{2.5}$$

where $\mathbf{v}_i$ is the i^{th} eigenvector corresponding to λ_i, the i^{th} eigenvalue. In the context of the present BSS algorithm, the largest r eigenvalues are retained in addition to the corresponding eigenvectors/principal components, where r is the number of expected sources beyond to the p known references. The principal components from PCA, rather than loadings as used by Haaland and Melgaard, are used for the algorithm introduced in this chapter [57]. Principal components have unit scaling, whereas loadings do not, and so principal components were chosen as "standardized" initial guesses for the subsequent MCR-ALS algorithm. The principal components, once computed, are squared so that the spectra are nonnegative for MCR-ALS (step three). This process is shown in Equation 2.6, where the

elements of each principal component are squared ($\circ 2$ represents the elementwise square or Hadamard power of a vector/matrix). This follows the methodology of Maggioni's two-step BSS algorithm that squares independent components from ICA (which may also have negative components) before inputting into MCR-ALS, which has a nonnegativity constraint applied [20]. Squaring the principal components changes the direction of the principal components away from the directions that maximize variance, but the squared principal components are more physically realistic since nonnegativity is implied by the Beer-Lambert Law in spectroscopy; this trade-off is made with this algorithm.

$$\mathbf{U} = \mathbf{V}^{\circ 2} = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1^{\circ 2} & \mathbf{v}_2^{\circ 2} & \dots & \mathbf{v}_r^{\circ 2} \\ | & | & & | \end{bmatrix} \tag{2.6}$$

The third step of the BSS algorithm, shown in Figure 2.1c, uses the known references and estimated unknown sources from PCA as source estimates for MCR-ALS. MCR-ALS is a bilinear model that decomposes mixture spectra into a concentration matrix, $\mathbf{C}$, and a reference spectra matrix, $\mathbf{S}^{\mathrm{T}}$, as is shown in Equation 2.7 [61, 62, 63]. In Figure 2.1c, the reference spectra are divided into known references (K) and estimated references (U) that are the squared principal components from the PCA step. Lastly, MCR-ALS is used, as shown in the following equation.[3]

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E}_{\text{MCR-ALS}} \tag{2.7}$$

MCR-ALS solves the bilinear problem shown in Equation 2.7 for the contributions ($\mathbf{C}$) and component sources ($\mathbf{S}^{\mathrm{T}}$) that comprise a mixture signal ($\mathbf{A}$) [64, 65, 66]. The alternating least-squares algorithm finds a local solution given initial conditions; for MCR-ALS, initial guesses can be supplied for either $\mathbf{C}$ or $\mathbf{S}^{\mathrm{T}}$. In this chapter, reference spectra ($\mathbf{S}^{\mathrm{T}}$) are supplied which MCR-ALS iterates to match the process spectra. MCR-ALS is

[3] $\mathbf{S}^{\mathrm{T}}$ is shown as concatenated matrices $\mathbf{K}$ and $\mathbf{U}$ in Figure 2.1. This split is shown in Equation 2.8.

calculated with a nonnegativity constraint in this chapter to eliminate negative or physically unrealistic spectra. The initial guess for spectra, $\mathbf{S}^{\mathrm{T}}_{\mathrm{Guess}}$ is given by a vector of concatenated columns of p known reference spectra and r spectral estimates provided by PCA. The number of spectral estimates, r, is determined by the MCR-ALS algorithm.

The MCR-ALS step can determine the number of guesses to provide in Equation 2.8. When given a large number of sources, the MCR-ALS algorithm may calculate some sources to be zero at every wavenumber, which indicates that the proposed source does not contribute significantly to the mixture spectra. The number of spectral guesses from CLS, r, was incrementally increased until a source was calculated as zero by MCR-ALS.[4] Thus, the BSS algorithm can provide a systematic way to determine the number of additional sources to include. Other methods could be utilized to determine the number of expected sources. Previously, singular value decomposition has been used to determine the number of expected sources [20].

Equation 2.8 shows the initial guess provided for $\mathbf{S}^{\mathrm{T}}$. In this chapter, a constraint is applied to the MCR-ALS algorithm so that the target species remain identical to the user-supplied references; this is a default functionality of the MCR-ALS algorithm from the National Institute of Standards and Technology (NIST) and denoted by the bar ($^{-}$) over species held constant in Equation 2.8 [66]. The constraint on target species may be advantageous for a real-time monitoring scenario since the constraint gives the BSS algorithm predictable behavior for the target species, improving model interpretability and reliability. In this chapter, references are provided for $\mathrm{NO_3^-}$, $\mathrm{NO_2^-}$, $\mathrm{C_2H_3O_2^-}$, and $\mathrm{H_2O}$. The MCR-ALS algorithm constrains the calculated spectra to match the user-supplied reference spectra for $\mathrm{NO_3^-}$, $\mathrm{NO_2^-}$, and $\mathrm{H_2O}$.

[4] A tolerance could be prescribed for this method, but there was a clear (i.e. greater than 8 orders of magnitude) demarcation in the data when relevant sources were no longer found.

$$\mathbf{S}_{\text{Guess}}^{\text{T}} = \begin{bmatrix} | & | & | & | & | & | & & | \\ \bar{\mathbf{k}}_{\text{NO}_3^-}^{\text{T}} & \bar{\mathbf{k}}_{\text{NO}_2^-}^{\text{T}} & \mathbf{k}_{\text{C}_2\text{H}_3\text{O}_2^-}^{\text{T}} & \bar{\mathbf{k}}_{\text{H}_2\text{O}}^{\text{T}} & \mathbf{u}_{\text{Guess 1}}^{\text{T}} & \mathbf{u}_{\text{Guess 2}}^{\text{T}} & \cdots & \mathbf{u}_{\text{Guess r}}^{\text{T}} \\ | & | & | & | & | & | & & | \end{bmatrix} \tag{2.8}$$

The MCR-ALS algorithm produces a model that is capable of decomposing the process spectra into estimated source spectra. Subsequently, the sources corresponding to non-targets are subtracted. In this chapter, glycolate is a non-target species and is subtracted, along with any other detected sources. Water, while not a target, is the solvent and not subtracted in this chapter since it does not contribute significant spectral variation between measurements. Equation 2.9 shows the subtraction of non-target sources using the model produced by MCR-ALS.

$$\mathbf{A}_{\text{Preprocessed}} = \mathbf{A} - \begin{bmatrix} — & \mathbf{C}_{\text{C}_2\text{H}_3\text{O}_2^-} & — \\ — & \mathbf{C}_{\text{Guess 1}} & — \\ & \vdots & \\ — & \mathbf{C}_{\text{Guess r}} & — \end{bmatrix} \begin{bmatrix} | & | & | \\ \mathbf{S}_{\text{C}_2\text{H}_3\text{O}_2^-}^{\text{T}} & \mathbf{S}_{\text{Guess 1}}^{\text{T}} & \cdots & \mathbf{S}_{\text{Guess r}}^{\text{T}} \\ | & | & | \end{bmatrix} \tag{2.9}$$

After source subtraction, the resulting preprocessed data can be robust to infrared-active species outside of the training dataset and spectral changes that occur in complex mixtures. Process implementation of BSS preprocessing can enable more accurate quantification of target analytes. However, concurrently operating fault detection algorithms should, in general, utilize preprocessed spectra cautiously, since the BSS preprocessing discussed in this chapter may mask equipment failures or significant spectral changes that may be valuable for detecting sensor or process faults. This idea is revisited in Chapter 7 where preprocessing is performed prior to statistical fault detection.

2.2.5 Quantification

Partial least squares regression (PLSR) (see Chapter B) with four latent variables was used to estimate concentrations from all spectra in this chapter, and is performed after BSS-preprocessing described in Section 2.2.4. Since concentration is restricted to nonnegative values, a nonnegativity constraint is applied to all PLSR results for physical accuracy. Prior to being input into PLSR models but following BSS-preprocessing from Section 2.2.4, all spectra are filtered with the Savitzky-Golay method utilizing seven filter points, a second order polynomial, and a first order derivative to reduce the effect of shifting baseline. Spectra and concentrations are standard normal variate (SNV)-scaled by mean-subtraction and scaling to unit variance immediately prior to PLSR quantification (see Chapter A in the Appendix). More detail on PLSR quantification can be found in Chapter B.

2.2.6 Computation

Python 3.9 was used for all computation and data analysis in this chapter, with the code and experimental dataset from this chapter published on GitHub[5]. The scikit-learn implementations of PLSR, PCA, and FastICA were used, while SciPy was used for Savitzky-Golay Filtering. The NIST package, pyMCR, was used to perform MCR-ALS [66].

2.3 Results and Discussion

ATR-FTIR reference spectra of measurable solution analytes are shown in Section 2.3.1. The ATR-FTIR spectrum of glycolate ($C_2H_3O_3^-$) was observed to display nonlinear peak-shifting as a result of variable process parameters in Section 2.3.2. In Section 2.3.3, the developed BSS algorithm was used to remove a glycolate source from spectra of Run 1 and was used to identify and remove two glycolate sources from Run 2. Lastly five spectra from Run 1 were quantified in Section 2.3.4 and 3902 spectra from Run 2 were quantified in Section 2.3.5.

[5]https://github.com/magrover/Blind_Source_Separation_CLS.PCA.MCR-ALS

Table 2.6: Concentrations of reference spectra shown in Figure 2.2, representative of expected process concentrations from Table 2.1 and Table 2.2. Measured pH was collected at 25°C and calibrated at 22°C.

Reference	Formula	pH	Concentration [M]
Water	H_2O	5.93	55.494
Nitrate	NO_3^-	6.53	0.946
Nitrite	NO_2^-	7.49	0.281
Glycolate	$C_2H_3O_3^-$	7.86	0.795
Carbonate	CO_3^{2-}	11.34	0.093
Sulfate	SO_4^{2-}	5.79	0.037
Formate	CHO_2^-	6.97	0.076
Oxalate	$C_2O_4^{2-}$	7.42	0.024
Phosphate	PO_4^{3-}	7.86	0.021

2.3.1 Reference Spectra

ATR-FTIR reference spectra were collected to determine spectral signatures of possible analytes in the SRAT and SME vessels. While nitrate and nitrite represent the target species for the slurry, the slurry may include additional anions at low concentrations (see Table 2.1 and Table 2.2). Reference data for all known infrared (IR)-active species at approximate SRAT/SME concentrations are shown in Figure 2.2 with reference spectra concentrations listed in Table 2.6. The target species are highly IR-active, while some of the non-target species (such as oxalate and formate) have weak IR signals at process-relevant concentrations. Since the non-target species, excluding glycolate, are not present at high concentrations in the waste, it is possible that they do not significantly interfere with the signals of the targets. The peaks of non-target species (shown in Figure 2.2b) other than glycolate were not included as references in the BSS algorithm. Carbonate and sulfate are 2.58 and 2.70 times more concentrated, respectively, in the references of Figure 2.2b than in the typical feed to the SRAT (Table 2.1). Carbonate, sulfate, and other non-targets were not reliably observed in the mixture spectra, and so the BSS algorithm is tasked with identifying these species if they are significantly present in the solution.

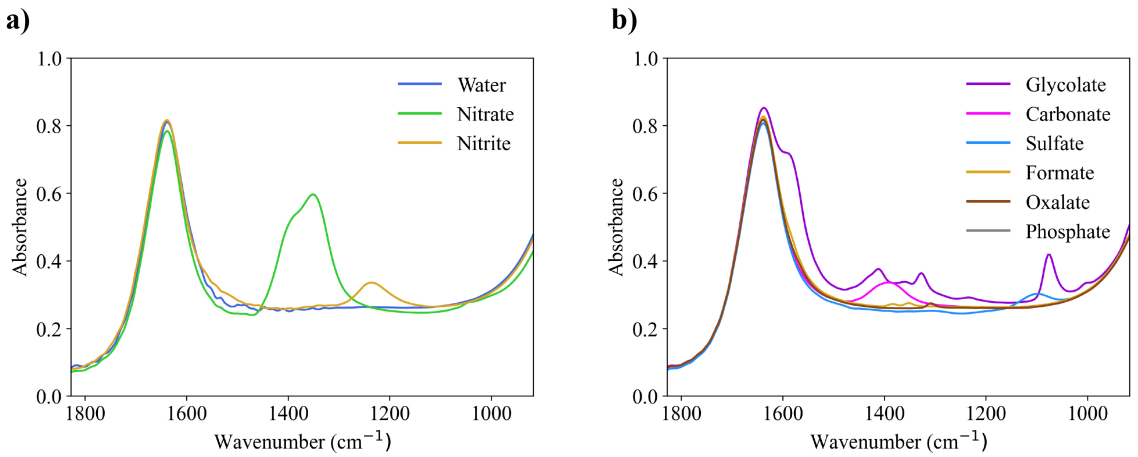


Figure 2.2: Reference spectra of aqueous **a)** target anions (including water) and **b)** non-target anions found in the DWPF waste during the SRAT and SME processes collected at 25°C. Corresponding concentrations are listed in Table 2.6.

2.3.2 Processing Variables: Temperature and pH

Processing variables can affect ATR-FTIR signals, which affect both interpretation and reliability of signal-to-composition models. Processing variables of interest for the SRAT/SME process included temperature and pH. Temperature started at room temperature (approximately 25°C) but was near-boiling (approximately 93°C) for much of the process. Likewise, pH began at 13 and dropped to 4 through acid addition. In this section, variations in spectra with processing parameters will be discussed to determine what non-linearities, if any, were present in the ATR-FTIR spectra of the SRAT and SME processes.

Low-temperature (25°C) and high-temperature (80–93°C) 1 M reference spectra of nitrate (NO_3^-), nitrite (NO_2^-), glycolate ($C_2H_3O_3^-$), and glycolic acid ($C_2H_4O_3$) are shown in Figure 2.3[6]. Nitrate, shown in Figure 2.3a, had a temperature-dependent peak shift that can be seen on the lower-wavenumber side of the 25°C nitrate peak. The nitrate anion (NO_3^-) has previously been shown to display peak-shifting in the Raman spectrum resulting from ion association [67]. In addition, temperature-dependent ion association has been demonstrated by Yu et al. [68]. At elevated temperatures (40°C $\rightarrow$ 80°C) and high concentrations

[6]These experiments were conducted at Georgia Tech by the author of this thesis.

of sodium nitrate, Yu et al. attributed shifts in the Raman spectrum to complex aggregated contact ions, indicating that sodium (Na^+) and nitrate (NO_3^-) tended to aggregate at higher temperatures. There was a decrease in the total peak area and a shifted location of the nitrate peak, possibly indicating affected sub-peaks for the nitrate anion. The peak location for nitrite, glycolate, and glycolic acid all displayed subtle temperature variations with our instrument, shown in Figure 2.3b–d. However, pH had a greater impact than temperature on Fourier transform infrared (FTIR) spectra at the Savannah River Site.

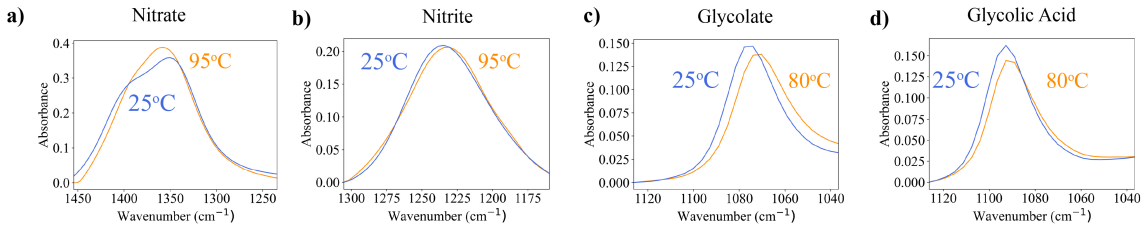


Figure 2.3: Comparison of 1 M ATR-FTIR absorbance spectra at low and high temperatures for **a)** nitrate (NO_3^-), **b)** nitrite (NO_2^-), **c)** glycolate ($C_2H_3O_3^-$), and **d)** glycolic acid ($C_2H_4O_3$).

The feed stream to the SRAT typically had a pH of 13, which decreased to a pH around 4 after the addition of both nitric acid and glycolic acid. This pH shift affected the speciation of weak acids, such as glycolic acid, in the solution phase. Glycolate began SRAT processing in a deprotonated form, due to the high solution pH, and shifted to a protonated form as the solution pH decreased. Glycolate speciation, shown in Figure 2.4a, was calculated by using the pKa of glycolic acid and Equation 2.10. Based on values available in the literature, a pKa value of 3.83 for glycolic acid at 25°C was used [69].

$$mole\ fraction\ of\ protonated\ acid = \frac{10^{-pH}}{10^{-pH} + 10^{-pKa}} \tag{2.10}$$

The ATR-FTIR spectra of glycolate, after the spectra of water had been subtracted, are shown as a function of pH in Figure 2.4b. As can be seen from the magnified region in Figure 2.4b, the glycolate peak matched the peak seen in the training data at high pH. As pH was lowered, another peak appeared on the higher-wavenumber side of the high-pH

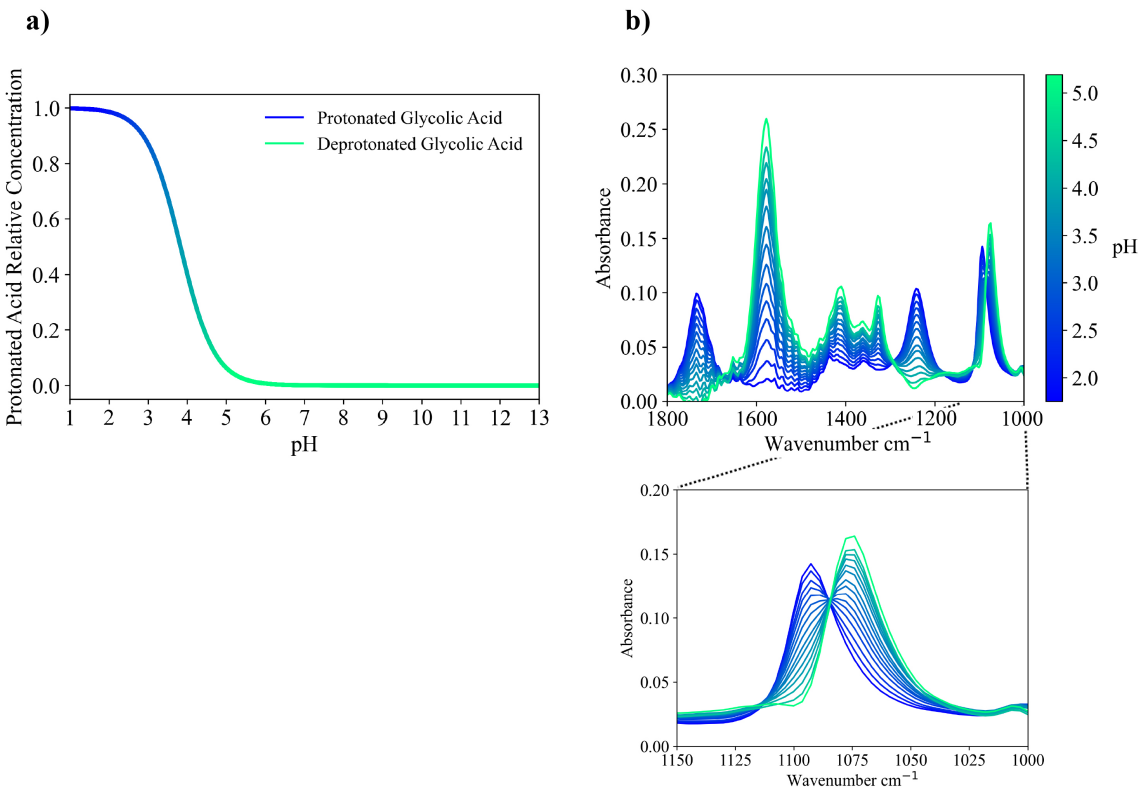


Figure 2.4: **a)** Protonation of glycolic acid calculated as a function of pH using a pKa of 3.83 [69] and **b)** reference spectra (water subtracted) for glycolate at 25°C as a function of pH.

glycolate peak while the original peak decreased. The result was a 15 cm^{-1} shift in the 1078 cm^{-1} glycolate peak to 1093 cm^{-1}. The acidic form of nitrite (NO_2^-), nitrous acid (HNO_2), is a weak acid with a pKa of 2.3 [70]. However, nitrite-destruction reactions decreased nitrite below the limit of detection for ATR-FTIR before nitrous acid was expected to be detectable in solution.

Since the key measurements during DWPF processing occurred at a constant temperature of 93°C, the training data for the system were collected at 93°C and are shown in Figure 2.5a. The process data, shown in Figure 2.5b (Run 1) and Figure 2.5c (Run 2), were sampled during runs of the SRAT/SME cycle. The spectra baseline and the shape of the glycolate ($C_2H_3O_3^-$) peak differed between the training and process spectra shown in Figure 2.5. The effect of the variable baseline was minimized by taking the first derivative

of the spectra using the Savitzky-Golay filter (not shown). The glycolate peaks in both process runs deviated from that of the training data, as is shown by the magnified sections in Figure 2.5. Specifically, the glycolate peak from Run 1 (Figure 2.5b) has a shoulder on the higher-wavenumber side that was not present in the training data (Figure 2.5a). Run 2 (Figure 2.5c) had, by comparison to Run 1, a sharper glycolate peak that had a smaller shoulder on the higher-wavenumber side and shoulder on the lower-wavenumber side as well. The peak-shifting from Figure 2.4 may have accounted for much of the difference observed between training data and process data in Figure 2.5. These changes in the glycolate peak location and shape may have affected a spectra-to-concentration model fit using the training data from Figure 2.5a, where the glycolate peak was in a single location.

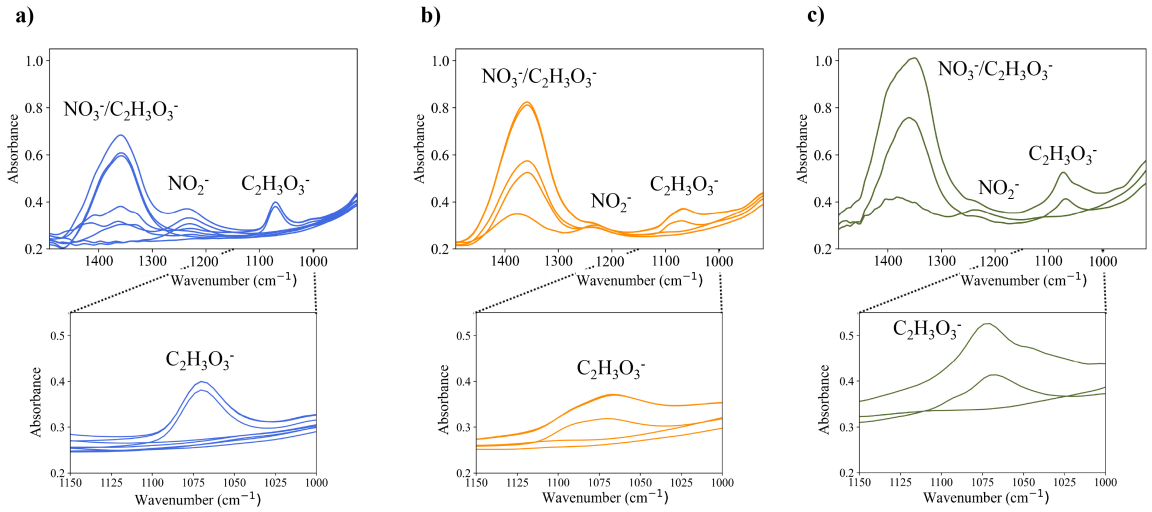


Figure 2.5: **a)** PLSR model training data, **b)** process data from Run 1 collected from five different conditions in the SRAT/SME process, and **c)** process data from Run 2 from three different conditions in the SRAT/SME process. The magnified region (1000-1150 cm^{-1}) highlights the glycolate ($C_2H_3O_3^-$) peak.

2.3.3 BSS Source Matching

In this chapter, nitrate and nitrite were target species to be quantified. Therefore, it was the objective of the BSS algorithm to remove non-target species: known sources that

deviate from training conditions (glycolate) and unknown sources with unknown spectra (potentially oxalate, carbonate, formate, phosphate, and sulfate). While the presence and approximate concentration ranges of non-target components may have been known, much of their chemistry had not been fully elucidated in the complex mixtures present in the DWPF.

The BSS algorithm proposed in Section 2.2.4 was run separately on the process spectra from Run 1 (Figure 2.5b) and Run 2 (Figure 2.5c). The BSS algorithm was provided the target reference spectra (including water) from Figure 2.2a and an initial guess for glycolate from the high-pH reference in Figure 2.2b. The MCR-ALS step determined the number of additional sources, beyond glycolate, to be zero for Run 1 and one for Run 2.

Run 1 BSS source matching is shown in Figure 2.6a, where the BSS-determined glycolate source (pink) was compared to the high-pH glycolate reference (dashed black) and the mixture spectra (blue) of Run 1. BSS, particularly the MCR-ALS step, estimated the glycolate contribution to be broader than the supplied reference, as can be seen by the pink curve having greater peak width at the 1080 cm^{-1} peak than the dashed black curve. The MCR-ALS algorithm produced this "widened" glycolate source by altering the provided reference spectra of non-targets to match the calculated bilinear model (Equation 2.7) with the experimental mixture spectra from Figure 2.6a (shown in blue). From Figure 2.6a, the BSS-determined source (pink) better matched the qualitative shape of the measured (blue) peak at 1080 cm^{-1} than the supplied high-pH glycolate reference (dashed black).

Run 2 BSS source matching is shown in Figure 2.6b, where continuous run data (3902 spectra) of the SRAT and SME processes were analyzed. A source (brown) was identified in Figure 2.6b that did not match any user-input reference sources. The proposed source resembled a baseline shift, in addition to model mismatch in the region of glycolate (1080 cm^{-1}) and nitrate, carbonate, and glycolate (1410 cm^{-1}). The discovered peak centered on 1410 cm^{-1} may have corresponded to glycolate or carbonate, both of which have associated peaks in that location. The combined baseline and model mismatch may have demonstrated

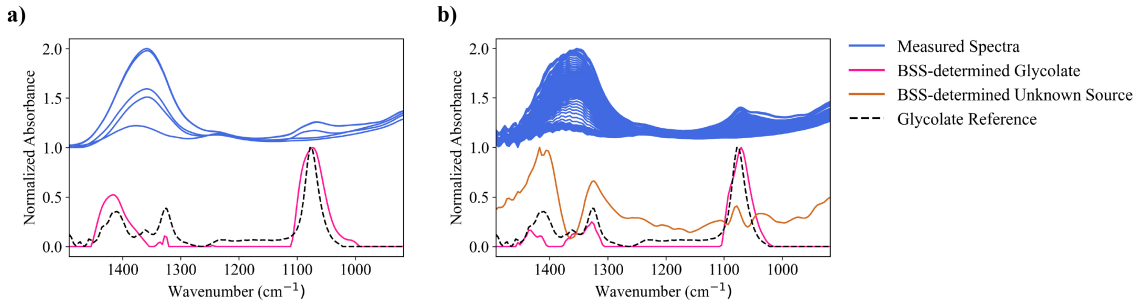


Figure 2.6: Comparisons between measured spectra (93°C), high-pH glycolate reference (25°C), BSS-estimated glycolate reference, and a calculated source from the BSS algorithm for **a)** Run 1 and **b)** Run 2.

a limitation of using PCA (or other latent variable methods) to identify sources; principal components may not have a single physical interpretation and may instead be a combination of sources. During Run 2 (Figure 2.6b), the glycolate peak (pink) was again "widened" to match the appearance of glycolate in the mixture spectra (blue).

2.3.4 Quantification of Anions in SRAT/SME Samples (Run 1)

Figure 2.7a shows BSS preprocessing and its removal of estimated glycolate contributions from process spectra of Run 1. Peaks were subtracted in the vicinity of 1080 cm^{-1} and 1450 cm^{-1}. The parity plots comparing the concentration predictions for PLSR models using either original spectra or BSS-preprocessed spectra are shown in Figure 2.7b. Quantifiable improvements were achieved with BSS preprocessing applied and are shown in Table 2.7. Nitrate quantification was improved from an R^2 value of 0.934 to an R^2 value of 0.988 with BSS preprocessing applied to the spectra. Nitrite quantification was also improved, but the quantification at process concentrations was limited for both original and preprocessed spectra. This error was likely due to low concentration of the nitrite anion in solution and corresponding low intensity of the nitrite peak in the FTIR spectrum. In Table 2.7, mean percent error (MPE) was not quantified for nitrite since it had IC-measured concentrations of zero, which caused a division-by-zero error.

Applying BSS to the process data improved the quantification of nitrate and possibly ni-

Table 2.7: Table of error metrics corresponding to Figure 2.7b.

Original	**Nitrate**	**Nitrite**
Coefficient of Determination (R^2)	0.934	0.605
Root Mean Squared Error (mol/L)	0.0130	0.0023
95% Confidence Interval (mol/L)	0.183	0.081
Mean Percent Error (%)	11.7	–

Preprocessed	**Nitrate**	**Nitrite**
Coefficient of Determination (R^2)	0.988	0.661
Root Mean Squared Error (mol/L)	0.0025	0.0020
95% Confidence Interval (mol/L)	0.081	0.076
Mean Percent Error (%)	9.91	–

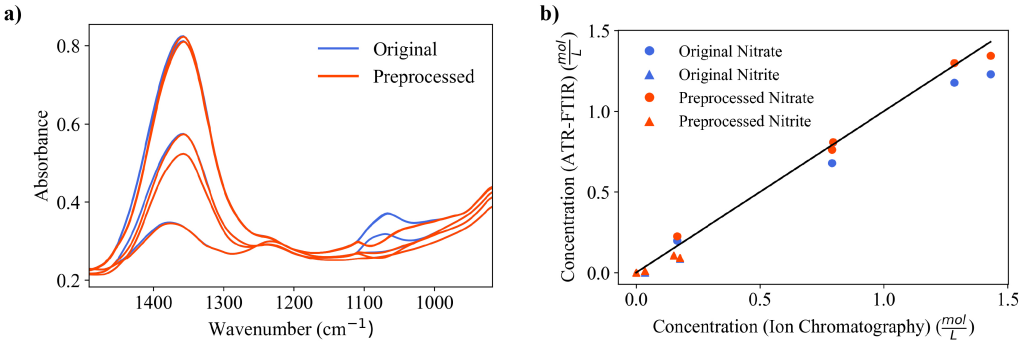


Figure 2.7: **a)** Overlay of Run 1 spectra before and after BSS-preprocessing; **b)** parity plot comparing Run 1 concentration predictions for a PLSR model (blue) and PLSR model with BSS-preprocessing applied (red).

trite while also providing visually interpretable results through peak subtraction. The peak that was removed at 1080 cm^{-1} corresponded to glycolate based on process knowledge and peak location. However, this glycolate peak had a different shape than the supplied glycolate reference and the spectra that appeared in the training dataset. The BSS algorithm was able to improve the initial guess of the glycolate spectrum to better match the components observed in the mixture spectra, which resulted in an accurate bilinear mixture model that allowed glycolate contributions to be removed for improved model quantification.

Table 2.8: Table of error metrics corresponding to Figure 2.8b.

Original	**Nitrate**	**Nitrite**
Coefficient of Determination (R^2)	0.267	0.772
Root Mean Squared Error (mol/L)	0.2959	0.0022
95% Confidence Interval (mol/L)	0.831	0.074
Mean Percent Error (%)	59.4	–

Preprocessed	**Nitrate**	**Nitrite**
Coefficient of Determination (R^2)	0.978	0.703
Root Mean Squared Error (mol/L)	0.0087	0.0029
95% Confidence Interval (mol/L)	0.139	0.084
Mean Percent Error (%)	20.2	–

2.3.5 Quantification of Anions in Continuous SRAT/SME Samples (Run 2)

Figure 2.8a shows BSS preprocessing and its removal of estimated glycolate contributions from 3902 process spectra. The parity plots comparing the concentration predictions for PLSR models using both original spectra and BSS preprocessing are shown in Figure 2.8b. Quantifiable improvements were achieved with BSS preprocessing for nitrate quantification. The model error is shown in Table 2.8. Nitrate quantification was improved from an R^2 value of 0.267 to an R^2 value of 0.978 with BSS preprocessing applied to the spectra. Nitrite quantification was less accurate with an initial R^2 of 0.722 which decreases to 0.703 with BSS preprocessing. A limitation of nitrite quantification results for Run 2 were that only three timepoints had corresponding IC-reported concentration values, and two of the three timepoints reported the nitrite IC concentration at zero (measured below 100 ppm). Therefore, there was a single measurement containing nitrite for quantification for Run 2. Nitrate, however, was present in significant amounts in all three timepoints in Figure 2.8b.

ATR-FTIR has the capability for real-time measurements, whereas IC measurements often incur measurement delays. Figure 2.9, which reported concentrations at every timepoint for Run 2 along with the three timepoints with IC data, highlighted the distinction between time resolution provided by ATR-FTIR and IC for a typical SRAT/SME cycle.

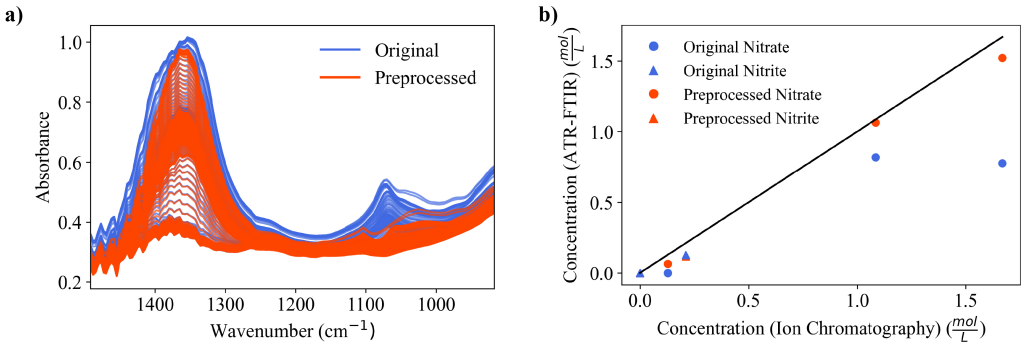


Figure 2.8: **a)** Overlay of 3902 process spectra from Run 2 before and after BSS preprocessing; **b)** parity plot comparing Run 2 concentration predictions for a PLSR model (blue) and a PLSR model with BSS preprocesssing applied (red).

Glycolate contributions were removed using BSS preprocessing in Figure 2.9a, whereas no BSS preprocessing was applied in Figure 2.9b. In Figure 2.9a, only two species were quantified: nitrate and nitrite. However, these two species matched the reported IC concentrations (denoted by the circles and triangles) more closely throughout processing than the equivalent quantification with no preprocessing in Figure 2.9b. Most notable is the difference in nitrate quantification, which was greatly improved when BSS preprocessing is applied. The better agreement of BSS-preprocessed ATR-FTIR prediction with IC measurements in Figure 2.9a had the drawback of not quantifying glycolate, even if Figure 2.9b suggested that glycolate was quantified poorly given the available training data. In future cases where it is desirable to improve quantification of nitrate and nitrite via BSS and additionally quantify glycolate, two separate models could be constructed for providing original glycolate estimates (still including any errors of the original spectra) while improving nitrate and nitrite estimates through source subtraction.

Based on the results for quantifying nitrate and nitrite, ATR-FTIR was able to measure the concentration of target anions as they underwent additions and reactions in the SRAT/SME processes. Nitrate concentration, monitored via ATR-FTIR in Figure 2.9, sharply increased 8 hours into Run 2. The measured nitrate increase corresponded to the addition of nitric acid in the SRAT, which indicated ATR-FTIR can verify nitric acid addi-

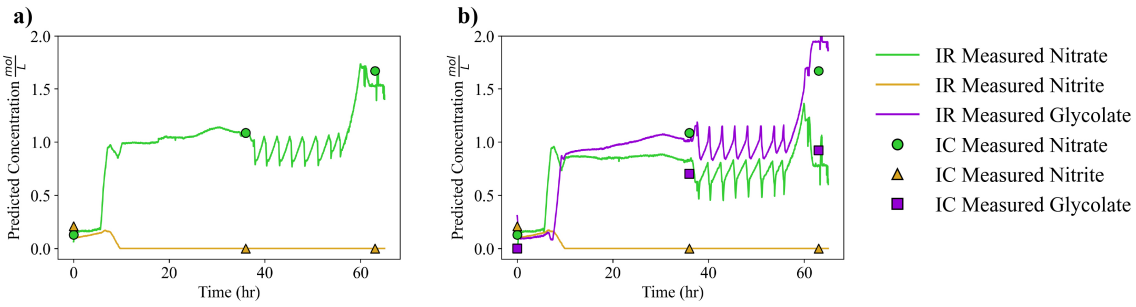


Figure 2.9: Continuous SRAT and SME data (Run 2) quantified using **a)** BSS to remove glycolate contributions and a PLSR quantification model and **b)** only a PLSR quantification model where glycolate is quantified and not removed.

tion. Additionally the saw-tooth pattern appearing in Figure 2.9 around 40 hours into the process corresponded to the repeated addition of water followed by dewatering. Dewatering was observed as an increase in concentration as the solvent evaporates, concentrating the remaining solution. In SRAT/SME processing, solution-phase concentration information during the dewatering step can be used to verify that the expected mass is evaporated. By measuring nitrate concentration, unexpected changes in heating efficiency, changes in specific heat of the feed stream, or clogs in the vapor outlet could be detected and undergo further troubleshooting. Similarly, the nitrite anion can be monitored to ensure that all nitrite is destroyed through acid addition, at least to the limit that is detectable with ATR-FTIR. In Figure 2.9b, qualitative information was provided for the glycolate anion despite quantitative inaccuracy. The glycolate concentration was seen to sharply increase shortly after the nitrate concentration increased. This is from the glycolic acid addition, which followed the nitric acid addition and introduced glycolate to the SRAT. The acid additions also lowered the solution pH to about 4, which caused some glycolate to exist in its protonated form for the remaining duration of SRAT/SME processing.

The presence of nitrate, an abundant analyte that is active in the IR spectrum, may allow for other nonvolatile and nonreactive solution species to be estimated with a mass-balance during the chemical additions, dewatering process, and up until the slurry is transported to the melter. As the final control point before the slurry is melted, close monitoring of the

solution-phase in the SRAT and SME could accelerate and support decisions made by the DWPF Analytical Laboratory regarding waste batch approval [47].

2.3.6 Comparison with Maggioni's and Kocevska's Blind Source Separation Model

A comparison was performed of the BSS method proposed by Maggioni et al. [20] for spectra preprocessing and subsequently adapted for real-time spectra quantification by Kocevska et al. [21]. To conduct the comparison, data from *Run 2* were preprocessed and subsequently quantified by PLSR as described in Section 7.2.6. Hyperparameters were selected for the MCR-ALS step of both BSS algorithms. Hyperparameter tuning consisted of a grid-search of the best regularization coefficient (α value) and L1 ratio for the elastic net that performs the alternating-least-squares in this implementation of MCR-ALS [66]. To perform hyperparameter optimization, 755 spectra collected in the same manner as Run 1 (but without concentration information and not used elsewhere in this thesis) were used to identify the best-performing parameters based on the five Run 1 spectra with reported concentrations. After hyperparameter tuning, Run 2 spectra were plotted and quantified after preprocessing from both BSS methods.

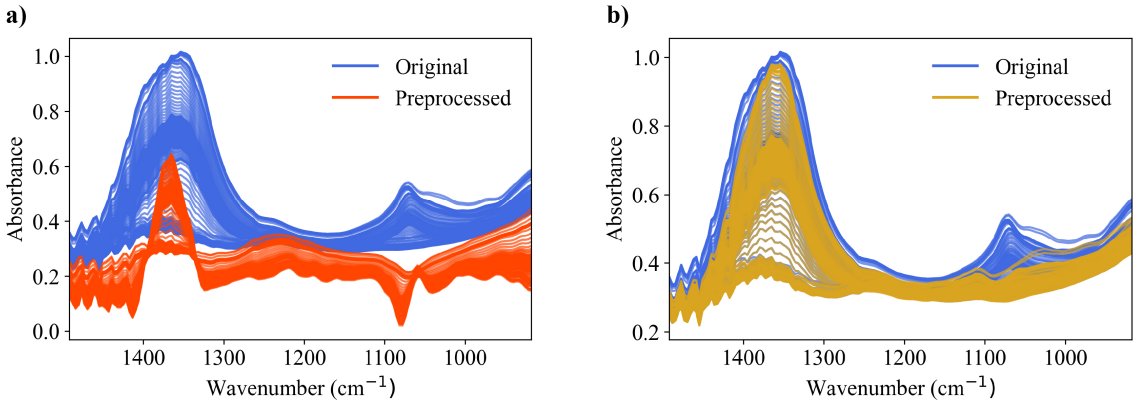


Figure 2.10: ATR-FTIR spectra from Run 2 that has been preprocessed with **a)** the previously reported BSS algorithm from [20] and [21] and **b)** the proposed BSS algorithm in this chapter. The blue spectra in both **a)** and **b)** correspond to measured spectra with no preprocessing.

Figure 2.10 shows the qualitative differences between the two BSS methods. The

method of Maggioni and Kocevska, using ICA to initially find component guesses, produced spectra with a physically unrealistic appearance (even of the known targets). The proposed method, shown in Figure 2.10b, demonstrated a less substantial alteration to the structure of the entire spectra. This resulted in spectra that retained physical realism, especially for the target species. The proposed algorithm constrained the target species to match their reference spectra, so only non-target species were identified and subtracted by the BSS algorithm.

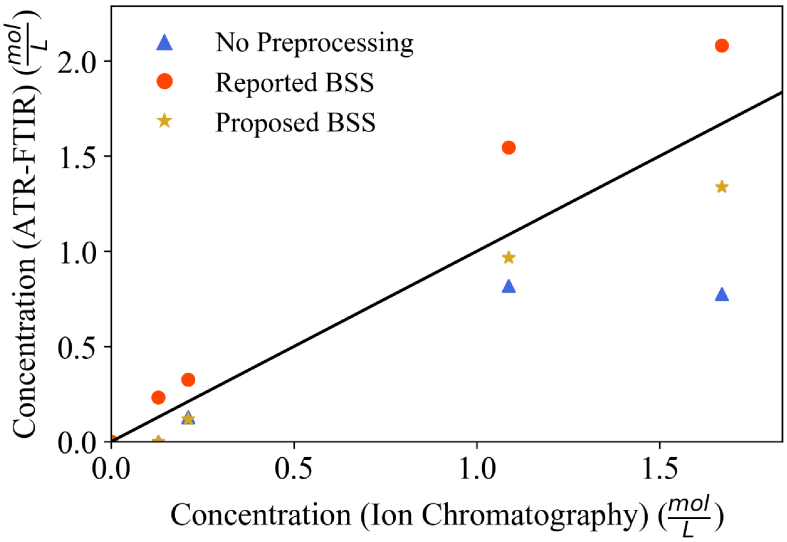


Figure 2.11: Nitrate and nitrite quantified from *Run 2* using the methodology described in Section 7.2.6 (including a first derivative Savitzky-Golay filter and PLSR quantification). Previously reported BSS (red circles) was preprocessed with the methodology of Maggioni [20] and Kocevska [21] while Proposed BSS (gold stars) was preprocessed with the BSS methodology of this chapter (Section 2.2.4).

Table 2.9: Table of error metrics corresponding to Figure 2.11

Root Mean Squared Error (mol/L)	**Nitrate**	**Nitrite**
No Preprocessing	0.296	**0.00223**
Reported BSS of Maggioni et al.	0.133	0.00445
Proposed BSS of this thesis	**0.047**	0.00277

Figure 2.11 and accompanying Table 2.9 quantitatively demonstrated the effect of the BSS preprocessing in Figure 2.10. The proposed BSS methodology improved nitrate quantification, outperforming both the previously reported BSS algorithm from Kocevska and Maggioni and spectra with no BSS preprocessing.

2.3.7 Including Fewer and Additional Glycolate Sources

Additional analyses were performed to test the proposed BSS algorithm performed in two additional scenarios. The first is when no glycolate reference is available, and the algorithm is truly "blind." The second is when both low- and high-pH glycolate references are available. Results from both of these tests are shown in Table 2.10.

BSS preprocessing was performed where the algorithm had neither the low-pH nor high-pH glycolate references available as a non-target species *a priori*. The algorithm was applied to Run 1 and Run 2 identically as was done in the previous sections. The results are displayed in Figure 2.12a, Figure 2.12b, Figure 2.13a, and Figure 2.13b. There were subtle differences in the preprocessed spectra in Figure 2.12a (no glycolate sources) and Figure 2.7a (high-pH glycolate source). As expected, the glycolate source identification was not as accurate without any glycolate source information (Figure 2.12), since the source subtraction affected a wider range of the spectrum. Removal was still effective, but did not fully match the curvature of the observed glycolate peak. Changes were seen for Run 2 as well. The nitrate peak at 1350 cm^{-1} was observed to decrease more substantially in Figure 2.13a compared to Figure 2.8a.

Table 2.10: Table of error metrics for Run 1 and Run 2 quantified with PLSR. BSS preprocessing is performed with different levels of glycolate information provided *a priori*.

Root Mean Squared Error ($\frac{mol}{L}$)	**Run 1**		**Run 2**	
	Nitrate	Nitrite	Nitrate	Nitrite
No Preprocessing	0.0130	0.0023	0.2959	0.0022
BSS + No Glycolate Source	0.0303	0.0026	0.0272	0.0191
BSS + High-pH Glycolate Source (Section 2.3.4)	0.0025	**0.0020**	0.0087	0.0029
BSS + High- and Low-pH Glycolate Sources	**0.0021**	0.0022	**0.0036**	**0.0019**

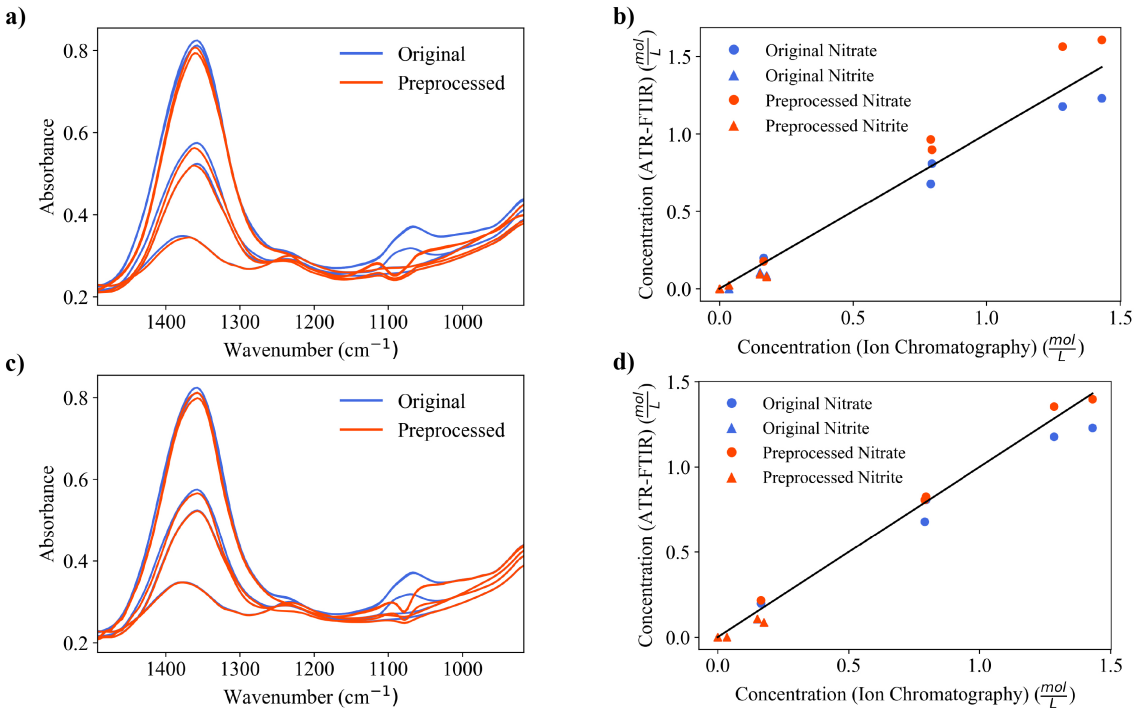


Figure 2.12: Run 1 BSS preprocessed with **(a, b)** no glycolate references included and **(c, d)** both high- and low-pH glycolate references included.

BSS preprocessing was performed where the algorithm additionally had a low-pH glycolate reference available in addition to the high-pH glycolate reference included as a non-target species. The algorithm was applied to Run 1 and Run 2 identically to Section 2.2.4, except for Run 1 where the regularization in MCR-ALS was decreased by a factor of 10 to improve algorithm convergence. The convergence issue may have arisen because an extra source was included (for six total) with the small dataset used in Run 1 (eight training data and five testing data). The results are displayed in Figure 2.12c, Figure 2.12d, Figure 2.13c, and Figure 2.13d. There were subtle differences in the preprocessed spectra in Figure 2.12c and Figure 2.7a. There was more noticeable removal in the area of the nitrite peak at 1250 cm^{-1}, corresponding to the glycolate reference for low pH. Therefore, the low-pH source was more effectively incorporated into BSS preprocessing when it was provided as a reference source. In addition, the glycolate removal at 1080 cm^{-1} was effective but resulted in a preprocessed spectrum with different curvature. The curvature was indicative of subtracted

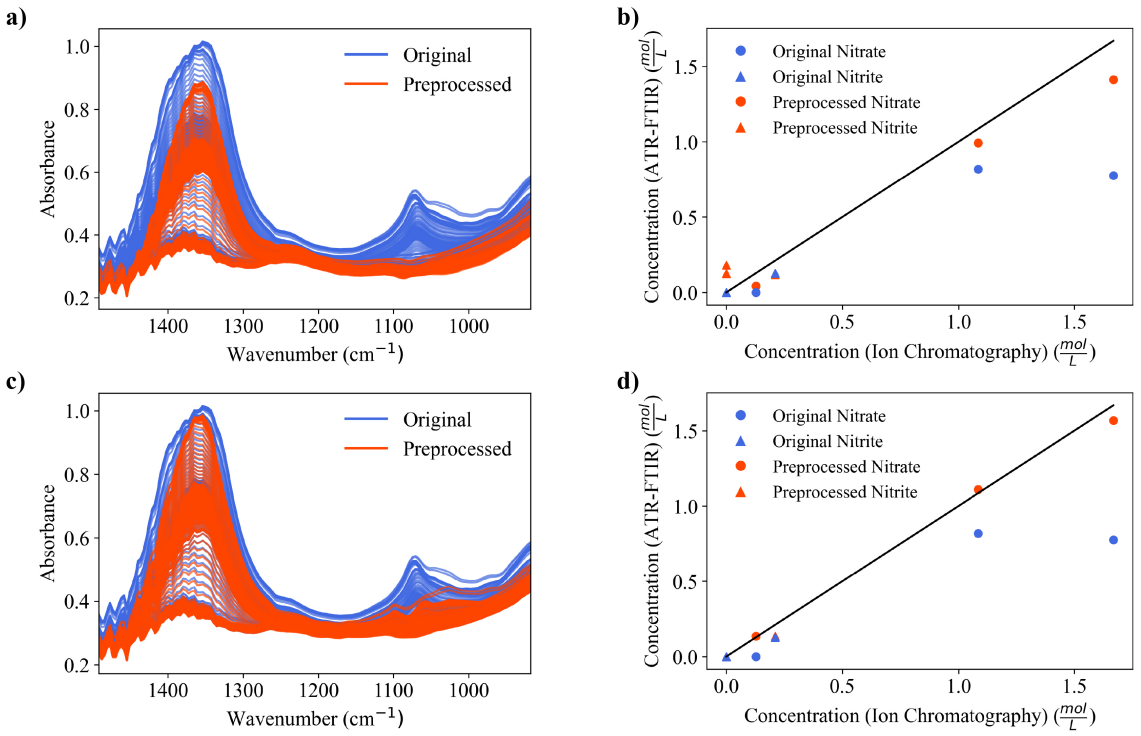


Figure 2.13: Run 2 BSS preprocessed with **(a, b)** no glycolate references included and **(c, d)** both high- and low-pH glycolate references included.

glycolate peaks that were sharper than that shown previously in this chapter (Section 2.3.4 and Section 2.3.5). Figure 2.13c (both high- and low-pH glycolate sources) did not show any different trends than that of Figure 2.8a (high-pH glycolate source). However, the quantification was slightly better as measured by root mean squared error (RMSE) with both sources included, as shown in Table 2.10.

2.4 Conclusion

The work of this chapter serves as both a proof of concept on the use of ATR-FTIR spectroscopy for monitoring slurry samples at the Savannah River Site and an example of a novel BSS algorithm for improving quantification of complex mixture spectra. ATR-FTIR spectroscopy performs rapid process measurements compared to typical IC sample analysis, enabling real-time monitoring and decision-making. However, the complex chemistry and variable process parameters in vessels at the Savannah River Site necessitate ATR-

FTIR spectra-to-composition models that are robust to changing chemical and process conditions. In this chapter, measured ATR-FTIR spectra were combined with a BSS algorithm to overcome limited training spectra that do not match process spectra. Specifically, the glycolate anion ($C_2H_3O_3^-$) was observed to change spectroscopic behavior in slurries typical of the Savannah River Site, which was attributed to shifting pH from nitric and glycolic acid additions during processing. To address the behavior of the glycolate anion in future Savannah River Site monitoring tasks, a spectroscopic training set may be constructed that probes the full range of process-relevant pH's so that acidic and basic forms of glycolate are included in spectroscopic training data.

The concentrations of target species, nitrate (NO_3^-) and nitrite (NO_2^-), were predicted by PLSR using both raw and BSS-preprocessed spectra. For two different runs of the SRAT/SME processes, nitrate quantification improved from an R^2 of 0.934 to 0.988 and from 0.267 to 0.978 when subtracting overlapping BSS-estimated glycolate peaks from measured spectra. BSS preprocessing may be useful even when reference spectra are available, since process conditions can stray from well-controlled bounds where quantification models are typically designed, impacting the spectral signatures of key species. Beyond nuclear-waste processing, applications of the presently discussed BSS technique may be found in instances where spectral quantification of complex mixture spectra is necessary, but process information is limited *a priori* or process conditions vary.

CHAPTER 3
FEATURE SELECTION USING NONNEGATIVELY CONSTRAINED CLASSICAL LEAST SQUARES

This chapter deals with the issue raised by the case study presented in the last chapter (Chapter 2): how can unanticipated chemical species be removed so that target species can be accurately quantified? The method of this chapter, in contrast to Chapter 2, is adept at removing non-target components in real-time. Methods exist in the literature (and the prior chapter) that are able to accomplish this task. However, no existing methods are able to accurately remove non-target contributions from spectra with a single process spectrum; this is highlighted in Figure 3.9 in this chapter. To address this, a new method is introduced called nonnegatively constrained classical least squares (NCCLS) that is able to accurately remove non-target contributions with nothing other than reference spectra of target species, estimation of sensor error, and an assumption that non-target species have spectral *nonnegativity*.[1]

3.1 Introduction

Unknown chemical species (i.e. non-targets, adulterants, interferents, etc.) can disrupt quantification models when applying spectroscopy to monitor industrial processes. Methods exist for identifying and removing non-target spectral sources so that a regression model can accurately quantify spectra with non-target spectral contributions.

Using a model for a new task outside of the domain of its training data is referred to in the broader machine learning literature as *transfer learning*. Transfer learning is a broad class of techniques that allow for models or algorithms trained on one set of data to be effectively applied to another set of data that may violate the common assumption of inde-

[1] Much of the content and many of the figures in this chapter are reproduced from Crouse et al. [71]

pendent and identically distributed (i.i.d.) data between training and application datasets. In the context of industrial spectroscopy, this includes using lab-created spectroscopic datasets to quantify a process that may have changing conditions or product requirements. Using the transfer learning classes distinguished by Pan et al., different model training and application domains correspond to transductive transfer learning [72]. In transductive transfer learning, the training and application goal is the same (regression of target species), while the domains may be different (the presence of non-target components). In the spectroscopic literature, this type of problem has seen substantial work, although not always under the guise of transfer learning. Haaland et al. developed spectral residual augmented classical least squares (SRACLS) to increase spectroscopic quantification accuracy when unmodeled components are included in measured spectra [57, 56, 73]. In SRACLS, references for unknown sources are first estimated and then included in a classical least squares (CLS) prediction scheme [74, 75]. More recently, work by Maggioni et al. and Kocevska et al. used blind source separation (BSS) algorithms to identify and remove non-target species from spectra [20, 21], which were improved in the prior chapter (Chapter 2). Neural networks have also been applied to "denoising," where spectra have been processed via autoencoders [76, 77, 78, 79].

In this chapter, the removal of non-target species from spectra is referred to as *preprocessing*. This definition of preprocessing is distinct from other processing steps that spectra may undergo prior to quantification such as scaling or baseline correction. The preprocessing methods introduced in the previous paragraph have primarily been examined in *batch* processing scenarios, which here refers to multiple spectra being analyzed simultaneously. Batch processing contrasts with the potentially industrially relevant scenario of *real-time* processing, where a single spectrum containing non-target species must be preprocessed for timely decision-making. Preprocessing to remove unwanted spectral signals can be demonstrated by CLS, a physically insightful projection method.

CLS begins by projecting typically high-dimensional spectroscopy data into a space

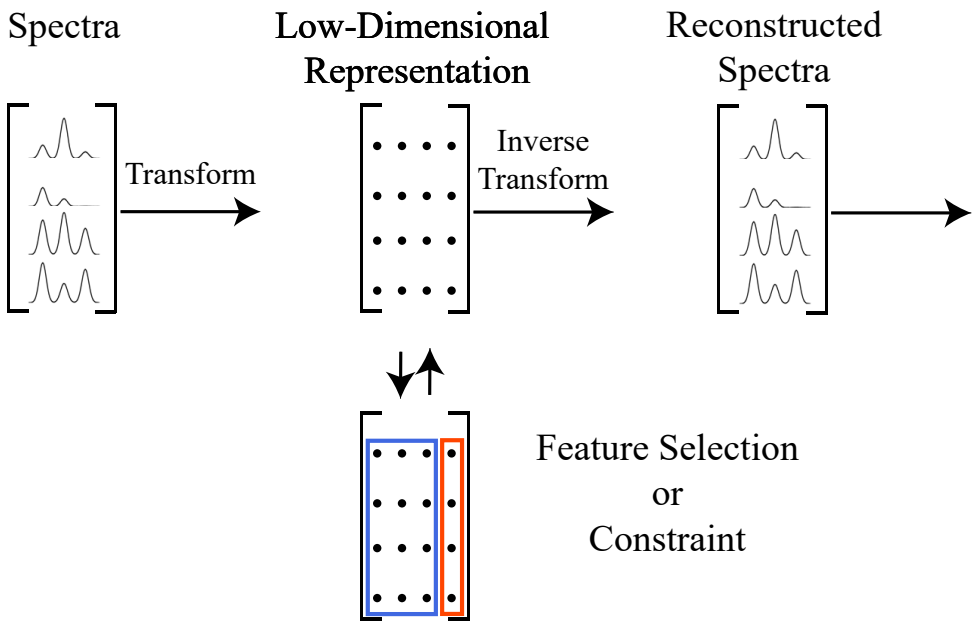


Figure 3.1: The methods investigated in this chapter consist of a transformation to a low-rank space, feature selection or an applied constraint, and then an inverse transformation. Reconstruction error from the low-rank approximation may remove non-target species.

of typically low-dimensional concentrations. The high-dimensional spectroscopy data can then be reconstructed from the low-dimensional, concentration representation. Using CLS, non-target species are likely to be removed because the projected space (concentrations of target species) has fewer degrees of freedom that capture information of only the target species, since reference spectra are typically only available for target species. All of the methods studied in this chapter operate on a similar principle of transforming data to a low-dimensional representation, possibly selecting or constraining a subset of features in the reduced space (this was not done in the CLS example), and then reconstructing the high-dimensional spectra from the low-dimensional representation.

A limitation of using CLS for preprocessing is that the least squares solution may not always correspond to a physically realistic solution when non-target species are present. Some constraints have been used in spectroscopic analysis that produce more meaningful results: nonnegativity of concentrations and unimodality, among others. Baek et al. used

a linear programming constraint on spectra to predict the varying baseline of spectra [80]. Hyperspectral images have been analyzed using nonnegative abundance constraints [81, 82]. Iterative optimization technology (IOT) has been applied in the pharmaceutical industry so that mixture constraints, such as mole fraction summation to unity and upper- and lower-bounds on variables, are satisfied [83, 84, 85]. The works on IOT have additionally used an extended Beer-Lambert Law to model spectroscopically-active mixtures with non-target species (as well as other factors) present.

In addition to constraining concentrations to be nonnegative as has been done elsewhere, a nonnegativity constraint for non-target species spectral contributions is proposed in this chapter for removing overlapping non-target peaks and producing physically realistic spectra based on a linearity assumption (i.e. the Beer-Lambert law in absorbance spectroscopy). The resulting NCCLS model can operate in real-time and is adept at preprocessing spectra with unknown species present. Additionally, NCCLS does not require any prior information about the unknown species aside from their adherence to physically-motivated nonnegative spectral contributions. The preprocessing algorithm is a constrained minimization problem with a quadratic (least-squares) objective function and two linear constraints: nonnegative concentrations and nonnegative non-target spectral contributions. NCCLS does not require estimates of unknown component spectra nor multiple time-series measurements to operate. NCCLS may have bearing on many industrial and field-monitoring spectroscopic tasks where non-target species are present.

To the author's knowledge, the application of a nonnegative constraint on unknown non-target species spectral contribution for the purpose of preprocessing mixture spectra has not been previously reported. In this study, both *in silico* and experimental Raman and attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectra are preprocessed in batch and real-time monitoring scenarios to remove non-target peak contributions and the spectra are subsequently quantified using partial least squares regression (PLSR). Preprocessing to remove non-target species is achieved by the novel NCCLS method and

additionally five spectra preprocessing methods based on methods existing in the literature: principal component analysis (PCA), SRACLS, a convolutional denoising autoencoder (CDAE), and two BSS methods that are distinguished by their use of independent component analysis (ICA) or PCA[2] as a feature identification step, followed by multivariate curve resolution - alternating least squares (MCR-ALS). With the *in silico* spectra, the amount of non-target peak overlap, noise, and test data are varied and the performance of the preprocessing methods is reported. Finally, experimental Raman and ATR-FTIR training spectra of aqueous sodium salts are collected and used to quantify a previously reported dataset of spectra containing four additional non-target species not present in the training data.

3.2 Spectra Preprocessing Methods for Non-Target Removal

The following sections describe the use of preprocessing methods for spectroscopy to deal with overlapping target and non-target peaks. The following notation is used for the spectroscopic systems studied in this chapter: there are n training experiments, m testing experiments[3], r target species, s non-target species, and w wavenumbers. An accent over a variable, $\acute{\text{o}}$, corresponds to training data, while the absence of an accent, o, corresponds to test data. A hat above a variable, $\hat{\text{o}}$, corresponds to an estimated quantity (i.e. concentration data estimated from mixture data). The first method described is NCCLS, the method introduced in this chapter.

3.2.1 Nonnegatively Constrained Classical Least Squares (NCCLS)

NCCLS has use in real-time spectral analysis because of its accuracy, intuitive behavior, requirement of only a single measurement spectra, and low computational cost. This method is agnostic to the type of spectral interferents (except the requirement that they

[2]BSS PCA is the method introduced in Chapter 2 of this thesis, while BSS ICA is the method introduced by Maggioni et al. and Kocevska et al. [20, 21].

[3]Testing data in this study are distinguished from training data by the inclusion of non-target species.

have nonnegative spectral contributions), making the method well-suited to maintaining accurate quantification in systems that deviate from calibration data. NCCLS also assumes target species follow a linear relationship between concentrations and spectra (i.e. the Beer-Lambert Law). Lastly, it is assumed that measurement noise from the sensor itself can be estimated and is indicative of measurement noise expected for future samples.

Classical Least Squares and Unknown Mixture Components

CLS is a well-studied model where mixture spectra are linear combinations of reference spectra. The CLS model is motivated by the structure of the Beer-Lambert Law (in the case of absorbance spectroscopy) and can be used to quantify spectra in single-phase solutions with low analyte concentration. The CLS model is

$$\mathbf{A} = \mathbf{C}\mathbf{K} + \mathbf{E}_\mathbf{A} \tag{3.1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times w}$ is a matrix of mixture spectra, $\mathbf{C} \in \mathbb{R}^{r \times m}$ is a matrix of concentrations, $\mathbf{K} \in \mathbb{R}^{r \times w}$ is a matrix of reference spectra, and $\mathbf{E}_\mathbf{A} \in \mathbb{R}^{m \times w}$ is a matrix of model error.

To solve for unknown concentrations, $\hat{\mathbf{C}}$, the least-squares minimization problem can be solved with respect to concentrations, $\hat{\mathbf{C}}$. This unconstrained minimization problem is

$$\min_{\hat{\mathbf{C}}} \quad ||\mathbf{A} - \hat{\mathbf{C}}\mathbf{K}||_F^2 \tag{3.2}$$

where $||\cdot||_F^2$ is the squared Frobenius or Euclidean norm of a matrix. A constraint to the minimization problem of Equation 3.2 will be introduced that improves CLS fitting when non-target species are present.

To model the case when additional species (non-targets) are present in a mixture to be quantified, $\mathbf{C}_\mathrm{u}$ and $\mathbf{K}_\mathrm{u}$ can be included in Equation 6.27, as is shown in the same form by Saeys et al. [86] but has been shown by other authors as well [87, 88, 84, 73]:

$$\mathbf{A} = \mathbf{CK} + \mathbf{C}_\mathrm{u}\mathbf{K}_\mathrm{u} + \mathbf{E}_\mathbf{A} \tag{3.3}$$

where $\mathbf{C}_\mathrm{u} \in \mathbb{R}^{m\times s}$ and $\mathbf{K}_\mathrm{u} \in \mathbb{R}^{s\times w}$ are the concentration and pure-component reference matrices, respectively, of additional species that are unknown to the CLS model based on available reference spectra. In the cited works for Equation 3.3, the quantity $\mathbf{C}_\mathrm{u}\mathbf{K}_\mathrm{u}$ is estimated so that it can be subtracted. The contribution of this chapter is using physical assumptions of $\mathbf{C}_\mathrm{u}$ and $\mathbf{K}_\mathrm{u}$ to motivate a spectra preprocessing algorithm where $\mathbf{C}_\mathrm{u}$ and $\mathbf{K}_\mathrm{u}$ do not need to be estimated. Assuming Equation 3.3 represents the spectroscopic system to be quantified and the Beer-Lambert Law (or related linearity assumption) applies, then physical constraints dictate that every element of $\mathbf{C}_\mathrm{u}$ and $\mathbf{K}_\mathrm{u}$ are nonnegative. The Beer-Lambert law prohibits nonnegative concentrations and nonnegative absorptivity, with analogous statements preventing a nonnegative calibration matrix for Raman spectroscopy [89, 90]. From the matrix multiplication of two nonnegative matrices, every element in the resulting matrix of non-target species will be nonnegative. Therefore,

$$\mathbf{C}_\mathrm{u}\mathbf{K}_\mathrm{u} \geq \mathbf{0} \tag{3.4}$$

where $\mathbf{0}$ is an $m \times w$ null matrix and $\geq$ represents an element-wise inequality. Inserting this inequality into Equation 3.3 yields

$$\mathbf{CK} - \mathbf{A} + \mathbf{E}_\mathbf{A} \leq \mathbf{0} \tag{3.5}$$

Physically, the constraint in Equation 3.5 ensures that CLS-fit spectra of target species $(\mathbf{CK})$ is less than the mixture spectra $(\mathbf{A})$ while accounting for experimental error $(\mathbf{E}_\mathbf{A})$. In practice, enforcing this constraint may prevent the least-squares solution of target species from capturing contributions of non-target species. Unlike the unconstrained minimization problem of Equation 3.2, an estimate of instrument error, $\mathbf{E}_\mathbf{A}$, must be provided to use the

constraint of Equation 3.5 on real systems. The distribution of $\mathbf{E_A}$ can be estimated from training data, as is shown in Section 3.2.1.

Gaussian Error

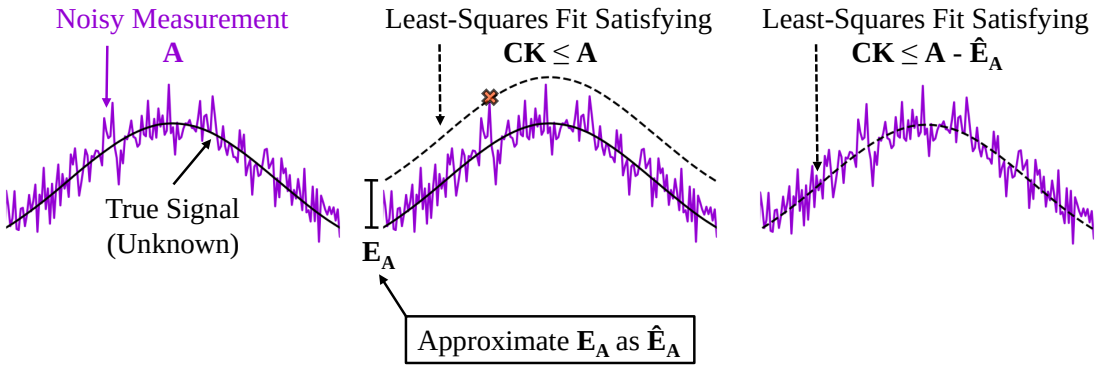


Figure 3.2: Measurement noise must be estimated ($\hat{\mathbf{E}}_\mathbf{A}$) to use the NCCLS constraint accurately (Equation 3.5). Without accounting for measurement error, $\mathbf{E_A}$, least-squares fit spectra will be offset from the true spectra by the largest positive residual in the spectrum (the red "$\times$" in the figure).

To use Equation 3.5 as a constraint to a least squares minimization problem, $\mathbf{E_A}$ must be estimated ($\hat{\mathbf{E}}_\mathbf{A}$) as shown in Figure 3.2. Different mechanisms exist for estimating and incorporating error into constraints like that seen in Equation 3.5 [91]. In this chapter, $\mathbf{E_A}$ is approximated via a probabilistic approach by assuming that measurement error of the training data and measurement error of the test data are taken from the same Gaussian distribution; the error for the training data and test data are i.i.d.. The standard deviation is found of the residuals of training data ($\acute{\mathbf{E}}_\mathbf{A}$). Then, a confidence interval is chosen based on the standard deviation of training data to determine $\hat{\mathbf{E}}_\mathbf{A}$. In this chapter, the confidence interval is selected based on an analytical expression shown in Chapter C. The assumption of i.i.d. and Gaussian error may be relaxed by using other distributions or a separate distribution for each wavenumber, in the case of a wavenumber-dependent error. Additionally, the magnitude of $\mathbf{E_A}$ can be reduced through filtering or averaging spectra.

NCCLS

The quadratic programming problem of NCCLS is

$$\begin{aligned} \min_{\hat{\mathbf{C}}} \quad & ||\mathbf{A} - \hat{\mathbf{C}}\mathbf{K}||_F^2 \\ \text{s.t.} \quad & \hat{\mathbf{C}}\mathbf{K} - \mathbf{A} + \hat{\mathbf{E}}_\mathbf{A} \leq \mathbf{0} \\ & \hat{\mathbf{C}} \geq 0 \end{aligned} \tag{3.6}$$

Equation 3.6 has an identical objective function to the unconstrained least-squares optimization shown in Equation 3.2, but with the additional constraint of nonnegative non-target spectral contributions introduced in Equation 3.5 and a nonnegative concentration constraint. The objective is constrained so that the least squares spectra fit, $\hat{\mathbf{C}}\mathbf{K}$, is always less than or equal to the true mixture spectra, $\mathbf{A}$, and the expected uncertain region of measurement error, $\hat{\mathbf{E}}_\mathrm{A}$. The reconstructed nonnegatively constrained CLS spectra will have a least squares fit except in locations where non-target species, $\mathbf{C}_\mathrm{u}\mathbf{K}_\mathrm{u}$, are present and the constraint is enforced. Where non-target species ($\mathbf{C}_\mathrm{u}\mathbf{K}_\mathrm{u}$) do appear, their spectral contributions will be removed from reconstructed spectra by enforcing nonnegativity of non-targets species (refer to Equation 3.4 and Equation 3.5). With the new NCCLS method introduced, the comparison methods for this chapter will now be described.

3.2.2 Principal Component Analysis (PCA)

PCA is an unsupervised feature reduction method that reduces the dimensionality of data by projecting it onto a space of lower dimension while maximizing the variance in the projected data [92, 93]. In the context of spectral preprocessing, test data can be transformed into the reduced-dimension space found by performing PCA on training data. This highlights features of the test data that match high-variance components of the training data (i.e. where spectral peaks in the training data exist). Transforming the projected test data back into its original space using only select (usually the first few) principal compo-

nents will yield preprocessed spectra that have only contributions in directions given by the principal components of the training data, thereby eliminating contributions of non-target species that are not in the directions of the principal components [59, 60].

PCA preprocessing is performed through several steps. First, the covariance matrix ($\mathbf{\Phi}$) is found from mean-subtracted training spectra ($\mathbf{A}_{\text{train}}$). The eigenvalues ($\lambda_i$) and eigenvectors ($\mathbf{v}_i$) are found of the covariance matrix, and sorted based on the magnitude of the eigenvalues. The largest r eigenvalues (along with their eigenvectors) are kept, where r is the number of target species (or the minimum number of principal components that can accurately describe the training data). The retained eigenvectors are the principal components of PCA, and provide a pathway to transform additional data into this reduced space.

$$\acute{\mathbf{\Phi}} = \frac{1}{n-1}(\acute{\mathbf{A}} - \acute{\bar{\mathbf{A}}})^{\mathrm{T}}(\acute{\mathbf{A}} - \acute{\bar{\mathbf{A}}}) \tag{3.7}$$

$$\acute{\mathbf{\Phi}}\acute{\mathbf{v}}_i^{\mathrm{T}} = \lambda_i \acute{\mathbf{v}}_i^{\mathrm{T}} \tag{3.8}$$

where $\acute{\bar{\mathbf{A}}} \in \mathbb{R}^{n \times w}$ is a matrix of the mean training spectrum, $\acute{\bar{\mathbf{a}}} \in \mathbb{R}^{1 \times w}$, stacked n times.

In order to maintain consistency with matrix representation in the spectroscopic literature and the other methods used in this chapter, $\mathbf{A}$ is a matrix of spectra represented by row vectors, $\mathbf{a}_j$, and $\mathbf{V}^{\mathrm{T}}$ is a matrix of right eigenvectors represented by column vectors, $\mathbf{v}_i^{\mathrm{T}}$. Reconstructing a single spectra ($\tilde{\mathbf{a}}_j$) can be accomplished with Equation 3.11 using the principal components found from training data ($\acute{\mathbf{v}}_1$) and the mean spectrum ($\acute{\bar{\mathbf{a}}}$).

$$\mathbf{A} = \begin{bmatrix} \text{—} & \mathbf{a}_1 & \text{—} \\ \text{—} & \mathbf{a}_2 & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_m & \text{—} \end{bmatrix} \tag{3.9}$$

$$\acute{\mathbf{V}}^{\mathrm{T}} = \begin{bmatrix} | & | & & | \\ \acute{\mathbf{v}}_1^{\mathrm{T}} & \acute{\mathbf{v}}_2^{\mathrm{T}} & \dots & \acute{\mathbf{v}}_r^{\mathrm{T}} \\ | & | & & | \end{bmatrix} \tag{3.10}$$

$$\tilde{\mathbf{a}}_j = \bar{\mathbf{a}} + \sum_{i=1}^{r} (\mathbf{a}_j \acute{\mathbf{v}}_i^{\mathrm{T}} - \bar{\mathbf{a}} \acute{\mathbf{v}}_i^{\mathrm{T}}) \acute{\mathbf{v}}_i \tag{3.11}$$

$$\tilde{\mathbf{A}}^{\mathrm{PCA}} = \begin{bmatrix} — & \tilde{\mathbf{a}}_1 & — \\ — & \tilde{\mathbf{a}}_2 & — \\ & \vdots & \\ — & \tilde{\mathbf{a}}_m & — \end{bmatrix} \tag{3.12}$$

In practice, Equation 3.7 through Equation 3.12 can be implemented through matrix operations or using an available PCA package, such as is available from scikit-learn.

3.2.3 Spectral Residual Augmented Classical Least Squares (SRACLS)

SRACLS is ordinarily a quantification model similar to classical least squares, except that available reference spectra are augmented by additional sources found through a factor analysis (PCA is used here) of residual spectra [57]. In this chapter, SRACLS is adapted to be a preprocessing algorithm rather than a spectra quantification method. To perform SRACLS, the residual spectra from classical least squares fitting on test data have PCA performed on them. In contrast to PCA preprocessing (Section 3.2.2), the PCA of SRACLS is performed on the error matrix of CLS fitting the test data, $\mathbf{E}_{\mathbf{A}}$, rather than the training data, $\acute{\mathbf{A}}$. The similar structure of SRACLS and CLS suggests that SRACLS behaves similarly to unconstrained CLS. Therefore, SRACLS is a control experiment for NCCLS without a nonnegative spectral contributions constraint.

To perform SRACLS, the eigenvectors from PCA (v_i^{T}) are used to supplement available reference spectra. This is shown by the augmented reference spectra matrix, shown in

Equation 3.16.

$$\mathbf{E_A} = \mathbf{A} - \mathbf{CK} \tag{3.13}$$

$$\mathbf{\Phi} = \frac{1}{n-1}(\mathbf{E_A} - \bar{\mathbf{E}}_\mathbf{A})^{\mathrm{T}}(\mathbf{E_A} - \bar{\mathbf{E}}_\mathbf{A}) \tag{3.14}$$

$$\mathbf{\Phi}_{\mathbf{E_A}}\mathbf{v}_i^{\mathrm{T}} = \lambda_i \mathbf{v}_i^{\mathrm{T}} \tag{3.15}$$

$$\mathbf{K} = \begin{bmatrix} — & \mathbf{k}_1 & — \\ — & \mathbf{k}_2 & — \\ & \vdots & \\ — & \mathbf{k}_r & — \end{bmatrix} \quad , \quad \mathbf{K}_{\mathrm{aug}} = \begin{bmatrix} — & \mathbf{k}_1 & — \\ — & \mathbf{k}_2 & — \\ & \vdots & \\ — & \mathbf{k}_r & — \\ — & \mathbf{v}_1 & — \\ — & \mathbf{v}_2 & — \\ & \vdots & \\ — & \mathbf{v}_s & — \end{bmatrix} \tag{3.16}$$

After augmenting classical least squares with eigenvectors from PCA (Equation 3.16) concentration ($\hat{\mathbf{C}}$) values can be found using $\mathbf{K}_{\mathrm{aug}}$ and the CLS relationship giving Equation 3.17:

$$\hat{\mathbf{C}} = \mathbf{A}\mathbf{K}_{\mathrm{aug}}^{\mathrm{T}}(\mathbf{K}_{\mathrm{aug}}\mathbf{K}_{\mathrm{aug}}^{\mathrm{T}})^{-1} = \mathbf{A}(\mathbf{K}_{\mathrm{aug}})^{\dagger} \tag{3.17}$$

If desiring quantitation from SRACLS, as was shown by Haaland and Melgaard, the first r rows from $\hat{\mathbf{C}}_{\mathrm{process}}$ (Equation 3.17) will yield estimated concentrations of target species [73]. However, SRACLS can be adjusted slightly to create a spectra preprocessing algorithm. Replacing rows $r+1$ through $r+s$ with $\mathbf{0}$ vectors yields the corresponding SRACLS estimated spectra *without* contributions of estimated non-target species, $\hat{\mathbf{C}}_{\mathrm{preprocessed}} \in \mathbb{R}^{r \times m}$. This is shown in Equation 3.18. Lastly, the final preprocessed spectra can be found by matrix multiplication between $\hat{\mathbf{C}}_{\mathrm{preprocessed}}$ and the original reference

spectra, $\mathbf{K}$. Note that $\mathbf{K}_{\text{aug}}$ is not necessary for the final preprocessing step, but is an intermediate for calculating $\hat{\mathbf{C}}_{\text{preprocessed}}$ if quantification, rather than preprocessing, is desired.

$$\left[\begin{array}{c:c} \hat{\mathbf{C}}_{\text{preprocessed}} & \mathbf{0}^{r\times s} \\ \hdashline \mathbf{0}^{s\times r} & \mathbf{0}^{s\times s} \end{array}\right] = \left[\begin{array}{c:c} \mathbf{I}^{r\times r} & \mathbf{0}^{r\times s} \\ \hdashline \mathbf{0}^{s\times r} & \mathbf{0}^{s\times s} \end{array}\right] \hat{\mathbf{C}} \tag{3.18}$$

$$\tilde{\mathbf{A}}^{\text{SRACLS}} = \hat{\mathbf{C}}_{\text{preprocessed}}\mathbf{K} \tag{3.19}$$

3.2.4 Blind Source Separation (BSS)

BSS refers to a series of techniques that estimate pure component source signals and amplitudes (concentrations in this context) from mixture spectra alone [20, 54, 94]. These techniques have been applied to vibrational spectroscopy problems using MCR-ALS with constraints applied, such as nonnegativity of sources and concentrations [20, 21, 30]. Because of the nonnegativity constraint, MCR-ALS is a subset of a broader class of nonnegative matrix factorization. The different implementations for MCR-ALS arise from the requirement of initial guesses for either concentrations or reference spectra. Reference spectra for target species are often readily available in spectroscopic contexts, but accurate reference spectra for non-target species may not always be available. Thus, data-driven methods have been used to identify reference spectra from spectra alone for use as initial guesses for the MCR-ALS algorithm.

Details of how to apply BSS to spectroscopic data exist elsewhere, with references for BSS with ICA identifying sources found in Maggioni et al. [20] and Kocevska et al. [21]. BSS with PCA identifying sources can be found in Crouse et al. [30] as well as in Section 2.2.4 of this thesis.

MCR-ALS is the underlying algorithm used in BSS algorithms in the spectroscopic literature:

$$\mathbf{A} = \mathbf{C}_{\text{MCR}}\mathbf{S}_{\text{MCR}} + \mathbf{E}_{\text{MCR}} \tag{3.20}$$

where $\mathbf{A}$ is a matrix of mixture spectra as previously defined, $\mathbf{C}_{\text{MCR}} \in \mathbb{R}^{n\times(s+r)}$ is a matrix of latent "concentrations" with $s + r$ (sum of target and non-target species) latent variables, and $\mathbf{S}_{\text{MCR}} \in \mathbb{R}^{(s+r)\times w}$ is a matrix of latent variable "spectra". Notably, MCR-ALS (Equation 3.20) is similar to CLS, except that $\mathbf{C}_{\text{MCR-ALS}}$ and $\mathbf{S}_{\text{MCR-ALS}}$ correspond to latent variables rather than physical reference spectra and concentrations.

Once $\mathbf{C}_{\text{MCR-ALS}}$ and $\mathbf{S}_{\text{MCR-ALS}}$ are fit to mixture data, feature selection of peaks that overlap in the original domain become possible in the space of latent concentrations ($\mathbf{C}_{\text{MCR-ALS}}$). One of the challenges of BSS is identifying which features (sources) to subtract. Methods that have been explored in the literature include matching latent sources and reference spectra via correlation [20, 21] and constraining MCR-ALS to match reference spectra [30]. After feature selection in the latent concentration space ($\mathbf{C}_{\text{MCR-ALS}}$) has reduced the $r + s$ dimensional space down to r, the final preprocessing equation is

$$\tilde{\mathbf{A}}^{\text{BSS}}_{\text{preprocessed}} = \hat{\mathbf{C}}^{r}_{\text{MCR}}\hat{\mathbf{S}}^{r}_{\text{MCR}} \tag{3.21}$$

Latent variables corresponding to non-target species are not included in the recreation of the mixture spectra ($\tilde{\mathbf{A}}^{\text{BSS}}_{\text{preprocessed}}$) and so are effectively removed from spectra.

3.2.5 Convolutional Denoising Autoencoder (CDAE)

Artificial neural networks have been used for vibrational spectral analysis abundantly in the literature [95, 96, 97, 76, 98]. Autoencoders use a type of neural network architecture that is particularly suited for feature reduction and extraction. An autoencoder is similar to PCA in that it maps data from its original space into a low-dimensional latent space and then back into the original space. The ability of neural networks to map arbitrary, nonlinear functions has been well-documented [99]. However, neural networks typically have a large number of parameters that are time- and data-intensive to train, which may preclude them from some spectroscopy tasks. However, because of their prevalence in current literature,

a CDAE architecture is included in this chapter.

A CDAE framework is chosen based loosely on the CDAE used by Fan et al. because of the similar application of denoising autoencoders [76]. Whereas Fan et al. used a CDAE for removing noise, it is used in this chapter to remove unknown species. The CDAE network of this chapter begins by standard-scaling data. There are three one-dimensional convolutional layers each followed by a rectified linear unit (ReLU) activation layer and a one-dimensional max-pooling layer. Following the last max-pooling layer, there are three additional convolutional layers that are preceeded by ReLU activation layers which are further preceeded by up-sampling layers. This creates a dimensional "bottleneck" structure that is typical of autoencoders. The last layer is a densely connected layer with a linear activation function with size equal to the dimension of the input spectra. The number of nodes (given a 1000-wavenumber spectra input) follows the following structure: $1000 \rightarrow 500 \rightarrow 250 \rightarrow 125$ before up-sampling back into the original dimension: $125 \rightarrow 250 \rightarrow 500 \rightarrow 1000$. The smallest layer in the CDAE is a reduced-dimension space that learns a condensed representation of the original features that describe the training data. Therefore, the CDAE is designed to only recreate features that are present in the training data, thereby removing contributions of non-target species.

Hyperparameter tuning for the CDAE in this chapter was performed with 35 training spectra and 15 validation spectra, all with only target species contributions. The entire CDAE contained 189,928 parameters for Computational Studies 1 and 2, 86,920 parameters for experimental Raman spectra, and 71,320 parameters for experimental ATR-FTIR spectra. The author acknowledges that there are many more parameters than data in this case; this is a feature of the study and the same data are used by all methods in this chapter.

The goal of a CDAE is to map the input back to itself, or to solve minimization problem of Equation 3.22 where $\boldsymbol{\theta}$ are the parameters of the CDAE and $\mathbf{f}(\mathbf{A};\boldsymbol{\theta})$ is a function representing the output of the CDAE as a function of mixture spectra ($\mathbf{A}$) and dependent on parameters ($\boldsymbol{\theta}$) [60]. After optimization, the trained CDAE can be applied as shown in

Equation 3.23:

$$\min_{\hat{\boldsymbol{\theta}}} ||\mathbf{A} - \mathbf{f}(\mathbf{A}; \hat{\boldsymbol{\theta}})||_2^2 \tag{3.22}$$

$$\tilde{\mathbf{A}}^{\text{AE}}_{\text{preprocessed}} = \mathbf{f}(\mathbf{A}; \hat{\boldsymbol{\theta}}) \tag{3.23}$$

3.3 Materials and Methods

The feature selection of NCCLS was compared with other available preprocessing methods in this chapter. Figure 3.3 shows the information flow for all experiments conducted. Spectra underwent a signal-to-noise ratio (SNR) based feature selection scheme that removed features (wavenumbers) from training and test spectra that had an SNR less than 10 at that feature. Preprocessing (non-target removal) occured after feature selection and consisted of one of PCA, SRACLS, CDAE, NCCLS, BSS ICA, or BSS PCA. After preprocessing (non-target removal), the spectra was scaled so that each feature had zero mean and unit variance (shown in Equation A.3) before being quantified by a PLSR model trained on data that also had the same feature selection and scaling applied. See Chapter B for additional detail on PLSR models.

Standard normal variate (SNV) scaling was prior to all PLSR quantification but *after* all preprocessing methods (the preprocessing methods are applied to the spectra with feature selection applied). More detail on SNV scaling can be found in Chapter A.

3.3.1 Computational Experiments

Two computational studies were performed to compare the six preprocessing methods studied in this chapter. *In silico* data were generated using Gaussian curves and additive white noise. The reference spectra were calculated assuming a Gaussian peak using:

$$f_i(\lambda) = CIe^{\frac{-(\lambda-\bar{\lambda}_i)^2}{2\sigma_i^2}} \tag{3.24}$$

where λ was the domain of the spectra (wavenumber), I was the maximum reference in-

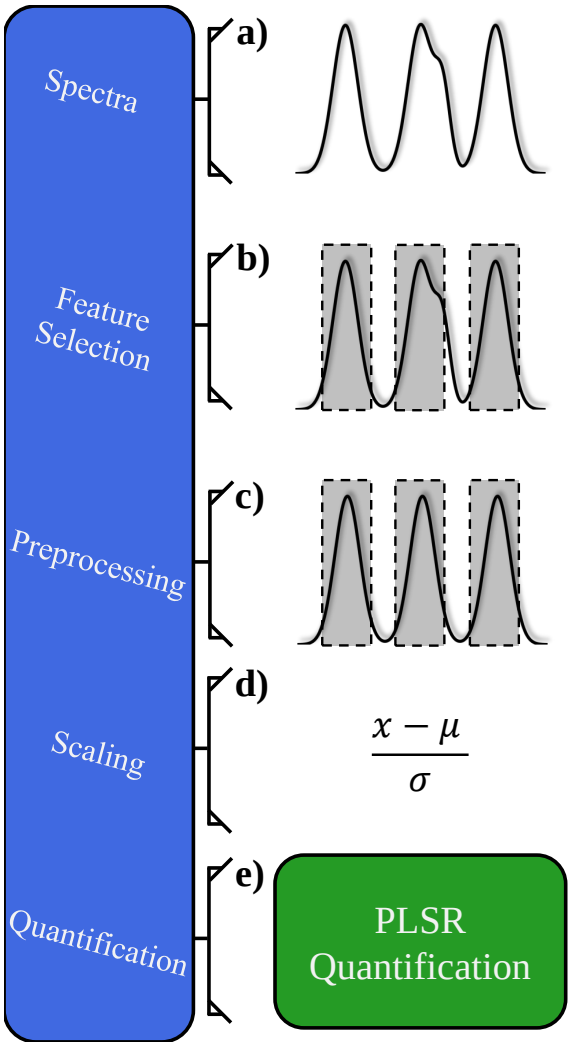


Figure 3.3: Information flowchart for quantification used in all studies in this chapter. The process begins with **a)** test spectra that **b)** undergo feature selection where features (wavenumbers) with SNR of target species greater than 10 are selected, then **c)** the preprocessing methods discussed in this chapter are employed (PCA, SRACLS, CDAE, NCCLS, BSS ICA, and BSS PCA), then **d)** features are SNV scaled by subtracting the mean of each feature and dividing by that feature's standard deviation, then **e)** a trained PLSR model is used to predict output concentrations.

tensity (at 1 M for this study), $\bar{\lambda}_i$ was the center of peak i (cm^{-1}), and σ_i was the standard deviation describing the width of peak i.

After computational references have been created, the training and testing concentration data were created by using Latin hypercube sampling. The dimension of the Latin hypercube for constructing the training data is r (the number of target species) and the

dimension of the Latin hypercube for constructing the testing data is $r + s$ (the number of target and non-target species). The concentration data were then used to create linear superpositions of the reference spectra to create mixture spectra. Training data are created using Equation 3.25 and testing data are created using Equation 3.26.

$$F_{training}(\lambda) = \sum_{i=1}^{r} f_i(\lambda) + v(\sigma) \tag{3.25}$$

$$F_{testing}(\lambda) = \sum_{i=1}^{r+s} f_i(\lambda) + v(\sigma) \tag{3.26}$$

Computational Study 1 compares the different preprocessing methods in two idealized scenarios distinguished by batch preprocessing vs. real-time preprocessing. *Computational Study 2* modifies the scenario of Computational Study 1 with batch preprocessing by changing the amount of non-target overlap, the amount of measurement noise, and the number of testing data. The methods are compared on a basis of root mean squared error (RMSE) of test data predictions, which is

$$RMSE = \sqrt{\frac{1}{mr} \sum_{i=1}^{m} \sum_{j=1}^{r} (\mathbf{C}_{ij} - \hat{\mathbf{C}}_{ij})^2} \tag{3.27}$$

For all computational experiments, there were three target species and a single non-target species, all with dimensionless maximum peak heights of one. Computationally-generated spectra had an arbitrary x-axis ranging from 0 – 1000 referred to as wavenumbers in this chapter to match vibrational spectra. The amount of non-target peak overlap was quantified from the normalized target and normalized non-target references. A quantitative measure of peak overlap was produced by dividing the area of the peak overlap by the target peak area. By this measure, two non-overlapping peaks have 0% overlap, while two identical peaks (in mean, width, and shape) have 100% overlap. Percent overlap is shown

in Equation 3.28, which was calculated numerically using a trapezoid integration scheme.

$$\% \text{ Overlap} = \frac{\text{Overlapping Area}}{\text{Target Area}} = \frac{\int_{-\infty}^{\infty} \min(f_{target}(\lambda), f_{nontarget}(\lambda))d\lambda}{\int_{-\infty}^{\infty} f_{target}(\lambda)d\lambda} \times 100\% \quad (3.28)$$

Computational Study 1

For Computational Study 1, 35 computational spectra were created as training data and 15 computational spectra were created as test data. Each training and test spectrum was a sum of three target reference spectra multiplied by pseudorandom concentrations taken from a non-centered Latin hypercube sampling scheme. The test spectra additionally have a non-target species peak added at concentrations determined by another non-centered Latin hypercube sampling scheme. An additive Gaussian noise was added to all training, reference, and test spectra that was defined by a standard deviation of 0.2% of the maximum peak height of the references. The target and non-target reference spectra for Computational Study 1 are shown in Figure 3.4 with details of the experiment listed in Table 3.1. With the non-target species centered at 560 cm^{-1} and the next closest target species (Target 2) centered at 600 cm^{-1}, there was 31.7% peak overlap.

The test spectra for Computational Study 1 were analyzed using a batch preprocessing methodology and a real-time preprocessing methodology. Batch preprocessing emulated historical data analysis so that all of the test spectra were analyzed at once; each experiment had 35 training data and 15 test data. Real-time preprocessing emulated time- or data-limited analysis so that each test spectrum was analyzed individually; each experiment had 35 training data and one test spectrum but preprocessing was repeated for 15 individual test spectra so that the same exact data were preprocessed and quantified using both batch and real-time preprocessing methodologies.

Table 3.1: Values used to create the *in silico* data in Computational Studies 1 and 2. Boxed values are parameters varied in Computational Study 2. SD refers to standard deviation.

Peak Property	**Target 1**	**Target 2**	**Target 3**	**Non-Target**
Maximum Intensity	1	1	1	1
Mean	300	600	650	560
SD (Peak Width)	10	20	15	20
Spectrum Property	**Value**			
Training Spectra	35			
Testing Spectra	15			
SD (Noise)	0.002			
Wavenumber Range	0–1000			

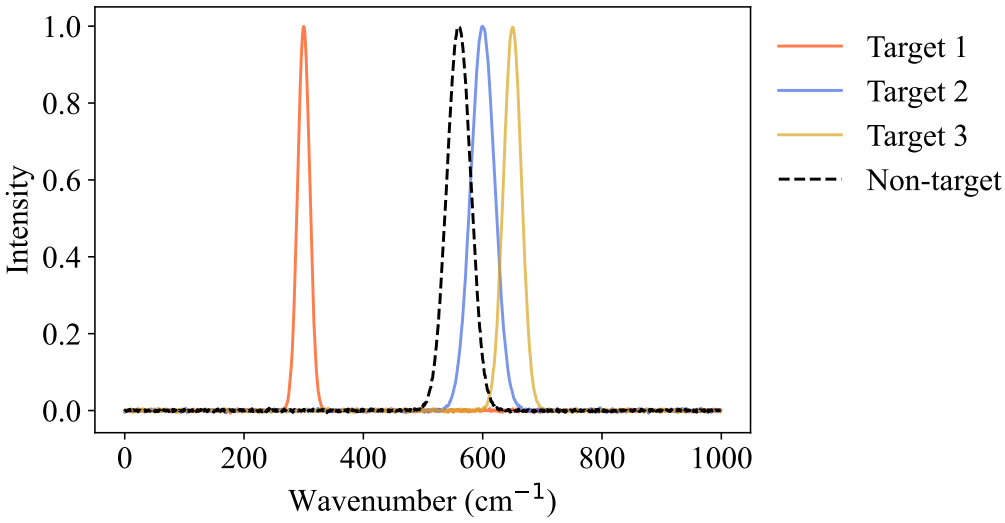
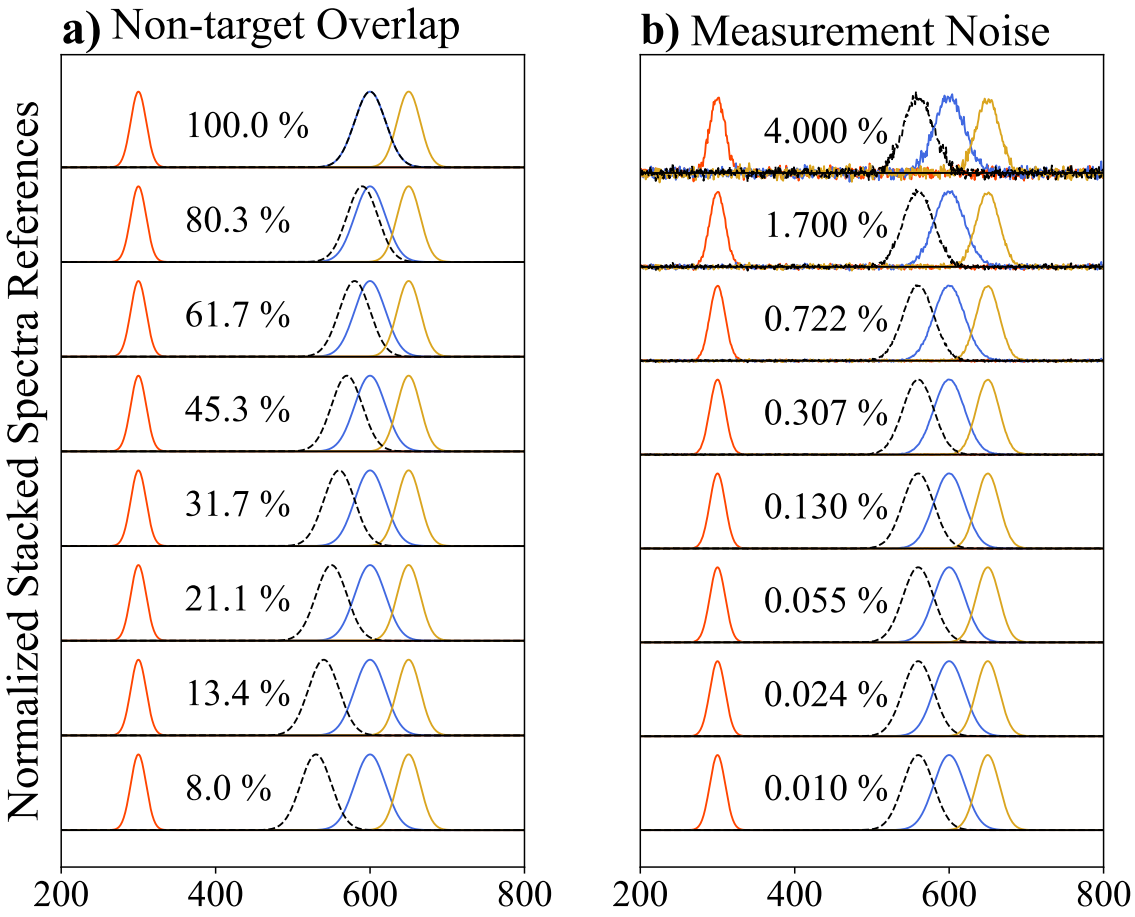


Figure 3.4: References for Computational Study 1 and Computational Study 2. The non-target species is included only in test data. The goal of the preprocessing methods is to remove the non-target species without prior information of the non-target species.

Computational Study 2

Computational Study 2 is identical to Computational Study 1 with a batch preprocessing methodology except five replicates (compared to one) are performed with different initialization states and the system is adjusted along three axes shown in Figure 3.5 and Table 3.1 (the boxed quantities): the amount of non-target peak overlap, the amount of measurement noise, and the number of testing data available. The system conditions from Computational Study 1 still applied for Computational Study 2 — there was 31.7% non-



c) Test Data

Exp	1	2	3	4	5	6	7	8
Data	1	4	14	52	193	720	2683	10000

Figure 3.5: The four manipulated axes investigated in Computational Study 2: **a)** spectral overlap (8.0% – 100%), **b)** standard deviation of Gaussian measurement noise (0.01% – 4.00%), and **c)** the number of test data that contain the non-target peak (1 – 10,000).

target overlap with Target 2; there was 0.2% additive measurement noise; there were 15 testing data — except when an experimental condition was varied in agreement with Figure 3.5a, Figure 3.5b, or Figure 3.5c. Additionally, Computational Study 2 uses RMSE of Target 2 (rather than RMSE of all target species) as an error metric, since this study focuses

primarily on how the methods perform with target and non-target overlap.

3.3.2 Physical Experiment

In addition to the *in silico* experiments performed, experimental Raman and ATR-FTIR spectra measuring soluble anions in solutions typical of nuclear waste were preprocessed and quantified [5, 24, 7]. Nine training experiments were prepared and measured via Raman and ATR-FTIR spectroscopy. Measurements were collected in a Mettler Toledo OptiMax vessel at 25°C and stirred at 400 RPM. *In-situ* Raman spectroscopy measurements were collected with a Mettler Toledo ReactRaman 785 using a 785 nm laser and 300 mW laser power at 0.75 s exposure time and 10 averaged scans. *In-situ* ATR-FTIR measurements were collected using a Mettler Toledo ReactIR 10 with one-minute sample collection time corresponding to 256 averaged scans. All spectra used for this study were collected after complete dissolution of all chemical species, verified by the previously mentioned *in-situ* Raman and ATR-FTIR probes. There were 42 experimental spectra (with no prior preprocessing) containing additional non-target species collected using the same exact instrumentation and settings and published by Kocevska et al. in 2021 [21]; these data were used as test data in this study. To handle possible shifting baselines, the experimental Raman and ATR-FTIR spectra are Savitzky-Golay filtered prior to standard normal variate scaling (Figure 3.3d) with seven filter points, a second order polynomial, and a first order derivative [100].

In this chapter, the training data included the target anions of nitrate (NO_3^-), nitrite (NO_2^-), and sulfate (SO_4^{2-}). The dataset published by Kocevska et al. includes the same target species in addition to non-target anions of carbonate (CO_3^{2-}), phosphate (PO_4^{3-}), acetate $(C_2H_3O_2^-)$, and oxalate $(C_2O_4^{2-})$. Of the non-target species, carbonate is the primary non-target species studied in this chapter because it overlaps significantly with nitrate in both the Raman spectrum and infrared spectrum, and therefore cannot be effectively subtracted by feature selection in the original domain of the spectra. In the Raman spec-

trum, carbonate appears as a shoulder on the higher-wavenumber side of the nitrate peak at 1050 cm^{-1} giving 24.6% peak overlap. In the Fourier transform infrared (FTIR) spectrum, carbonate appears as a shoulder on the higher-wavenumber side of the nitrate peak at 1340 cm^{-1} giving 79.8% peak overlap. Reference spectra for the experimental system are shown in Figure 3.6.

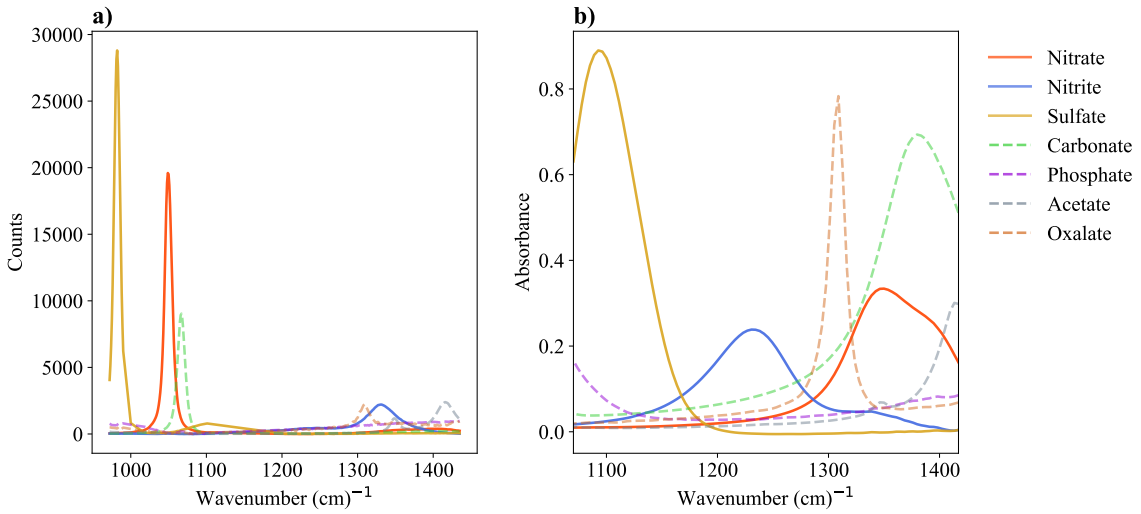


Figure 3.6: Experimental references at a concentration of 1 M for **a)** Raman and **b)** ATR-FTIR spectra. Target species are displayed in solid lines, while non-target species are displayed in the dashed lines.

3.3.3 Partial Least Squares Regression

PLSR is a well-studied quantification model that has been applied to quantifying multicollinear data, as is commonly seen in spectroscopic applications [101, 102, 103, 24]. PLSR quantification models with four components were used to quantify all spectra in this chapter. All spectra in Computational Study 1, the Raman experimental study, and the ATR-FTIR experimental study were quantified by a single PLSR model for each respective set of data. The PLSR model for Computational Study 2 was retrained for each experiment on new pseudorandom training data, but the same model and training data was used between the six preprocessing methods. Additional information on PLSR can be found at Chapter B.

The PLSR models were fit to training data that included only target species; there were no non-target species present in the training data. The data used to train the PLSR quantification model are described in Section 3.3.1 for the *in silico* experiments and in Section 3.3.2 for the physical experiments. The PLSR model outputs had a nonnegative filter applied so that no reported concentrations are less than zero.

3.3.4 Computation

NCCLS relies on solving Equation 3.6, which is a quadratic programming problem that is solved in this chapter numerically using IBM's CPLEX solver (v22.1.1) [104] within the Pyomo optimization framework (v6.4.0) developed at Sandia National Laboratory [105]. In this implementation, the optimization problem can be performed in real-time, although any quadratic program solver would suffice. All computations for this study were performed using Python (v3.9.12), NumPy (v1.21.5), Scikit-learn (v1.1.1), TensorFlow (v2.8.0), and Keras (v2.8.0). Code for NCCLS, all other preprocessing methods used, and data can be found on GitHub[4].

All experiments used a "cold start" initialization with pseudorandomly generated initial conditions determined by the respective method with a set pseudorandom seed for reproducibility. Additionally, the computational experiments used a set pseudorandom seed to generate data; identical data were preprocessed via the different preprocessing methods within the same experiment in the computational and physical experiments.

3.4 Results: Computational Spectra

In this section, a spectroscopic system was created computationally with additive Gaussian noise, and the efficacy of six preprocessing algorithms were compared.

[4]https://github.com/magrover/Spectral-Preprocessing-NCCLS/tree/main

3.4.1 Computational Study 1

Batch Preprocessing Methodology

The performance of the preprocessing methods were compared qualitatively through visual comparisons of the spectra. Results from Computational Study 1 are shown during a batch preprocessing methodology (15 test spectra are available at once). In Figure 3.7, the yellow dotted line corresponds to original spectra that have not undergone preprocessing to remove non-target species, the solid red lines correspond to the reconstructed spectra after preprocessing, and the dashed blue lines correspond to the true target peaks (i.e. the spectra if the non-target peak was not present). An SNR feature selection was performed prior to the preprocessing methods shown, so discontinuities appeared in wavenumber ranges that do not overlap with target species (this is most prominantly seen in Figure 3.7a). In comparing the methods, the plot with no preprocessing (Figure 3.7a) shows that the non-target peak was very prominent on the lower-wavenumber side of the target peak located at 600 cm^{-1}, even after the SNR feature selection. PCA (Figure 3.7b), SRACLS (Figure 3.7c), and possibly CDAE (Figure 3.7d) all removed some contributions of the non-target peak as can be seen by the correct peak width and location at 600 cm^{-1}, but the the reconstructed peaks over-predicted the target peak at 600 cm^{-1}. NCCLS (Figure 3.7g), BSS ICA (Figure 3.7e), and BSS PCA (Figure 3.7f) reconstructed the target peaks with the correct peak width, location, and height of the target peak at 600 cm^{-1}.

Reconstructing spectra of the original dimension before quantification, rather than regressing in the projected space, allows for any already-trained quantification model to be used for quantification without potentially expensive and resource-intensive model recalibration while also allowing for human interpretation of non-target species removal. Parity plots of quantification after preprocessing are shown in Figure 3.8. NCCLS had the lowest RMSE of 0.016, followed by BSS PCA (0.017), BSS ICA (0.031), PCA (0.082), SRACLS (0.083), and CDAE (1.127). Based on the results of Figure 3.7 and Figure 3.8, the meth-

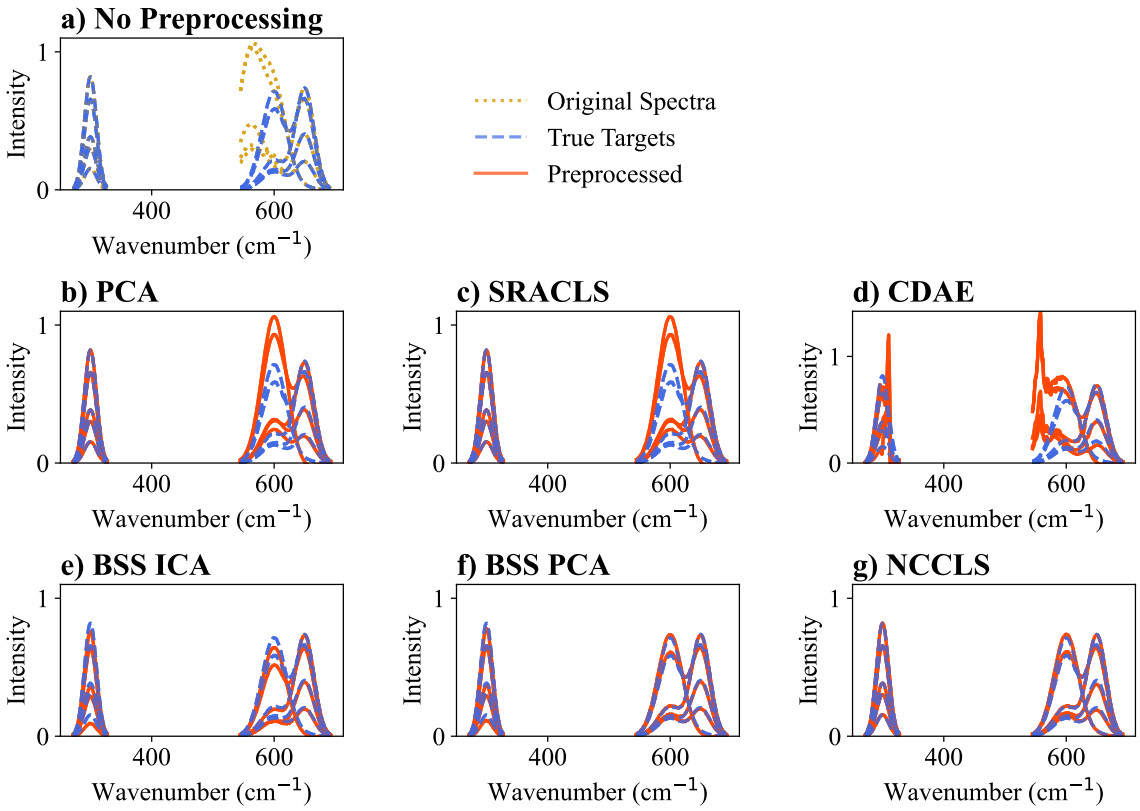


Figure 3.7: Comparison of five batch preprocessed *in silico* spectra with the true target peaks from those spectra (perfect preprocessing will cause "preprocessed" and "true" spectra to be indestinguishable) with different preprocessing: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

ods that remove unimportant features (PCA, SRACLS, and CDAE) reconstructed data that resembled training data in peak width and location, compared to the case with no preprocessing. However, the methods that explicitly identified or constrained the non-target contributions (BSS ICA, BSS PCA, and NCCLS) reconstructed data that resembled training data and additionally matched the true target peak heights. Therefore, this study suggests that specifically identifying or constraining the non-target species is valuable to removing its spectral contribution via a preprocessing algorithm.

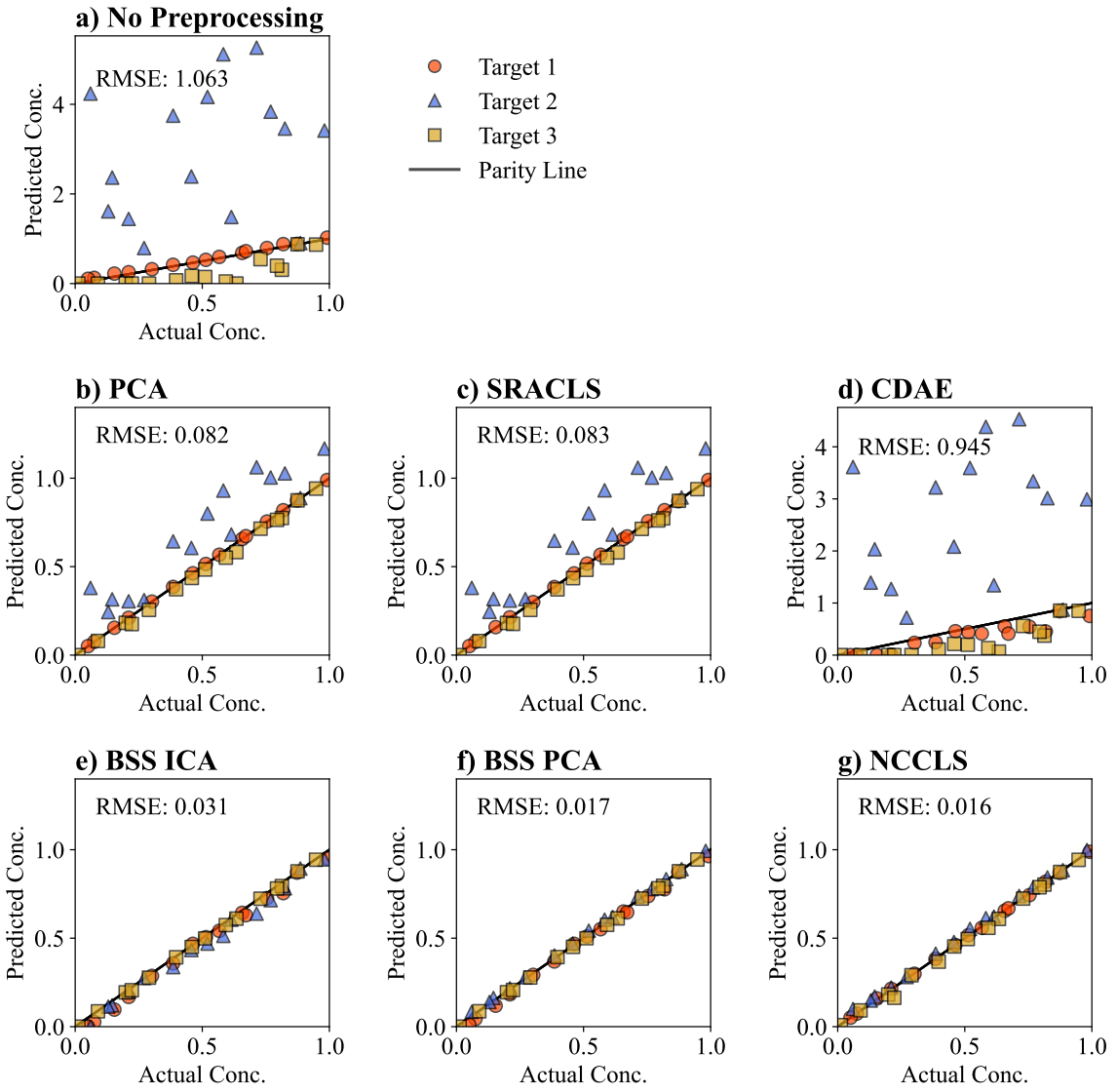


Figure 3.8: Parity plots quantifying batch preprocessed *in silico* spectra from Computational Study 1 with different preprocessing: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

Real-Time Preprocessing Methodology

Figure 3.9 shows the results from Computational Study 1 in a real-time monitoring scenario where the same data from Computational Study 1 was quantified, but each test spectrum was preprocessed in *isolation*. This was to distinguish from the results shown in Figure 3.8, where all 15 test spectra were preprocessed simultaneously. Analyzing a single spectrum at a time is more challenging for methods that extract information from

test data for source separation, such as in the BSS methods studied in this chapter. Figure 3.9 shows that PCA, BSS PCA, and BSS ICA were negatively impacted as preprocessing methods when preprocessing a single test spectrum. However, SRACLS, CDAE, and NCCLS had unaffected performance, suggesting that these methods are robust for real-time or data-limited applications. It is shown in a study with replicates (Section 3.4.2) that the performance of SRACLS in a real-time preprocessing scenario may not be consistently reproducible.

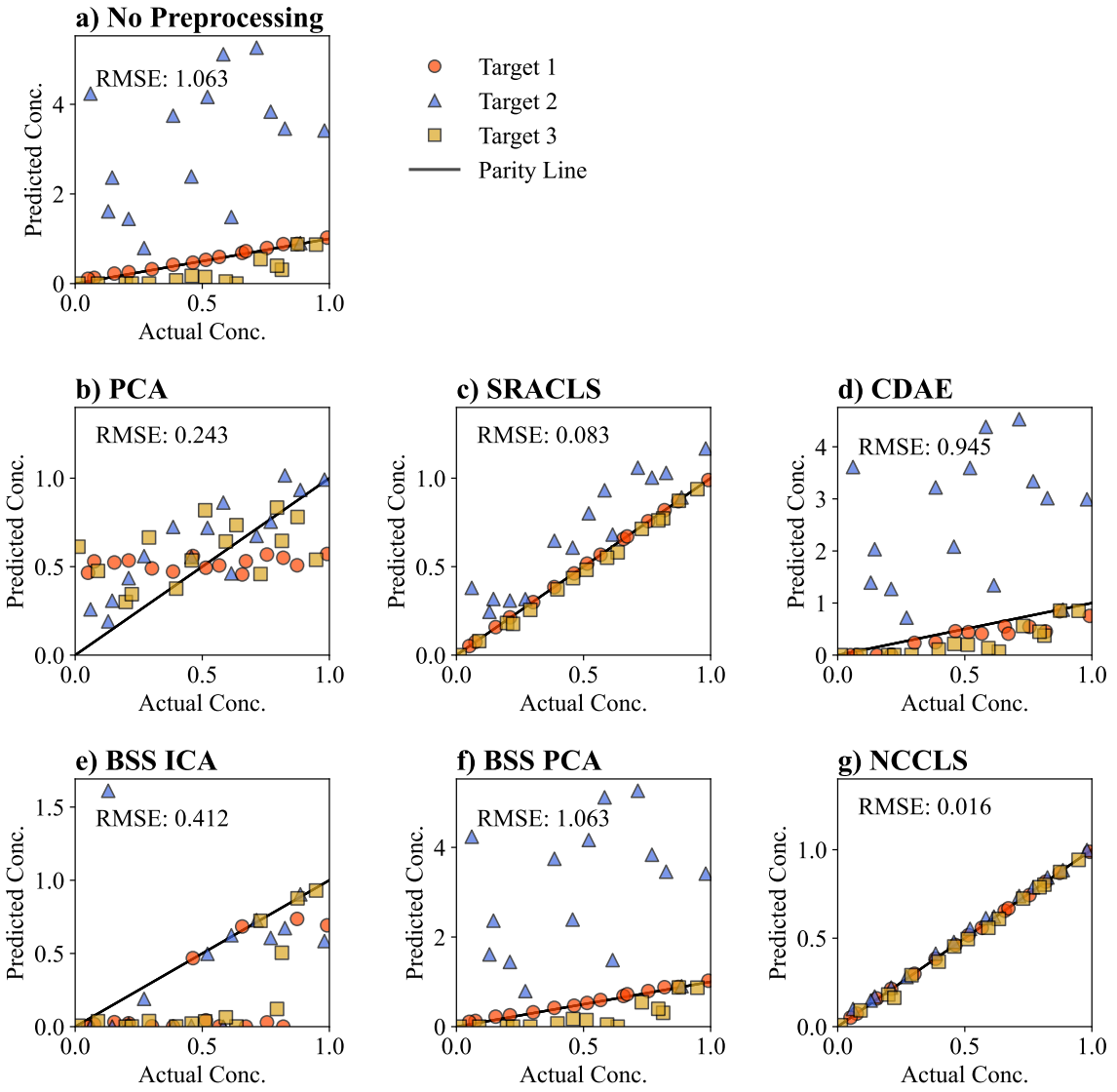


Figure 3.9: Parity plots quantifying real-time preprocessed *in silico* spectra from Computational Study 1 with different preprocessing: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

3.4.2 Computational Study 2

Computational Study 2 performs replicates of Computational Study 1 with a batch preprocessing methodology. First, the different preprocessing methods were compared as the amount of non-target peak overlap was adjusted from 8.0% – 100.0%. In Figure 3.10, the effect of different amounts of overlap can be seen for the different preprocessing methods. The RMSE for the *no preprocessing* scenario increased and then decreased as the amount

of non-target overlap was increased. This behavior was due to SNV scaling that was applied to all of the data prior to being input into the PLSR quantification model. While 100% overlap caused the most overlap between the target and non-target peaks, the greatest quantification interference occured between 20% – 30% peak overlap. The training data in this study had a lower standard deviation far from the target peaks. Since the non-target peak was in the region of the domain which had a low standard deviation in the training data, the non-target peak was divided by a relatively small standard deviation (scaled to be relatively large) when SNV-scaled (shown in Equation A.3). This scaling behavior increased the impact of the SNV-scaled non-target peak.

In Figure 3.10, all of the methods except CDAE improved quantification compared to no preprocessing at small and intermediate amounts of peak overlap. However, the performance of all methods converged at high peak overlap amounts and did not offer appreciable quantification improvement at 100% peak overlap. This result was expected, because at 100% peak overlap, there were no distinguishing features between the target and non-target peaks in this study. At large overlaps (50% or more), the best performing methods were BSS ICA, NCCLS, and then BSS PCA, which all outperformed PCA, SRACLS, and CDAE. The BSS methods and the autoencoder (BSS PCA, BSS ICA, and CDAE) had greater variances across different runs compared to the other methods in Figure 3.10 (PCA, SRACLS, and NCCLS). This result was consistent with the performance of MCR-ALS, ICA, and network layers being variable run-to-run with different initializations (as was done in this chapter).

Figure 3.11 shows how the preprocessing methods were affected by varying the standard deviation of Gaussian instrument noise from 0.01% – 4.00% of the maximum peak height. The results with no preprocessing (red circles) indicated a counter-intuitive result; increased instrument noise resulted in better quantification at high noise levels. This counter-intuitive behavior can be attributed to the scaling applied to all spectra prior to PLSR quantification in this chapter. Increased instrument noise resulted in an increased

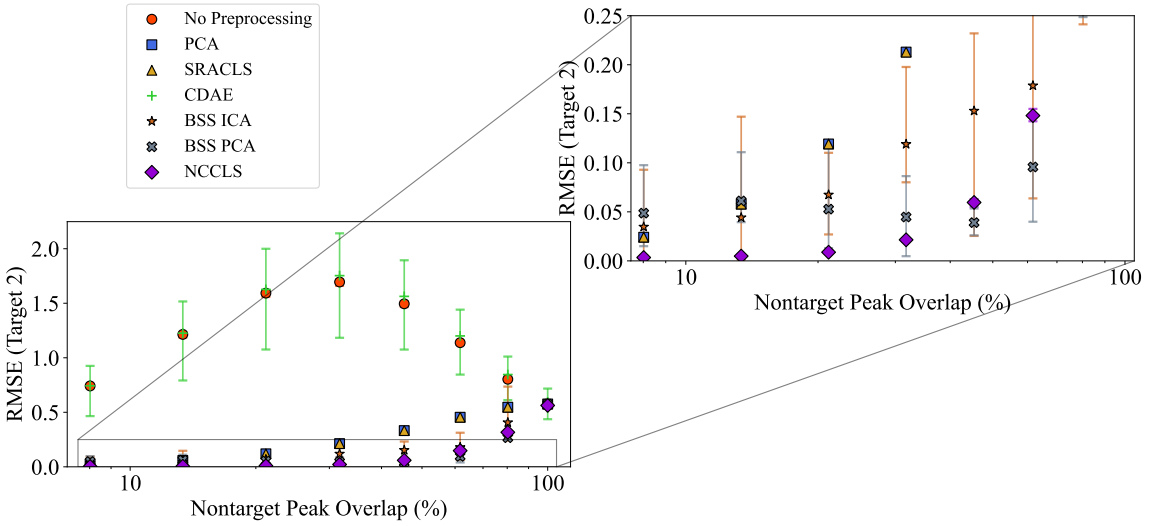


Figure 3.10: Comparison of Target 2 RMSE for different preprocessing methods as a function of non-target peak overlap percentage. Error bars correspond to full range of five replicates used for this study with pseudorandom spectra and model initializations (see Section 3.3.1 for experiment description). The inset plot is distinguished by a "zoomed in" scaling.

standard deviation of training data (refer to Equation A.3 for SNV scaling). An increased standard deviation decreased the scaled peak height of the non-target peak in the scaled space prior to quantification. Therefore, increased noise in the training data resulted in better quantification by the PLSR model in this case.

Figure 3.11 also shows that all of the preprocessing methods, with the exception of CDAE, improved PLSR quantification at all studied noise levels. NCCLS preprocessing resulted in less accurate quantification as noise levels were increased. BSS PCA and BSS ICA did not have a definite trend with increased noise. PCA, SRACLS, and CDAE improved as noise levels were increased. The improved quantification of these methods with increasing noise was due to fewer features providing meaningful signal when the spectrum was noisy. Only features that provided large amounts of target signal relative to the noise were identified as important features by the algorithms, and therefore projected into the methods' subspaces. Thus, the non-target peak did not have as much of its information projected at higher noise levels with PCA, SRACLS, and CDAE.

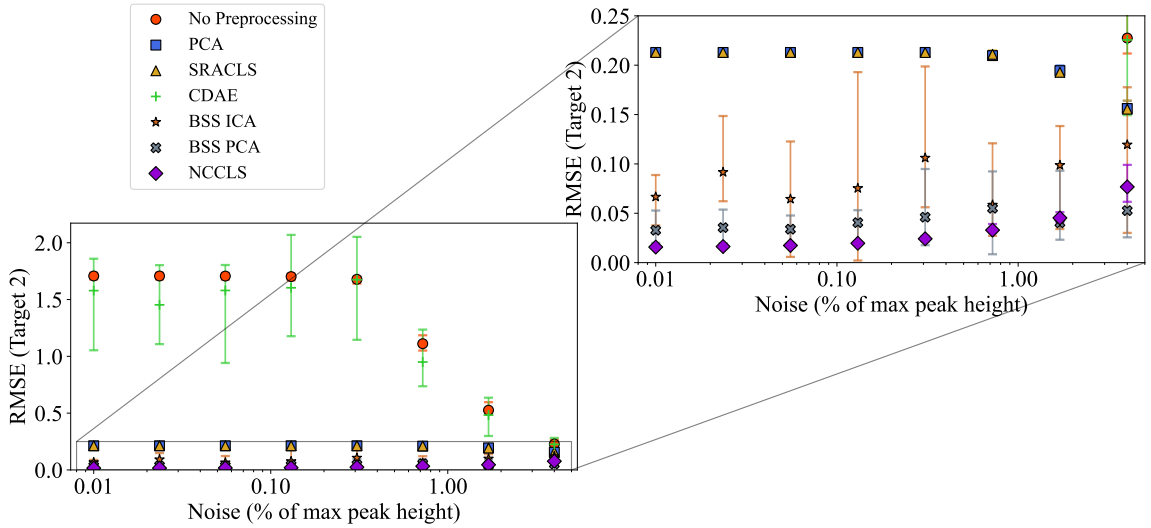


Figure 3.11: Comparison of Target 2 RMSE's for different preprocessing methods as a function of instrument (spectra) noise. Error bars correspond to full range of five replicates used for this study with pseudorandom spectra and model initializations (see Section 3.3.1). The inset plot is distinguished by a "zoomed in" scaling.

Figure 3.12 shows the effect different amounts of test data (1–10,000) have on the preprocessing methods studied in this chapter. As the amount of test data were increased, BSS ICA and BSS PCA showed better performance. At low amounts of test data (i.e. one test spectrum available at a time), all of the methods had a relatively large variance except NCCLS. With just one test spectrum, PCA, SRACLS, and BSS ICA all have at least one run that outperformed the worst performing run of NCCLS. However, NCCLS is the only method that consistently performed accurately during all five *in silico* trials with a Target 2 RMSE of 0.012. PCA was the closest in average performance to NCCLS with a Target 2 RMSE of 0.092 representing a 695% increase over NCCLS when provided a single test spectrum. The performance of SRACLS given five replicates suggests that the good real-time performance shown by SRACLS in Computational Study 1 (Figure 3.9) may not be reproducible.

NCCLS offered appreciable improvement to spectra quantification for all studied conditions of non-target overlap, noise level, and amount of testing data except when the non-target and target species were indistinguishable at 100% overlap. Additionally, NCCLS is

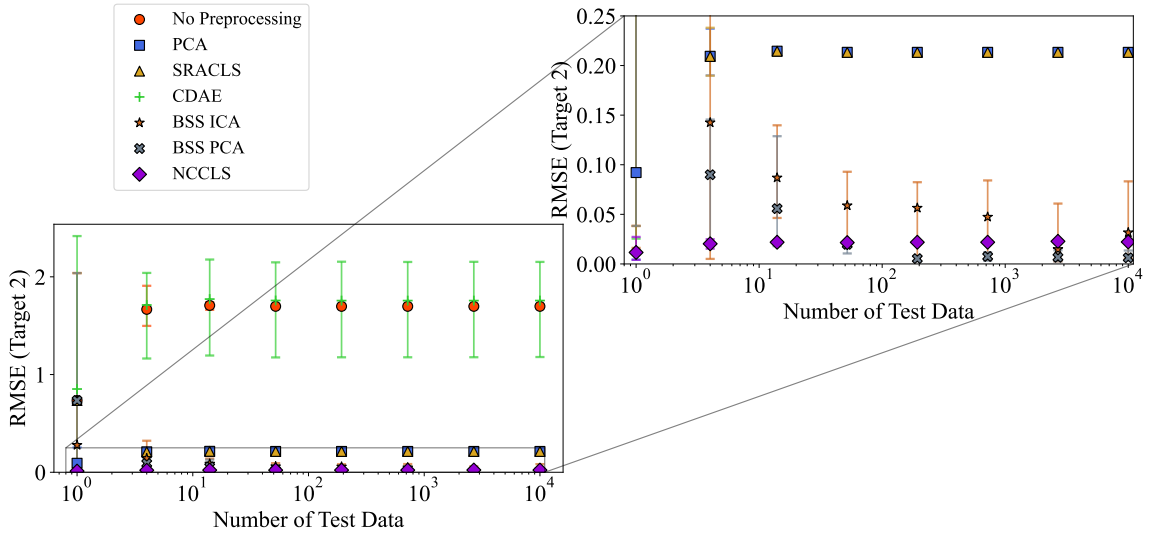


Figure 3.12: Comparison of Target 2 RMSE for different preprocessing methods as a function of number of test data. Error bars correspond to full range of five replicates used for this study with pseudorandom spectra and model initializations (see Section 3.3.1). The inset plot is distinguished by a "zoomed in" scaling.

the only method that consistently enabled accurate quantification when only one test spectrum was available for analysis. The effective and reproducible non-target removal using NCCLS in this *in silico* study supported its implementation on experimental data.

3.5 Results: Experimental Raman and ATR-FTIR Spectra

Experimental Raman and ATR-FTIR spectra were preprocessed with the preprocessing methods described in Section 3.2 in a real-time scenario where only one test spectrum at a time was available for preprocessing. The information flowchart from Figure 3.3 also applied to the experimental spectra in this section.

3.5.1 Raman Spectra (Real-Time)

The primary visual indicator of successful non-target removal in the Raman spectrum was the removal of the carbonate shoulder (1067 cm^{-1}) on the higher-wavenumber side of the nitrate peak (1050 cm^{-1}) without affecting the target peaks. Figure 3.13 shows the different preprocessing methods removing non-target spectral contributions. PCA, SRA-

CLS, and NCCLS display a subtraction of the non-target carbonate peak centered at 1067 cm^{-1} while leaving the nitrate peak at 1050 cm^{-1} unaffected. CDAE and BSS PCA did not appreciably subtract the carbonate peak, suggesting that these methods were not effective in identifying and subtracting the non-target carbonate peak in a real-time monitoring scenario. BSS ICA displayed the significant subtraction of the target nitrate peak in addition to the non-target carbonate peak, possibly indicating an incorrectly identified source. Target peak subtraction was attributed by Kocevska et al. to subtle peak-shifting that occurs with the nitrate peak as the solution phase becomes more concentrated [21, 67]. The effect of peak shifting for target species is not accounted for in this chapter for any of the methods presented.

The resulting preprocessed Raman spectra were also quantified by a PLSR model, with the results being included in the parity plots of Figure 3.14. SRACLS was the best performing preprocessing method followed by NCCLS, PCA, BSS PCA, CDAE, and BSS ICA. The more accurate quantification of SRACLS, NCCLS, and PCA agreed with the qualitative non-target removal shown in Figure 3.13, where the peak height of the nitrate peak was largely unaffected but the shoulder on the higher-wavenumber side of the peak was removed. Nitrate in the PCA preprocessed spectra was accurately quantified, as can be seen by the red circles in Figure 3.14b. However, sulfate and nitrite quantification were significantly impacted (peaks not shown in Figure 3.13), explaining the discrepancy between the apparent effective carbonate removal in Figure 3.13b and apparent ineffective quantification in the parity plots of Figure 3.14b. In Figure 3.14e, BSS ICA was underpredicting nitrate for most predictions, which agreed with the visual subtraction shown in Figure 3.13e.

3.5.2 ATR-FTIR Spectra (Real-Time)

ATR-FTIR spectra have some differences from the Raman spectra of the previous section. Notably, the peaks examined in this chapter were more tightly packed in the

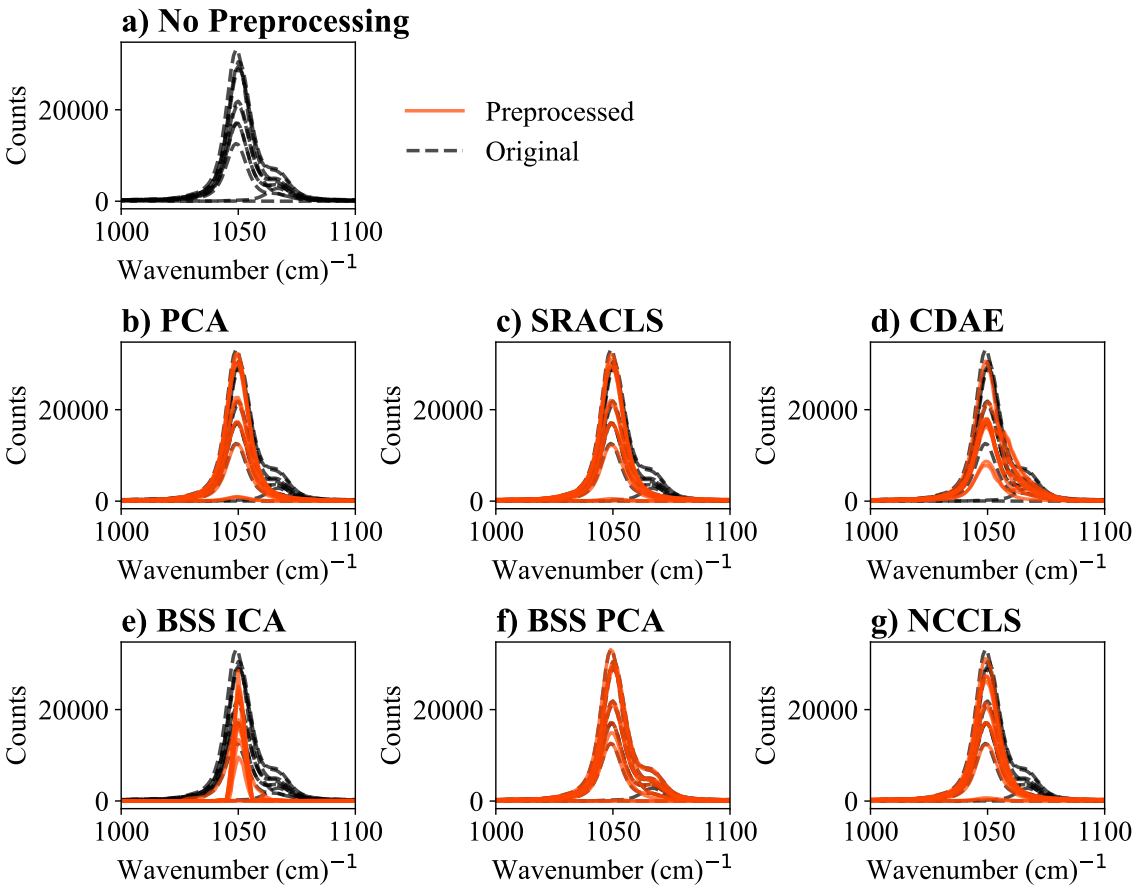


Figure 3.13: Experimental Raman spectra of sodium salts (dashed black) and preprocessed spectra (solid red) highlighting the nitrate peak at 1050 cm^{-1} and carbonate peak at 1067 cm^{-1}. Non-targets are removed in a real-time scenario using: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

FTIR spectrum, causing more peak overlap between target species and non-target species. Because of the large degree of peak overlap in this system, traditional feature selection methodologies are more likely to fail with ATR-FTIR spectra in this real-time monitoring scenario.

The primary visual indicator of successful non-target removal in the FTIR spectrum was the removal of the carbonate shoulder (1380 cm^{-1}) on the higher-wavenumber side of the nitrate peak (1350 cm^{-1}). Figure 3.15 shows the different preprocessing methods removing non-target spectral contributions. All of the preprocessing methods subtracted from the overlapping nitrate and carbonate peak at 1350 cm^{-1} with the possible exception of CDAE.

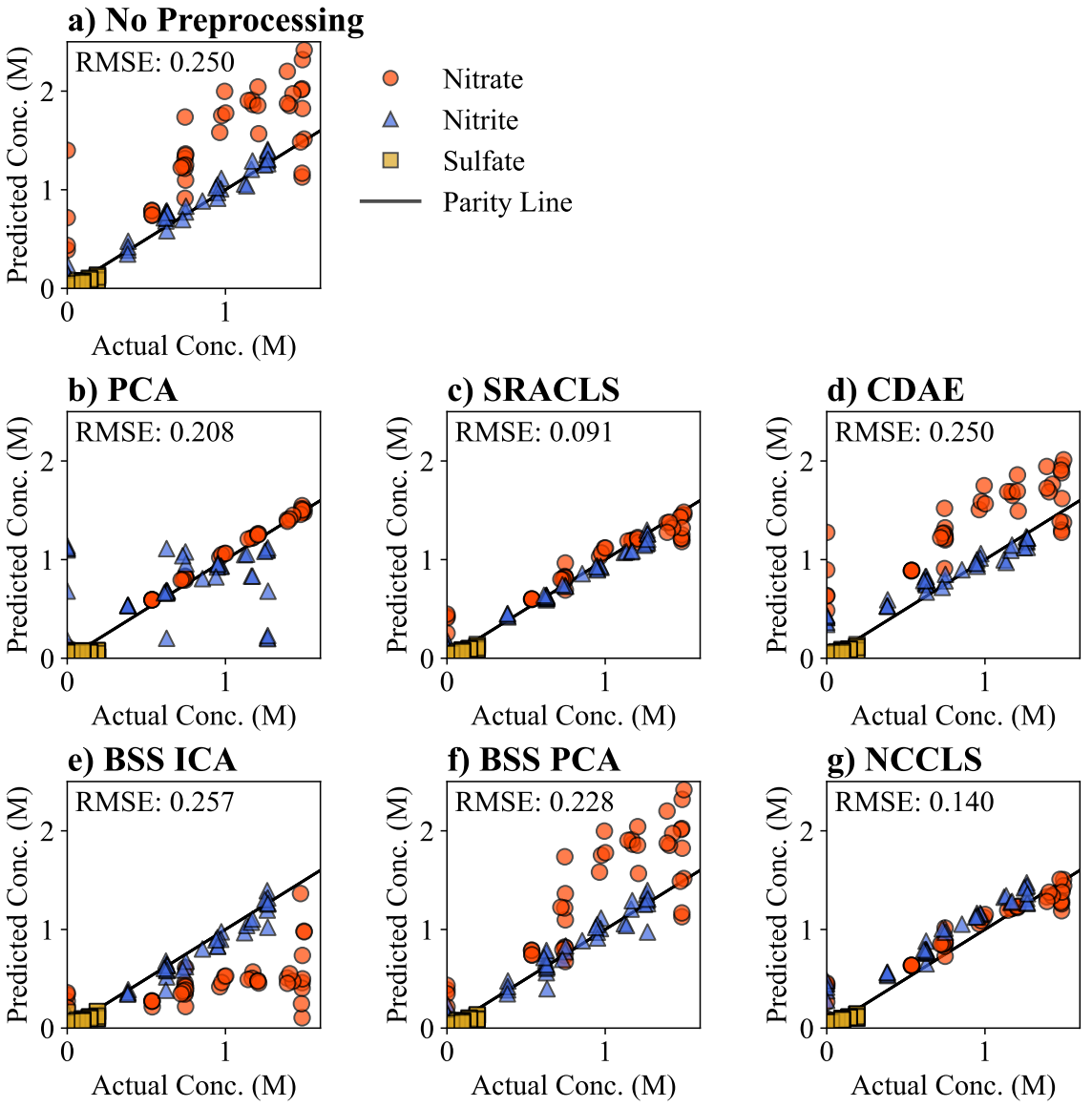


Figure 3.14: Parity plots quantifying nitrate, nitrite, and sulfate from Raman spectra preprocessed in a real-time scenario using: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

All of the other studied methods appeared to subtract the carbonate peak such that the nitrate peak shape was retained. However, further visual inspection is hindered by the large amount of overlap between the nitrate and carbonate peaks, and so the spectra were quantified via a PLSR model for a quantitative indicator of non-target removal accuracy.

The preprocessed FTIR spectra were quantified by a PLSR model, with the results being included in the parity plots of Figure 3.16. The lowest RMSE observed for the ATR-FTIR

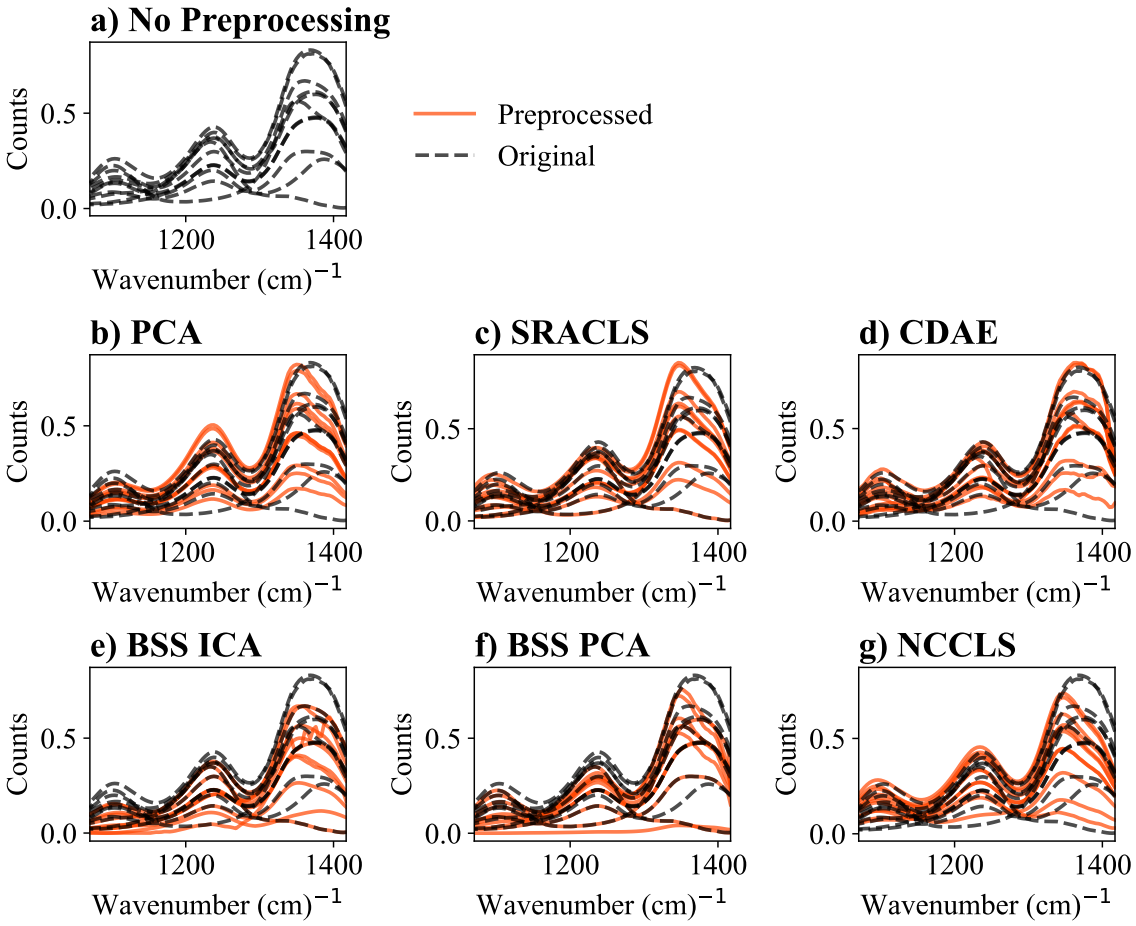


Figure 3.15: ATR-FTIR spectra of sodium salts (dashed black) and preprocessed spectra (solid red) where non-targets are removed in a real-time scenario using: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

quantification is NCCLS, which is the only method that improved quantification compared to quantification with no preprocessing. The remaining methods were most accurate in the order: SRACLS, BSS ICA, BSS PCA, PCA, and CDAE. Visual inspection of Figure 3.15 agrees with the quantitative results of Figure 3.16; NCCLS retained the shape of the nitrate peak unlike CDAE and BSS ICA, more accurately predicted nitrate peak heights than PCA and SRACLS, and did not affect other target peaks like BSS PCA.

The lack of effective quantification for the ATR-FTIR spectra contrasted with the relative success observed in the computational studies and experimental Raman spectra. The deteriorated preprocessing and quantification performance is in agreement with similar per-

formance observed in Computational Study 2 (Section 3.4.2) with a small amount of testing data and large amount of peak overlap. In Computational Study 2, all of the methods except NCCLS led to reduced quantitative accuracy in low data regimes (i.e. one test spectrum); NCCLS had the best preprocessing performance. Additionally, at large amounts of non-target overlap, calculated as 79.8% for ATR-FTIR spectra, all of the methods that improved quantification through preprocessing had less significant performance improvements in Computational Study 2. The poor quantification can be attributed to a combination of a high-degree of peak overlap, a single testing spectrum, and the possibility of other nonlinearities that appear in real spectra. Analogous results for physical spectra preprocessed in a batch preprocessing manner are shown in Section D.2 of the Appendix. These results are qualitatively similar to the results presented in this section, with NCCLS being the only method that improved quantification RMSE compared to PLSR with no preprocessing. Additional results that investigated computational efficiency as well as the experimental results presented here analyzed in a "batch" manner can be found in Chapter D.

In cases where non-target removal is ineffective, as is suggested by the imperfect removal of carbonate from the nitrate peak in the ATR-FTIR data, the possibility of including non-target species as additional target species may warrant consideration. Inclusion of additional target species will increase the size of training data. However, this method of handling non-target species has been demonstrated for overlapping species previously. For example, Kocevska et al. treat nitrate and carbonate both as target species in a similar chemical system and instead remove only non-target species with less significant overlap [21].

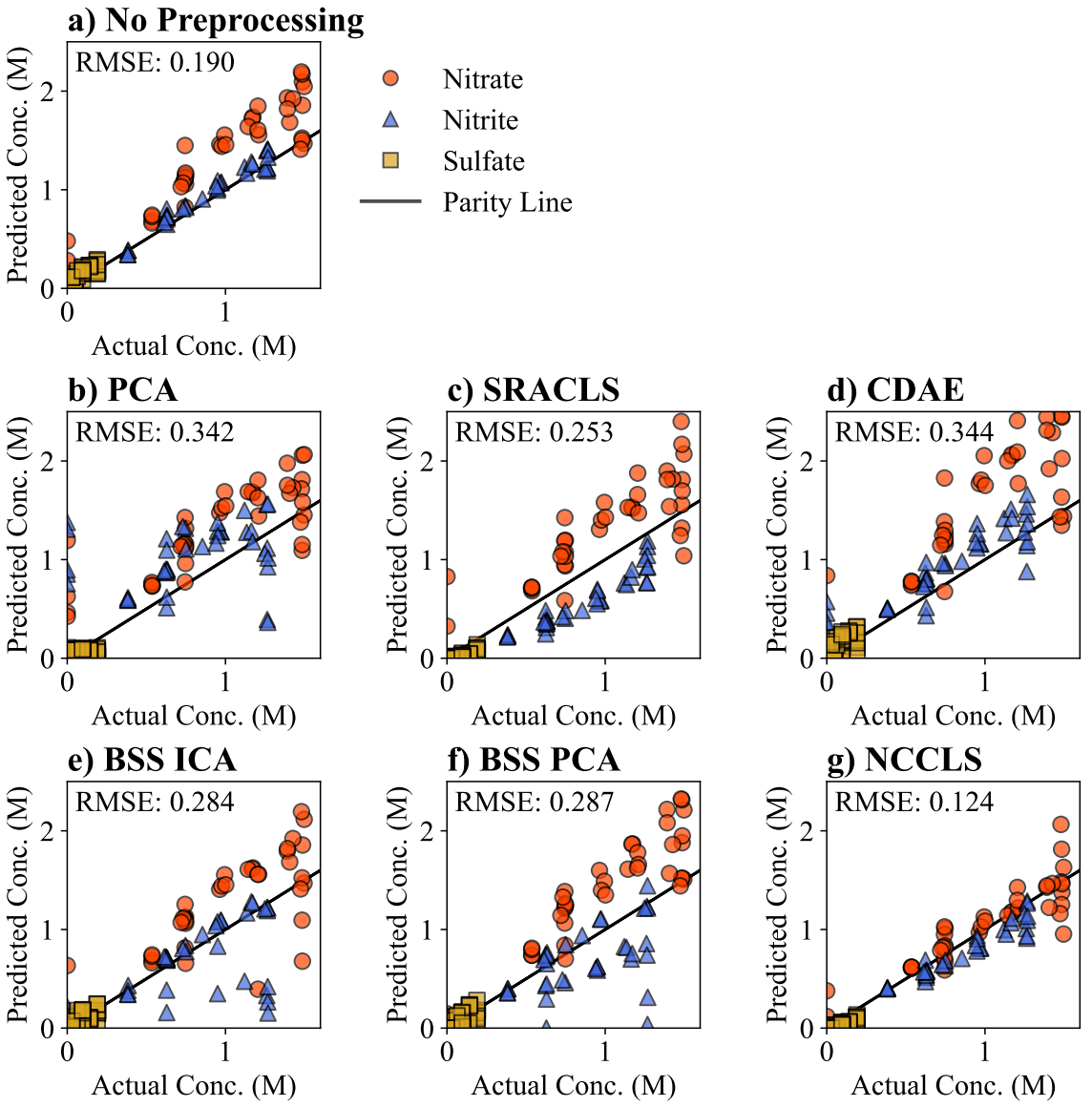


Figure 3.16: Parity plots quantifying nitrate, nitrite, and sulfate from ATR-FTIR spectra preprocessed in a real-time scenario using: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

3.6 Conclusion

Establishing robust quantification models is vital to industrial processes utilizing spectroscopy where non-target species may appear. This chapter introduces a new spectral preprocessing technique, nonnegatively constrained classical least squares (NCCLS), that leverages the nonnegativity of non-target spectral contributions to preprocess solution-

phase vibrational spectra. In the present chapter, NCCLS was shown to improve preprocessing in comparison to other available methods, including the BSS method introduced in Chapter 2. The comparison was with both idealized spectra in a computational study and real spectra in an experimental study using Raman and ATR-FTIR spectra. For *in silico* spectra, NCCLS was shown to be the best performing method for both batch preprocessing and real-time preprocessing scenarios. Additionally, it was demonstrated that NCCLS may have particular use in industrial real-time monitoring scenarios, with an average (over five replicates) overlapping-target RMSE of 0.012 representing a 7.95 times improvement over the next-best method in an *in silico* real-time example. The robustness of NCCLS was demonstrated by testing the effects of varied peak overlap, noise level, and data quantity.

Additionally, the effectiveness of NCCLS was demonstrated in an experimental study utilizing Raman and ATR-FTIR spectra in a real-time transfer learning scenario; spectroscopically active species appear in the test data that do not exist in the training data. In the Raman spectrum, NCCLS performed comparably to other methods, only being outperformed by SRACLS. In the ATR-FTIR spectrum, NCCLS was the only preprocessing method that was able to improve quantification compared to spectra with no preprocessing. This chapter explored NCCLS as a single method for improving the robustness of vibrational spectra quantification models. The physical constraints utilized in NCCLS may have bearing to additional preprocessing methods and applications beyond the currently explored study.

CHAPTER 4

MEASURING DENSE SLURRIES WITH SPECTROSCOPIC SENSORS

The prior chapters in this thesis have been primarily concerned with measuring solutions; i.e. systems where all components fully dissociate in a solvent. However, nuclear waste is often a *slurry*. All slurries — with some examples being mud, paint, toothpaste, and cake batter — have two parts: a solution phase and an insoluble phase. The multiphase nature of slurries can make them difficult to process and monitor in real time due to having variable composition, particle size, and particle shape. The nuclear-waste slurries present at the Hanford site in Washington State are multicomponent, multiphase, and inhomogeneous.[1]

As mentioned in the introduction of this thesis, the state-of-the-art for analyzing radioactive waste at Hanford requires laboratory results from an on-site analytical laboratory (grab-sampling), which can delay processing speed and create exposure risks for workers. In-line probes may provide an alternative route to collecting necessary composition information. However, there is limited published literature showing quantification of spectra in dense multicomponent slurries typical of nuclear waste.

In this chapter, Raman spectroscopy and attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy are tested on simulants of nuclear-waste slurries containing up to 23.2 wt% solids. It was observed that ATR-FTIR spectroscopy was effective in measuring the solution phase of the studied slurry systems (3.52% mean percent error), while Raman spectroscopy also provided information about the suspended solids in the slurry system (18.21% mean percent error). Measurement of multicomponent solids typical of nuclear-waste processing is previously unreported. The composition of both the solution and solid phases are vital in ensuring stable glass formulation and effective dis-

[1]Much of the content and many of the figures in this chapter are reproduced from Prasad et al. [19]

posal of nuclear waste at Hanford. Raman and ATR-FTIR spectroscopies may provide a safer, faster alternative for acquiring compositional information on nuclear-waste slurries.

4.1 Introduction

The use of ATR-FTIR and Raman spectroscopy is poorly studied in multicomponent systems with concentrated solids [19]. Because of the variety of real systems that contain solid particulate matter, the results presented here have importance for operations at Hanford and have bearing on real-time monitoring in many fields [106, 107]. Optical spectroscopy has been applied as a process analytical technology (PAT) tool in pharmaceutical, food, and mining industries [19]. ATR-FTIR spectroscopy primarily measures the solution phase due to a short laser path length, while Raman spectroscopy (not in an attenuated total reflectance (ATR) configuration) interrogates both the solution phase and suspended solid particles. However, most applications of these spectroscopies have been limited to systems with concentrations of suspended solids up to 5 wt% or to binary components at higher solid loadings [108]. Although other spectroscopies, such as X-ray Fluorescence and near-infrared spectroscopy have been used in the mining industry to investigate systems containing 21 wt% solids [109], the application of vibrational spectroscopy for monitoring multi-component slurries with high solids content is scarce. Specifically in the nuclear field, Raman spectroscopy and ATR-FTIR spectroscopy have been shown effective in identifying and quantifying molecular species in solution. Despite complex solution behavior and overlapping spectral bands, progress has been made in the analysis of streams of nuclear waste through preprocessing techniques [67], physical modeling [110], and use of multiple excitation wavelengths [22]. However, much of this work has been done analyzing optically transparent solutions without high concentrations of suspended solids.

For high-level waste (HLW) processing at Hanford, the proposed process is designed for 20 wt% insoluble solids [111]. For the currently planned low-activity waste (LAW) process, the addition of glass-forming chemicals (GFCs) will create a slurry with roughly

22–33 wt% insoluble solids [112]. Since the application of PAT to monitor nuclear-waste slurries has received scant attention, this work investigates the applicability of in-line probes to analyze slurries representative of nuclear waste containing up to 400 g of insoluble solids/kg solvent (23.2 wt%). Measuring the solids concentration of slurries with Raman spectroscopy presents challenges because of the strongly absorbing and scattering properties of most slurries, in addition to competing fluorescence effects [113, 90].

The present chapter focuses on a specific scenario in waste processing at Hanford. Specifically, monitoring the composition of nuclear-waste feeds within the Melter Feed Preparation Vessel (MFPV) as a case study of nuclear waste containing suspended solids. The MFPV will have a similar purpose in both LAW and HLW processing: that is, GFCs are added to nuclear waste in the MFPV tank before eventually being transported to the melter. Compositional measurements of the MFPV are planned to verify proper compositions for a durable glass form before being melted [7]. The analytical measurements of the slurry in the MFPV vessel represent a hold point during HLW vitrification, indicating the waste will not be further processed until concentration measurements are obtained. Therefore, the implementation of in-line probes may offer advantages by facilitating faster downstream decision making and mitigating the risk associated with grab-samples. In addition to the specific study of nuclear waste entering the MFPV, much of the work presented here is also applicable to other processing instances that have multicomponent suspended solids [114].

4.2 Methods

4.2.1 Instrumentation

Measurements were collected at a 100 mL scale in a Mettler Toledo (MT) OptiMax reactor (250 mL) installed with a pitched-blade agitator (Alloy C-22, downward, ∅ 45 mm) fitted with the following in situ devices: a Raman probe, a pH probe, a temperature probe, and an ATR-FTIR probe (see Figure 4.1). A Teflon vessel was used due to the high pH

of the solution. Raman spectra were recorded with a Mettler Toledo ReactRaman 785 instrument using a 785 nm laser at 300 mW power, 1 s exposure time, 10 averaged spectra, and a spectral resolution of $6\,\mathrm{cm}^{-1}$. The Raman probe tip is a ball–probe configuration leading to a focal point about 200 µm from the probe surface [115, 116]. Infrared spectra were recorded with a Mettler Toledo ReactIR 10 instrument with a diamond probe tip and a spectral resolution of $8\,\mathrm{cm}^{-1}$. The ATR design limits the penetration depth of the infrared radiation to $\sim 2\,\mu\mathrm{m}$ [53]. Data were acquired using iC Raman and iC IR software from Mettler Toledo. Fouling of the ATR-FTIR probe was observed in the presence of silicates (see Figure 4.2). To minimize the effect of fouling, the ATR-FTIR probe was cleaned before each measurement. No fouling was observed on the Raman probe; the difference in fouling behavior between the Raman and ATR-FTIR probe tips may be caused by a difference in material (sapphire for Raman and diamond for ATR-FTIR) or geometry (convex for Raman and concave for ATR-FTIR).

4.2.2 Composition

The system is based on the 5.6 M Na^+ low-activity waste (LAW) pretreatment system simulant (sodium salts) [21] combined with simulated glass-forming chemical (GFC) recipes composed predominantly of metal oxides and silicates [112]. Figure 4.1b shows a slurry obtained after the addition of GFCs with the simulants (900 g GFCs/kg water). The LAW simulants comprise water, sodium hydroxide, and seven sodium salts: nitrate, nitrite, carbonate, sulfate, phosphate, oxalate, and acetate. The GFC compositions were formulated and provided by Savannah River National Laboratory (SRNL) [112]. The GFC simulants comprise five insoluble silicates: silica, kyanite, wollastonite, olivine, and zircon; four metal oxides: hematite, rutile, tin oxide, and zinc oxide; and four additional soluble species: vanadium pentoxide, boric acid, sucrose, and lithium carbonate. The concentration ranges for all species are listed in Table 4.1 – Table 4.3. All experiments were conducted at 3 m NaOH to simulate the basic conditions expected at the Hanford Waste

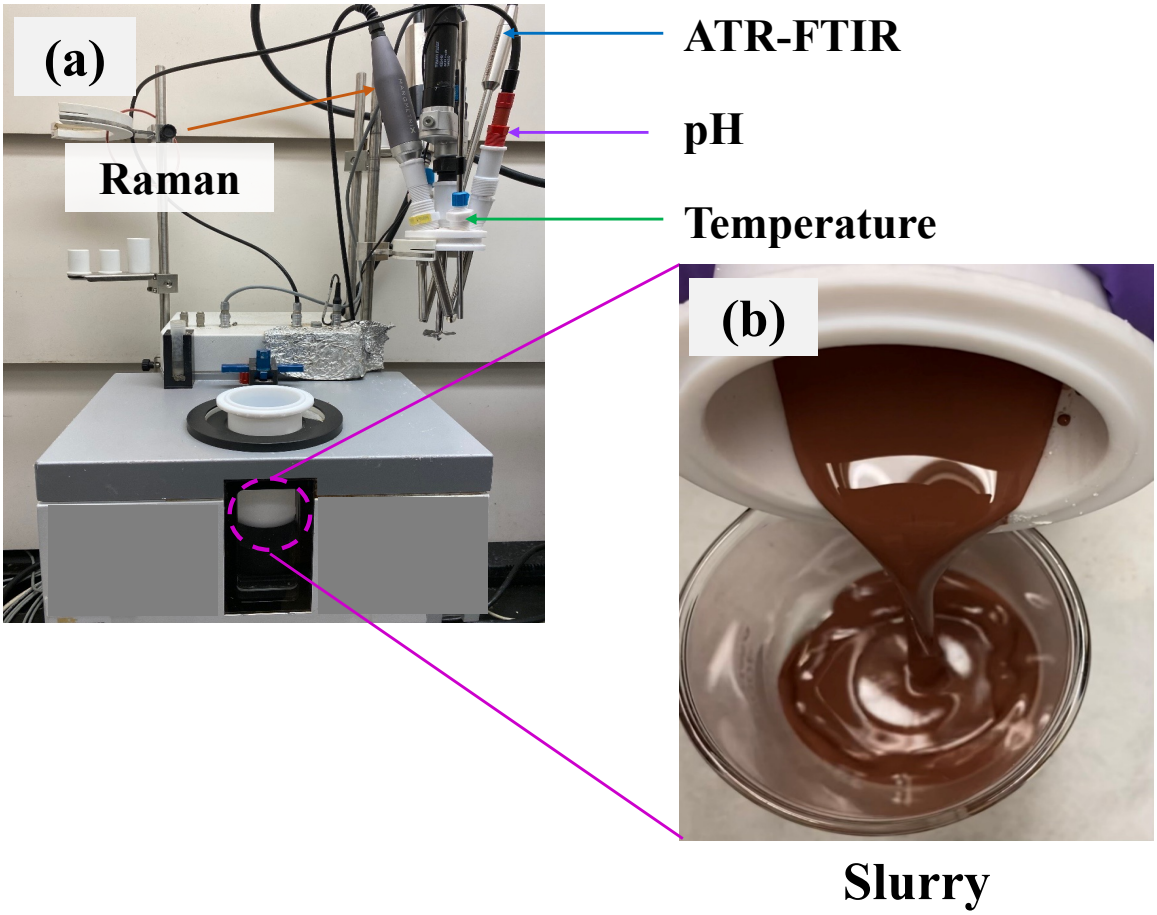


Figure 4.1: **a)** Photograph of the OptiMax reactor setup with in-situ probes: Raman, ATR-FTIR, temperature, and pH, **b)** a close-up view of the dense slurry studied in this work.

Treatment Plant (WTP). At this molality of sodium hydroxide, the pH remained above 13 for all experiments, which was verified with an in situ pH probe. The experiments were temperature-controlled to 25 °C and stirred at 400 rpm to maintain a suspension of solids inside the apparatus. The experiments were designed using the MATLAB (2022b) random number generator to randomly design compositions within the bounds of each species. Additional details about the validity of the experimental design are discussed in Chapter F.

4.2.3 Design of Experiments

The experimental space was designed by creating component-wise lower and upper bounds based on reported nuclear-waste simulants [117, 112]. Species were bounded by

Table 4.1: Range of solid concentrations tested in experiments by addition of GFCs (insoluble components in 3 m NaOH).

Species	Formula	Maximum (g/kg water)	Minimum (g/kg water)
Kyanite	Al_2SiO_5	99.6	5.9
Wollastonite	$CaSiO_3$	99.8	0.0
Olivine	Mg_2SiO_4	45.3	0.0
Silica	SiO_2	249.9	18.0
Zircon	$ZrSiO_4$	42.4	0.0
Hematite	Fe_2O_3	9.4	0.0
Rutile	TiO_2	1.2	0.0
Tin Oxide	SnO_2	12.7	0.0
Total (Single Experiment)	–	401.8	53.1

Table 4.2: Range of solids composing GFCs with intermediate solubility measured in 3 M NaOH.

Species	Formula	Maximum (g/kg water)	Minimum (g/kg water)
Vanadium Pentoxide	V_2O_5	31.9	0.0
Zinc Oxide	ZnO	22.5	0.0
Sucrose	$C_{12}H_{22}O_{11}$	53.7	4.7

Table 4.3: Range of dissolved anions present in the nuclear-waste simulants studied (soluble components in 3 M NaOH).

Species	Formula	Maximum (mol/kg water)	Minimum (mol/kg water)
Hydroxide	OH^-	2.98	2.94
Nitrate	NO_3^-	1.44	0.75
Nitrite	NO_2^-	1.16	0.56
Soluble Carbonate	CO_3^{2-}	0.96	0.14
Sulfate	SO_4^{2-}	0.13	0.03
Borate	$B(OH)_4^-$	2.84	0.15
Phosphate	PO_4^{3-}	0.07	0.02
Oxalate	$C_2O_4^{2-}$	0.01	0.00
Acetate	$C_2H_3O_2^-$	0.19	0.02
Sucrose	$C_{12}H_{22}O_{11}$	0.24	0.01

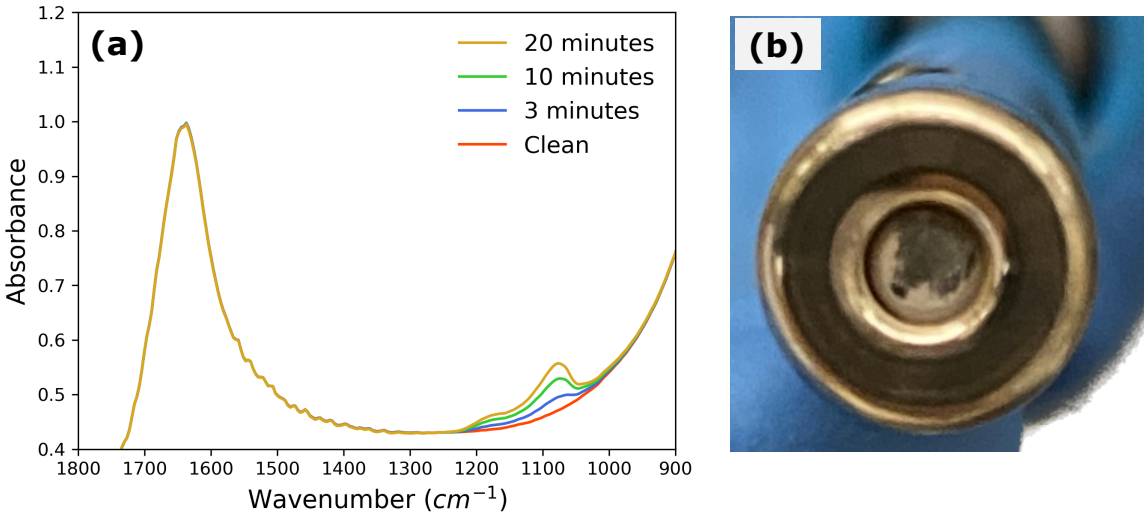


Figure 4.2: **a)** Deposition of solids on the ATR-FTIR probe tip when placed in a 3 m NaOH solution with 10 g silica/kg solvent suspended silica, and **b)** image of the probe tip showing gradual buildup on the probe tip.

concentrations expected in nuclear-waste streams at the Hanford WTP. Individual experiments were conducted by randomly sampling from this experimental space using a pseudo-random uniform distribution in MATLAB 2021a. These pseudo-random experiments were then ordered in batches of 6–12 step-wise samples to allow serial addition; this enabled more data collected with the same materials. There were 48 unique samples (at different concentrations) collected for ATR-FTIR, and there were 66 unique samples (at different concentrations) collected for Raman. When silicates were present, solid particles deposited on the ATR-FTIR probe during data collection. To eliminate the spectral bands caused by depositing solids, the ATR-FTIR probe was cleaned before every measurement containing silicates to ensure reliable spectra. Every sample was allowed to equilibrate before the ATR-FTIR probe was cleaned and a measurement was taken. Data showing this deposition are provided in Section 4.2.3. All solid species were nominally 325-mesh or less (corresponding to a diameter of 45 μm or less). The ranges of added concentrations are listed below for insoluble and partially soluble species in Table 4.1 and Table 4.2, and soluble species in Table 4.3. The solubility for these species in 3 m NaOH is reported by Prasad et al. as measured by inductively coupled plasma (ICP) [19].

ATR-FTIR Probe-Tip Cleaning

The silicates (silica, kyanite, wollastonite, olivine, and zircon) were observed to deposit on the ATR-FTIR probe at the basic conditions studied. This deposition appeared as a broad combination of peaks centered around 1100 cm^{-1} and spread from 1000 cm^{-1} to 1200 cm^{-1}. In Figure 4.2a, these peaks can be seen as they appear in the Fourier transform infrared (FTIR) spectrum and in Figure 4.2b, images of the solid substance on part of the probe tip. Notably, the solids provided a peak but did not interfere with quantification of the solution phase (as shown by the unaltered water peak centered at 1640 cm^{-1}).

While the solid deposition was not observed to interfere with the solution phase measurements and could potentially be subtracted as a baseline in practice, the experimental procedure included cleaning the probe tip before every measurement to minimize unanticipated sources of variation and to ensure experimental consistency. Because of the slow buildup of solids on the probe tip, measurements were taken with 15 s scan time and immediately after reinserting probe into the solution. Detection and removal of solid contributions may be achieved through computational approaches presented elsewhere [21, 118] and appearing in Chapter 2 and Chapter 3 of this thesis.[2] In addition, different materials or probe geometries may result in a less favorable surface for deposition. Solids were not found to attach to the Raman probe tip during experiments.

4.2.4 Data Preprocessing

All preprocessing steps were performed with Python 3.9.12. The Python code and the dataset can be found on Github[3]. ATR-FTIR data were narrowed to a range of 900 to 1800 cm^{-1} for quantification. This range contains all solution peaks provided by the instrument. A Savitzky-Golay filter [119, 100] was used to differentiate the ATR-FTIR spectra with respect to wavenumber with five filter points, a second order polynomial, and first derivative.

[2] Computational removal of probe fouling is done in Chapter 7.

[3] https://github.com/magrover/multicomponent-slurry-quantification

The publicly available package, SciPy, was used to perform Savitzky-Golay filtering. The Raman data were narrowed to a range of 100 to 1700 cm^{-1} for quantification. The range contains all observable peaks for insoluble species. No derivative was taken for the Raman spectra because of the correlation between baseline and some solid species, although the effect of first-derivative Savitzky-Golay filtering may improve model performance and reproducibility and is shown in Section 4.3.5.

4.2.5 Regression Model

A linear spectra-to-concentration relationship is expected with a Raman probe in an ideal system [53, 120]. The quantitative basis for Raman spectroscopy is given as:

$$L = P_D \, \beta \, D \, K \tag{4.1}$$

where L is a measure of intensity $[\frac{\text{photons}}{\text{sr}\cdot\text{cm}\cdot\text{s}}]$, P_D is power of the laser $[\frac{\text{photons}}{\text{cm}^2\cdot\text{s}}]$, β is the differential Raman cross section $[\frac{\text{cm}^2}{\text{mol}\cdot\text{sr}}]$, D is the density of Raman-scattering molecules $[\frac{\text{molecules}}{\text{cm}^3}]$, and K is a dimensionless geometric factor accounting for detection angle [90, 120]. Equation 4.1 shows that, when $P_D \, \beta \, K$ are constants, the detected Raman scattering depends linearly on the density of scattering molecules. Attenuation or absorption can affect the Raman bands of optically dense samples, which might be seen in a slurry with suspended solids and is investigated in Chapter 5 [90, 120, 121].

For the ATR-FTIR probe, the Beer–Lambert Law is expected to apply to solution-phase measurements, with little interference from suspended solids [53]. The Beer–Lambert Law is given as:

$$A_\lambda = \varepsilon_\lambda \, l \, c \tag{4.2}$$

where A_λ is absorbance at a particular wavelength, ε_λ is the molar absorptivity at that wavelength $[\frac{\text{L}}{\text{mol}\cdot\text{cm}}]$, l is the effective path length $[\text{cm}]$, and c is the species' concentration

$[\frac{\text{mol}}{\text{L}}]$.

Partial least squares regression (PLSR) [101, 102] was chosen as the spectra-to-composition model in this work. PLSR has been used for analysis of both nuclear waste solutions [21, 24, 9, 122] and pharmaceutical slurries [123, 124, 120]. The scikit-learn package (version 1.0.2) implementation of PLSR was used for all quantification in this work. Additional description of PLSR is given in Chapter B of this thesis. No scaling was applied to the data prior to data quantification in this chapter, since scaling had a negative impact on Raman quantification of the insoluble species presented here.

Training and testing datasets were determined using a leave-one-out cross validation scheme. A single sample (test spectra) was estimated using a PLSR model trained on all other samples (training spectra), and then the process is repeated. Minimum Akaike information criterion (AIC) was used for determining the number of latent variables in each PLSR model [125]. The Raman PLSR quantification model was determined to use 10 latent variables, while the ATR-FTIR PLSR quantification model was determined to use 15 latent variables.

As part of the analysis in Section 4.3.4 and Section 4.3.5, mixture spectra were predicted from the gravimetrically measured concentrations of both dissolved salts and suspended solids. An indirect classical least squares (ICLS) method was used for visualizing what spectra "should" look like for Raman or ATR-FTIR based on Equation 4.1 and Equation 4.2 [126]. ICLS is distinguished from classical least squares (CLS) by calculating reference spectra (linear basis functions) from mixture spectra with known composition; in CLS, reference spectra are known. From the estimated reference spectra from ICLS and known composition information, linearly-predicted mixture spectra were estimated and compared to observed spectra. For ICLS predictions, all spectra were mean-centered. Linear references were determined from the mixture data by fitting the experimental spectra of mixtures with known concentrations. The calculated references were used in conjunction with Equation 4.1 and Equation 4.2 to predict spectra with gravimetrically measured

concentrations. In Section 4.3.4 and Section 4.3.5, predicted spectra were used to show deviations from the assumed linear models.

Error Metrics

Four error metrics were used to quantify the quantitative accuracy of the spectra-to-composition PLSR models. These are root mean squared error (RMSE), mean absolute error (MAE), the one-sided 95% confidence interval ($\mathrm{CI}_{95\%}$), and mean percent error (MPE). The metrics are defined below in Equation 4.3 – Equation 4.6 for single species:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4.3}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4.4}$$

$$\mathrm{CI}_{95\%} = \frac{z\,s}{\sqrt{N}} \tag{4.5}$$

$$\mathrm{MPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{4.6}$$

where y_i is the true concentration in experiment i, $\hat{y}_i$ is the predicted concentration in experiment i, N is the total number of experiments, s is the standard deviation of measured concentrations, and z is the confidence level value (±1.96 for 95%).

4.3 Results and Discussion

4.3.1 Solution Measurements with Probes

The analysis of the studied slurry system is enabled by the complementary capabilities of the Raman probe and the ATR-FTIR probe. Figure 4.3 demonstrates a ternary system consisting of soluble sodium nitrate, insoluble silica, and 3 m NaOH solution. Response profiles can be seen for Raman (Figure 4.3a) and the ATR-FTIR (Figure 4.3b) probes when

the concentration of suspended solids is increased. Notably, the ATR-FTIR instrument can detect soluble NO_3^- anions with no apparent dependence on solids concentration in the solution (Figure 4.3b). This result matches other published research indicating that the infrared laser does not appreciably contact suspended solids given the shallow penetration depth of the ATR mode of operation [53, 127]. The Raman probe, however, is in a ball probe configuration with a sapphire lens and has a path length approximately two orders of magnitude greater than a probe in ATR configuration [115]. Because of the increased path length, the Raman probe may be affected by the optical density of the slurry. In Figure 4.3a and Figure 4.3c, the Raman probe shows a reduced NO_3^- signal intensity with the increase in solids concentration, suggesting that Raman spectroscopy may not be effective at measuring the solution phase at high solids concentrations. However, this does not preclude Raman spectroscopy from providing information about the solid phase of the slurry, which is not provided by probes in ATR configuration.

4.3.2 ATR-FTIR and Raman Spectroscopy Analysis of Individual GFC Components Dispersed in Alkaline Media

The components of GFCs were grouped into three categories based on their solubility in basic solution (results shown in the Supplemental Information of Prasad et al. [19]): insoluble species (silica and other silicates such as kyanite, wollastonite, olivine, and zircon), metal oxides (hematite, rutile, tin oxide, zinc oxide), and soluble species (vanadium pentoxide, boric acid, and lithium carbonate). Solubilities were estimated using ICP over 10 days [19]. Peaks of each component are identified via reference spectra in Chapter E.

4.3.3 Quantification of Slurries

ATR-FTIR was used to quantify the concentration of dissolved molecular species in slurries comprising GFCs in nuclear-waste simulants. Based on the system studied, the most abundant (and therefore process-relevant) soluble species were quantified with ATR-

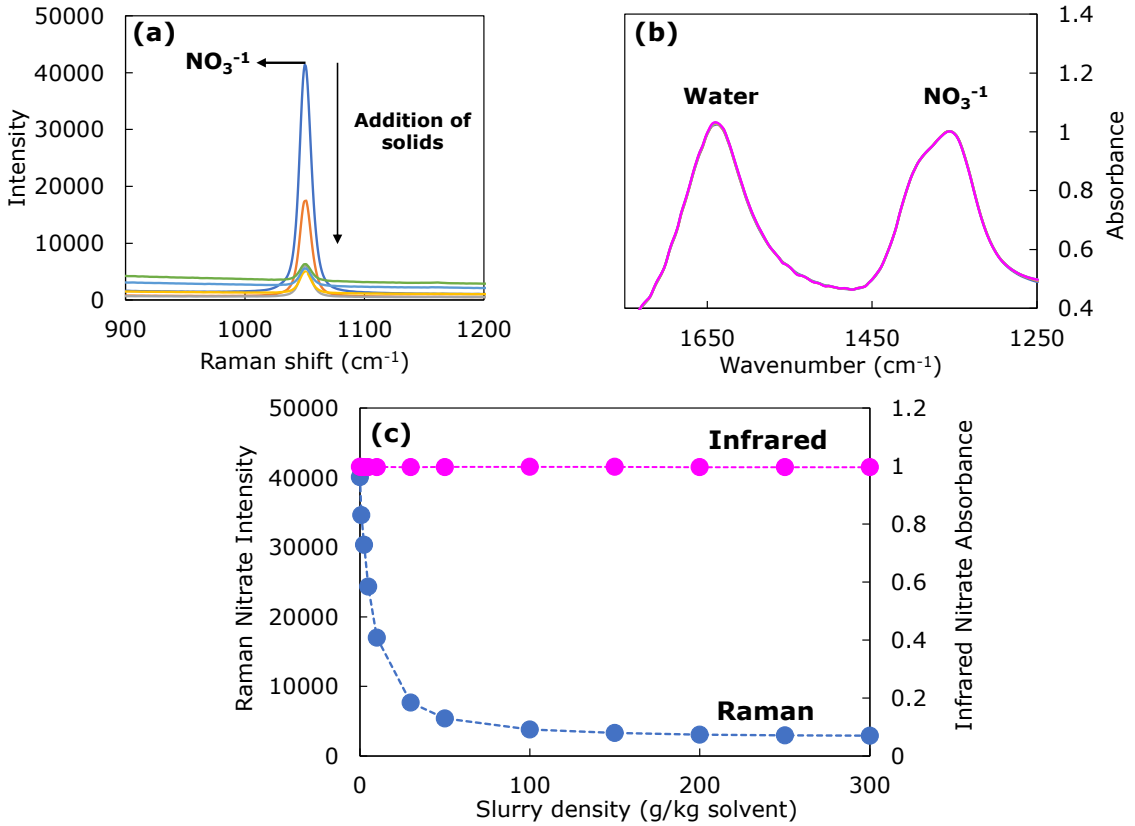


Figure 4.3: Spectra for **a)** Raman and **b)** ATR-FTIR spectroscopy at different concentrations of suspended solid particles and **c)** attenuation of the nitrate peak with increasing solids concentration.

FTIR [21, 117]: NO_3^-, NO_2^-, CO_3^{2-}, and SO_4^{3-}. In addition, borate ($B(OH)_4^-$) was chosen to quantify in the solution phase based on boric acid having high solubility and the distinguishable FTIR peak intensity shown in Figure E.1a. Overlapping spectra of all 48 ATR-FTIR experiments are shown in Figure 4.4a, along with the peaks of the soluble species quantified. Two of the studied soluble species were contributed from components in the GFCs: lithium carbonate (yielding a soluble carbonate anion) and boric acid (yielding a soluble borate anion). The dissolution and dissociation of lithium carbonate implies that the carbonate anion has two sources in the studied slurries: sodium carbonate (from waste simulants) and lithium carbonate (from solid GFCs). Lithium carbonate dissolution was approximated as 0.304 m (pure lithium carbonate solubility in a 3 m sodium hydroxide solution at 25 °C) to calibrate the PLSR model with gravimetric measurements of total

carbonate concentration. The carbonate dissolution model resulted in improved carbonate quantification as demonstrated in the following section; the carbonate dissolution model was used to quantify carbonate in this work.

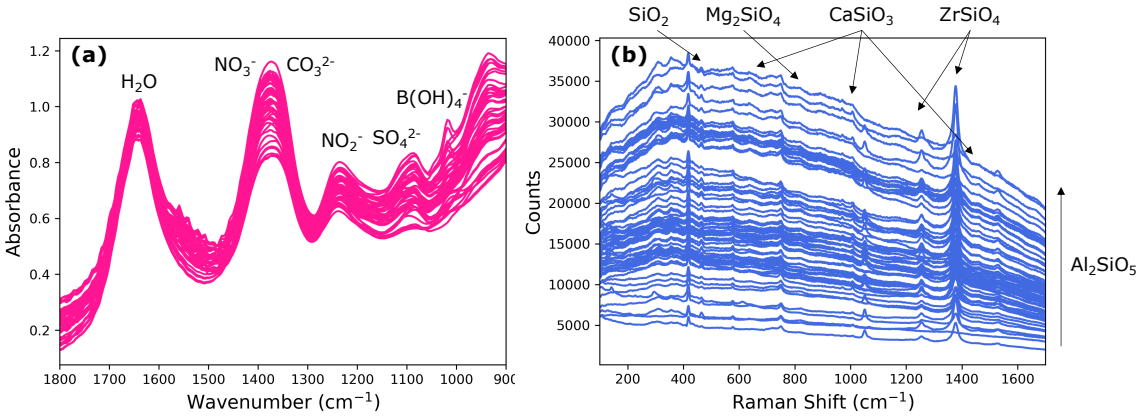


Figure 4.4: Overlapped **a)** ATR-FTIR, and **b)** Raman spectra before separating into training and testing sets.

Carbonate Quantification

The carbonate anion has two sources in the studied slurries: sodium carbonate (from waste simulants) and lithium carbonate (from solid GFCs) as indicated by Equation 4.7 – 4.9.

$$[CO_3^{2-}]_{aq} = [Na_2CO_3]_{dissociated} + [Li_2CO_3]_{dissociated} \tag{4.7}$$

$$[Na_2CO_3]_{dissociated} = [Na_2CO_3]_{added} \tag{4.8}$$

$$[Li_2CO_3]_{dissociated} \neq [Li_2CO_3]_{added} \tag{4.9}$$

A single peak, corresponding to CO_3^{2-}, was shared by both components in FTIR. Experiments confirmed that CO_3^{2-} anions in solution contributed from either salt were identical. Given only ATR-FTIR spectra as the model input, predictions cannot be made that differ-

entiate the contributions of sodium carbonate and lithium carbonate: only their sum total. However, sodium carbonate and lithium carbonate had different solubilities at the conditions tested. In the studied system, the experimental solubility of lithium carbonate was exceeded, limiting a gravimetric calibration procedure.

Because the PLSR model calibration required knowledge of the total carbonate dissolved, a solubility model was used to predict the amount of dissolved lithium carbonate (as observed by our ATR-FTIR probe). In lieu of a more complete dissolution model (accounting for all present species), lithium carbonate solubility was estimated by a ternary system comprising sodium carbonate, lithium carbonate, and 3 m sodium hydroxide solution (NaOH and H_2O). However, the dissolution of lithium carbonate did not have a significant trend in the process-relevant ranges of sodium carbonate (0.1–1 m). Therefore, a constant lithium carbonate solubility of 0.304 m (the pure component solubility of lithium carbonate in a 3 m sodium hydroxide solution at 25 °C) was applied to training and testing data for the PLSR model where $[Li_2CO_3\ ]_{added} > [Li_2CO_3\ ]_{soluble}$. This resulted in the application of the solubility model substitution in 17 of 48 FTIR samples.

The effect of the solubility model was tested in Figure 4.5. Use of the solubility model (Figure 4.5a) improves the PLSR model performance over quantification lacking a solubility model (Figure 4.5b) in terms of both R^2 and visual fit. This indicates that the constant solubility model better matches the data (Figure 4.5a) than an assumption of complete dissolution (Figure 4.5b).

Raman spectroscopy was used to quantify the concentration of suspended solids in slurries of GFCs in simulant solution. Solid silicate species: silica, kyanite, olivine, wollastonite, and zircon were studied due to their abundance in the solid GFC mixtures compared to other solid compounds and their limited dissolution. Overlapping Raman spectra of all 66 unique experiments at different solids concentrations are shown in Figure 4.4b.

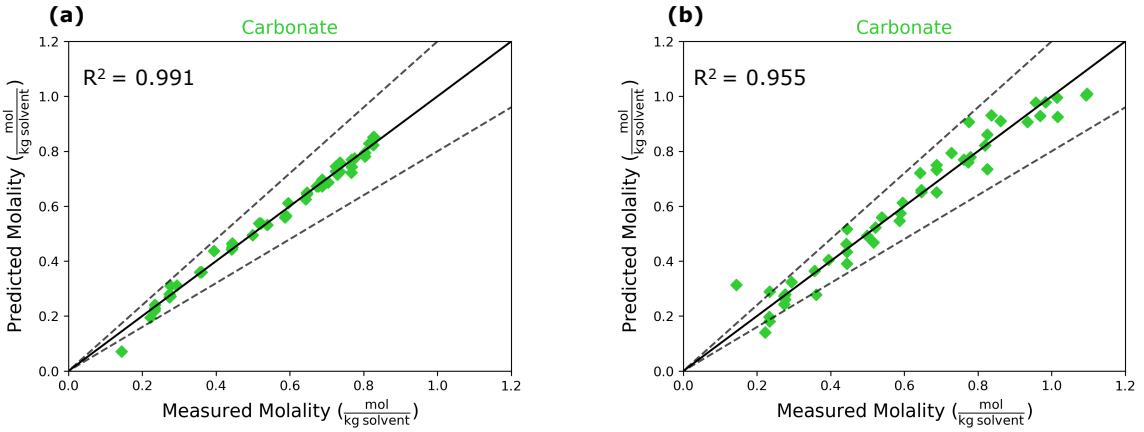


Figure 4.5: **a)** Parity plot showing quantification of carbonate with solubility model applied to concentration data, and **b)** parity plot showing quantification of carbonate with no solubility model applied.

4.3.4 Solution Phase Quantification with ATR-FTIR Spectroscopy

Analysis of the FTIR spectra was performed by comparing the measured spectra to predictions from the Beer-Lambert Law (Equation 4.2). In Figure 4.6a and Figure 4.6b, two measured experimental spectra (shown in red) were fit with ICLS (least squares fit shown in blue). Based on the known solution concentrations and fit references, application of the Beer-Lambert Law predicted mixture spectra (shown in yellow). Table 4.4 quantified the differences in the spectra that may be difficult to observe based on the spectra alone (Figure 4.6).

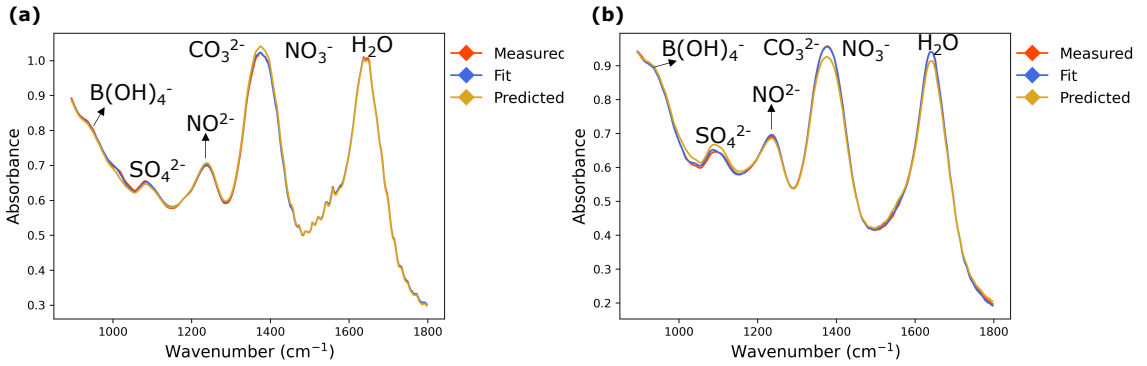


Figure 4.6: Two different ATR-FTIR experiments **a)** and **b)** fit using ICLS and predicted spectra using gravimetrically measured masses.

From Figure 4.6a, the predicted spectrum (based on known masses added) over-predicted the measured spectrum at the nitrate/carbonate peak (1400 cm^{-1}), suggesting that the measured spectrum had a negative deviation from the Beer-Lambert Law. A negative deviation was reflected in the quantification of the measured spectrum of Figure 4.6a in Table 4.4, as both nitrate and carbonate were underpredicted based on the measured spectra. Another deviation from the Beer-Lambert Law was observed in Figure 4.6b. The predicted spectrum overpredicted the measured spectrum at the sulfate peak (1100 cm^{-1}), suggesting the measured spectrum had a negative deviation from the Beer-Lambert Law. Table 4.4, again, shows that a PLSR model underpredicted sulfate based on the experimental spectrum in Figure 4.6b. The size of the deviations shown in the two experiments in Table 4.4, with a mean deviation of ±0.114 mol/kg solvent, showed general agreement between the Beer-Lambert Law and measured ATR-FTIR spectra in slurry conditions.

Table 4.4: Predicted concentrations in mol/kg solvent from ATR-FTIR spectra from Figure 4.6 using a PLSR model

	Nitrate	Nitrite	Carbonate	Sulfate	Borate
Fig. 5a Predicted	0.863	0.711	0.517	0.025	0.305
Fig. 5a Gravimetric	0.883	0.719	0.538	0.026	0.304
Fig. 5b Predicted	0.877	0.827	0.560	0.073	0.862
Fig. 5b Gravimetric	0.879	0.818	0.522	0.078	0.852

The quantification of soluble anions is shown in Figure 4.7, with each data point having a unique concentration of dissolved analytes. Bounds of ±20% were motivated by measurement accuracy specifications of the WTP process control data quality objectives [128] and are shown in Figure 4.7 by dashed lines for reference. Nitrate, nitrite, and borate were quantified most accurately with prediction R^2 values over 0.995 for the tested concentration ranges. Notably, nitrate and nitrite did not exceed a ±20% limit over process-relevant concentrations. Carbonate, sulfate, and borate exceeded the ±20% bound, primarily at lower concentrations. One sulfate measurement produced a particularly poor prediction (shown by the sulfate prediction at 0 mol/kg solvent). This particular experiment had near the

minimum amount of both boric acid and sulfate. Both boric acid and sulfate being at low concentrations may have resulted in the poor sulfate quantification at that point, particularly since sulfate and boric acid have overlapping peaks in that wavenumber range and sulfate has a relatively large limit of detection compared to its tested concentrations. Four accuracy metrics — MAE, RMSE, $CI_{95\%}$, and MPE — are listed in Table 4.5.

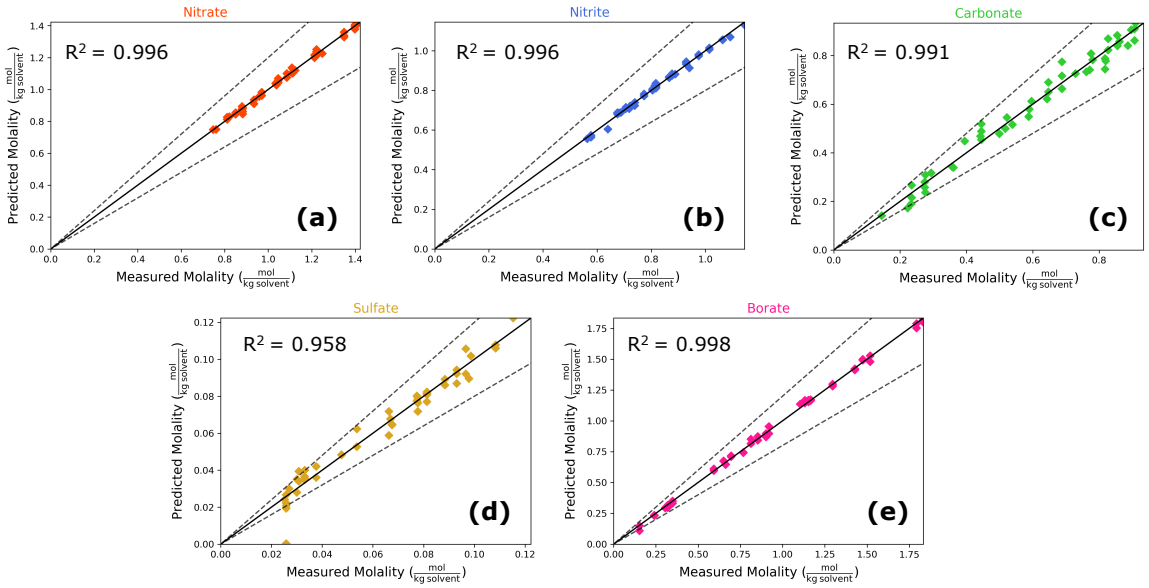


Figure 4.7: Parity plots (showing ±20% bounds) for soluble anion quantification in the simulant slurry using ATR-FTIR spectra input into a PLSR model with 15 latent variables for **a)** nitrate, **b)** nitrite, **c)** carbonate, **d)** sulfate, and **e)** borate.

Prediction residuals are shown in Figure 4.8. Due to the comparatively little sulfate included in these waste simulants, low molality quantification may be hindered by detection limits of sulfate with our ATR-FTIR probe. The calculated limit of detection (LOD) is 0.01 mol/kg solvent using the method described by Harris [129] with a sulfate reference and 12 repeated measurements. From Figure 4.8, it can be seen that the samples with lower sulfate concentration were close to the LOD. In addition, the LOD may have been further impacted by the overlapping borate peak in all samples, since borate was not included when calculating LOD. For this reason, signal-to-noise ratio (SNR) may become more important for low-concentration analytes that have overlapping spectral signatures, such as sulfate.

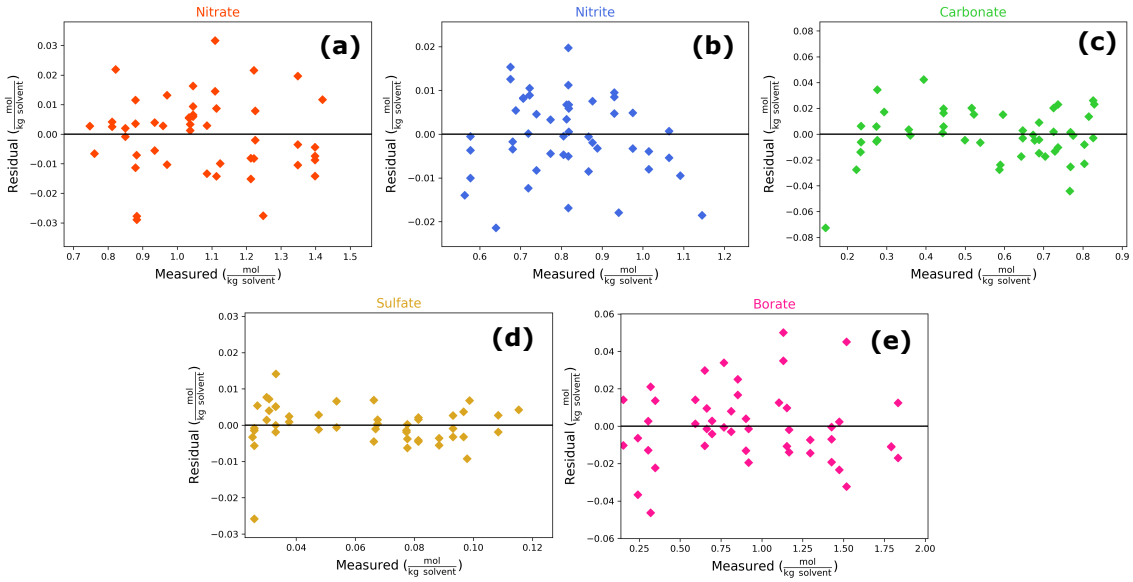


Figure 4.8: Residual plots of soluble anion quantification using ATR-FTIR spectra input into a PLSR model with 15 latent variables for **a)** nitrate, **b)** nitrite, **c)** carbonate, **d)** sulfate, and **e)** borate.

Table 4.5: Anion quantification accuracy with ATR-FTIR.

Metric	Nitrate	Nitrite	Carbonate	Sulfate	Borate
Mean Absolute Error (mol/kg solvent)	0.0101	0.0074	0.0148	0.0040	0.0150
Root Mean Squared Error (mol/kg solvent)	0.0127	0.0091	0.0201	0.0058	0.0196
95% Confidence Interval (mol/kg solvent)	0.0277	0.0183	0.0362	0.0087	0.0422
Mean Percent Error (%)	0.96	0.94	3.78	9.37	2.54

4.3.5 Solid Phase Quantification with Raman Spectroscopy

Analysis of the Raman spectra was performed by comparing measured spectra to spectra that a linearity assumption predicted by Equation 4.1. In Figure 4.9a and Figure 4.9b, two measured experimental spectra (shown in red) were fit with ICLS (least squares fit shown in blue) to compute linear references for kyanite, wollastonite, silica, olivine, and zircon. Predicted spectra (shown in yellow) were calculated from known, gravimetrically-measured slurry concentrations and fit references. Table 4.6 quantifies the differences in spectra that are evident in Figure 4.9.

Figure 4.9 shows that the experimental (measured) spectra and fitted spectra matched

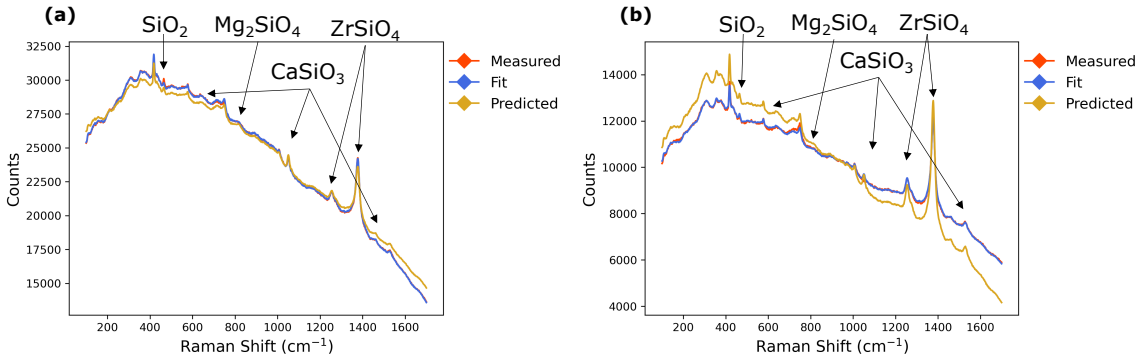


Figure 4.9: Two different Raman Experiments **a)** and **b)** fit using ICLS and predicted using gravimetric measurements.

Table 4.6: Prediction accuracy of Raman spectra from Figure 4.9 using a PLSR model (concentrations in g/kg solvent).

Method	Kyanite	Wollastonite	Olivine	Silica	Zircon
Fig. 7a Predicted	81.58	25.57	33.91	146.61	19.71
Fig. 7a Gravimetric	81.41	16.97	41.67	142.03	18.67
Fig. 7b Predicted	57.96	39.33	18.31	120.80	25.50
Fig. 7b Gravimetric	64.69	41.96	11.09	135.68	32.10

each other closely. In Figure 4.9b, the predicted spectrum did not match the measured spectrum as closely as is the case in Figure 4.9a. The predicted spectrum overpredicted the slope of the background in Figure 4.9b, which correlated closely with kyanite concentrations. This indicated a negative deviation from the Beer-Lambert Law for kyanite in this experiment. The deviation resulted in under-prediction of kyanite in this sample, as shown in Table 4.6. Another example was seen with the sharp peak at 1400 cm^{-1} corresponding to zircon, which was more prominent in the linearly-predicted (yellow curve) spectra than in the experimental spectra (red curve) in Figure 4.9b. This similarly suggested a negative deviation from linearity for zircon, and likewise appeared as a under-prediction in Table 4.6.

Parity plots showing the quantification of silicate solids concentration with ±20% bounds in dashed lines are shown in Figure 4.10. Notably, kyanite and wollastonite showed the best prediction performance with R^2 values of 0.932 and 0.912, respectively. Silica and zircon

had slightly less-accurate quantification with R^2 values of 0.885 and 0.837, respectively. Olivine was predicted with the least accuracy, showing an R^2 value of 0.527. A pattern can be seen in the parity plots of Figure 4.10, where the PLSR model underpredicted solid densities at high solids content. This effect was most prominent in zircon quantification, though it may have been present in the predictions of other species as well. One possible explanation for this apparent patterning may have been the loss of quantification linearity at high solids content, explored in Chapter 5.

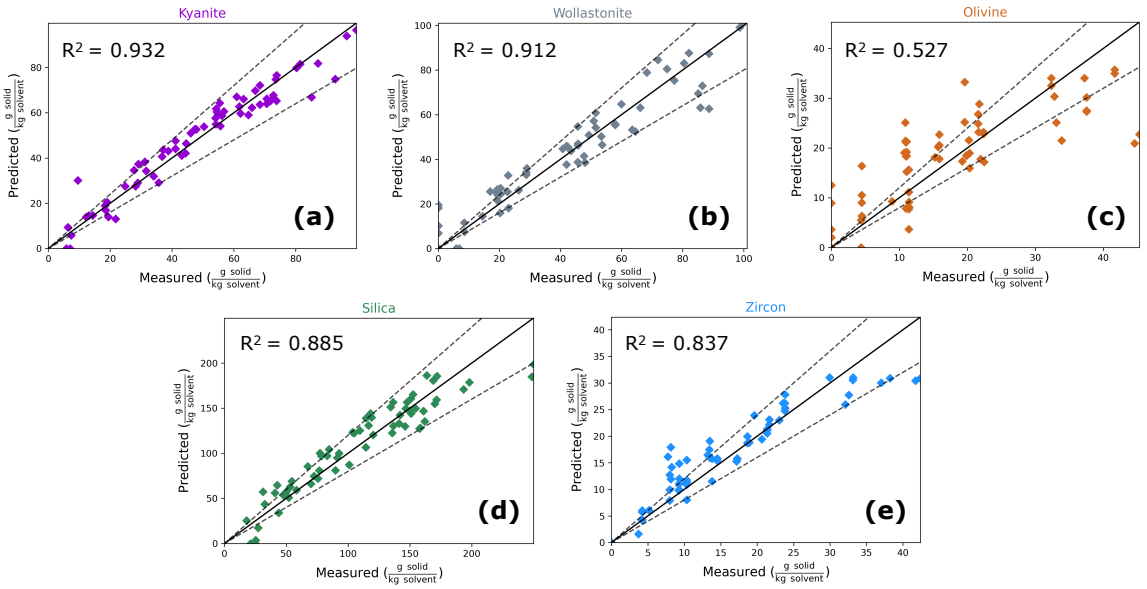


Figure 4.10: Parity plots (showing ± 20% bounds) of major insoluble silicate quantification using Raman spectra input into a PLSR model with 10 latent variables for **a)** kyanite, **b)** wollastonite, **c)** olivine, **d**) silica, and **e)** zircon.

From residual plots for solids (Figure 4.11), patterned residuals can be seen at high solids concentrations for all quantified silicates. This indicated that a PLSR model did not adequately model the relationships between inputs (spectra) and outputs (concentrations) at higher slurry concentrations. Figure 4.11 also suggested the nonlinear behavior may have been component-specific. For example, the zircon residuals appeared to undergo a transition around 25 g solid/kg solvent, where the model began to underpredict solids concentrations. In contrast, kyanite appeared to undergo an analogous transition around 60 g solid/kg solvent. Further analysis was limited by model noise obscuring residual trends.

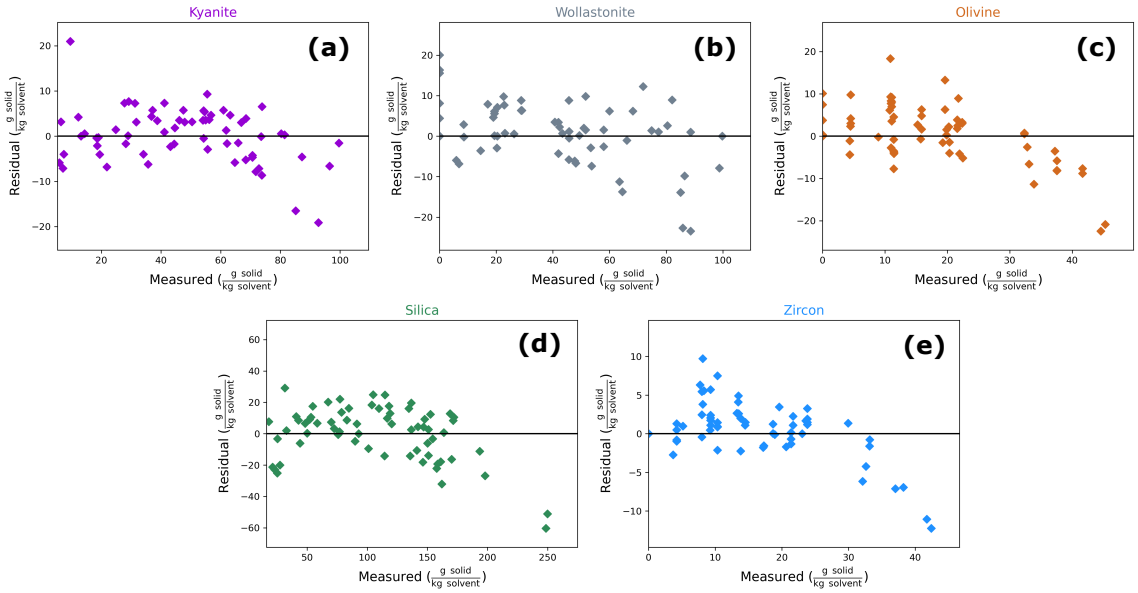


Figure 4.11: Residual plots of insoluble silicate quantification using Raman spectra input into a PLSR model with 10 latent variables for **a)** kyanite, **b)** wollastonite, **c)** olivine, **d)** silica, and **e)** zircon.

The poor quantification of olivine may have been attributed to its low abundance in the slurry and obscured spectral features. Zircon, despite its similarly low slurry concentration, was highly Raman-active in the region studied with a prominent peak at 1378 cm^{-1}. Due to low SNR of Raman in these slurries, olivine quantification may have been improved by including measurements with higher proportions of olivine. Four accuracy metrics — MAE, RMSE, $CI_{95\%}$, and MPE — are listed in Table 4.7. Independence of the quantified chemical species is depicted in Chapter F.

Table 4.7: Solid quantification accuracy with Raman.

Metric	Kyanite	Wollastonite	Olivine	Silica	Zircon
Mean Absolute Error (g/kg solvent)	4.57	5.66	5.20	13.53	2.52
Root Mean Squared Error (g/kg solvent)	6.04	7.88	6.95	17.29	3.68
95% Confidence Interval (g/kg solvent)	9.15	16.18	12.80	28.58	7.40
Mean Percent Error (%)	16.5	16.7	39.4	18.2	21.4

Solid Phase Quantification with Raman Spectroscopy after Savitzky-Golay Filtering

Optionally, the first derivative could be taken of Raman spectra to minimize baseline shifts caused by fluorescence. This was done as a proof-of-concept with a Savitzky-Golay filter with a second-order polynomial, 19 filter points, and a first derivative. For the spectra utilized in this work, 11 filter points or fewer resulted in noisy derivative spectra. The spectra after filtering are shown below in Figure 4.12 and quantification results are shown in Table 4.8. As can be seen in Table 4.8, quantification was improved after Savtizky-Golay filtering for wollastonite, olivine, silica, and zircon. Kyanite quantification, however, was less accurate after filtering. These results show promise for the Savitzky-Golay filter to preprocess noisy and varied Raman spectra for more accurate quantification.

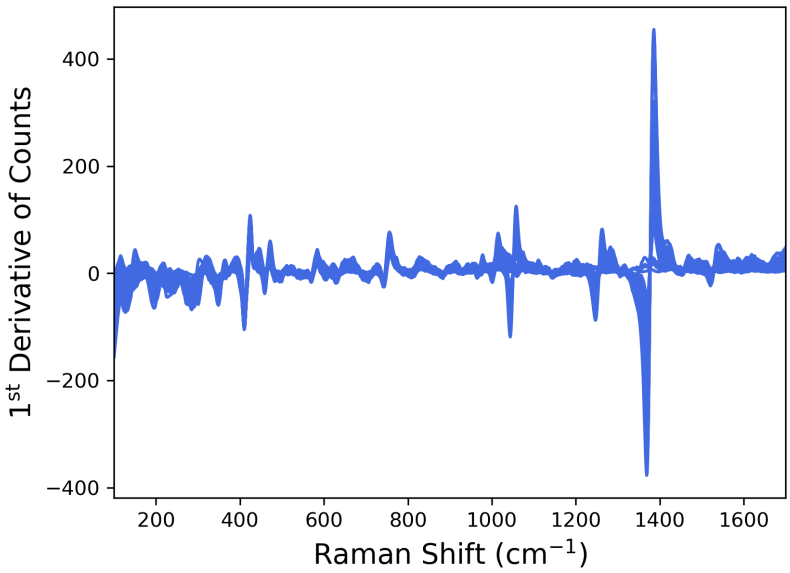


Figure 4.12: All Raman spectra of training and testing data when Savitzky-Golay filter has been applied.

Table 4.8: R^2 values of predictions for quantified insoluble species with and without Savitzky–Golay filtering on Raman spectra.

Concentrations (g/kg solvent)	Kyanite	Wollastonite	Olivine	Silica	Zircon
R^2 (No filtering)	**0.932**	0.912	0.527	0.885	0.837
R^2 (Filtering, 1st derivative)	0.912	**0.916**	**0.770**	**0.905**	**0.863**

Experimental Reproducibility for Raman Spectra

A total of 34 replicate measurements were collected using slurries that were constructed and measured at least 11 (and no more than 15) days after the original data appearing in the main text (Figure 4.4b). As can be seen in Figure 4.13a, there was a baseline offset in some measurements, with the greatest difference in a measurement being 1495 counts (Figure 4.13b). The average difference between the original measurement and its replicate across the 34 measurements was 239.6 counts (or 1.24% difference from original measurements). The difference between replicate measurements also showed high-frequency noise in addition to baseline offset. The replicate data were also quantified via the same PLSR model as used earlier in this section with and without Savitzky-Golay spectra preprocessing, shown in Figure 4.14. Quantification with Savitzky-Golay preprocessing appeared to be more precise since a vertical offset appeared in the predictions of the raw spectra. This result agreed with the offset that can be seen in the replicate spectra. Based on these data, we conclude that both our experimental methodology and our Raman apparatus were robust to within a relatively small baseline offset for experimental replicates.

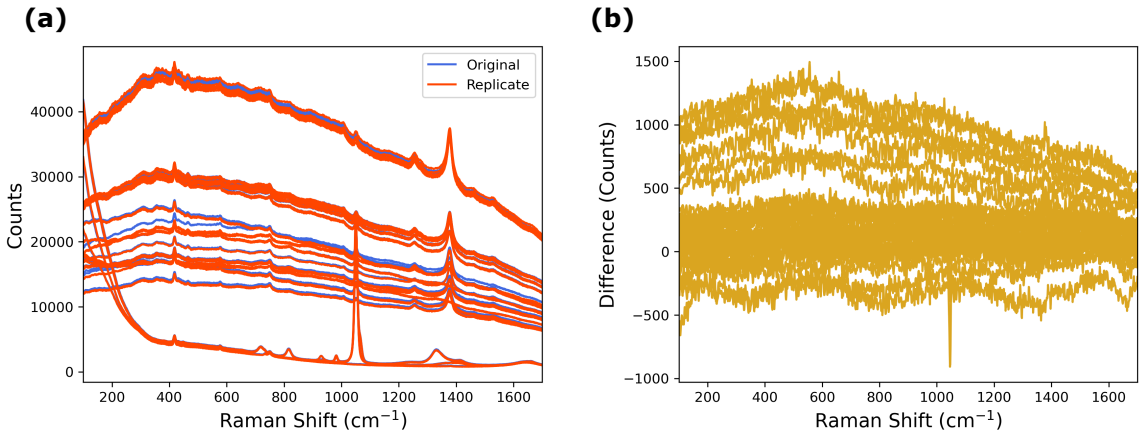


Figure 4.13: **a)** Original spectra and replicate spectra collected for 34 replicate measurements ranging from 0 to 25 wt% solids, and **b)** the difference between original and replicate spectra.

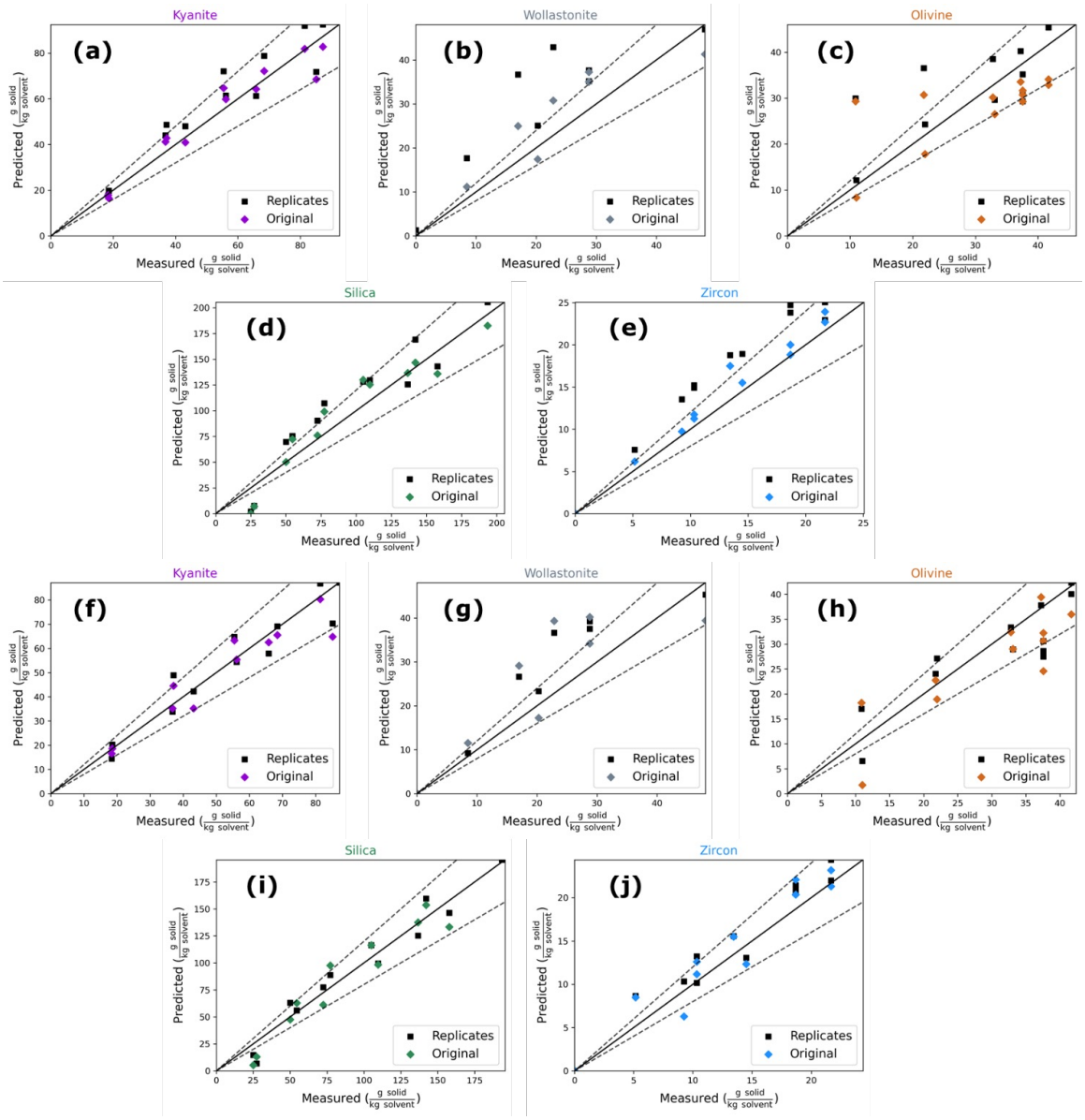


Figure 4.14: Parity plots showing PLSR predictions of original and replicate **(a–e)** raw spectra and **(f–j)** spectra passed through a 19-point Savitzky-Golay filter.

4.4 Conclusion

Establishing accurate models and conceptual feasibility of monitoring dense slurry solutions is important for nuclear-waste management and many other systems. The present work shows that instrumentation employing ATR-FTIR technology can accurately estimate the concentrations of key solutes in the liquid (solution phase) portion of a slurry contain-

ing a high concentration of suspended solids with a mean accuracy of 3.52%. Companion results using Raman spectroscopy facilitate an ability to distinguish and quantify different species suspended in the solid phase with a mean accuracy of 18.21% for the four most abundant and spectroscopically active silicates. While this work demonstrates the feasibility of Raman and ATR-FTIR spectroscopies for monitoring slurries typical of nuclear-waste processing, the wide application breadth of these instruments does not limit the bearing of these results to the highlighted application of this thesis.

The results of this chapter can also be extended. For example, the complex data observed by the Raman spectrometer in this study may be able to be understood better by using complementary information from the ATR-FTIR probe so that known solution-phase peaks can be subtracted from the multiphase Raman spectra. Additionally, the results presented in this chapter are limited to linear methods primarily because of the many features and few data collected, which is limited by the experimental collection procedure; nonlinear models may require more data. Lastly, the suspended insoluble phase of the slurries presented in this chapter maintain a constant particle size distribution; this assumption may not hold true in actual nuclear-waste processing and may affect the measurements observed by the instruments.

CHAPTER 5

NONLINEAR METHODS FOR MEASURING OPTICALLY-DENSE SLURRIES AND ATTENUATING MEDIA WITH RAMAN SPECTROSCOPY

In the previous chapter, Raman spectroscopy and attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy were combined with linear models (partial least squares regression (PLSR)) to quantify dense multicomponent slurries; the data seemed to suggest nonlinearities were present despite linear methods providing satisfactory quantification. Therefore, this chapter investigates physics-based nonlinear models to predict concentration from slurry spectra. *Attenuation* — reduced signal because of the interference of turbid media — is well-reported in the literature. However, there seems to be no accepted trend between concentration and attenuation, especially in multicomponent slurries. This gap is an issue for modeling attenuation behavior.[1]

In this chapter, an analysis is performed using a photon balance on media containing attenuating particles. However, the resulting model has an unknown dependence on concentration with essentially no literature investigating the concentration dependence of attenuation (nonlinearities). Therefore, two simplified models are created: 1) an optically saturated model and 2) a variable pathlength model. The two models rely on 1) using mass fraction of the suspended phase rather than of the entire slurry and 2) transforming spectra and concentration data using the natural logarithm, respectively. The two models are demonstrated on a Raman spectral data set from a system consisting of suspended solids in solution that mimic conditions expected in nuclear-waste processing at Hanford.

[1] Some of the content and some of the figures in this chapter are reproduced from conference proceedings from Crouse et al. [28]

5.1 Introduction

A Waste Treatment Plant (WTP) is being constructed by the United States Department of Energy in Hanford, Washington as detailed in Section 1.2. Difficulties from suspended solids challenge the implementation of real-time in-line monitoring in slurry streams that may have substantial solid content at Hanford. Raman spectroscopy has been shown to measure both the solid and liquid phases in many systems. However, at high concentrations of suspended solids, Raman spectroscopy becomes less effective at measuring both the liquid and solid phases of slurries (see Chapter 4). Suspended solids tend to scatter and absorb photons, rather than acting as an optically transparent medium that is typically assumed with linear models. Two models are proposed in this chapter that may model photon transport in slurries and enable more accurate quantification of suspended solids in a studied slurry system. These are: 1) an optically saturated model and 2) a variable pathlength model. The optically saturated model makes the assumption that a fixed number of Raman-scattered photons are detected at high densities of suspended solids. Under this scenario, relative concentrations ($\frac{C_i}{\sum_{j=1}^{n} C_j}$) better correlate to spectral features than absolute concentrations (C_i). The second model, the variable pathlength model, is motivated under the assumption that the Raman laser will not penetrate as deeply into the solution when an optically dense sample (consisting of suspended particles) is probed, thus creating a laser pathlength that has a power-law dependence on the concentration of suspended particles. This scenario will violate the common assumption of constant interrogation volume that is required for a linear relationship between Raman spectral intensity and solids concentration. These ideas are developed and tested in the subsequent sections.

5.2 Method and Materials

5.2.1 Experimental Procedure

This system is identical to the data used in Chapter 4, but is summarized here again. The system is based on a variation of the 5.6 M Na low-activity waste (LAW) pretreatment system simulant (solution) [21] combined with simulated glass-forming chemical (GFC) recipes (mostly metal oxides and silicates) [112]. This combined system serves as a high-level waste (HLW) simulant, since the suspended solids mimic particles in HLW processing and HLW will have a similar solution-phase composition with LAW. The solution-phase simulants are comprised of water, sodium hydroxide, and seven sodium salts: nitrate, nitrite, carbonate, sulfate, phosphate, oxalate, and acetate. The GFC compositions were formulated by M.J. Siegfried and M.E. Stone and were provided by the Savannah River National Laboratory (SRNL) [112]. Glass-forming chemicals (GFCs) are comprised of five silicates: silica, kyanite, wollastonite, olivine, and zircon; five oxides: hematite, tin oxide, rutile, vanadium pentoxide, and zinc oxide; and three additional species: boric acid, lithium carbonate, and glucose. All experiments were conducted at 3 m NaOH to simulate the basic conditions expected at the Hanford WTP. At this molality of sodium hydroxide, the pH remained above 13 for all experiments, which was verified with an in-situ pH probe. All experiments were temperature-controlled to 25 °C and stirred at 400 rpm to maintain suspension of solids inside a 250 mL Teflon reactor.

5.2.2 Instrumentation

Raman spectra were recorded with a Mettler Toledo ReactRaman 785 using a 785 nm laser at 300 mW power, 1 s exposure time, 10 averaged spectra, and a spectral resolution of 6 cm^{-1}. The probe tip is constructed of sapphire, leading to the three peaks present in every measurement located at: 419 cm^{-1}, 577 cm^{-1}, and 751 cm^{-1}. The probe tip was a "ball probe" design [116]. This setup allows solutions and slurries to be measured without

the requirement of a flow cell.

5.2.3 Partial Least Squares Regression

This chapter uses a linear model, PLSR, for quantification after linearizing data using the optically saturated model or variable pathlength model developed. PLSR is well-developed as a viable model in nuclear-waste spectroscopic monitoring literature and is described in Chapter B. No data scaling (other than noted to linearize data) was applied prior to quantification in this chapter.

5.3 Theory

In this section, prior literature on attenuation is used to develop a general model for attenuation in slurries. This model assumes the slurry is a continuum, and therefore treats the slurry medium as having constant attenuating properties everywhere. A conceptual model of attenuation behavior is presented in Figure 5.1.

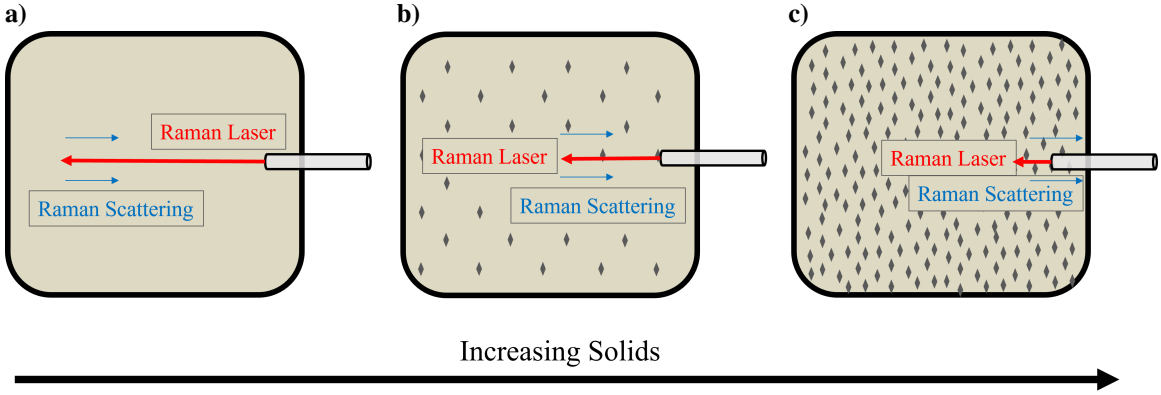


Figure 5.1: Visualization of attenuation as the density of solid particles increase in a slurry. Solid density increases as viewed from **(a)** to **(c)** while the effective laser pathlength decreases.

5.3.1 Derivation of Optical Attenuation Model

Attenuation has been introduced elsewhere and is usually taken as the sum of photon absorption and scattering (elastic scattering rather than Raman scattering) [130, 131, 90].

These two terms are often combined into a single attenuation parameter.

$$\alpha_{absorbed} + \alpha_{scattered} = \alpha \tag{5.1}$$

From McCreery, the equation for determining the intensity of a Raman signal in optically transparent solutions is [90]:

$$S(e^-) = CP_0\beta K A_D \Omega_D T Q t_s \tag{5.2}$$

where $S(e^-)$ is the detector signal in electrons, C is the density of molecules in $\frac{\text{molecules}}{\text{cm}^3}$, P_0 is incident laser photon flux in $\frac{\text{photons}}{\text{cm}^2\text{s}}$, β is the differential Raman cross-section in $\frac{\text{cm}^2}{\text{molecule}\times\text{sr}}$, K is a geometric factor (pathlength) in cm, A_D is sample area monitored by the spectrometer in cm^2, Ω_D is solid angle observed by the spectrometer in sr, T is a transmission factor in $\frac{\text{photon}}{\text{photon}}$ (unitless), Q is quantum efficiency of the detector in $\frac{\text{e}^-}{\text{photon}}$, and t_s is integration time in s.

For any Raman-scattered photon in a turbid solution, the photon may be absorbed or scattered before Raman scattering (α_L) or after Raman scattering (α_R). Below, a photon balance is performed in the z-direction in a box with width W and height H for photon flux in the z-direction, P.

$$\text{In} - \text{Out} - \text{Laser Attenuation} - \text{Raman Attenutation} = 0 \tag{5.3}$$

$$\alpha_L = \alpha_{L,\ absorbed} + \alpha_{L,\ scattered} \quad , \quad \alpha_R = \alpha_{R,\ absorbed} + \alpha_{R,\ scattered} \tag{5.4}$$

$$WHP|_{z=z+\Delta z} - WHP|_{z=z} = -WH\Delta z(\alpha_L + \alpha_R)P \tag{5.5}$$

$$\lim_{\Delta z \to 0} \left(\frac{P|_{z=z+\Delta z} - P|_{z=z}}{\Delta z} \right) = \frac{dP}{dz} = -(\alpha_L + \alpha_R)P = -\alpha P \tag{5.6}$$

$$\int \frac{dP}{P} = \int -\alpha dz \tag{5.7}$$

$$P(z) = \mathcal{B}e^{-\alpha z} \quad , \quad P(0) = P_0 \tag{5.8}$$

$$P(z) = P_0 e^{-\alpha z} \tag{5.9}$$

where P is the photon flux through any cross section $W \times H$ at depth z from the laser, P_0 is the photon flux from the tip of the laser, and $\mathcal{B}$ is an integration constant. The photon shell balance is shown in Figure 5.2.

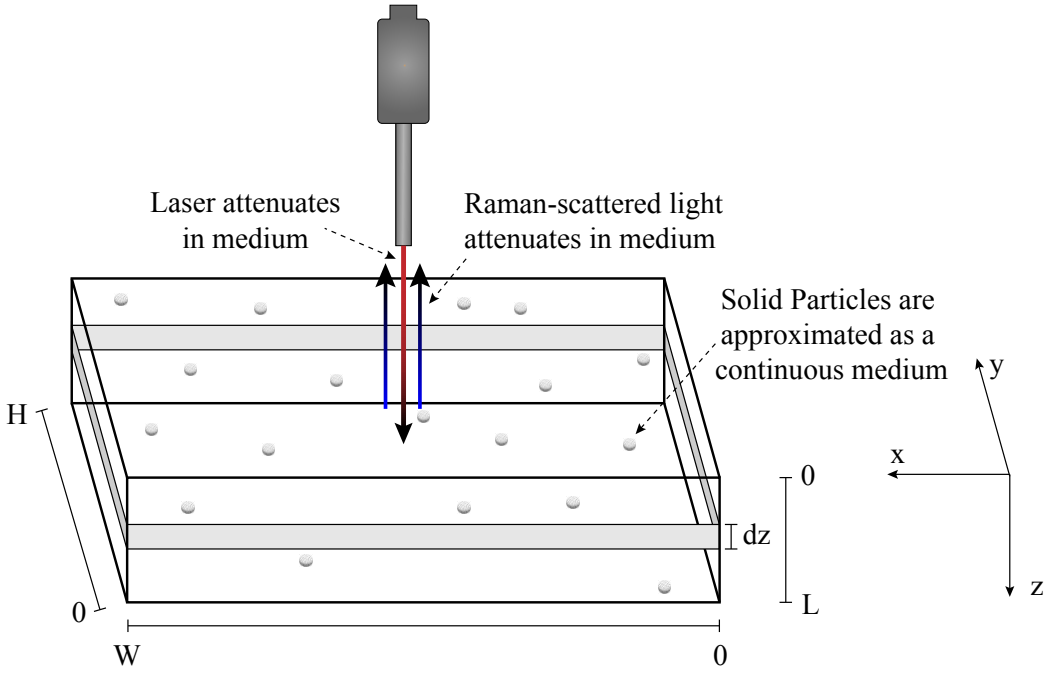


Figure 5.2: Photon balance for a Raman spectrometer. Both the Raman laser and the Raman-scattered light are attenuated as a function of depth in the medium.

Attenuation in Observed Signal

Attenuation can be inserted into the Raman quantification equation (Equation 5.2) by introducing effective pathlength, K, as is shown by McCreery [90]:

$$S(e^-) = \int_0^L P_0 C \beta A_D \Omega_D T Q t_s e^{-\alpha z} dz \tag{5.10}$$

$$K(\lambda) = \int_0^L e^{-\alpha z} dz = \frac{(1 - e^{-\alpha(\lambda)L})}{\alpha(\lambda)} \tag{5.11}$$

In Equation 5.10, Raman-scattered photons per unit area are integrated over the pathlength, 0 to L, in the direction of the laser. The expression can be simplified by combining the pathlength-dependent terms (attenuation) into a single variable representing the pathlength, K (units of cm). Then, the integral can be evaluated to give effective pathlength in attenuating media, as shown in Equation 5.11. This pathlength, K, can be verified by computing the limits in the case of infinite (Equation 5.12) and no attenuation (Equation 5.13) with L'Hôpital's rule.

$$\lim_{\alpha(\lambda)\to\infty} K(\lambda) = \lim_{\alpha(\lambda)\to\infty} \frac{(1 - e^{-\alpha(\lambda)L})}{\alpha(\lambda)} = 0 \tag{5.12}$$

$$\lim_{\alpha(\lambda)\to 0} K(\lambda) = \lim_{\alpha(\lambda)\to 0} \frac{(1 - e^{-\alpha(\lambda)L})}{\alpha(\lambda)} \xrightarrow[w.r.t.\ \alpha(\lambda)]{\text{L'Hôpital}} \lim_{\alpha(\lambda)\to 0} L e^{-\alpha(\lambda)L} = L \tag{5.13}$$

As the model approaches infinite attenuation, the predicted pathlength approaches zero. The opposite scenario, no attenuation, predicts the pathlength to approach L. Both of these results match intuition.

The product of transparent pathlength and attenuation ($\alpha(\lambda)L$) that appears in Equation 5.11 is a dimensionless measure of attenuation; a large product indicates that few photons travel the full pathlength (L) of the uninhibited laser. Additionally, if $\alpha(\lambda)L >> 1$

(i.e. the system is strongly attenuating), then an assumption can be made that $(e^{-\alpha(\lambda)L}) \approx 0$ in Equation 5.11 [90]. The assumption of strongly attenuating media will made for now; this gives the following two equations:

$$K(\lambda) = \frac{1}{\alpha(\lambda)} \tag{5.14}$$

$$S(e^{-}) = P_0 K C \beta A_D \Omega_D T Q t_s \tag{5.15}$$

Complete Attenuation Model

Equation 5.16 show the signal reaching the Raman detector as a function of the density of particles, C_i, and an unknown attenuation parameter, $\alpha(\lambda)$; all other variables are assumed to not be functions of concentration.

$$S(\lambda) = \frac{\sum_{i=1}^{n} w_i C_i}{\alpha(\lambda)}, \qquad w_i = P_D \beta(\lambda) A_D \Omega_D T Q(\lambda) t_s \tag{5.16}$$

The key point of this analysis is that the denominator term, $\alpha(\lambda)$, is expected to have a concentration dependence: $\alpha(\lambda) = f(\mathbf{C}, \lambda; \boldsymbol{\theta})$ where $\mathbf{C}$ is a $1 \times n$ vector of component densities and $\boldsymbol{\theta}$ is a vector of model parameters. However, the form of concentration dependence is unknown and is not part of McCreery's analysis [90]. The contribution of this chapter is the development of simplifying assumptions that can be made based on physical understanding of the system that account for concentration-dependent attenuation. Additionally, experimental data collected of slurry samples may provide insight into which of these models (if any) may be valid for modeling spectra of slurry systems. Equation 5.16 will be the starting point for both the optical saturation model and variable pathlength model. The following assumptions are made so far:

- A continuum assumption is made in the analysis; individual particles are approx-

imated as contributing to bulk properties of the slurry, such as optical attenuation $\alpha(\lambda)$.

- The system is strongly attenuating, $\alpha(\lambda)L >> 1$.
- Photons may be scattered or absorbed as part of the excitation laser, $\alpha_L(\lambda)$ or as part of Raman-scattered photons, $\alpha_R(\lambda)$. These effects can be combined as a combined attenuation term as shown elsewhere, $\alpha_L(\lambda) + \alpha_R(\lambda) = \alpha(\lambda)$ [90].
- The variables captured by w in Equation 5.16 do not vary between sample measurements.

5.3.2 Optical Saturation Model

Assume that a constant number of photons reach the detector; this can be modeled by assuming attenuation, $\alpha(\lambda)$, has a linear dependence on concentration. In this case, increasing the concentration of all species by a proportional amount should not change the acquired signal; the total returned signal (photons) is constant. However, changing the proportion of solid particles would change the detected signal. Near-constant signal was observed in the previous chapter, Chapter 4, when increasing the density of solid particles past a certain point had negligible changes in measured intensity. Physically, this may occur since the interrogated volume decreases with a proportional increase in concentration; i.e. all of the Raman laser is already contacting solid particles and so increased concentration will move the mean pathlength shorter, but the same number of solid particles (surface area) will be detected. Therefore, it is expected that mass fraction rather than density would be responsible for the observed signal.

Assume that attenuation parameter depends linearly on concentration, as described by Baker and Lavelle [132], and that attenuation effects are additive (which follows from the assumption of linear attenuation in Equation 6.3). Summing the attenuation of multiple

species attenuation with a linear dependence on concentration, where there are n attenuating species:

$$\alpha(\lambda) = f(\mathbf{C}, \lambda; \boldsymbol{\theta}) = \sum_{i=1}^{n} \theta_i(\lambda) C_i \tag{5.17}$$

where $\theta_i(\lambda)$ is a wavelength-dependent parameter that depends on the optical properties of the material being interrogated. If the assumption is made that particles have roughly equivalent attenuation per unit mass (as may be the case in a geometric "shadowing" of particles), $\theta_1 \approx \theta_2 \approx ... \approx \theta_n \approx \theta$, θ can be isolated as its own variable and mass fraction (Y_i)[2] rather than density (C_i) is related to signal:

$$S(\lambda) = \sum_{i=1}^{n} w_i^{\star}(\lambda) Y_i \,, \qquad w_i^{\star}(\lambda) = P_D \beta(\lambda) A_D \Omega_D T Q(\lambda) t_s \theta^{-1} \tag{5.18}$$

This linear model is a function of mass fraction, Y_i, which has units $\frac{g\ component\ i}{g\ total\ solids}$. Thus far, Equation 5.18 relies on two new assumptions:

- Attenuation varies linearly with concentration.
- Attenuation is species-independent: $\theta_1 = \theta_2 = ... = \theta_n = \theta$.

5.3.3 Variable Pathlength Model

A variable pathlength model was also investigated that assumed an empirical power law for the concentration dependence of attenuation. A power-law form is motivated by analogous power-laws derived for monosized, spherical particles [17], [18]. Shavishankar et al. propose a form for suspended particles where inter-particle distance is a function of particle volume fraction to the $-\frac{1}{3}$ power. In the currently studied system, the particles

[2]The denominator in Equation 5.16 becomes $\theta \sum_{i=1}^{n} C_i$, and so each component's density, C_i, in the numerator of Equation 5.16 is divided by the total density to yield mass fraction, Y_i.

are non-spherical and polysized. An identical relationship (power-law) is assumed here. However, the assumptions of developed models (monosized, spherical particles) do not apply to the currently studied system, so the parameters will be determined from the data. A power law for the attenuation coefficients are shown as

$$\alpha(\lambda) = \sum_{i=1}^{n} \theta_i(\lambda) C_i^{\gamma_i} \tag{5.19}$$

where γ_i is an exponent to be determined from data. This gives the quantification equation (beginning oncemore with Equation 5.16):

$$S(\lambda) = \frac{\sum_{i=1}^{n} w_i(\lambda) C_i}{\sum_{i=1}^{n} \theta_i(\lambda) C_i^{\gamma_i}} \tag{5.20}$$

If the additional assumption is made (motivated in the bullet points below) that a single component is present in the insoluble phase ($n = 1$), Equation 5.20 simplifies to

$$S(\lambda) = \frac{w_i(\lambda) C_i}{\theta_i(\lambda) C_i^{\gamma_i}} = w_i^{\dagger}(\lambda) C_i^{1-n} \tag{5.21}$$

which can be simplified via logarithms to yield a linear equation:

$$\ln(S(\lambda)) = \ln(w_i^{\dagger}(\lambda) C_i^{1-n}) = (1-n)\ln(C_i) + \ln(w_i^{\dagger}) \tag{5.22}$$

Equation 5.22 suggests that a log-log plot may be linear in the case of an attenuating mixture where attenuation has a power-law dependence on concentration. In this section, two additional assumptions are made:

- Attenuation has a power-law relationship with concentration (i.e. $\alpha(\lambda) = \sum_{i=1}^{n} \theta_i(\lambda) C_i^{\gamma_i}$)
- A single component is present in the insoluble phase.
  - While seemingly prohibitive to *multicomponent* slurries, this assumption may

be valid when peak selection or feature selection is performed and so a single component primarily contributes to the signal/attenuation at a particular wavenumber.

5.4 Results and Discussion

The two model structures proposed in Section 5.3.2 and Section 5.3.3 were tested using the slurry dataset collected in Chapter 4.

5.4.1 Optically Saturated Model

Figure 5.3 (bottom row) shows the quantification of the slurry system with an optically saturated (Equation 5.18) quantification model, while Figure 5.3 (top row) shows quantification of the original data (Equation 5.2). The optical saturation model used relative concentrations $\frac{\text{g solid i}}{\text{g total solids}}$ rather than density.[3] Figure 5.3 shows that the optically saturated model performed better than a PLSR model with no data transformation. The R^2 values of the optically saturated model (Figure 5.3, bottom row) were higher than the R^2 values of the untransformed data (Figure 5.3, top row), with the exception of wollastonite. This may suggest that the assumptions made in the optically saturated model were more applicable to quantification in this system than that of linear Raman quantification. In this system, composition appeared to better correlate with Raman spectra than absolute concentrations correlated with Raman spectra. However, direct comparison between the optically saturated model and the untransformed model was limited because the optically saturated model output relative concentration (mass fraction), whereas the untransformed model output absolute concentration. For nuclear-waste processing, other measurements (volume, density, etc.) may allow the use of mass fraction measurements since they appeared to correlate better with observed Raman signal for most species of interest.

[3] $\frac{\text{g solid}}{\text{kg solvent}}$ was used for linear quantification and correlated very closely with density in these slurries; this quantity is typically used rather than density ($\frac{\text{mass}}{\text{volume}}$).

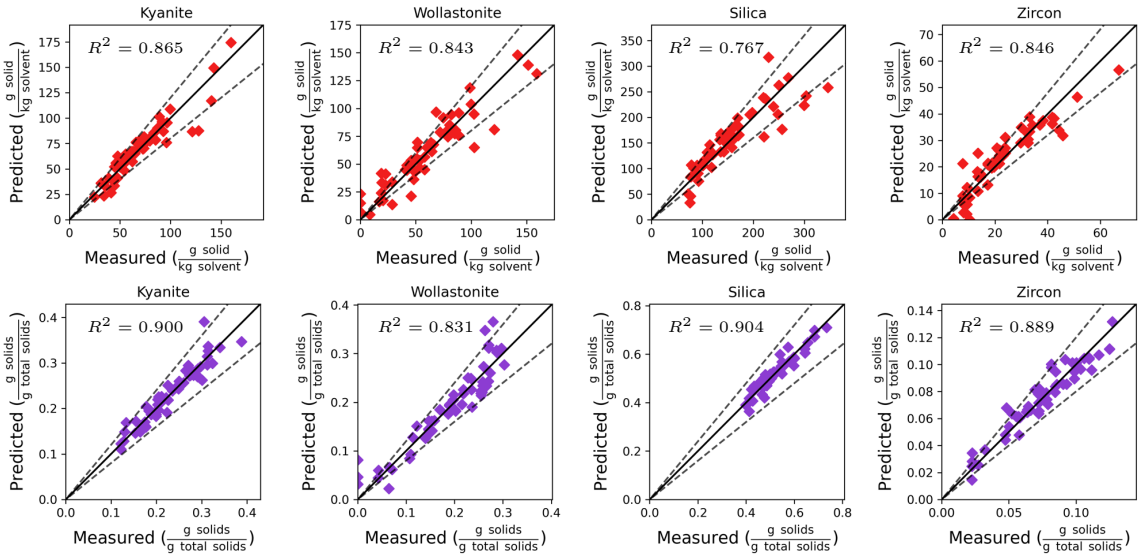


Figure 5.3: Parity plots of PLSR quantification of Raman spectra with **(top row)** original data (absolute concentration) and **(bottom row)** mass-fraction data (relative concentration).

5.4.2 Variable Pathlength Model

The variable pathlength model (Equation 5.22) was implemented by logarithmically transforming spectra and concentrations for model fitting and prediction. Figure 5.4 shows the quantification of four silicates: kyanite, wollastonite, silica, and zircon. Figure 5.4 (top row) shows quantification using the original Raman and concentration data. Figure 5.4 (bottom row) logarithmically transforms both the Raman spectra and the concentrations, representing the variable pathlength model from Equation 5.22. Overall, no significant improvement or detriment to quantification accuracy was observed by logarithmically scaling data prior to quantification. Specifically, quantification was improved for kyanite, while less effective quantification was observed for silica and zircon. Wollastonite had marginal quantification changes between the standard model and the logarithmically transformed model.

Figure 5.5 shows residual scatter as a function of measured concentration for both the standard model (original data) and a variable pathlength model (log-log data). At low measured concentrations of solids, the variance of prediction error decreased when applying a

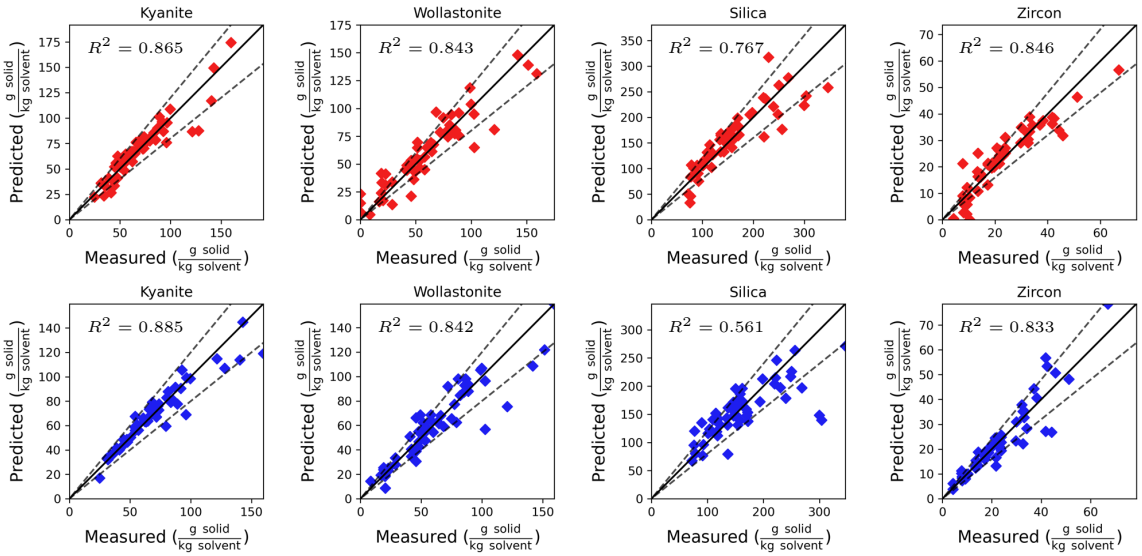


Figure 5.4: Parity plots of PLSR quantification of Raman spectra with **(top row)** original data and **(bottom row)** data after undergoing a log-log transformation.

logarithmic transform to the data. This may suggest the applied variable pathlength model was most applicable in the low-concentration regime, but did not correct for model error at high solids loading. Zircon, quantified using original data (Figure 5.5 (top row)), showed patterned residuals that under-predicted at low concentrations, over-predicted at intermediate concentrations, and under-predicted at high concentrations. As shown by Figure 5.5 (bottom row), a logarithmic transform removed this patterned model bias, but did not improve overall model performance. Silica was notably quantified less accurately after the logarithmic transformation in Figure 5.5 (bottom row).

5.4.3 Analysis

The improved quantitative performance when using the optically saturated model indicated that there may be approximately a fixed number of detected photons at high densities of suspended solids (signal is near-constant with increasing slurry concentrations). While the concentration (density) of particles increases, the pathlength of light also decreases as the suspended particles become more closely packed and scatter/absorb more light, leading to a near-constant number of detected photons. The optically saturated model also assumes

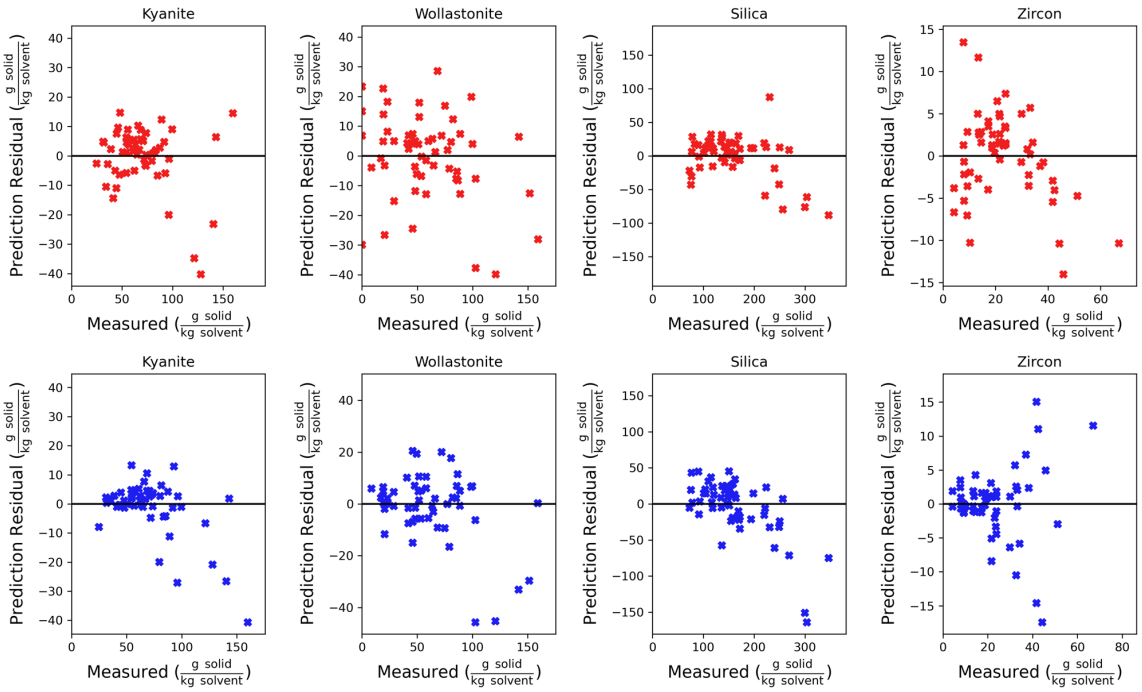


Figure 5.5: Residual plots of PLSR quantification of Raman spectra with **(top row)** original data and **(bottom row)** data after undergoing a log-log transformation.

that the particles contribute equally to optical opacity, and this does not seem to be a poor assumption based on the model performance on the system studied. Particle size and morphology were held approximately constant and not studied in this chapter, but particle size may have a significant effect on the Raman spectra. Further work in this area may determine whether constant particle size and morphology effects are important for an optically saturated model.

The author of this thesis performed several tests attempting to fit attenuating data to a general attenuation model including component-specific attenuation (i.e. $\theta_1 \neq \theta_2 ... \neq \theta_n$). This model has the form:

$$S(\lambda) = \frac{\sum_{i=1}^{n} w_i C_i}{\sum_{i=1}^{n} \theta_i C_i^{\gamma_i}}, \qquad w_i = P_D \beta(\lambda) A_D \Omega_D T Q(\lambda) t_s \tag{5.23}$$

The goal with fitting this equation to data is that the data would suggest the best fit. With the limited number of slurry experiments available (66 experiments), the fitting was un-

satisfactory and a richer dataset is likely required. Equation 5.23 can serve as a guide for data-driven schemes that are better suited for quantification in attenuating/turbid media than linear models. However, such a model might be unwieldy for practical use.

5.5 Conclusion

Nuclear-waste slurries present challenges for implementing traditional process analytical technology (PAT) for process monitoring, such as high-solid content and multicomponent solutions. In this chapter, two new quantification models, an optically saturated model and a variable pathlength model, have been developed and tested on Raman spectra of a model nuclear-waste system. Based on model performance on experimental slurry data, it was determined that the assumptions of an optically saturated (relative density) model outperformed both traditional models and a variable pathlength model. However, the variable pathlength model had accurate predictions at relatively low amounts of total solids studied in this system. This result suggests that a linear attenuation approximation correlates better to spectra at high concentrations, while a power-law approximation of attenuation may hold better at low concentrations. A hybrid model combining these two models may have physical significance and engineering use. The modeling results of this chapter may improve the analysis of slurries typically seen in nuclear-waste processing and related industries.

CHAPTER 6

MODEL-BASED FAULT DETECTION

The previous chapters in this thesis have focused on the measurements obtained from vibrational spectroscopy. In this chapter, measurement is related to the process that it is observing; a mass-balance model of a three-tank system of vessels housing slurries (i.e. nuclear waste) is developed. By combining measurement with this model, both state estimation (similar to prior chapters) and parameter estimation (new) are possible. This chapter specifically focuses on a mixing fault that may occur when transferring slurries from one vessel to the next with the presence of (untransferred) heel masses. The dual-Kalman filter is motivated for this application and investigated as a means to monitor this previously unstudied, but plausible and important, fault.[1]

In chemical batch reactors (including non-radioactive processes), proper mixing is important for consistent process and product quality. Heel masses (mass that is left behind from batch to batch) can be a feature of process design, or may occur inadvertently. The current practice at the Hanford nuclear-waste processing site is to include heel masses (up to 30% by volume) to reduce batch to batch variation of chemical waste that will be vitrified. To monitor this process, the proposed process uses laboratory measurements with a mass balance to propagate composition estimates from tank to tank. Incorporating in-line monitoring tools allow for faults relating to the heel mass to be detected when they occur in processing. Observability of model states and parameters are shown and encourage the use of real-time sensors in each vessel at Hanford. A simulation study was performed based on reported data on expected tank concentrations, sensor accuracies, and analytical laboratory accuracies. By incorporating in-line sensors, the mean absolute error (MAE) for the

[1] Some of the content and some of the figures in this chapter are reproduced from conference proceedings from Crouse et al. [32]

heel mass of kyanite, an insoluble solid, is improved by a factor of 3.88 over mass-balance model approaches in a simulation study including a fault in heel mass.

6.1 Introduction

To create a stable and optimal glass form from a nuclear-waste slurry at the Hanford site, waste loading must be considered as a function of the waste's composition and physical properties. Currently, the process at Hanford is expected to utilize a Monte-Carlo mass-balance model that may be combined with laboratory measurements taken of the incoming waste stream (before processing) and prior to being melted [10, 39]. An effect that has to be incorporated into any mass balance at the Hanford Site are *heels*, which are material that remain in the original tank after some material is transferred to the next tank. Heels enable efficient processing with slurries that may be difficult to transport while also decreasing batch to batch variation [11]. At Hanford, heels are expected to be in excess of 30% by volume for some tanks [10]. The heels are susceptible to changes with time due to varying mixing performance or tank transfer efficiency. A feed-forward modeling structure, as has been proposed with a mass-balance approach, will not allow the process to adapt to changing processing conditions and to detect process faults [15]. Additionally, laboratory measurements can incur measurement delays, preventing prompt detection of abnormal process conditions. In this chapter, real-time monitoring (optical spectroscopy) probes are investigated as complementary tools to mass balance models for detecting changes in mixing performance in batch reactors with heels.

In order to detect some process faults, including faults in heel mass composition, real-time sensors at Hanford must simultaneously *estimate* uncertain process states and *identify* uncertain process model parameters. To solve this simultaneous state estimation and parameter identification problem, a dual Kalman filter framework was used, where chemical compositions (states) and mixing/transfer performance (model parameters) are estimated simultaneously from a mass-balance model and plant measurements. Three processing

structures were investigated within the dual Kalman filter framework: an entirely feed-forward mass-balance approach, a mass-balance approach with a confirmation laboratory measurement, and a mass-balance approach with a laboratory measurement or in-line spectroscopic measurements at each tank. A "sudden change" fault scenario was investigated with regards to mixing performance (model parameters). For the present chapter, results on kyanite (an insoluble component that accumulates in the heel) and sodium nitrate (a soluble component hypothesized to not accumulate in the heel) are presented.

6.2 Non-ideal Mixing in Batch Reactors

In the presence of poor mixing and dead-zones, slurries within a batch reactor may not be spatially uniform. Poor mixing can lead to error from a predictive mass balance model when mass is transferred from tank-to-tank. A mass balance model is constructed in this section that incorporates poor mixing and heel masses.

6.2.1 Mass Balance for Batch Reactors in Series

For batch chemical reactors in series, a mass balance can be constructed around each species in each tank as a function of time, $m^i_{k,t}$. Variable i is the chemical species, k is the tank number in the series, and t is the time index or batch number. There are s chemical species, v tanks, and T timepoints. The feed to the first of the batch reactors is $m^i_{0,t}$. Based on the measured composition of waste in the tanks, a control action, $g^i_{k,t}$, is taken that adds mass to each tank. The mass balance for all species in all tanks in a series of batch reactors is shown in Figure 6.1 and can be written as a mass balance including heels (Equation 6.1)

with nonnegative concentrations (Equation 6.2):

$$m^i_{k,t+1} = (1 - \alpha^i_{k-1,t})m^i_{k-1,t} + \alpha^i_{k,t}m^i_{k,t} + g^i_{k,t} \tag{6.1}$$

$$m^i_{k,t} \geq 0 \tag{6.2}$$

$$\text{for } i = 1, 2, ..., s$$

$$\text{for } k = 1, 2, ..., v$$

$$\text{for } t = 1, 2, ..., T$$

where $\alpha^i_{k,t}$ is a time-varying parameter representing the heel mass (for a single species, tank, and batch) that remains after transfer to the next tank. The heel mass fraction is bounded between zero and one:

$$\begin{aligned} 0 \leq \alpha^i_{k,t} \leq 1 \\ \text{for } i = 1, 2, ..., s \\ \text{for } k = 1, 2, ..., v \\ \text{for } t = 1, 2, ..., T \end{aligned} \tag{6.3}$$

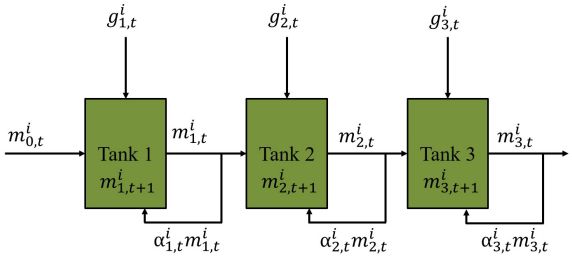


Figure 6.1: General mass balance of a batch system where each tank has a control input, g, and a heel mass (shown as a recycle stream) determined by a parameter α.

Equation 6.3 shows that $\alpha^i_{k,t}$ describes the fraction of mass (for a single species, tank, and batch) that remains in the tank after batch transfer is complete (Equation 6.1).

Equation 6.1 is linear with respect to $m^i_{k,t}$ and linear with respect to $\alpha^i_{k,t}$. The dual Kalman filter allows Equation 6.1 to be treated as a function of $m^i_{k,t}$ while $\alpha^i_{k,t}$ is held

constant. An analogous problem is then solved whereby $\alpha^i_{k,t}$ is allowed to vary while $m^i_{k,t}$ is held constant. A nonlinear model can be addressed by using the dual extended Kalman filter, or by using an unscented Kalman filter [133, 134, 135, 136].

Some assumptions are made to restrict the general series batch reactor system (Equation 6.1–6.3) to model the system at Hanford: there are three tanks in the series of batch reactors (Equation 6.4), there is no heel in the liquid supernate of Tank 1 (Equation 6.5), the only control actions are glass-forming chemical (GFC) additions in the second tank (Equation 6.6), and the control actions are nonnegative (Equation 6.7):

$$v = 3 \tag{6.4}$$

$$\alpha^i_{k,t} = 0 \text{ when } k = 1 \tag{6.5}$$
$$\text{for } i = 1, 2, ..., s$$
$$\text{for } t = 1, 2, ..., T$$

$$g^i_{k,t} = 0 \text{ when } k \neq 2 \tag{6.6}$$
$$\text{for } i = 1, 2, ..., s$$
$$\text{for } t = 1, 2, ..., T$$

$$g^i_{k,t} \geq 0 \text{ when } k = 2 \tag{6.7}$$
$$\text{for } i = 1, 2, ..., s$$
$$\text{for } t = 1, 2, ..., T$$

A mass balance is constructed assuming no chemical or nuclear reactions of the monitored species. The three tank system — Tank 1, Tank 2, and Tank 3 — corresponds to physical tanks at the Hanford Site directly prior to the melter that produces the glass. Representative concentrations and process vessel sizes are based on values from Vienna and Kim and modified to make each tank equivalent in process mass [39]. The nominal ves-

sel compositions used in the simulated study are shown in Table 6.1. The mass balance equations after the simplifying assumptions of Equation 6.4–6.7 are:

$$\begin{aligned} m^i_{1,t+1} &= m^i_{0,t} \\ m^i_{2,t+1} &= m^i_{1,t} + \alpha^i_2 m^i_{2,t} + g^i_{2,t} \\ m^i_{3,t+1} &= (1-\alpha^i_2) m^i_{2,t} + \alpha^i_3 m^i_{3,t} \\ &\text{for } i = 1, 2, ..., s \\ &\text{for } t = 1, 2, ..., T \end{aligned} \tag{6.8}$$

where the nonnegativity from Equation 6.7 still applies.

6.2.2 Simulation

A simulation study was performed where the species and compositions in Table 6.1 are used as initial conditions and subsequently propagated through the mass balance of Equation 6.8. The following sections motivate assumptions made in the simulation. The simulation is shown in Figure 6.2, with the role of the dual-Kalman filter in relation to the plant shown in Figure 6.1.

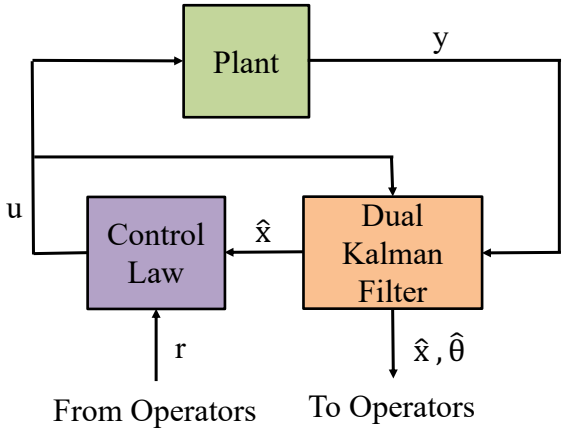


Figure 6.2: The role of the dual-Kalman filter in process feedback; the dual-Kalman filter relies on process measurements and a system model.

6.2.3 Optical Spectroscopy and Offline Measurements

To measure the species in Table 6.1, two analytical techniques are expected in the on-site analytical laboratory at the Hanford site. Inductively coupled plasma - atomic emission spectroscopy (ICP-AES) can measure elemental species and will be used to measure boron (boric acid), silicon (silica), aluminum (kyanite), and calcium (wollastonite) within a relative percent difference of $\leq$20% [137]. Ion chromatography (IC) can measure soluble molecular species and will be used to measure nitrate, nitrite, carbonate, and sulfate within a relative percent difference of $\leq$20%. In this chapter, a $\pm 20\%$ range about the mean is assumed to correspond to ± 3 standard deviations about the mean for that component. The analytical laboratory accuracy is shown in the "offline" column in Table 6.2.

The accuracy of optical in-line sensors is based on reported values for attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy and Raman spectroscopy [19]. Full covariance matrices are used for spectroscopic prediction accuracy. ATR-FTIR has shown the ability to measure solution phase compositions in slurries, and so will be simulated to measure soluble nitrate, nitrite, carbonate, sulfate, and borate anions. Raman spectroscopy, meanwhile, has been shown to measure solid species in slurries (Chapter 4). In this chapter, Raman will be simulated to measure insoluble kyanite, wollastonite, and silica. Standard deviations for measurement accuracies are shown in the "In-line" column in Table 6.2. It is assumed that there are no phase changes for any of the species, although the dual Kalman filter framework would allow for phase changes to be estimated as a process parameter as well (which could provide useful process information in the absence of an accurate solubility model).

Since an on-site analytical laboratory could have any capabilities of in-line probes at their disposal, the in-line measurement standard deviation is used for the laboratory whenever the in-line sensor is reported as more accurate than the corresponding offline sensor. The analytical laboratory measurement standard deviation is shown in the "Laboratory" column in Table 6.2. Planned in-line measurements at the Hanford high-level waste (HLW)

Table 6.1: Simulation vessel compositions in kg.

Species	Tank 1	Tank 2	Tank 3	Soluble
Sodium Nitrate	5742.2	5742.2	5742.2	Yes
Sodium Nitrite	2505.0	2505.0	2505.0	Yes
Sodium Carbonate	1603.3	1603.3	1603.3	Yes
Sodium Sulfate	207.7	207.7	207.7	Yes
Boric Acid	0.0	3764.3	3764.3	Yes
Kyanite	0.0	7424.1	7424.1	No
Wollastonite	0.0	7842.3	7842.3	No
Silica	0.0	7790.1	7790.1	No

Table 6.2: Standard deviations of prediction for in-line optical sensors, offline measurements performed via ICP-AES or IC, and the lower of online/offline that an analytical laboratory at Hanford is likely to utilize.

Species (kg)	In-line	Offline	Laboratory
Sodium Nitrate	30.9	191.4	30.9
Sodium Nitrite	18.0	83.5	18.0
Sodium Carbonate	60.8	53.4	53.4
Sodium Sulfate	23.5	6.9	6.9
Boric Acid	34.7	125.5	34.7
Kyanite	170.8	247.5	170.8
Wollastonite	221.7	261.4	221.7
Silica	489.4	259.7	259.7

Waste Treatment Plant (WTP) that are not the focus of this chapter include laser liquid-height measurements and GFC mass measurements [39].

6.2.4 Fault Scenario

A fault scenario is investigated in this chapter that may occur in series batch reactors with heel masses. For all insoluble species, the heel mass parameter begins at 0.20 at Batch 1. Let $\mathcal{I}$ denote the subset of species that are insoluble, shown in Table 6.1.

$$\begin{aligned} &\alpha_{k,1}^{i} = 0.20 \\ &\quad \text{for } i \in \mathcal{I} \\ &\quad \text{for } k = 1, 2, ..., v \end{aligned} \tag{6.9}$$

The first fault scenario is a gradual 2% per batch increase in the heel mass fraction in

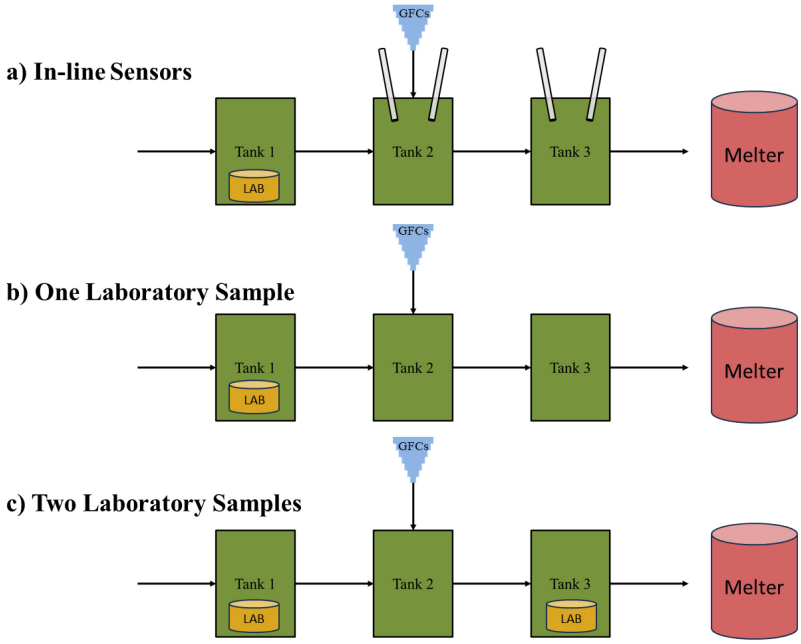


Figure 6.3: Measurement scenarios in this chapter are a laboratory sample of Tank 1 and gravimetric measurements of the control input (glass-forming chemicals (GFCs)), in addition to: **a)** in-line sensors at Tanks 2 and 3, **b)** no additional measurements, and **c)** an additional laboratory sample in Tank 3.

Tank 2 starting with the seventh batch. This may be caused by solid buildup restricting fluid flow, change in the settling properties of a specific component, a change in macroscopic slurry properties, mixing degradation, or pumping degradation. The "drift" fault rule is shown in Equation 6.10.

$$\begin{aligned} \alpha^i_{2,t} &= (1.02)\alpha^i_{2,t-1} \\ &\text{for } i \in \mathcal{I} \\ &\text{for } t = 7, 8, ..., T \end{aligned} \tag{6.10}$$

The second fault scenario is an abrupt change in heel mass fraction in Tank 2 that occurs during Batch 7. This may be caused by a mixing or pumping failure, different settling properties of a particular solid, agglomeration of suspended particles, or a substantially different nuclear-waste composition. The fault rule is shown in Equation 6.11.

$$\alpha^i_{2,7} = (1.5)\alpha^i_{2,6} \quad \text{for } i \in \mathcal{I} \tag{6.11}$$

There is also an additive white noise influencing the parameter values in addition to the fault rules. A dimensionless Gaussian noise with a standard deviation of 5×10^{-3} is applied to the parameter values at each batch, shown in Equation 6.12.

$$\alpha^i_{k,t} = \alpha^i_{k,t-1} + \mathcal{N}(0, \mathbf{\Sigma}_\theta) \quad \text{for } i \in \mathcal{I} \quad \text{for } k = 1, 2, ..., v \quad \text{for } t = 1, 2, ..., T \tag{6.12}$$

6.2.5 Sensor Models

Two analytical techniques are expected to be used at the on-site analytical laboratory at Hanford. ICP-AES can measure elemental species and is planned to be used to measure boron (boric acid), silicon (silica), aluminum (kyanite), and calcium (wollastonite). IC can measure soluble molecular species and will be used to measure the nitrate, nitrite, carbonate, and sulfate anions. For both of these instruments, measurement accuracy is expected to be within a relative percent difference of less than 20% (±20% of the true value). To convert ±20% into the form of covariance expected by the dual-Kalman filter, it is assumed that ±20% corresponds to ±3 standard deviations.

Optical in-line sensors are used as a complementary technique to laboratory measurements in this work. Raman spectroscopy and ATR-FTIR spectroscopy have representative accuracies taken from results measuring slurries in the work by Prasad et al. [19]. Raman spectroscopy is used to interrogate insoluble solids and ATR-FTIR is used to measure soluble anions. By charge balance, the measurement of anions is assumed to have a matching sodium cations that are included in the total mass, although ionic sodium could not be

measured directly by Raman or ATR-FTIR spectroscopies.

Three sensor models are investigated in this chapter and are qualitatively shown in Figure 6.3. All sensor models begin with an analytical laboratory measurement in Tank 1, since this is the current plan for processing nuclear waste at Hanford. The first sensor model additionally has in-line sensors at Tanks 2 and 3, which can measure concentrations and detect deviations from a mass-balance model. The second sensor model does not have any additional measurements beyond the laboratory sample at Tank 1. The second sensor model corresponds to a feed-forward process model, where control is based on the mass balance model. The third sensor model corresponds to an additional laboratory sample taken from Tank 3.

6.3 The Dual Kalman Filter

The Kalman filter is a recursive estimator in which a previous state estimate is propagated through a system model and subsequently updated with the most recent sensor measurements. The dual Kalman filter allows both states and parameters to be estimated using two separate filters, a state filter and a parameter filter. The conversion to state-space form from the model built in Equation 6.8 is shown in Chapter G.

The mass balance equations (Equation 6.8) of the system can be collected (concatenated) into discrete-time state-space form. Instead of m, α, and g used in Equation 6.1 and Equation 6.8, the following equations use $\mathbf{x}$, $\boldsymbol{\theta}$, and $\mathbf{u}$, respectively. In state-space form, the state and measurement equations are:

$$\mathbf{x}_{t+1} = \mathbf{F}(t, \boldsymbol{\theta})\mathbf{x}_t + \mathbf{B}(t, \boldsymbol{\theta})\mathbf{u}_t + \mathbf{w}_d \tag{6.13}$$

$$\mathbf{y}_{t+1} = \mathbf{H}(t, \boldsymbol{\theta})\mathbf{x}_t + \mathbf{w}_z \tag{6.14}$$

where the state of the three-tank system at Batch t, $\mathbf{x}_t \in \mathbb{R}^{n\times 1}$, is a vector of chemical masses within each tank, where $n = v \times s$. Similarly, the parameter vector at Batch t,

$\boldsymbol{\theta}_t \in \mathbb{R}^{p\times 1}$, describes the heel mass fractions for each species, where p is the number of process parameters. In this chapter, there is one parameter for each state, so $p = n$. The control vector at Batch t, $\mathbf{u}_t \in \mathbb{R}^{c\times 1}$, describes additions to the tanks, where c is the dimension of the control input. The measurement vector at Batch t, $\mathbf{y}_t \in \mathbb{R}^{q\times 1}$, describes in-line or laboratory measurements, where q is the dimension of measurements collected.

6.3.1 Stochastic Matrices

The covariance matrices of this chapter corresponding to process noise ($\boldsymbol{\Sigma}_x$), parameter noise ($\boldsymbol{\Sigma}_\theta$), and measurement noise ($\boldsymbol{\Sigma}_z$), are assumed to be known prior to dual Kalman filter implementation. The covariance matrices also dictate the distribution of additive noise used in simulations in this chapter. Methods exist for initializing these matrices when unknown *a priori* [138].

The vector of process noise, $\mathbf{w}_d \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x) \in \mathbb{R}^{n\times 1}$, is assumed to be zero-mean with covariance given by $\boldsymbol{\Sigma}_x \in \mathbb{R}^{n\times n}$, which is a diagonal covariance matrix[2], $\boldsymbol{\Sigma}_x = \text{Diag}((0.02)\bar{\mathbf{x}})^{\circ\frac{1}{2}}$ (standard deviation of 2% of the average or nominal state vector, $\bar{\mathbf{x}}$). This choice of process noise is somewhat arbitrary; knowledge of the Hanford mass balance (and deviations from such a mass balance) would be required to inform the vector of process noise. However, we are operating under the assumption that deviations from the mass balance would be minimal (2%). More process noise would cause the Kalman filter estimates to rely on measurement more than model prediction due to the higher model uncertainty.

The vector of measurement noise, $\mathbf{w}_z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_z) \in \mathbb{R}^{q\times 1}$ is assumed to be zero-mean[3] and a *non-diagonal* covariance matrix, $\boldsymbol{\Sigma}_z \in \mathbb{R}^{q\times q}$, estimated from sensor data shown in Table 6.2 and taken from the data shown in Chapter 4 [19]. For measurements using the analytical laboratory, covariances are assumed to be zero (only variance is considered since covariance is not reported). Parameter drift is shown by Equation 6.15 and characterized

[2] A Hadamard power (elementwise power) of $\frac{1}{2}$ is indicated by $^{\circ\frac{1}{2}}$.

[3] Sensor biases could result due to unmodeled instrument drift or poor measurement models; Chapter 2, Chapter 3, Chapter 4, and Chapter 5 address these issues for solutions and slurries.

by $\mathbf{w}_\theta$:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{w}_\theta \tag{6.15}$$

where the vector of parameter noise, $\mathbf{w}_\theta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\theta) \in \mathbb{R}^{p\times 1}$, is assumed to be zero-mean with a covariance given by $\boldsymbol{\Sigma}_\theta \in \mathbb{R}^{p\times p}$. In this chapter, parameter drift is assumed to obey $\boldsymbol{\Sigma}_\theta = \text{Diag}((0.025)\bar{\boldsymbol{\theta}})^{\circ\frac{1}{2}}$ (standard deviation of 2.5% of the average or nominal parameter value, $\bar{\boldsymbol{\theta}}$). Parameter drift would be dependent on the regularity of heel masses at the Hanford process; an estimate of 2.5% is used because it is substantially less than the condition of a fault (50% change).

6.3.2 Dual Kalman Filter

The dual Kalman filter was implemented in the method shown by Wan and Nelson [138, 135]. The four steps are parameter prediction, state prediction, state update, and parameter update. In Equation 6.16–6.26, $\hat{}$ denotes an estimated quantity and $^-$ denotes a model-predicted quantity (before measurement update).

Parameter prediction:

$$\hat{\boldsymbol{\theta}}_t^- = \hat{\boldsymbol{\theta}}_{t-1} \tag{6.16}$$

$$\boldsymbol{\Phi}_\theta^-(t) = \boldsymbol{\Phi}(t-1) + \boldsymbol{\Sigma}_\theta \tag{6.17}$$

State prediction:

$$\hat{\mathbf{x}}_t^- = \mathbf{F}(t, \hat{\boldsymbol{\theta}}_t^-)\hat{\mathbf{x}}_{t-1} + \mathbf{B}(t, \hat{\boldsymbol{\theta}}_t^-)\mathbf{u}_t \tag{6.18}$$

$$\boldsymbol{\Phi}_{\text{x}}^-(t) = \mathbf{F}(t, \hat{\boldsymbol{\theta}}_t^-)\boldsymbol{\Phi}_{\text{x}}(t-1)\mathbf{F}^{\text{T}}(t, \hat{\boldsymbol{\theta}}_t^-) + \boldsymbol{\Sigma}_{\text{x}} \tag{6.19}$$

State update:

$$\mathbf{K}_{\mathrm{x}}(t) = \mathbf{\Phi}_{\mathrm{x}}^{-}(t)\mathbf{H}_{\mathrm{x}}^{\mathrm{T}}[\mathbf{\Sigma}_{\mathrm{z}} + \mathbf{H}_{\mathrm{x}}\mathbf{\Phi}_{\mathrm{x}}^{-}(t)\mathbf{H}_{\mathrm{x}}^{\mathrm{T}}]^{-1} \tag{6.20}$$

$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^{-} + \mathbf{K}_{\mathrm{x}}(t)[\mathbf{y}_t - \mathbf{H}_{\mathrm{x}}\hat{\mathbf{x}}_t^{-}] \tag{6.21}$$

$$\mathbf{\Phi}_{\mathrm{x}}(t) = [\mathbf{I} - \mathbf{K}_{\mathrm{x}}(t)\mathbf{H}_{\mathrm{x}}]\mathbf{\Phi}_{\mathrm{x}}^{-}(t) \tag{6.22}$$

$$\tag{6.23}$$

Parameter update:

$$\mathbf{K}_{\theta}(t) = \mathbf{\Phi}_{\theta}^{-}(t)\mathbf{H}_{\theta}^{\mathrm{T}}(t,\hat{\mathbf{x}}_t)[\mathbf{\Sigma}_z + \mathbf{H}_{\theta}(t,\hat{\mathbf{x}}_t)\mathbf{\Phi}_{\theta}^{-}(t)\mathbf{H}_{\theta}^{\mathrm{T}}(t,\hat{\mathbf{x}}_t)]^{-1} \tag{6.24}$$

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_t^{-} + \mathbf{K}_{\theta}(t)[\mathbf{y}_t - \mathbf{H}_{\theta}(t,\hat{\mathbf{x}}_t)\hat{\boldsymbol{\theta}}_t^{-}] \tag{6.25}$$

$$\mathbf{\Phi}_{\theta}(t) = [\mathbf{I} - \mathbf{K}_{\theta}(t)\mathbf{H}_{\theta}(t,\hat{\mathbf{x}}_t)]\mathbf{\Phi}_{\theta}^{-}(t) \tag{6.26}$$

where $\mathbf{\Phi}_{\theta}(t) \in \mathbb{R}^{p\times p}$ and $\mathbf{\Phi}_{\mathrm{x}}(t) \in \mathbb{R}^{n\times n}$ are estimation error covariance matrices for parameters and states, respectively. The estimation covariance matrices are initialized in this work as $\mathbf{\Sigma}_{\theta}$ and $\mathbf{\Sigma}_{\mathrm{x}}$, and are updated with each iteration of the dual Kalman filter. $\mathbf{K}_{\theta}(t) \in \mathbb{R}^{p\times q}$ and $\mathbf{K}_{\mathrm{x}}(t) \in \mathbb{R}^{n\times q}$ are the Kalman gains calculated for the parameters and states, respectively.

The linear parameter and state update equations are $\mathbf{F}_{\theta} \in \mathbb{R}^{p\times p}$ and $\mathbf{F} \in \mathbb{R}^{n\times n}$, respectively. In this chapter, the parameter update is equivalent to the identity matrix since mixing parameters do not nominally change from batch to batch, $\mathbf{F}_{\theta} = \mathbf{I}^{p\times p}$. The state transition matrix for three batch reactors can be found by putting the general mass balance from Equation 6.8 into matrix format; this matrix represents the mass being transferred to

the subsequent tank (except for heel masses):

$$\mathbf{F}(\boldsymbol{\theta}, t) =$$

$$\left[\begin{array}{c|c|c} \mathrm{Diag}(\boldsymbol{\alpha}_{1,t}^{\mathrm{all}}) & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{I} - \mathrm{Diag}(\boldsymbol{\alpha}_{1,t}^{\mathrm{all}}) & \mathrm{Diag}(\boldsymbol{\alpha}_{2,t}^{\mathrm{all}}) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} - \mathrm{Diag}(\boldsymbol{\alpha}_{2,t}^{\mathrm{all}}) & \mathrm{Diag}(\boldsymbol{\alpha}_{3,t}^{\mathrm{all}}) \end{array}\right] \tag{6.27}$$

Important to this chapter, the linear parameter and state measurement equations are $\mathbf{H}_\theta \in \mathbb{R}^{q\times p}$ and $\mathbf{H} \in \mathbb{R}^{q\times n}$, respectively. $\mathbf{H}$ determines which states are measurable in different tanks; all of these matrices are diagonal since measurements of tank k only provide state information of tank k. Meanwhile, $\mathbf{H}_\theta$ determines which parameters (heel masses) are measurable in different tanks; this requires measurements of adjacent tanks (i.e. k and $k-1$). $\mathbf{H}_\theta$ and $\mathbf{H}$ are therefore different for the three measurement scenarios and are listed for each. $\mathbf{H}$ is shown in Equation 6.28–6.30 and $\mathbf{H}_\theta$ is shown in Equation 6.31–6.33 and calculated using expressions from Wan et al. [138].

$$\mathbf{H}^{\text{Scenario 1}} = \left[\begin{array}{c|c|c} \mathbf{I}^{s\times s} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}^{s\times s} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}^{s\times s} \end{array}\right] \tag{6.28}$$

$$\mathbf{H}^{\text{Scenario 2}} = \left[\begin{array}{c|c|c} \mathbf{I}^{s\times s} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right] \tag{6.29}$$

$$\mathbf{H}^{\text{Scenario 3}} = \left[\begin{array}{c|c|c} \mathbf{I}^{s\times s} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}^{s\times s} \end{array}\right] \tag{6.30}$$

$$\mathbf{H}_{\theta}^{\text{Scenario }1}(t,\hat{\mathbf{x}}_t) =$$

$$\left[\begin{array}{c|c|c} \text{Diag}(\hat{\mathbf{m}}_{1,t}^{\text{all}}) & \mathbf{0} & \mathbf{0} \\ \hline -\text{Diag}(\hat{\mathbf{m}}_{1,t}^{\text{all}}) & \text{Diag}(\hat{\mathbf{m}}_{2,t}^{\text{all}}) & \mathbf{0} \\ \hline \mathbf{0} & -\text{Diag}(\hat{\mathbf{m}}_{2,t}^{\text{all}}) & \text{Diag}(\hat{\mathbf{m}}_{3,t}^{\text{all}}) \end{array}\right] \tag{6.31}$$

$$\mathbf{H}_{\theta}^{\text{Scenario }2}(t,\hat{\mathbf{x}}_t) = \left[\begin{array}{c|c|c} \text{Diag}(\hat{\mathbf{m}}_{1,t}^{\text{all}}) & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right] \tag{6.32}$$

$$\mathbf{H}_{\theta}^{\text{Scenario }3}(t,\hat{\mathbf{x}}_t) =$$

$$\left[\begin{array}{c|c|c} \text{Diag}(\hat{\mathbf{m}}_{1,t}^{\text{all}}) & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & -\text{Diag}(\hat{\mathbf{m}}_{2,t}^{\text{all}}) & \text{Diag}(\hat{\mathbf{m}}_{3,t}^{\text{all}}) \end{array}\right] \tag{6.33}$$

6.4 Results

The simulation study was performed with all of the species included in Table 6.1 for the nuclear waste system at Hanford. However, the plotted results focus on a single species for clarity: kyanite. Kyanite is an insoluble GFC that is added in as a control input to the waste batch and can be reliably measured by in-line and offline measurement techniques.

6.4.1 Observability

The observability matrix ($\mathcal{O}$) of states and parameters are calculated for the system in Equation 6.34 and Equation 6.35, respectively. The rank of state and parameter observability matrices for the three measurement scenarios are shown in Table 6.3. The one laboratory sample measurement and two laboratory sample measurement scenarios have full observ-

Table 6.3: Observability of all system states and parameters of the three-tank system.

Measurement Scenario	$\mathcal{O}$(**States**)	$\mathcal{O}$(**Parameters**)
In-line Sensors	Yes	Yes
One Laboratory Sample	Yes	No
Two Laboratory Samples	Yes	No

ability rank for state measurements, indicating that the concentration of all species can be measured in each tank, given an accurate state model. However, the two measurement scenarios do not have full observability rank for parameters, indicating that they cannot take advantage of adaptive models with varying heel masses (such as the dual-Kalman filter in this study) nor detect model faults, such as heel mass changes. The in-line sensor model has full observability matrix rank for both states and parameters; the parameters, in addition to the states, can be used to build an adaptive model and detect faults that occur with tank heel masses.

$$\mathcal{O}(\mathbf{F},\mathbf{H}) = \begin{bmatrix} \mathbf{H} \\ \mathbf{HF} \\ \vdots \\ \mathbf{HF}^{n-1} \end{bmatrix} \tag{6.34}$$

$$\mathcal{O}(\mathbf{F}_\theta,\mathbf{H}_\theta) = \begin{bmatrix} \mathbf{H}_\theta \\ \mathbf{H}_\theta\mathbf{F}_\theta \\ \vdots \\ \mathbf{H}_\theta\mathbf{F}_\theta^{n-1} \end{bmatrix} \tag{6.35}$$

6.4.2 Gradual Parameter Drift

In this section, a gradual parameter drift was simulated in Tank 2 by increasing $\alpha_{2,t}^i$ by 2% for every batch starting with the seventh batch for all insoluble species, $i \in \mathcal{I}$

(Equation 6.10).

In Figure 6.4a, the true heel fractions began increasing above the mean value of 0.20 soon after the drifting fault occurred at batch seven. The in-line sensors more closely approximated the true heel as it drifted from its nominal steady state value compared to other monitoring schemes. Meanwhile, one laboratory sample combined with a mass-balance model was unable to adapt to changing downstream processing conditions. The poor performance of the two laboratory samples sensor model can be attributed to to the lack of feedback in the specific tank where the drift fault occurred, Tank 2. While Tank 2 had significantly different compositions with the drift fault (see Figure 6.5a), the input-output behavior of Tank 2 remained nearly constant and so the fault was unable to be detected without measurements directly of Tank 2. Figure 6.4b highlights how the different sensor models adjusted to white parameter noise in Tank 3 while Tank 2 underwent the drift fault.

In Figure 6.5, the error of estimated states (masses) of kyanite are shown for Tanks 2 and 3. In Tank 2 (Figure 6.5a), the in-line sensors were able to maintain a near-zero-mean error in the presence of changing process parameters. Meanwhile, the one laboratory sample and two laboratory samples sensor models underpredicted the true kyanite mass in the tanks as the observer model (Equation 6.26, a function of $\hat{\boldsymbol{\theta}}_t$) did not accurately adjust to the changing parameters, $\boldsymbol{\theta}_t$.

Figure 6.6 additionally shows predictions of Tanks 2 and 3 of sodium nitrate, a soluble species that is assumed to have zero retention in this study. Compared to the feed-forward sensor model of one laboratory sample, in-line sensors decreased the model prediction error in Tanks 2 and 3, and two laboratory samples decreased the model prediction error in Tank 3. Even with no component heel (as is the case for sodium nitrate in Figure 6.6) the additional measurements of in-line sensors and two laboratory samples reduced the impact of process noise when an identical mass-balance model was used in all three sensor model structures.

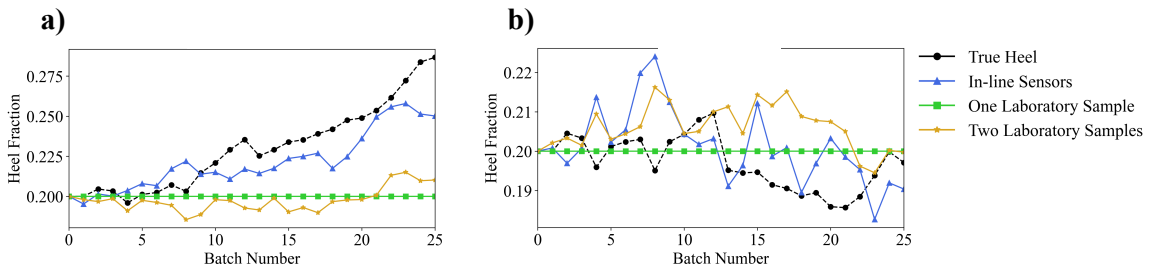


Figure 6.4: True and measured heel fraction of kyanite in **a)** Tank 2 and **b)** Tank 3 during a heel drift fault in Tank 2.

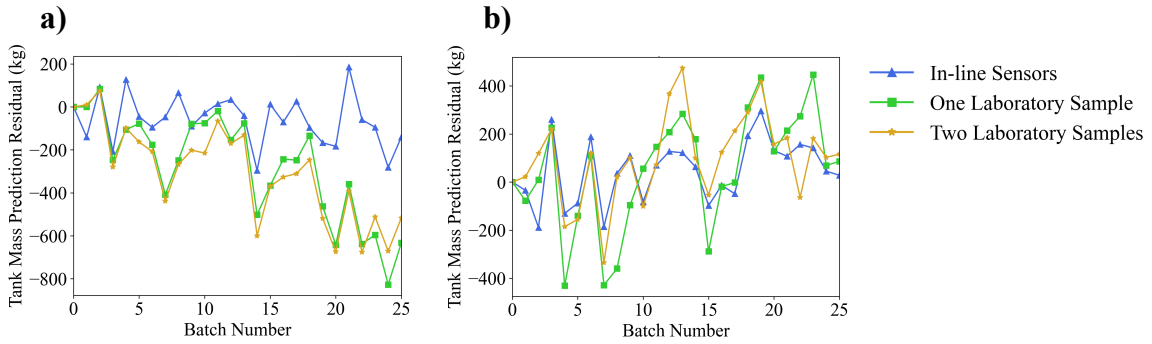


Figure 6.5: Deviation of measured states from true states of kyanite mass in **a)** Tank 2 and **b)** Tank 3 during a heel drift fault in Tank 2.

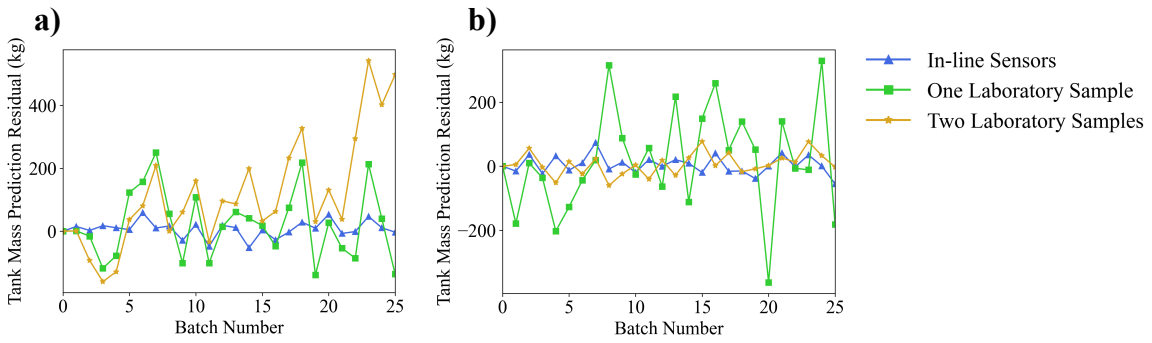


Figure 6.6: Deviation of measured states from true states of sodium nitrate mass in **a)** Tank 2 and **b)** Tank 3 during a heel drift fault in Tank 2.

6.4.3 Sudden Parameter Shift Fault

A sudden parameter shift was simulated in Tank 2 by increasing $\alpha_{2,t}^{i}$ by 50% at Batch 7 for all insoluble species, $i \in \mathcal{I}$ (Equation 6.11).

In Figure 6.7a the in-line sensor model was able to adjust estimated parameter values, $\hat{\boldsymbol{\theta}}_t$, to follow the true heel fraction, $\boldsymbol{\theta}_t$. This observed parameter shift could be used to detect a fault when it occurs. The other sensor models utilizing only laboratory samples and having no direct measurement of Tank 2 did not appreciably deviate from steady state estimates for $\hat{\boldsymbol{\theta}}_t$. Figure 6.7b shows the heel fraction prediction for Tank 3, the tank after the fault occurs. With no fault occurring, all of the measurement scenarios are able to predict the heel fraction accurately.

To show how heel mass changes may affect the amount of kyanite in the tanks, Figure 6.8a and Figure 6.8b show the errors of the estimated states (masses) for Tanks 2 and 3, respectively. An error of zero represents perfect measurement. In Figure 6.8a, the in-line sensor model adapts to the sudden shift in parameter values, and maintains a near zero-mean prediction error of the true tank state. An inaccuarate estimation of kyanite in Tank 2 (Figure 6.8a) may lead to increased mixer wear, deadzones, and the downstream slurry composition may be affected. The kyanite results of Figure 6.7–6.9 are tabulated in Table 6.4.

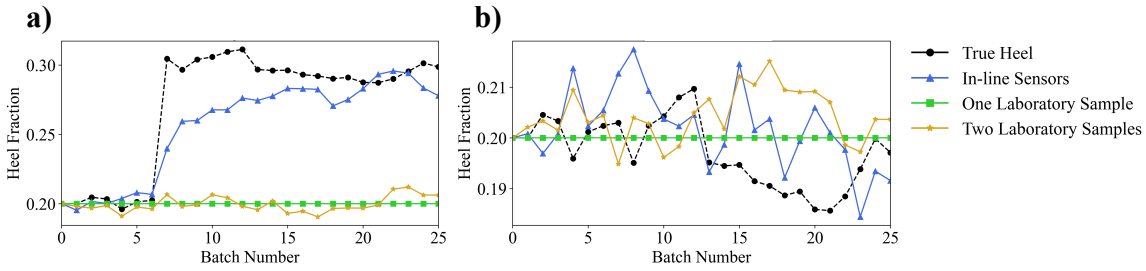


Figure 6.7: True and measured heel fraction of kyanite in **a)** Tank 2 and **b)** Tank 3 during a sudden heel shift fault in Tank 2.

6.4.4 Species with No Heel

These results can approximate the mixing behavior of *undetectable* species as well assuming their mixing parameters, $\boldsymbol{\theta}_t$, are similar to measurable species due to physical similarities (i.e. equivalent particle size distribution, material density, etc.) For example,

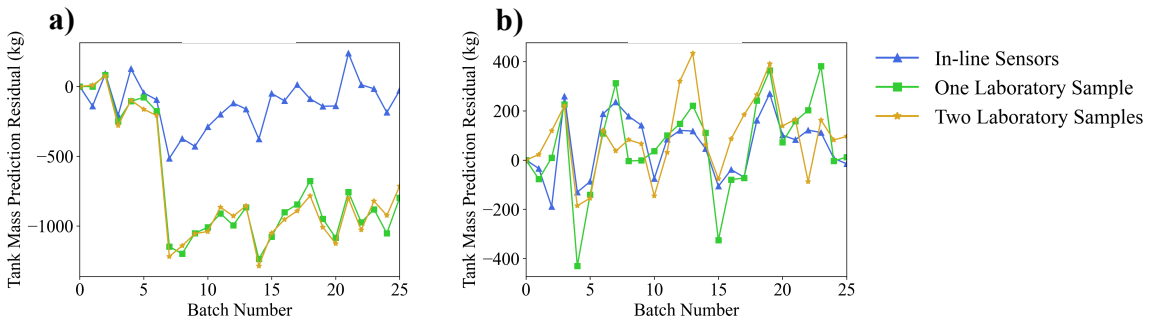


Figure 6.8: Deviation of measured states from true states of kyanite mass in **a)** Tank 2 and **b)** Tank 3 during a sudden heel shift fault in Tank 2.

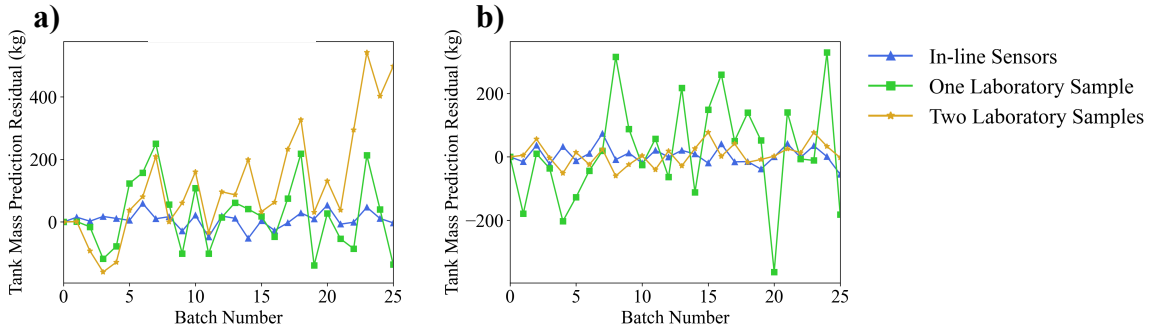


Figure 6.9: Deviation of measured states from true states of sodium nitrate mass in **a)** Tank 2 and **b)** Tank 3 during a sudden heel shift fault in Tank 2.

kyanite may be measurable with in-line sensors, but another solid species, such as gibbsite ($Al(OH)_3$), may not be measurable with in-line sensors. In such a scenario, if the two chemical constituents are determined to have similar solid-settling properties, the settling behavior of kyanite may suggest gibbsite settling.

Table 6.4: MAE and root mean squared error (RMSE) for estimated parameters ($\hat{\boldsymbol{\theta}}_t$) and states ($\hat{\mathbf{x}}_t$) of kyanite in Tank 2 when heel mass is suddenly increased by 50%.

Parameters ($\hat{\boldsymbol{\theta}}_t$, unitless)		
Kyanite	MAE	RMSE
In-line Sensors	**0.022**	**0.028**
One Laboratory Sample	0.097	0.097
Two Laboratory Samples	0.097	0.097
States ($\hat{\mathbf{x}}_t$, kg)		
Kyanite	MAE	RMSE
In-line Sensors	**183.3**	**235.0**
One Laboratory Sample	969.9	980.7
Two Laboratory Samples	974.3	985.7

6.5 Conclusion

This chapter highlights the importance of accurate and timely process measurements in a simulated nuclear-waste processing plant. Of the three studied sensor models, measurements in every tank enabled by real-time sensors provided the best real-time parameter estimation for the studied fault scenarios where heels changed the mass balance between tanks. For a sudden increase in heel mass, sensors in every tank (enabled by real-time sensors) had an MAE of 0.025, compared to 0.097 of a feed-forward "one laboratory sample" sensor model and and 0.097 of a "two laboratory samples" model. An observability analysis provided a rigorous underpinning of this result. Beyond parameter estimation, state estimation was also improved by measuring each individual tank. Including timely and accurate sample points in a chemical process, particularly of slurry processes with heel masses, may allow early detection of faults that may not be detected via standard process sampling and mass-balance prediction.

CHAPTER 7

DATA-DRIVEN FAULT DETECTION IN INDUSTRIAL SLURRIES

This chapter investigates fault detection, but from a *data-driven* paradigm rather than model-based as Chapter 6 had done. This chapter contributes to this thesis by investigating data-driven fault detection in industrial slurry processes with multiple instruments: Raman spectroscopy, attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy, and focused beam reflectance measurement (FBRM). This chapter then investigates these methods to detect sensor faults, mixing faults, and unanticipated chemistry in slurry systems typical of high-level waste (HLW).

Data from the three instruments were combined via a data fusion scheme utilizing principal component analysis (PCA), Hotelling T^2, and squared prediction error (SPE). Uncertainty was quantified yielding a three-sigma region of normal operation to identify faults. The three instruments allowed for the detection of a variety of process faults that may appear in either phase (solution or insoluble phase) of a slurry. This chapter represents a major step forward in 1) monitoring radioactive slurry processes by reducing the need for offline monitoring, 2) accurately detecting faults in the presence of probe fouling, and 3) enabling the use of commercially available spectroscopic sensors to detect a variety of faulty process states in real-time and remotely.

7.1 Introduction

Slurries are ubiquitous in process industries appearing in industrial mineral mining [139], industrial wastewaters [140, 141], municipal wastewaters [142], pharmacuetical crystallization [143], additive manufacturing [144], battery manufacturing [145], and nuclear-waste processing [146, 147, 148] among other industries. Despite their industrial relevance, slurries are often difficult to monitor in real-time because they comprise multiple phases.

Furthermore, many slurry processes are multicomponent and have poorly controlled particle size distributions. For industrial processes, there are many incentives to distinguish whether a process is operating normally or if a process fault has occurred. Abnormal process conditions can reduce economic productivity by producing a substandard product or by decreasing process efficiency, or threatening the safety of workers and surrounding communities.

As stated in Section 1.2, the use of real-time sensors (process analytical technology (PAT)) may have great benefit to waste processing of slurries (HLW) at Hanford. Despite the established literature on real-time sensors (see Section 1.1), little has been published on statistical fault detection using spectroscpic sensors in nuclear-waste systems.

This chapter will rely on multivariate statistical process monitoring (MSPM), which is often used for fault detection, diagnosis, and reconstruction. MSPM has historically been applied to industrial chemical processes [149], including nuclear-waste treatment. At the Savannah River Site, which has been processing HLW since 1996, measurements of waste composition have used statistical analysis including Hotelling T^2 to determine waste batch acceptability [150, 151]. At the Hanford site, MSPM methods have been used to monitor the radiological characteristics (alpha/beta/gamma emissions) of spent nuclear fuel [152]. However, real-time spectroscopic instruments have not been used for MSPM in nuclear-waste processing.

Outside of the nuclear field, MSPM is established with methods capable of inferring a faulty process state using measurement [153, 154]. Foundational methods that are commonly found in MSPM include multivariate cumulative sum, Hotelling T^2, and SPE. There have been enhancements to these traditional MSPM methods [155, 156, 157], but the base methods of Hotelling T^2 and SPE still have utility for fault detection in many systems including those monitored with spectroscopic sensors. Hotelling T^2 and SPE have been combined with spectroscopic sensors in pharmaceutical crystallization [158], coffee classification [159], polymer production [160], and pulsed-spray fluid-bed granulation [161].

Similar work using statistical process control with spectroscopic sensors (though not specifically Hotelling T^2 and SPE) has been used in pharmaceutical manufacturing [162], semiconductor manufacturing [163], and polymorphic transitions [164]. However, MSPM with spectroscopic sensors has not been demonstrated 1) in nuclear-waste processing and 2) in slurry systems with a variety of practical faults including sensor faults, mixing faults, and unanticipated chemistry.

In this chapter, the basic concepts behind feature-level data fusion and MSPM are reviewed, a fouling fault is classified as "expected" and its effect removed computationally, and experimental results are presented showing that the presently studied three instruments are robust in detecting a variety of sensor, mixing, and unanticipated chemistry faults in slurry systems. This chapter represents a major step forward in monitoring hazardous slurry processes by eliminating the need for offline monitoring and enabling the use of commercially available spectroscopic sensors to detect a variety of faulty process states remotely and in real-time.

7.2 Materials and Methods

The present study consists of three parts. First, a fouling fault on the surface of the ATR-FTIR probe appears in both normal slurries and slurries with other process faults; this fouling is removed using a technique established in Chapter 3: nonnegatively constrained classical least squares (NCCLS). Next, the Raman spectrometer, ATR-FTIR spectrometer, and FBRM instrument have their information concatenated and scaled. The dimensionality is reduced using PCA and the normal/faulty state is quantified using a three-sigma region in MSPM space (Hotelling T^2 and SPE). The effect of computationally removing the gibbsite fouling is investigated as it pertains to accurately classifying normal and faulty data. Lastly, the combination of the three sensors is compared to the ability of each individual instrument to distinguish faulty from normal slurries. Fault detection performance is measured in this study using accuracy, precision, recall, and F1-score.

7.2.1 Dataset Construction

Measurements were collected in a Mettler Toledo OptiMax vessel at 25°C and stirred at 400 RPM. *In-situ* Raman spectroscopy measurements were collected with a Mettler Toledo ReactRaman 785 using a 785 nm laser and 300 mW laser power at 1 s exposure time and 10 averaged scans. *In-situ* ATR-FTIR measurements were collected using a Mettler Toledo ReactIR 10 with one-minute sample collection time corresponding to 256 averaged scans. *In-situ* FBRM measurements were collected every minute with a Mettler Toledo ParticleTrack G-400 that had a laser speed of 2 $\frac{\text{m}}{\text{s}}$. All experimental conditions are listed in Table H.1, Table H.2, and Table H.3.

Gibbsite was the only chemical species that had significant partitioning into both soluble and insoluble phases; approximately 0.25 m dissolves at 25°C and high pH [165]. The starting solution was generated with water (solvent), sodium hydroxide (3.00 m), and gibbsite slightly below its equilibrium concentration (0.192 m); the starting solution was held at 37°C to increase the rate of dissolution. The gibbsite began as a white powder, and experiments only began after the starting solution became clear.

Three datasets (training, validation, and testing) were designed in Python and then constructed experimentally by adding pure components to create mixtures. Training data were representative of typical chemometric calibration and followed a Latin hypercube design-of-experiments; training data were used to determine directions of variation in the data (principal components). The validation data and test data were both constructed to be representative of a three-tank batch slurry system in series with masses (heels) of material computationally left behind to mix with subsequent batches. Validation data were analogous to historical process data that have been collected and are known to be fault-free; validation data were used to determine normal process statistics and develop a three-sigma uncertainty region for normal data. The test data were new data with an unknown state. A visual representation of the three datasets can be found in Figure 7.1.

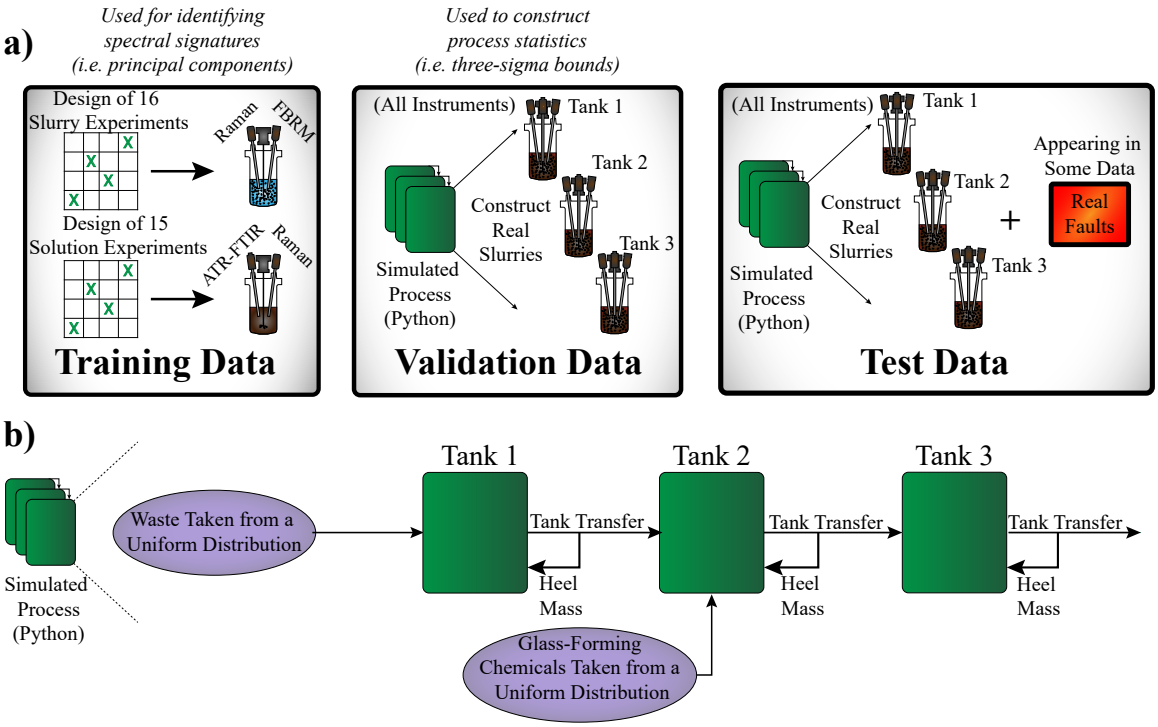


Figure 7.1: **a)** The distinction between training, validation, and testing datasets used in this chapter. **b)** The mass balance simulated in Python to inform experimental measurements.

Training Data

Training data were collected in two separate batches corresponding to the two different phases of the slurry: solution and suspended solids. Separating training data into two phases mimics a realistic scenario where limited resources or process knowledge may prohibit a comprehensive training dataset [21]. Principal components were found using the solution phase data (15 experiments) for ATR-FTIR, the insoluble phase data (16 experiments) for FBRM, and both datasets for the ball-probe Raman spectrometer that interrogated both phases. All training data were constructed using the bounds shown in Table 7.1 with the inclusion of glass-forming chemicals (GFCs). Scanning electron microscopy images of the amorphous solid particles at 1000x magnification indicated that particles were approximately 5 μm in diameter as shown in Figure 7.2. Notably, the particles are non-spherical and span a large range of characteristic particle lengths. Iron oxide (Fe_2O_3) appears as a fine powder at 1000$\times$ magnification.

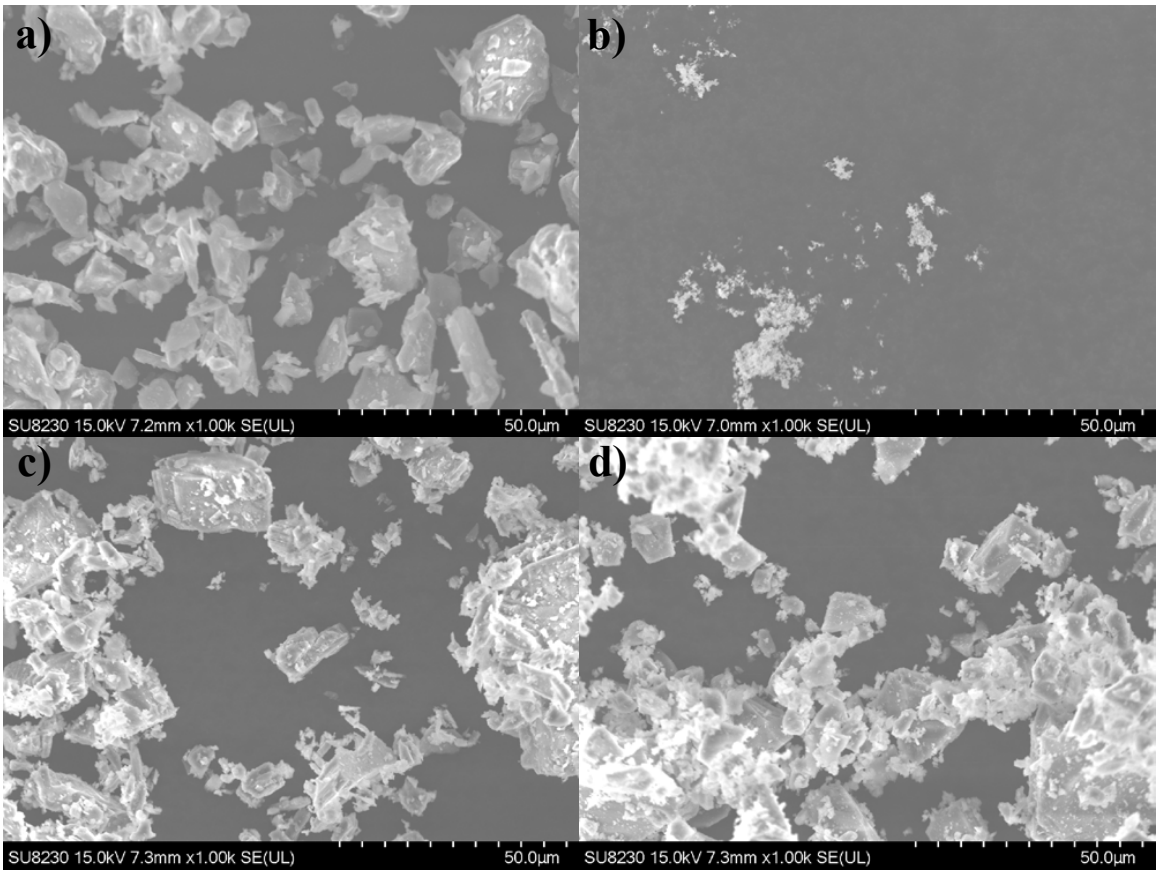


Figure 7.2: Images of the solid particles used in this study collected with a Hitachi SU8320 SEM at 1000× magnification: **a)** gibbsite, **b)** iron oxide, **c)** kyanite, and **d)** silica.

7.2.2 Process Data

Physical slurries were constructed and measured that matched results from a three-tank mass balance model constructed in Python as shown in Figure 7.1b. Waste was constructed from a uniform distribution and then GFCs from a separate uniform distribution are added in Tank 2; Tank 1 did not contain GFCs. One batch was followed from feed through the three-tank process. Unless explicitly tested as a fault, the simulation assumed that 80% of the mass was transferred during transfer to the next tank in series, while 20% of the mass remained behind as a heel mass [39, 32]. Bounds from Table 7.1 were used to construct an initial condition for each of the three tanks. The Tank 1 composition was randomly generated within bounds (Table 7.1), constructed, and measured as "Tank 1." Then, Tank 2 had its composition determined by taking 80% of the mass of Tank 1 and combining it with

Table 7.1: Bounds (in grams) used to construct both training and testing data.

Component	**Formula**	**Minimum (g)**	**Maximum (g)**	**Soluble**
water	H_2O	100.00	100.00	–
sodium hydroxide	NaOH	12.00	12.00	yes
sodium nitrate	$NaNO_3$	6.37	17.00	yes
sodium nitrite	$NaNO_2$	3.93	8.07	yes
sodium sulfate	Na_2SO_4	0.00	2.13	yes
iron oxide	Fe_2O_3	3.30	23.72	no
kyanite	Al_2SiO_5	1.28	8.41	no
silica	SiO_2	0.31	1.97	no
gibbsite	$Al(OH)_3$	3.29	23.65	partially
GFCs				
boric acid	H_3BO_3	5.00	13.42	yes
kyanite	Al_2SiO_5	5.00	15.00	no
silica	SiO_2	2.00	10.00	no

20% of the mass of the initial condition for Tank 2 and adding randomly generated GFCs (from Table 7.1). The result was physically constructed and measured as "Tank 2." The composition of Tank 3 was determined by taking 80% of the mass of Tank 2 and combining it with 20% of the mass of the initial condition for Tank 3. The result was constructed experimentally and measured as "Tank 3." This three-tank process was designed with 12 different initial conditions, leading to 36 total measurements. Of these 36 experiments, 19 were constructed with process faults (as detailed in the following section), leaving 17 experiments without process faults. All 19 faulty data were included in the test data. Three of the 17 normal experiments were set aside for test data (though not labeled as faulty) because of abnormal characteristics noticed during analysis (Experiments 21, 24, and 25 from Table H.3; discussed in Section 7.4.2). The remaining 14 normal experiments were randomly split to give a total of nine validation experiments and eight testing experiments.

7.2.3 Faults

Faults that may occur in slurry processing may cause significant deviation in in-line sensor readings, while others may be in the range of compositions listed in Table 7.1 while also producing no significant deviations in measured signal. The faults are listed below

with experiment numbers that correspond to data in Table H.3.

- Shifted pH (Experiment 8): May change the speciation of soluble species in solution. A single experiment was conducted where a pH shift from 13 to below the pKa of boric acid (approximately 9 at 25°C [166]) was observed. Lower pH causes a measureable shift in the ATR-FTIR spectrum for the borate anion [167].

- Unanticipated Species: May appear in a multicomponent system where the process feed stream is poorly characterized or if subsequent reactions occur during processing.

  - Three experiments were conducted that had 5.00 g, 4.00 g, and 3.20 g of soluble sodium carbonate (Na_2CO_3) added to otherwise typical slurries (Experiments 30 – 32).

  - Three additional experiments were conducted that had 5.00 g, 4.00 g, and 3.20 g of insoluble [19] zircon ($ZrSiO_4$) added to otherwise normal slurries (Experiments 33 – 35).

- Sensor Faults: The sensors monitoring the slurries can fail, producing abnormal readings.

  - Probe Fouling (All experiments to varying extent): May impact the signal observed by sensors in slurries. This behavior was observed on the ATR-FTIR probe in nearly all slurry measurements of this study and is investigated further in Section 7.3.1.

  - Sensor Failure: Spectroscopic sensors, like any sensor, may fail and produce abnormal measurements. This could be due to age, radiation exposure, incorrect sensor placement, or other factors. In this chapter, three sensor fault measurements were created by leaving individual sensors — Raman (Experiment 12), ATR-FTIR (Experiment 14), or FBRM (Experiment 10) — in the headspace of

the OptiMax vessel. Additionally, one additional sensor fault experiment was constructed by failing to shroud the Raman vessel during data collection (Experiment 11), which can increase sensor noise from sources such as cosmic rays [90, 20, 167].

- Stirrer Failure (Experiment 9): May be caused by mechanical failure or significant change in slurry flow properties. One experiment was conducted where the Optimax stirrer was stopped (400 rpm $\rightarrow$ 0 rpm) for five minutes and then measurements were collected.

- Change in Heel Composition (Experiments 16, 17, 19, and 20): May be caused by change in slurry rheology or unanticipated crystallization. Four experiments had their heel amount adjusted in the Python simulation from 80% to a value in the range 59% – 92%. The compositions of these experiments were still within the bounds defined by Table 7.1.

- Incorrect glass-forming chemical (GFC) Addition (Experiments 28 and 29): May appear as decreased GFC composition. Two experiments had no kyanite added as a GFC, though kyanite still appears in the waste feed. This resulted in slurries with less kyanite than expected after GFC addition, but that were still within the range expected in the tanks (Table 7.1).

- Unanticipated Recycle Behavior (Experiment 27): May appear as GFCs appearing in the feed stream. There was one experiment where 2.52 g of boric acid, which is only appearing as a GFC in this study, appeared prior to GFC addition (Tank 1). However, this experiment was within the bounds defined by Table 7.1.

7.2.4 Data Processing

High-dimensional data from Raman spectroscopy, ATR-FTIR spectroscopy, and FBRM were measured in the experiments performed. Raman spectroscopy measures the Raman

effect, where a photon (roughly 1 in 10^{10} incident photons) changes frequency depending on vibrational transitions of a molecule [90]; the Raman effect competes with fluorescence which can contribute significantly to the spectra of some materials, particularly slurries or powders. To reduce the effect of fluorescence when measuring slurries with our Raman spectrometer, signal smoothing and baseline subtraction were performed. The data were smoothed with a second-order Savitzky-Golay filter with a window size of 19 cm^{-1}. Baseline subtraction was achieved by taking the first derivative of the data with respect to wavenumber. Data were reported by our Raman spectrometer in the range 100 – 3200 cm^{-1} (dimensionality of 3101) and were reduced to the region where significant peaks were observed in the Raman training data: 200 – 1600 cm^{-1} (dimensionality of 1401).

ATR-FTIR data were reported in the range 649 – 2998 cm^{-1} (dimensionality of 631). For the ATR-FTIR instrument, feature selection was used to isolate the spectral region where known anions in the waste absorb in the Fourier transform infrared (FTIR) spectrum: 850 – 1450 cm^{-1} (dimensionality of 162). Then, computational removal of fouling on the probe tip was accomplished by applying NCCLS locally to the region of gibbsite appearance (847 – 1096 cm^{-1}) as shown in Section 7.3.1 [31]. Lastly, the ATR-FTIR spectra were processed with a second-order Savitzky-Golay filter with a window size of seven before having the first derivative taken with respect to wavenumber to eliminate potentially shifting spectral baselines.

FBRM data were constructed of bins where chord-length counts of solid particles were recorded. When measuring slurries of this study with particle diameters typically comprising 5 μm or less, the FBRM sensor recorded few measurements of large (greater than 100 μm) chord lengths. During data analysis, this effect became significant when scaling data; standard normal variate (SNV) scaling (as used in this chapter) introduced erroneous outlier data (greater than six standard deviations) in the sparsely populated features representing large chord lengths when a large measurement was recorded. This type of behavior may have been erroneously labeled a fault, when it was in fact a normal slurry measurement.

In practice, these large chord length features may provide valuable information about actual agglomeration or particle size changes. In the authors' opinions, three options may address these sparse measurements at some features: feature selection, data smoothing, or constructing a larger dataset. Feature selection was used in this chapter. The chord-length range 1 – 1000 μm (100 logarithmically-spaced features) was reduced to 1 – 89 μm (65 features).

The data from the three instruments at a single condition were concatenated and scaled according to SNV scaling. Following the matrix convention of the chemometric literature, the first dimension (row) contains experiments and the second dimension (column) contains the features. Concatenation was achieved by taking a sensor vector from Raman as $\mathbf{x}_{Raman} \in \mathbb{R}^{1\times 1401}$, a sensor vector from ATR-FTIR as $\mathbf{x}_{ATR-FTIR} \in \mathbb{R}^{1\times 162}$, and a sensor vector from FBRM as $\mathbf{x}_{FBRM} \in \mathbb{R}^{1\times 65}$ [168]:

$$\mathbf{x}_{unscaled} = \left[\mathbf{x}_{Raman} \frown \mathbf{x}_{ATR-FTIR} \frown \mathbf{x}_{FBRM}\right] \in \mathbb{R}^{1\times m} \tag{7.1}$$

where $\frown$ indicates concatenation and $m = 1628$ in this chapter but is left as m for generality. The data can then be SNV-scaled by subtracting the mean feature vector and variance matrix (both from the training data) as shown in Chapter A to give the concatenated and scaled input vector, $\mathbf{x} \in \mathbb{R}^{1\times m}$, for the fault detection algorithms used in this chapter.

7.2.5 Fault Detection Metrics

Hotelling T^2 and SPE have been studied extensively for analyzing chemical processes and are closely related to PCA [169, 170]. PCA decomposes a matrix into principal components and loading vectors by finding the vector that maximizes the projected variance in the data [92, 171, 93, 172]. PCA first requires the data to be mean-centered and often is applied after setting feature variance to unity by dividing by feature-specific variance (SNV scaling). All data in this section are assumed to have already undergone SNV scaling. Then, the covariance matrix, $\mathbf{C} \in \mathbb{R}^{m\times m}$, can be computed from the training data

$\mathbf{X}_{train} \in \mathbb{R}^{n \times m}$ where n are the number of samples and m are the number of features:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}_{train}^T\mathbf{X}_{train} \tag{7.2}$$

The principal components are the eigenvectors of the covariance matrix, $\mathbf{C}$ (i.e. columns of eigenvector matrix $\mathbf{V} \in \mathbb{R}^{m \times m}$) [59, 60]. The variance in the principal component direction is given by the eigenvalues of the covariance matrix, $\mathbf{C}$ (i.e. elements of diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{m \times m}$). The eigen equation is

$$\mathbf{CV} = \mathbf{V\Lambda} \quad \text{or} \quad \mathbf{C} = \mathbf{V\Lambda V}^T \tag{7.3}$$

To use PCA for fault detection, a new measurement vector, $\mathbf{x}$, can be projected from the original space of the data $\mathcal{O}$ into principal-component subspace $\mathcal{S}$ by right-multiplication with a truncated eigenvector (principal component) matrix ($\mathbf{V}_r \in \mathbb{R}^{m \times l}$) where the first l components are used to yield a projection $\mathbf{x}_s \in \mathbb{R}^{1 \times l}$ (i.e. the score vector).

$$\mathbf{x}_s = \mathbf{xV}_r \tag{7.4}$$

Hotelling T^2 quantifies the deviation of new data in the directions of the principal components. Hotelling T^2 arises naturally from the score vector, $\mathbf{x}_s$, and a truncated diagonal matrix ($\mathbf{\Lambda}_r \in \mathbb{R}^{l \times l}$) of the l largest eigenvalues of the covariance matrix ($\mathbf{C}$) to yield Hotelling T^2 [169, 173, 174, 175, 172]:

$$T^2 = \mathbf{x}_s\mathbf{\Lambda}_r^{-1}\mathbf{x}_s^T = \mathbf{xV}_r\mathbf{\Lambda}_r^{-1}\mathbf{V}_r^T\mathbf{x}^T \tag{7.5}$$

SPE (seen elsewhere as SPE statistic or Q-statistic) is calculated based on the reconstruction error of PCA, or the information orthogonal to the subspace defined by the principal components, $\mathcal{S}$ [169, 93]. The information lost when projecting $\mathbf{x}$ to the subspace $\mathcal{S}$ can be quantified by calculating $\hat{\mathbf{x}} \in \mathbb{R}^{1 \times m}$, which is the vector $\mathbf{x}$ projected to the subspace

$\mathcal{S}$ and then projected back into the original space, $\mathcal{O}$. The reduced eigenvector matrix, $\mathbf{V}_r$, has orthonormal columns and therefore $\mathbf{V}_r^T\mathbf{V}_r = \mathbf{I}_l$ where $\mathbf{I}_l \in \mathbb{R}^{l\times l}$ is the $l \times l$ identity matrix. Making use of this identity, SPE and $\hat{\mathbf{x}}$ can be calculated as [170, 172, 156]

$$\hat{\mathbf{x}} = \mathbf{x}_s\mathbf{V}_r^T = \mathbf{x}\mathbf{V}_r\mathbf{V}_r^T \tag{7.6}$$

$$SPE = (\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^T \tag{7.7}$$

Hotelling T^2 and SPE were calculated for (historical) validation data and a three-sigma confidence region constructed. Then, Hotelling T^2 and SPE were calculated for test data and compared to the expected three-sigma region. Data outside of the three-sigma region were identified as faulty. The fault detection performance is assessed through four typical fault detection metrics: accuracy, precision, recall, and F1-score. These values quantify the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions from the fault detection scheme.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.8}$$

$$Precision = \frac{TP}{TP + FP} \tag{7.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{7.10}$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{7.11}$$

7.2.6 ATR-FTIR Quantification Model

Partial least squares regression (PLSR) is a supervised learning technique that accurately predicts collinear with few training experiments; additional information on the PLSR

model can be in Chapter B [103, 102, 101]. A PLSR model with four components was used to quantify all ATR-FTIR spectra. The model was calibrated using the training data shown in Figure 7.1a; the model then was used to predict the mass of anions as a fraction of the total slurry ($wt\%^{slurry}$) and as a fraction of the total solution ($wt\%^{solution}$). For an experiment j and chemical component i, these quantities were defined as

$$wt\%_{i,j}^{slurry} = \frac{m_{i,j}}{\sum_{k=1}^{a} m_{k,j}} \tag{7.12}$$

$$wt\%_{i,j}^{solution} = \frac{m_{i,j}}{\sum_{k=1}^{b} m_{k,j}} \tag{7.13}$$

where $m_{i,j}$ is the mass of component i in experiment j, a is the total number of components in the slurry, and b is the total number of components in the solution phase. Root mean squared error (RMSE) was used to quantify the performance of the PLSR quantification model:

$$\text{RMSE} = \sqrt{\frac{1}{pq}\sum_{i=1}^{q}\sum_{j=1}^{p}(wt\%_{i,j} - \hat{wt}\%_{i,j})^2} \tag{7.14}$$

where $\hat{wt}\%_{i,j}$ is the model-estimated solution phase weight percent, q is the number of species predicted by the model, and p is the number of experiments that are being predicted. $wt\%_{i,j}^{solution}$ is used for all predictions except for subsection 7.3.3 where $wt\%_{i,j}^{solution}$ and $wt\%_{i,j}^{slurry}$ are compared.

7.2.7 Computation

All data and code can be found on GitHub[1]. Computation was performed in Python 3.10.13. All packages used for computation include NumPy 1.26.4, scikit-learn 1.5.2, SciPy 1.14.1, and PyMCR 0.5.1.

[1] https://github.com/magrover/MSPM-Fault-Detection

7.3 An Expected Fault – Fouling

This study distinguishes the fouling fault from all other faults: it appeared in nearly all slurry measurements collected. Because this known fault appeared in nearly all measurements but did not represent a hazardous condition, the effect of the fouling was computationally removed so that other process faults could be accurately identified.

7.3.1 Gibbsite Fouling

Gibbsite fouling was suggested by analysis of the validation data. Some validation data (but not all) contained an additional peak around 1020 cm^{-1} (shown in Figure 7.3) that did not correspond to any of the known solution species: nitrate, nitrite, sulfate, borate, or gibbsite which has no observed solution-phase peak in the FTIR spectrum.

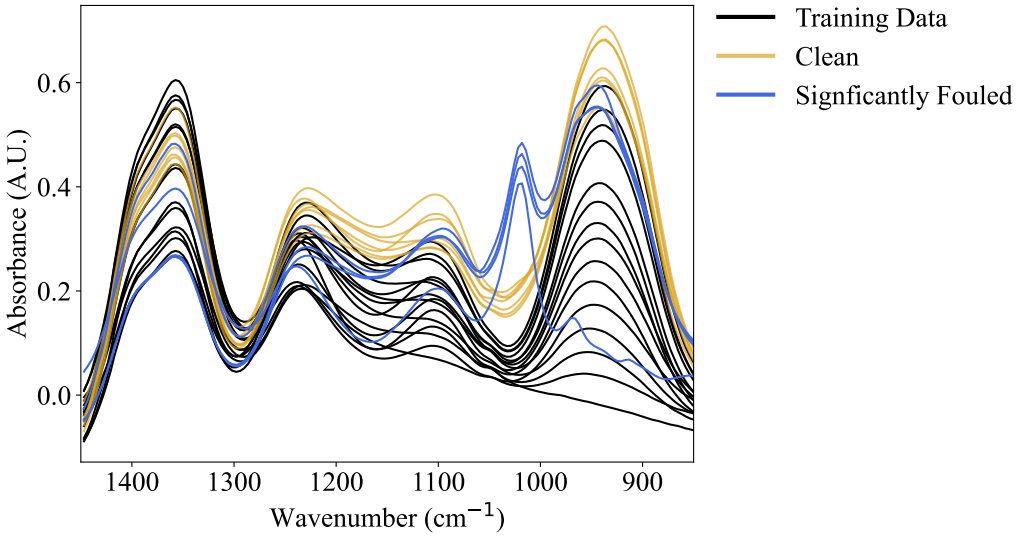


Figure 7.3: ATR-FTIR spectra of training data (black) and validation data that did not have a significant gibbsite fouling peak appear (yellow) and validation data that did have a gibbsite fouling peak appear (blue).

To isolate the observed fouling behavior, the ATR-FTIR spectrometer was used to analyze a slurry that comprised 100.00 g H_2O, 12.02 g NaOH, and 0.0247 g undissolved gibbsite ($Al(OH)_3$) at 25°C. A gibbsite peak was observed which increased with time, as seen in Figure 7.4. Notably, if the probe was cleaned and reinserted into the solution, the

peak in Figure 7.4 disappeared. Thus, this peak was inferred to be solid deposition on the probe surface. At the resolution of the ATR-FTIR instrument, the gibbsite fouling peak was observed with a peak maximum between 1018 cm^{-1} and 1022 cm^{-1}. A smaller peak overlapping with boric acid showed a peak centered between 966 cm^{-1} and 970 cm^{-1}. These peaks are reported elsewhere in studies measuring FTIR spectra of gibbsite and potassium bromide suspensions [176, 177]. The peaks located near 1020 cm^{-1} and 967 cm^{-1} are both attributed to OH bending vibrations of surface hydroxyl sites found in gibbsite ($Al(OH)_3$). It has been shown elsewhere that gibbsite particles display strong repulsive forces from other gibbsite particles in strong electrolyte solutions [178]. However, fouling of diamond-tip probes with gibbsite slurries appears previously unreported.

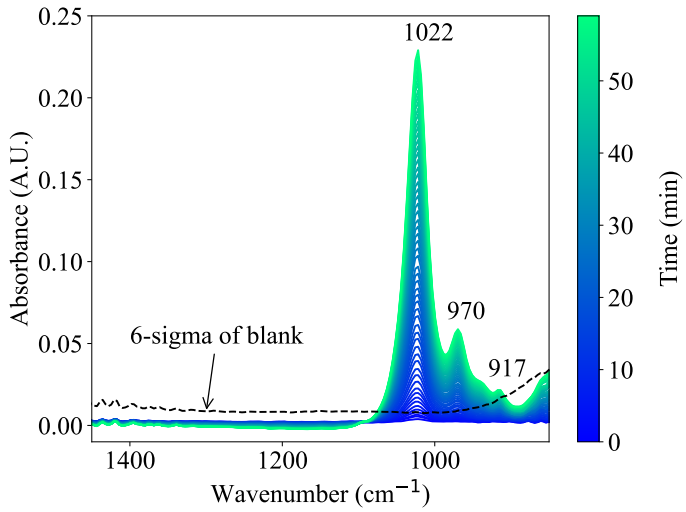


Figure 7.4: Gibbsite fouling in the ATR-FTIR spectrum as a function of time since introduction of gibbsite in minutes. The dashed black line indicates six standard deviations of 40 "blank" spectra consisting of 100.00 g H_2O and 12.02 g NaOH. The spectra shown are measured from the same solution with 0.0247 g undissolved gibbsite ($Al(OH)_3$) added.

7.3.2 Computational Removal of Gibbsite Fouling – NCCLS

Since gibbsite fouling appears in nearly all spectra, even those representing normal process conditions, the contribution of the gibbsite peak was removed using NCCLS for the remainder of this study except for comparison in Section 7.4.1. Justification for the

choice of NCCLS is provided in this section.

The three best-performing methods from Chapter 3 were chosen for removing gibbsite fouling computationally: NCCLS, blind source separation (BSS) with PCA, and BSS with independent component analysis (ICA) [31]. The performance of these methods can be seen in Figure 7.5. All three methods removed varying amounts of the aforementioned gibbsite peak. The effect of the gibbsite removal was quantified via the parity plots shown in Figure 7.6. Here, RMSE is used to quantify accuracy of the predictions. As shown in Figure 7.6b, NCCLS was the best-performing method in this study and improved quantification compared to that performed with no gibbsite removal (Figure 7.6a).

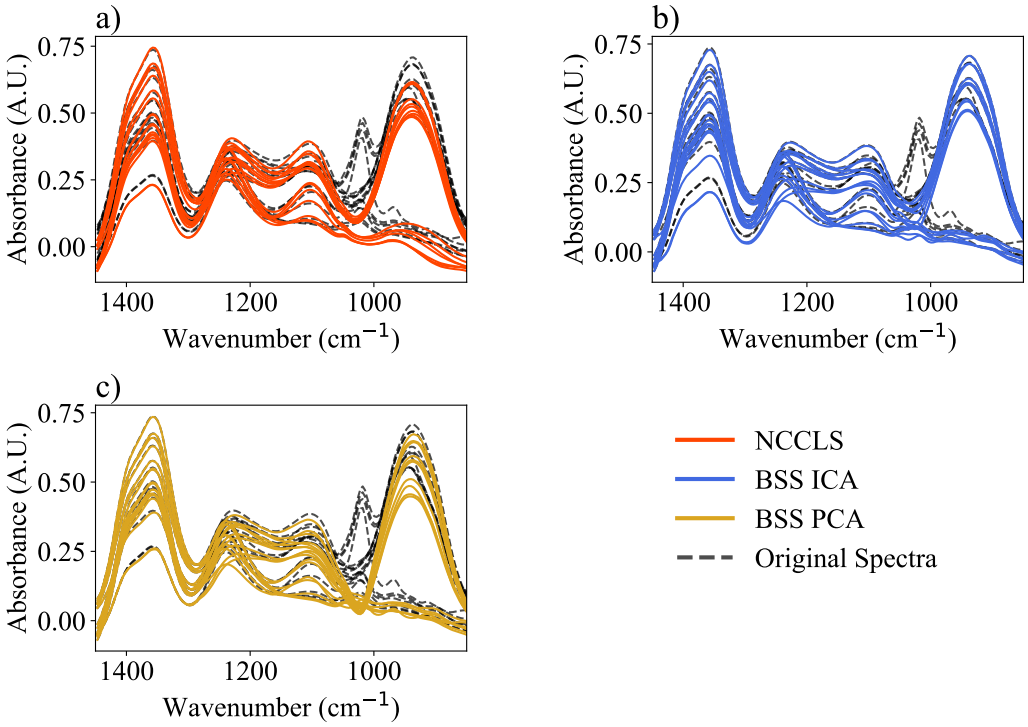


Figure 7.5: ATR-FTIR spectra of normal process data with the gibbsite fouling occurring around 1022 cm^{-1} where the fouling is removed with **a)** NCCLS, **b)** BSS ICA, and **c)** BSS PCA.

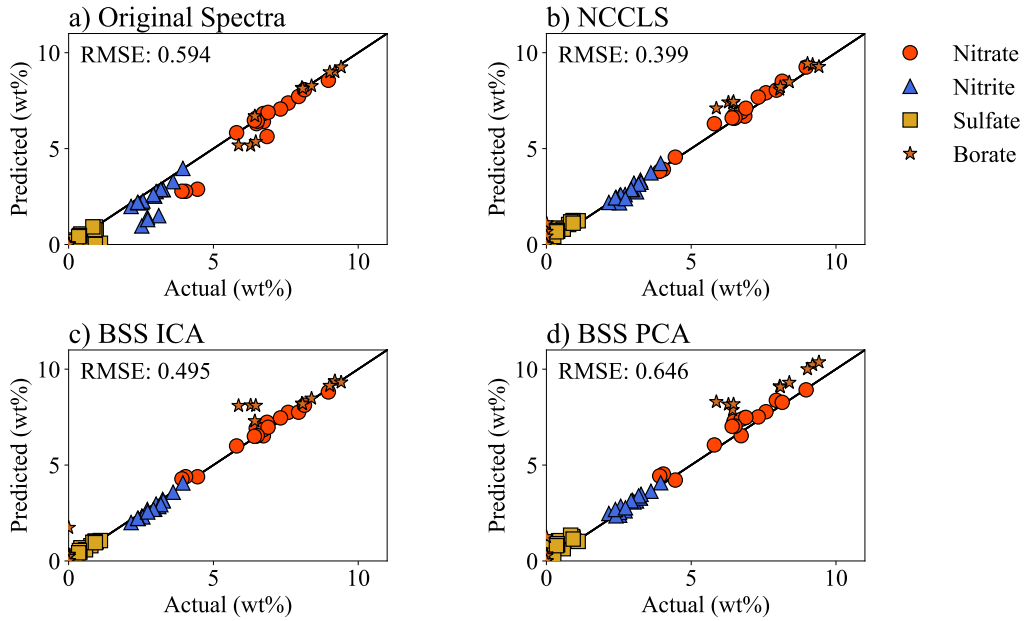


Figure 7.6: Quantification of four soluble anions in fault-free slurries (validation and test data) with a PLSR model with **a)** no preprocessing and computational removal of the gibbsite fouling provided by **b)** NCCLS, **c)** BSS ICA, and **d)** BSS PCA [31].

7.3.3 Solution-Phase Training Data

The ATR-FTIR PLSR quantification model was more accurate when predicting $wt\%^{solution}$ rather than $wt\%^{slurry}$. As shown in Section 7.2.1, training data for the ATR-FTIR instrument had no insoluble solids while the validation and testing data were multiphase slurries. Calibrating the sensor model using $wt\%^{solution}$ (Figure 7.7a) allowed ATR-FTIR training data to accurately predict compositions despite the differences between training data (solutions) and process data (slurries). This is supported by the decreased RMSE when using $wt\%^{solution}$ instead of $wt\%^{slurry}$ ($0.399 < 1.409$). Figure 7.7c shows that using $wt\%^{slurry}$ introduces a model bias that is not present when predicting just the solution phase ($wt\%^{solution}$). The results presented in Figure 7.7 were attributed to the short path length of the attenuated total reflectance (ATR) crystal; suspended solids were not effectively interrogated by the laser [179, 53]. This suggested that using $wt\%^{solution}$ should have made ATR-FTIR quantification independent of solids concentration. From the data presented however, there is no clear indication that model error increases with increasing

amounts of insoluble solids.

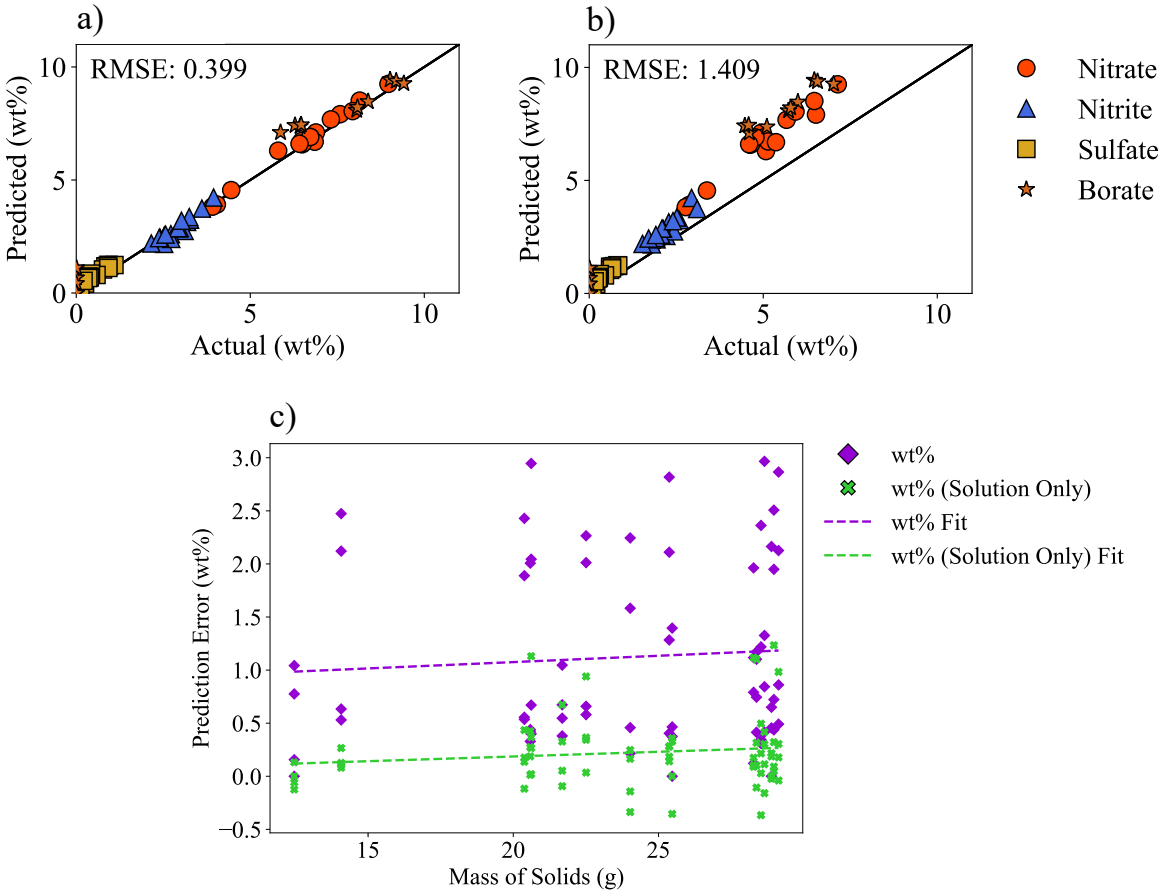


Figure 7.7: Parity plots of fault-free validation and testing slurries comparing predicted wt% from the ATR-FTIR spectrometer after NCCLS preprocessing to gravimetrically measured wt%. The value of wt% is calculated from **a)** only the solution phase ($wt\%^{solution}$) and **b)** from the entire slurry ($wt\%^{slurry}$). Also shown **c)** are model residuals (prediction error) as a function of solids concentration for the two parity plots shown.

7.4 Fault Detection with Multivariate Statistical Process Monitoring

Because training data were collected from two separate phases while validation and test data were collected from combined slurries, it was not apparent that training data would be representative of the validation and test data. The SNV-scaled training and validation data are presented for comparison in Figure 7.8 with NCCLS removal and Figure 7.9 without. The validation data differed from the training data primarily through the appearance of the gibbsite peak (1037 cm^{-1} in Figure 7.9) in the ATR-FTIR spectrum. In Figure 7.8e, the

gibbsite peak has been removed via computational methods and is no longer visible. The range of values observed in the validation data in Figure 7.8 (-3 to 3) was similar, though not identical, to that observed in the training data. This indicated that using disjoint training datasets for the two phases was sufficient for extracting sources of variation that appear in the full slurry.

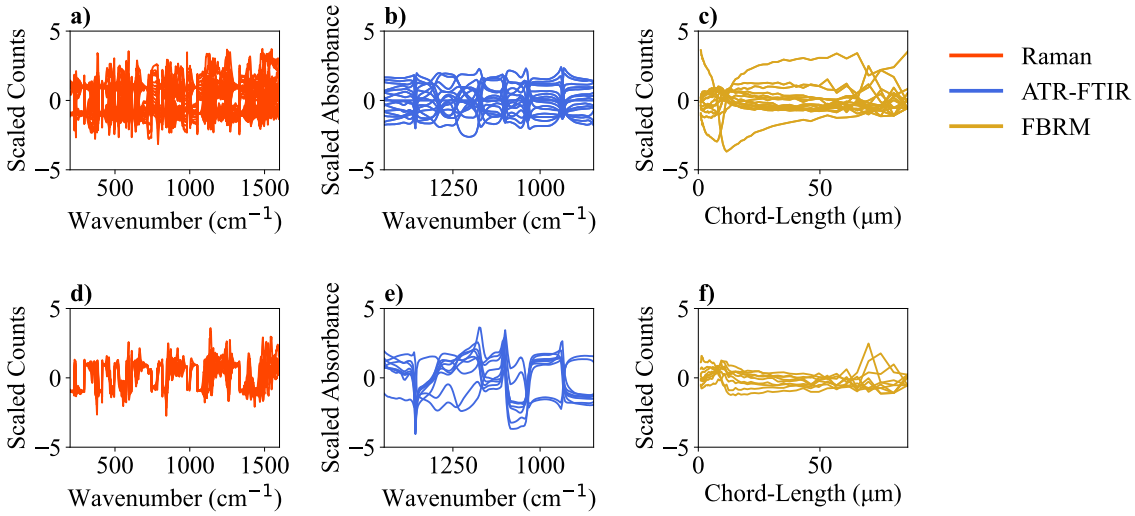


Figure 7.8: Training compared with validation data; no faults appear in either dataset except for probe fouling that has been removed using NCCLS in the validation data. The top row (**a**, **b**, and **c**) are training data while the bottom row (**d**, **e**, and **f**) are validation data. Raman (**a** and **d**), ATR-FTIR (**b** and **e**), and FBRM (**c** and **f**) data are shown.

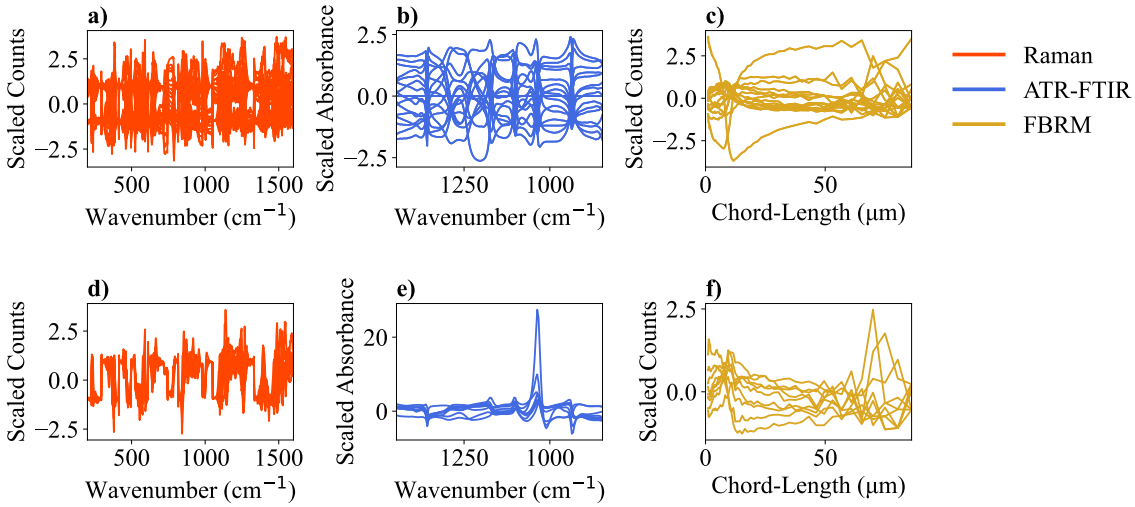


Figure 7.9: Training compared with validation data. No NCCLS was performed to remove the gibbsite fouling peak from the ATR-FTIR spectra. The top row (**a**, **b**, and **c**) are training data while the bottom row (**d**, **e**, and **f**) are validation data. Raman (**a** and **d**), ATR-FTIR (**b** and **e**), and FBRM (**c** and **f**) data are shown.

7.4.1 Statistical Identification of Faults

Figure 7.10a shows Hotelling T^2 plotted with SPE for a concatenated vector of Raman, ATR-FTIR, and FBRM data. A three-sigma region was created based on the validation (historical) data, and then all data were assessed as being inside or outside the range of normal operation. Training data may be used to generate expected process statistics instead (shown in Figure I.1) but did not affect the results of this study. Using (historical) validation data as was done here may be more robust for fault detection when significant differences exist between training and process data. For the present study, all normal data (blue) were within the three-sigma boundary except for two test data that exceeded the interval on the SPE axis. These anomalous points required further investigation into their false positive prediction, as investigated in Section 7.4.2.

When gibbsite was *not* removed from process data, the fault detection accuracy was negatively impacted with some faults — unanticipated zircon, changing pH, and failing to shroud the vessel — no longer detected as shown in Figure 7.10b. This result was quantified in Table 7.2; removing the fouling computationally improved the fault detection metrics of

accuracy, precision, recall, and F1-score for all non-fouling faults. While the placement of some data changed in Figure 7.10b compared to Figure 7.10a, the most prominent effect was the increased size of the three-sigma normal region; there was a greater degree of overlap between faulty and normal data without the removal of the gibbsite fouling fault.

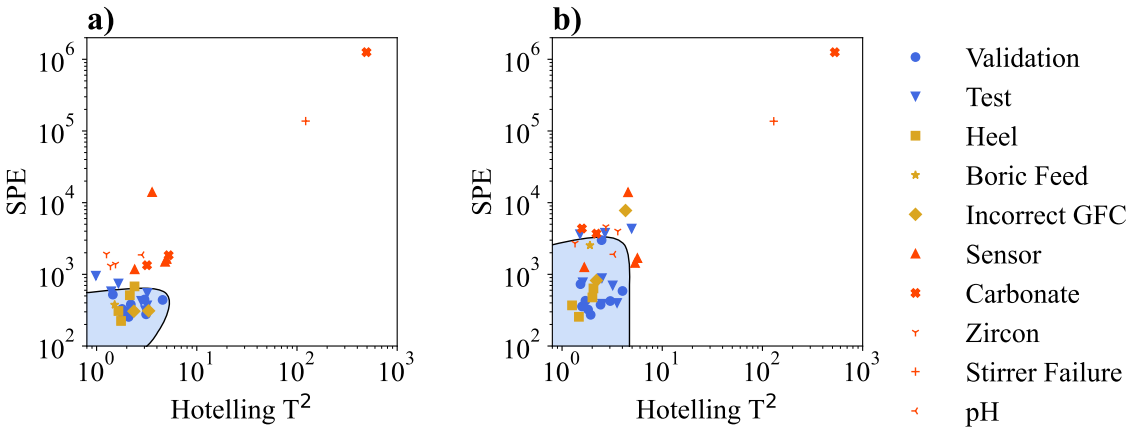


Figure 7.10: Hotelling T^2 and SPE of all process slurries **a)** with computational removal of the gibbsite fouling peak and **b)** without any removal are shown on logarithmically-scaled axes. A three-sigma confidence interval (ideally covering 98.9% of Gaussian data in two dimensions) based on validation data is shown in blue. Data without any known faults are displayed in blue. Data with process faults that have normal composition according to Table 7.1 and do not produce abnormal sensor measurements are shown in yellow. Data with the remaining process faults are shown in red.

Table 7.2: Table of error metrics comparing fault detection with the original data vs. data that has had the gibbsite fouling computationally removed from the ATR-FTIR spectra using NCCLS.

Metric	**Original**	**Gibbsite Removal**
Accuracy	0.611	**0.722**
Precision	0.769	**0.867**
Recall	0.476	**0.619**
F1-Score	0.588	**0.722**
Average	0.611	**0.732**

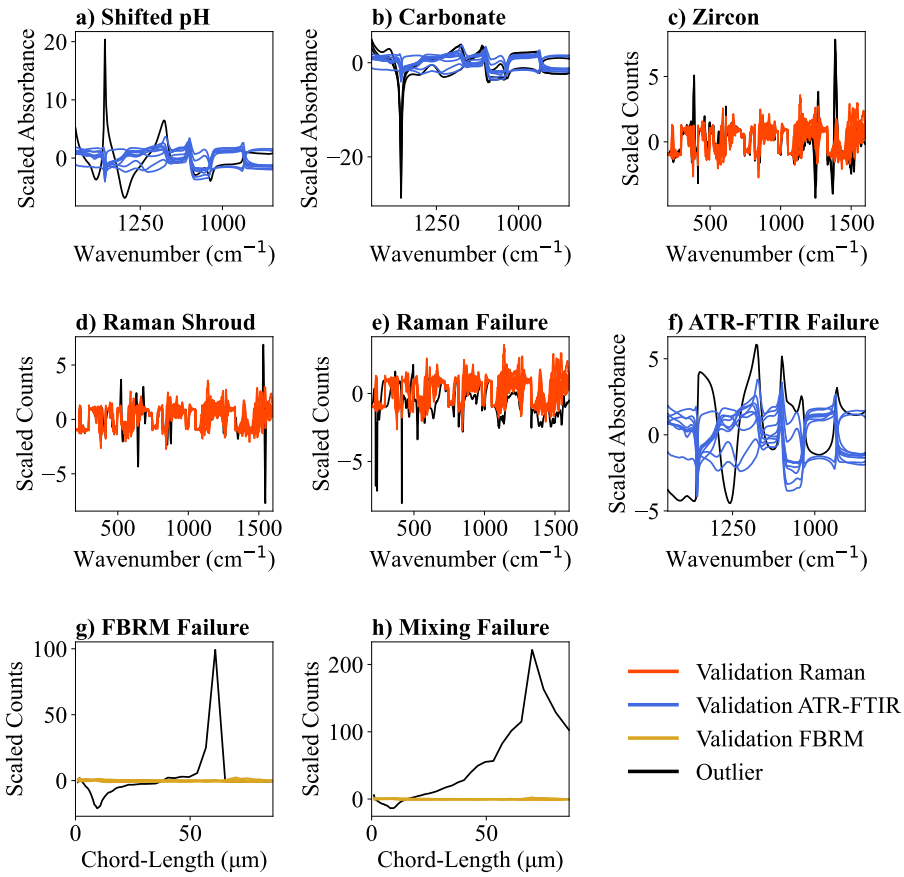


Figure 7.11: Sensor data (after SNV scaling) that leads to outliers in Hotelling T^2 and/or SPE. Normal validation data are shown for comparison. Faults are: **a)** the pH of the solution becoming more acidic, **b)** sodium carbonate in the waste, **c)** zircon in the waste, **d)** the Raman probe being unshrouded, **e)** the Raman probe being not immersed **f)** the ATR-FTIR probe not being immersed, **g)** the FBRM probe not being immersed, and **h)** the stirrer being stopped.

The faulty data are classified into two categories based on the sensors' ability to detect that a fault has occurred. Detected faults (red data) in Figure 7.10a were identified correctly and were located outside of the three-sigma confidence interval. "Faulty" data that did not produce abnormal compositions nor instrument measurements (yellow data) lie within the three-sigma confidence region in Figure 7.10a. Figure 7.11 shows the correctly identified faulty sensor data after feature selection, Savitzky-Golay filtering, and scaling. A list of the different faults, their process relevance, and their relationship to Hotelling T^2 and SPE is shown below.

- Shifted pH: Altered borate speciation at shifted pH was detectable with ATR-FTIR [167]. The pH of the slurry was not measured during the experiment; however, the effect of boric acid speciation as a function of pH is shown by Su and Suarez [180, 181]. Because the altered borate speciation affected a large region of the ATR-FTIR spectrum (800 – 1500 cm^{-1}), the effect did not appear in the training data and was classified as a statistical outlier via the SPE axis and is shown in Figure 7.11a.

  The experiment that was expected to have had a pH shift was not a designed fault and rather occured in a slurry that was expected to behave typically (Experiment 8 in Table H.3). The effect of shifted pH was not reproduced by any other experiment and was not reproduced when the exact experimental conditions were replicated using the same instrumentation. Therefore, the fault cannot be confirmed to be due to a pH shift. However, the disappearance of the borate peak is consistent with a decrease in pH beyond the pKa of boric acid. Our best hypothesis as to why this *particular* experiment had a pH shift is because the sodium hydroxide from the creation of a new batch of starting solution (water, sodium hydroxide, and gibbsite) may have not fully dissolved by the start of this experiment. This experiment was the first conducted with a new batch of starting solution.

  Further, pH-dependent ATR-FTIR spectra of borate and boric acid were collected and shown in Figure 7.12. It is shown that boric acid has its largest peak at 1410 cm^{-1}, while the borate anion has had its largest peak reported at 955 cm^{-1} [180] and 945 cm^{-1} [181]. In this study, it is measured at 936 cm^{-1}, which may be an effect of the concentrated electrolyte solution at 3 m NaOH.

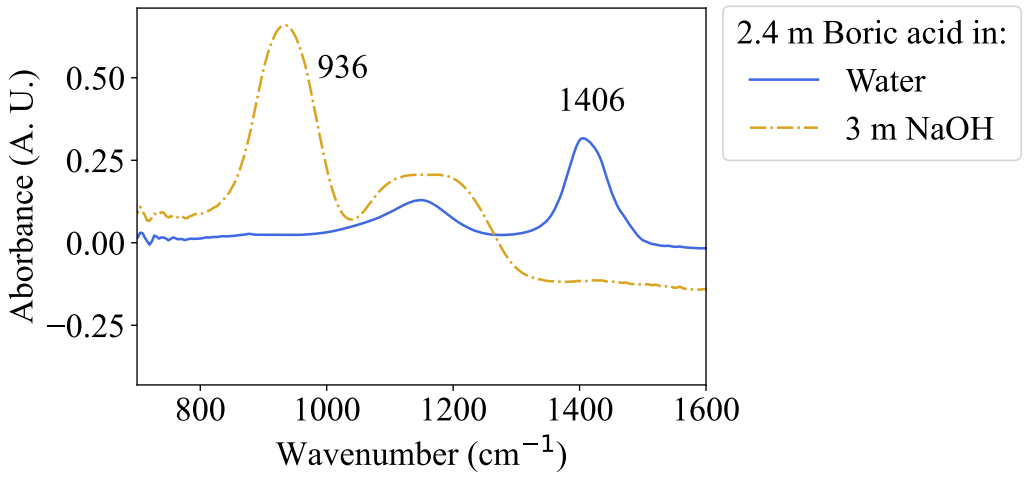


Figure 7.12: Background-subtracted (solvent-removed) references for boric acid in water and the borate anion in 3 m NaOH.

- Unanticipated species: All measurements of slurries containing new components were classified as faults on the SPE axis, owing to the appearance of new signals that are orthogonal to the training data.

  - Unanticipated carbonate: the carbonate peak appeared in the ATR-FTIR spectrum at 1384 cm^{-1}. This peak was not removed with the computational fouling removal because of the local (847 – 1096 cm^{-1}) application of the computational removal method in this chapter. In the first-derivative FTIR spectrum, the carbonate peak was observed at 1357 cm^{-1} in Figure 7.11b. Despite overlapping with the nitrate peak in the ATR-FTIR spectrum [21], the slurry with carbonate was correctly identified as a fault. The experiment that was also an outlier on the Hotelling T^2 axis (Experiment 30 in Table H.3) had an FBRM spectrum with larger chord-lengths detected (Figure 7.13). This experiment did not have an otherwise abnormal composition nor were larger particles observed via an in-line microscope.

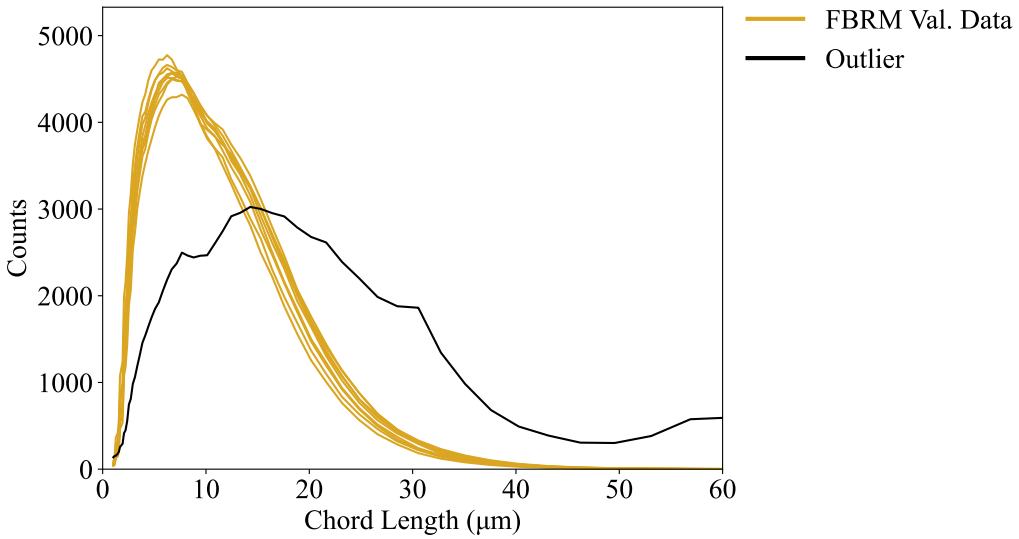


Figure 7.13: The carbonate fault with high Hotelling T^2 (Experiment 30 in Table H.3 was observed to have abnormal FBRM data for the duration of the experiment.

  - Unanticipated zircon: The zircon peak appeared in the Raman spectrum at 1377 cm^{-1} [19]. Zircon has prominent peaks in the Raman spectrum (Figure 7.11c), so it appeared as an outlier on the SPE axis since the signal did not appear in the training data and therefore did not appear in the principal components.

- Heel composition: This fault was not linearly separable from training data using Hotelling T^2 and SPE. These data produced statistically similar measurements to the training and validation data, so there was no distinguishing feature to separate these data from normal operation. In other words, this fault did not lead to out-of-spec composition nor measurement, so it was correctly not identified as a fault.

- Boric acid in feed stream: This fault was not classified as an outlier from normal data on the Hotelling T^2 and SPE axes. However, the composition of this tank was out-of-spec for Tank 1, which has no boric acid present in this study. However, since the data from all three tanks were not distinguished, the appearance of boric acid in Tank 1 was not flagged as an outlier. If more data of each tank were available, a separate normal operating region could be constructed for each tank. Additionally, it

may be feasible to utilize a model-based approach utilizing spectra quantification to detect faults from composition information [32].

- Incorrect GFC addition: This fault was not linearly separable from training data using Hotelling T^2 and SPE. These data were not statistical outliers based on measurements alone, nor were the compositions of the resulting experiments outside of the bounds of Table 7.1. However, if the Raman sensor of this study estimated slurry composition, this fault may have been detectable with a model-based approach as done in Chapter 6 in this thesis [32].

- Sensor failure: All sensor failures appeared outside the three-sigma confidence region by significant margins on the SPE axis and not the Hotelling T^2 axis. This suggested that the instrument failure types studied here were orthogonal to the training data, rather than having been significant variation in the direction of the principal components. This conclusion was supported by the instrument data shown in Figure 7.11d–g. Given the currently studied axes, these sensor faults appeared in similar regions as significant process faults, which may make fault diagnosis more challenging based on the methods presented here.

- Stirrer failure: A significant change in FBRM measurement with fluid velocity at the probe tip affected the measured chord-length distribution from the FBRM's spinning laser. The sensor observed a significant increase in counts at large chord-lengths, as shown in Figure 7.11h. Our stirrer had an outer diameter of 38 mm and an angular velocity of 400 rpm, giving a linear velocity at the impeller tip to be 0.80 $\frac{\text{m}}{\text{s}}$. The scan speed of the FBRM laser for our instrument was 2 $\frac{\text{m}}{\text{s}}$. For the particle sizes studied (approximately 5 μm), a Stoke's law study shown in Figure 7.14 suggested that particles in the system were still suspended during mixing failure measurement, taken here five minutes after the stirrer is stopped. Published literature conflict on the dependence of FBRM on fluid velocity with some authors reporting negligible changes

and others reporting significant changes [182]. This is likely a system-dependent phenomenon, but was detectable here as a significant outlier on the Hotelling T^2 and SPE axes. Exceeding normal operating regions on both axes indicated that unanticipated variation was observed in the direction of principal components *and* in the direction orthogonal to the principal components.

In Figure 7.14, the settling velocity and time are shown for a spherical silica particle as a function of particle radius. In addition to this calculation, experimental results showed that the altered FBRM result occurred in under 60 s of the stirrer being stopped, and did not change significantly when waiting five minutes. The FBRM data after five minutes was used in Experiment 9 in Table H.3. The calculated time for a 5 µm silica particle to fall 2 cm (approximation of sensor depth in slurry) was calculated to be 790 s (13.2 min).

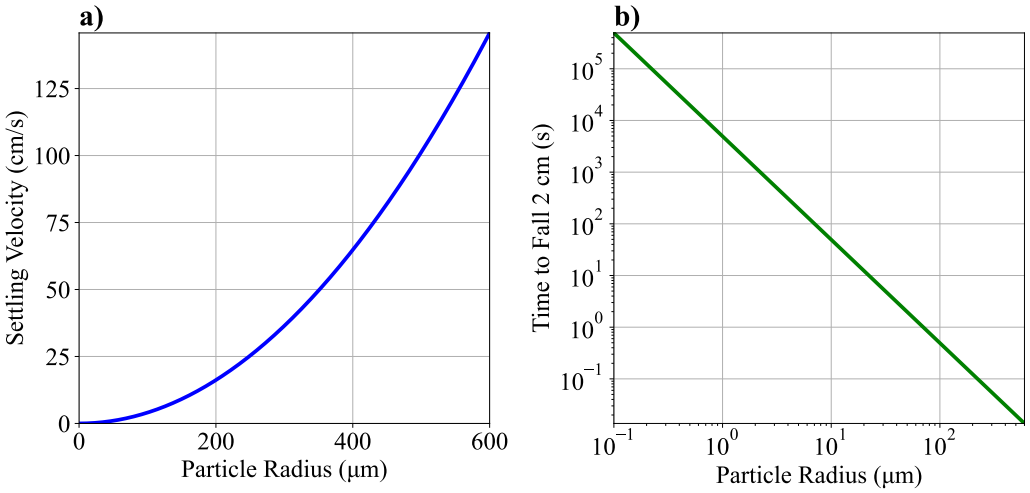


Figure 7.14: Stoke's Law for a spherical silica particle in water at 25°C. The **a)** settling velocity and **b)** 2 cm settling time are shown as a function of particle radius.

Based on the faults studied here, the SPE axis appeared more useful for detecting faults in sensors and unanticipated chemistry. The Hotelling T^2 axis provided no substantial benefit in this study, though it may have been flagged if compositions outside of the bounds provided in Table 7.1 were studied (i.e. an experiment has more sodium nitrate than typical).

A confusion matrix showing the number of correctly and incorrectly classified data is shown in Figure 7.15. Intuitively, all false negatives (type II errors) were a result of faults that did not produce outliers in the composition nor measurement. There were two unanticipated false positives (type I errors). In addition, the test point in Figure 7.10a that has part of its marker outside of the three-sigma confidence region was investigated as well since it was almost classified as a fault. Both false negatives and false positives are discussed in the following two sections.

Figure 7.15: Confusion matrix showing classes of predicted and actual faults in process data based on all process data from the study and corresponding to Figure 7.10a.

7.4.2 False Positives (Type I Errors)

As shown in the previous section (Figure 7.10a), there were two experiments that were classified as outliers, but had no fault associated with them. These data were false positives. In the context of nuclear-waste processing, these events did not represent hazardous conditions, but may have delayed processing until the cause of the abnormal reading was successfully identified using other techniques, such as grab-sampling. These data were outliers on the SPE axis but not the Hotelling T^2 axis, suggesting that a signal orthogonal to (i.e. not in the plane of) the training data appeared in these data. Simply, the chemical mixture was not outside normal bounds, but something new occurred. The data from these

two outliers were investigated to determine how the faulty classification resulted from the sensor measurement.

The two outlier data (Experiments 24 and 25 in Table H.3), as well as one datum that was on the edge of the normal region (Experiment 21 in Table H.3), are plotted in Figure 7.17 along with the validation data. The scaled Raman data in Figure 7.17a suggested that there were new peaks appearing in the range 380 – 390 cm^{-1} and at 415 cm^{-1}. In the unfiltered and unscaled Raman spectrum shown in Figure 7.16, there were no peaks in the range 380 – 390 cm^{-1} and 415 cm^{-1} corresponded to the sapphire probe window. However, the first derivative Raman spectra without SNV scaling (Figure 7.17b) showed that these outliers were caused by subtle shifts in signal *slope* in these regions. The validation data had little variation, shown by the closely overlapping signals, causing the fault detection algorithm to be sensitive to spectral differences in this region. To address this situation, feature selection in the Raman spectrum could be employed to remove these sensitive features from flagging a faulty condition when the process condition is normal, or by constructing a richer training set.

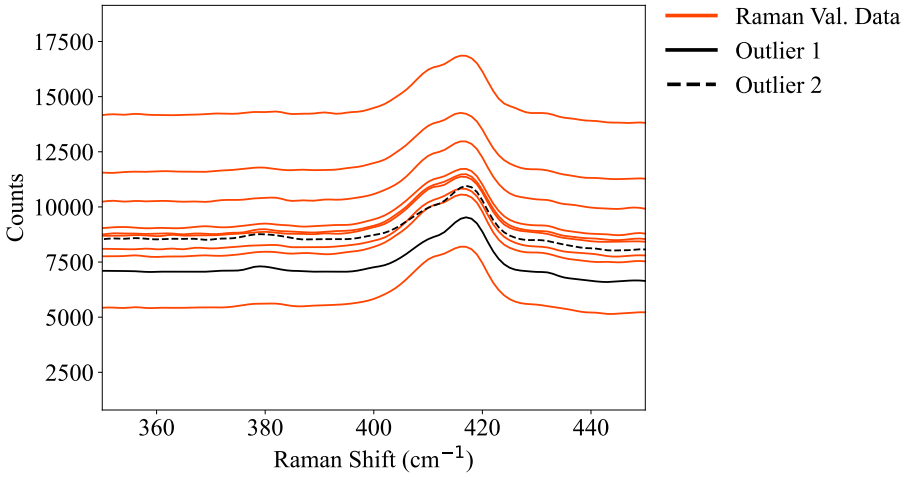


Figure 7.16: Unscaled and unfiltered Raman spectra of all validation data with the two outliers from the main text highlighted in black.

The one datum that was almost an outlier in the SPE axis in Figure 7.10a is referenced as "near outlier" in Figure 7.17c and Figure 7.17d. The data in Figure 7.17c associated

with the outlier was caused by a shift in the FBRM data to larger chord-lengths for this particular experiment (shown in Figure 7.17d). Out of the experimental space studied, this measurement had the second-lowest amount of total solids, while also having a relatively large amount of insoluble gibbsite and relatively small amount of silica. This experiment was nearly detected as a fault because of its composition was on the edge of the experimental space studied leading to a unique FBRM signal.

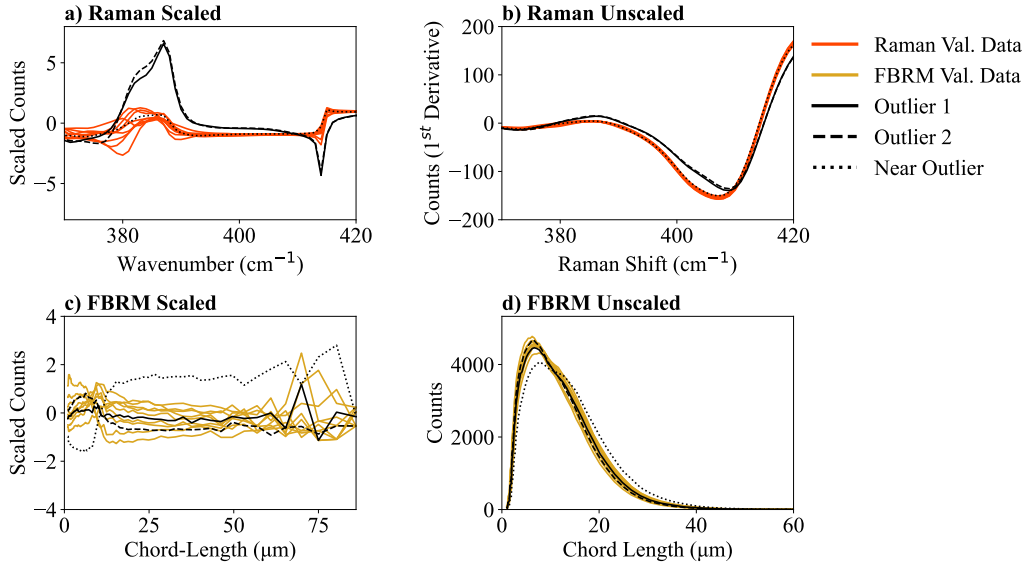


Figure 7.17: Sensor data corresponding to two false positives and one near-false positive. Validation (normal) data are plotted with the outliers for reference. The first column shows SNV-scaled data for **a)** Raman and **c)** FBRM. The second column shows **b)** unscaled Raman spectra after taking the first derivative with respect to wavenumber and **d)** unscaled FBRM data.

7.4.3 False Negatives (Type II Errors)

In nuclear-waste processing, false negatives represent significant process hazards; an unidentified faulty process may lead to significant process downtime or hazardous conditions. There were three conditions that were undetectable using the instrumentation and methods of this study: change in heel composition, boric acid in the feed stream, and incorrect GFC addition. Since these conditions did not produce measurement outliers in the statistical sense, a non-data-driven approach, i.e. a process model or soft sensor, may pro-

vide another axis in which to separate these data. Interestingly, all three of these process faults did not create abnormal chemical compositions. However, composition estimation through spectra-to-composition models or mass balance models may allow these events to be detected. Companion results and error metrics that do not include the undetectable faults of this chapter are shown in Figure I.2.

7.5 Data Fusion vs. Single-Instrument Fault Detection

Fault detection with individual instruments was compared to the array of sensors presented in the previous section. These results are presented in Figure 7.18 and visualized in confusion matrices in Figure 7.19. It was apparent that the combination of all three instruments better discerned a faulty state (Figure 7.19a, bottom right quadrant) compared to any individual instrument.

The effect of fusing the information of the three sensors for fault detection was quantified in Table 7.3. The combination of all instruments outperformed single sensors with respect to accuracy, recall, and F1-score. However, both the ATR-FTIR sensor and FBRM sensor individually outperformed the combination of all instruments with respect to precision. As mentioned in Section 7.4.2, the Raman spectrometer collected measurements listed as "normal," but that appeared faulty compared to the three-sigma normal region. The identification of false positives reduced the precision for the Raman spectrometer, and therefore the precision of the combination of instruments was impacted.

Of the individual instruments, FBRM had the highest precision, Raman had the best re-

Table 7.3: Table of error metrics corresponding to Figure 7.18.

Metric	**All Instruments**	**Raman**	**ATR-FTIR**	**FBRM**
Accuracy	**0.722**	0.611	0.611	0.556
Precision	0.867	0.818	0.889	**1.000**
Recall	**0.619**	0.429	0.381	0.238
F1-Score	**0.722**	0.562	0.533	0.385
Average	**0.732**	0.605	0.604	0.547

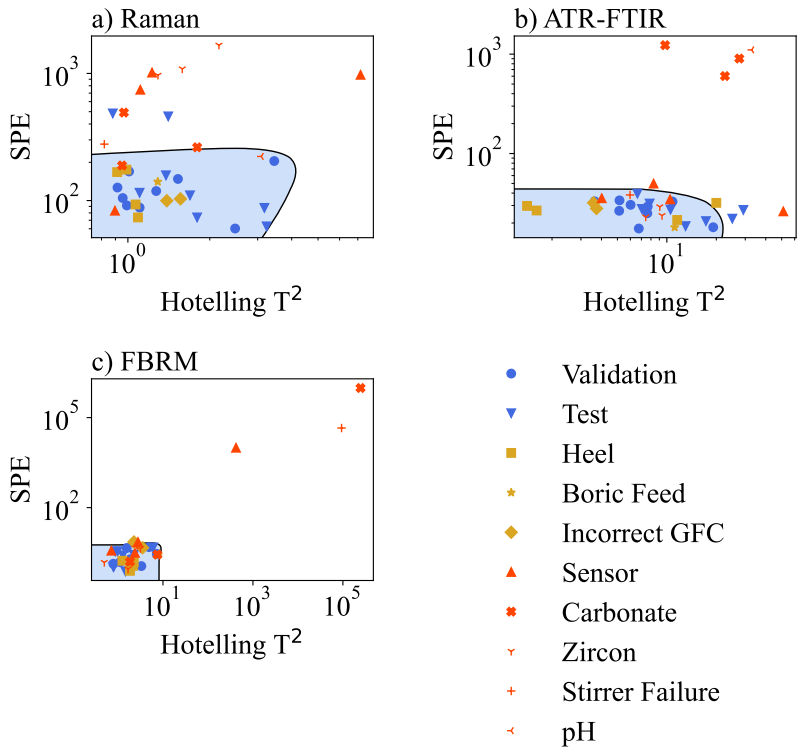


Figure 7.18: Hotelling T^2 and SPE on logarithmically-scaled axes with a three-sigma confidence interval (ideally covering 98.9% of data in two dimensions) based on validation data. Plot **a)** uses only Raman data, plot **b)** uses only ATR-FTIR data, and plot **c)** uses only FBRM data. Data without any known faults are displayed in blue. Data with process faults that have normal composition according to Table 7.1 and do not produce abnormal sensor measurements are shown in yellow. Data with the remaining process faults are shown in red.

call and F1-score, and Raman and ATR-FTIR had the best accuracy. In the dataset studied, the FBRM sensor model detected the fewest false positives due to its stable interrogation of chord-lengths providing little unanticipated variation. This led to the FBRM having the highest precision in comparison to the other sensors. Meanwhile, the Raman and ATR-FTIR sensors, with their ability to interrogate the chemical behavior of the slurry, were able to correctly place more data into their correct categories (higher accuracy than FBRM). This suggests that either the ball-probe Raman spectrometer or ATR-FTIR instrument may be useful as a single instrument for fault detection in slurries, depending on faults of greatest concern. However, the Raman instrument did correctly identify the most faults (highest recall) but also suffered from a relatively large number of false positives in this dataset

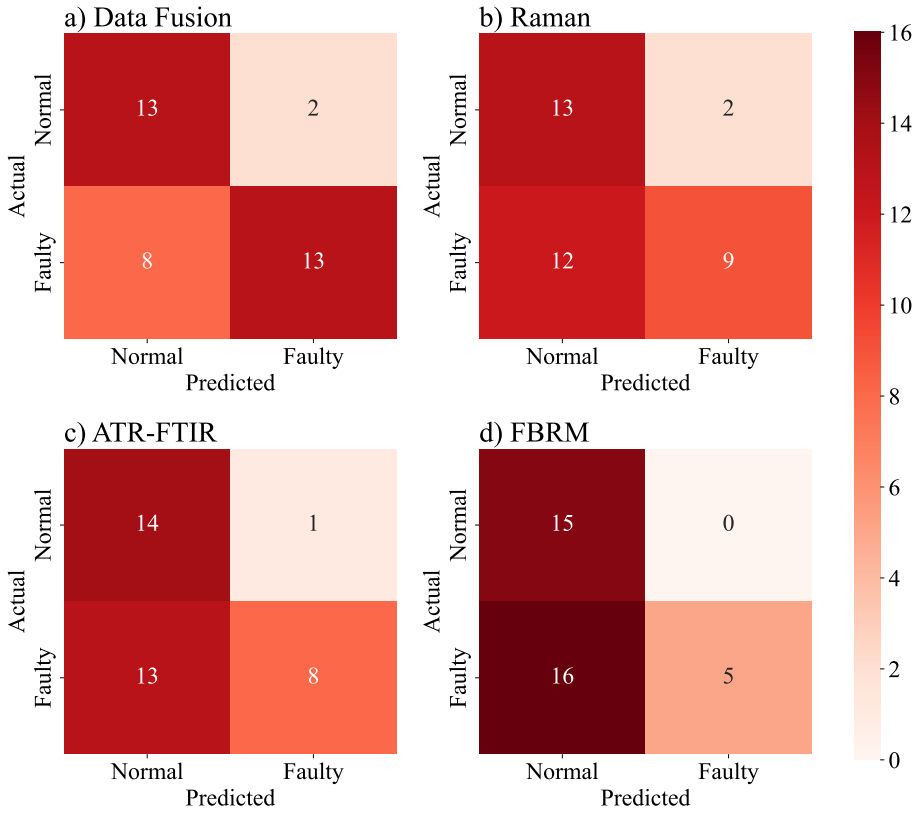


Figure 7.19: Confusion matrices showing classes of predicted and actual faults in process data based on Figure 7.10a and Figure 7.18. Shown are **a)** data fusion, **b)** Raman, **c)** ATR-FTIR, and **d)** FBRM schemes used for fault detection.

(low precision). The FBRM as a single sensor was limited in its ability to distinguish the faults studied in this chapter (low recall). However, mixing changes, as well as changes in particle size distribution (PSD) (not studied here), may make the FBRM sensor valuable in a fault detection platform. In this chapter, the combination of three sensors allowed for robust fault detection of both phases of the slurries studied.

Despite the array of three sensors (Raman spectroscopy, ATR-FTIR spectroscopy, and FBRM) performing better than the individual instruments according to most fault detection metrics, the individual instruments still provided value in their *specificity* of faults detected. This suggests that the individual instruments may be used for fault diagnosis to a particular phase or instrument in slurry systems. One can also envision the complementary use of other sensors for real-time and remote fault detection in nuclear-waste processing. For

example, laser induced breakdown spectroscopy (LIBS) may provide elemental information and has been investigated in the context of nuclear waste [183, 184] and combined with Raman spectroscopy [185]. Other measurements of temperature, pH, or radiological profile may also provide complementary and important real-time process information.

7.6 Conclusion

A combination of three instruments — Raman, ATR-FTIR, and FBRM — was used in examining the potential for real-time fault detection in hazardous multicomponent slurries typical of those found in nuclear-waste processing at the Hanford site. Faults that may occur during chemical processing were tested via an experimental dataset that was constructed based on a simulated three-tank system with heel masses. Studied faults include shifting pH, unanticipated species in the solution or insoluble phase, sensor failures, and mixing failure. An additional fault, gibbsite fouling on the surface of the ATR crystal, was observed and subsequently removed computationally using NCCLS. The computational removal resulted in better fault detection in terms of accuracy ($0.611 \rightarrow 0.722$), precision ($0.769 \rightarrow 0.867$), recall ($0.476 \rightarrow 0.619$), and F1-score ($0.588 \rightarrow 0.722$).

Compared to each individual instrument, the combination of the three instruments improved the fault detection capabilities as measured by accuracy (0.895), recall (1.000), and F1-score (0.923). Precision (0.857) of the combined data sources did not exceed that of the ATR-FTIR instrument alone (0.857) nor the FBRM instrument alone (1.000). These results were caused by the Raman instrument reporting anomalous readings on experiments designed to be normal, suggesting the importance of training set design, scaling, and feature selection when performing MSPM with these sensors.

The three-sensor platform studied here provides useful fault detection information remotely and in real-time to supplement or replace offline monitoring in hazardous chemical processes, such as nuclear-waste processing. The methods studied here are data-driven and therefore do not rely on the development of a process model to operate. The results

presented here illustrate the potential for timely fault detection without unnecessary radiological or chemical exposure to human workers operating hazardous slurry processes.

CHAPTER 8
FUTURE DIRECTIONS AND PAST REFLECTIONS

In this chapter, the contributions of this thesis are summarized as they pertain to the scientific knowledgebase. Areas where additional work remains to be done are also discussed.

8.1 Chapter 1

Chapter 1 motivates this thesis and provides contextual background for the studies presented. However, there are still issues beyond technical that may limit the implementation of process analytical technology (PAT) in actual nuclear-waste processing. Foremost among these is the political landscape around waste remediation. In recent memory, nuclear-waste remediation has been a bipartisan issue with support by both major political parties in the United States [186, 187]. However, individuals and organizations still disagree about the best technical path forward for immobilizing the waste [15, 188]. For example, the decision to vitrify all of the waste at Hanford was largely a political decision made with all stakeholders[1] in agreement (some scientists/engineers question the need for expensive glass storage when cementitious forms are cheaper [188]). In addition, a long-term storage location for high-level waste (HLW) in the United States is still unknown, with the planned Yucca mountain repository being cancelled [186, 40]. Despite these potential obstacles, nuclear-waste vitrification is proceeding as planned at the Hanford site with the first radioactive batches at Hanford expected in the next few months following the submission of this thesis. Current plans do not include in-line PAT at Hanford, but there is still

[1]Key stakeholders at Hanford include the Washington State Department of Ecology, The U.S. Environmental Protection Agency, the United States Department of Energy (US DOE), the Confederated Tribes of the Umatilla Indian Reservation, Yakama Nation, the Nez Perce Tribe, the Wanapum Tribe, the Confederated Tribes of the Colville Reservation, the Oregon Department of Energy, the Washington State Department of Health, the Washington Department of Fish and Wildlife, Tri-City Development Council, Central Washington Building Trades Council, Hanford Atomic Trades Council, Benton Community, Franklin Community, Grant County Governments, Richland City, West Richland City, Pasco City, and Kennewick City [189].

intellectual support and funding from the US DOE to enable sensor technology that may be implemented in either low-activity waste (LAW) or HLW vitrification.

8.2 Chapter 2

Chapter 2 examined waste simulants from the Savannah River Site and proposed a new blind source separation (BSS) algorithm that effectively removed glycolic acid contributions that did not have training data available. BSS is well-established for use in general signal processing, and it works well on spectroscopic signals typical of nuclear waste. Where larger amounts of data exist, nonlinear component estimation techniques (artificial neural networks) may be useful. However, for the nuclear-waste system studied in Chapter 2, the linear methods operated well and were functionally transparent.

The spectroscopic literature in nuclear waste has primarily utilized BSS to model multicomponent mixtures without addressing other processing variables (i.e. temperature, pressure, etc.) Therefore, there is opportunity to study BSS methods in the context of processes that have shifting process parameters, such as the shifting pH that was shown in Chapter 2. The option exists to integrate further constraints (beyond concentration nonnegativity and known target spectra) to improve the performance of multivariate curve resolution - alternating least squares (MCR-ALS)[2] in finding accurate and physically meaningful sources.

8.3 Chapter 3

Chapter 3 solved the problem of computationally removing (unknown) non-target species using a novel method: a constrained quadratic programming problem that leverages the nonnegativity of *all* spectral components, even non-targets. A result of this constrained optimization approach is that non-target removal can be performed with only a single spectrum; no historical data are required.

This chapter has several avenues for continued development, as the method introduced

[2]Much literature with MCR-ALS overlaps with the broader field of nonnegative matrix factorization.

was introduced with the content of this thesis and its related publication [31]. Firstly, the nonnegative constraint on non-target components may be hypothesized to be applicable to other types of spectral fitting beyond classical least squares (CLS) (i.e. partial least squares regression (PLSR), MCR-ALS, independent component analysis (ICA), etc.) Secondly, nonnegatively constrained classical least squares (NCCLS) performs better when feature selection has been applied since the hard constraint of NCCLS is sensitive to regions with low signal-to-noise ratio (SNR); this may be addressed by reformulating the optimization as a soft constraint rather than a hard constraint. This reformulation may increase the space of feasible solutions, when there may be no solutions that satisfy the hard constraint at every wavenumber (i.e. the optimization fails). Thirdly, there is the unstudied issue of algorithm convergence. While the quadratic objective function is convex and can be solved with a local solver, the hard constraints of NCCLS mean that there may be no feasible solutions. In the context of comparison methods (BSS), there is opportunity to study the conditions of a feasible solution for NCCLS. Lastly, feature transformations that further distinguish overlapping peaks, such as differentiating spectra with respect to wavenumber, may allow for more accurate non-target removal. However, spectral nonnegativity from the Beer-Lambert Law no longer applies after differentiating spectra; negative first-derivative spectral contributions are allowed. Rather than using nonnegativity, one may be able to use the fact that non-target spectral contributions should not be in directions opposite in sign to the target species; this behavior/constraint would be wavenumber-dependent.

8.4 Chapter 4

Chapter 4 demonstrated that a ball-probe Raman spectrometer can be used to analyze the insoluble phase of a slurry[3] while attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy can be used to interrogate the solution-phase of a slurry. It was shown that locally-linear methods (i.e. PLSR and indirect classical least squares

[3]Soluble species appear in the Raman spectrum in the presence of insoluble solids. However, their peak intensities were greatly reduced and depended on the amount of solids present.

(ICLS)) were effective in predicting the composition of waste from these spectra, but there were unmodeled nonlinearities in the Raman spectrum.

The quantity of experimental data presents a challenge for modeling nonlinearities that were present in the slurries measured. The reproducibility of the data suggest that there is deterministic map from the slurry composition to the observed measurement; the measurement is not dominated by stochastic behavior or noise. However, slurry experiments are time- and resource-intensive to conduct compared to solution-phase only experiments, which limits the number of practically obtainable data. In the context of this thesis, a ball-probe Raman spectrometer has not been used to measure time-series experimental data; this was done with an ATR-FTIR probe in Chapter 2. Lastly, there has not been thorough investigation into parsing two phases from a nuclear-waste slurry measurement with a ball-probe Raman spectrometer. An outstanding issue is that the solution phase has low signal when measuring a slurry. The only solution-phase peak that the author has observed to be significantly present in nearly every slurry measurement (at conditions similar to what may be expected at Hanford) is that of the nitrate anion (NO_3^-), which is highly Raman-active and at high concentrations in much Hanford waste.

8.5 Chapter 5

Chapter 5 used data transformations from physically-motivated variable pathlength and optically saturated models to analyze the slurry data presented in Chapter 4. It was found that using mass fraction of the insoluble phase (motivated by the assumption of a constant number of photons returning to the detector) rather than concentration/density of the entire slurry somewhat improved quantitative accuracy as measured by R^2. While these results were promising, the assumptions leading to the models used in this chapter were difficult to verify. Consequently, reducing the number of assumptions to work with an attenuation model with fewer assumptions had too many parameters to effectively identify a model based on available experimental data. It is plausible to imagine that an accurate model of

optical attenuation in multicomponent slurries could be identified if an abundance of experimental data were available. Additionally, neither this chapter nor Chapter 4 considered variable particle size, which would be another process variable that would impact optical attenuation of the Raman laser and Raman-scattered photons.

8.6 Chapter 6

Chapter 6 used a dual-Kalman filter to estimate the states (mass) and parameters (heel mass fraction) of three tanks in series. This chapter introduced a way to detect the change in heel mass composition in a single tank based on measurements of all tanks enabled by real-time sensors. While the proof-of-concept has been demonstrated in this computational study, these ideas have not been verified with an experimental system. Notably, the sensor accuracy demonstrated in Chapter 4 would suggest that the sensor measurements may be noisy in a multicomponent slurry. This makes accurate parameter estimation challenging if heel-mass changes result in sensor deviations on the same scale as instrument noise. The inherent sensor noise with the slurry spectra-to-composition models of this thesis represent a practical bottleneck for providing accurate and timely indication of heel-mass changes; though it was demonstrated to be possible with a 50% change in heel mass in the computational case study in this chapter. There may be an opportunity to follow up this proof-of-concept with a quantitative study of heel-mass change detection given spectra-to-composition accuracy.

8.7 Chapter 7

Chapter 7 used principal component analysis (PCA), Hotelling T^2, and squared prediction error (SPE) to combine the information from Raman spectroscopy, ATR-FTIR spectroscopy, and focused beam reflectance measurement (FBRM). Then, multivariate statistical process monitoring (MSPM) was performed and unseen abnormal process conditions were identified. Not explored in this chapter was the specificity of the three instruments

used in the data fusion study; i.e. if a fault was identified with the ATR-FTIR sensor, it likely resulted from something impacting the solution phase. Additionally, this study did not explore faults that would lead to an abnormal reading on the Hotelling T^2 axis. This could be accomplished by introducing samples that had out-of-spec compositions based on the training data (i.e. samples with abnormally high concentrations of sodium nitrate). Additionally, the specificity of the three individual tanks were not considered; for example, the first tank would not have glass-forming chemicals (GFCs) present. Lastly, this chapter was connected to Chapter 6 and a dual-Kalman filter was used to identify altered heel masses in the study. However, the sensor noise was greater than the change in composition provided by the changed heel masses (as mentioned in the previous section) in this particular study for both the Raman and ATR-FTIR sensors, so this study was not pursued further.

8.8 Final Remarks

In summary of this thesis, vibrational spectroscopy can be used to obtain useful information from solutions and slurries, and this thesis presents methods for analyzing sensor data and making use of the sensor data in the context of chemically processing radioactive slurries.

Appendices

APPENDIX A

SCALING: THE STANDARD NORMAL VARIATE TRANSFORM

Standard normal variate (SNV) scaling (also seen as standard scaling) is achieved by subtracting the mean sample from a dataset and dividing by the sample's standard deviation so that the sample has unit variance. This type of scaling is prevelant in many data science applications [59], and is used in Chapter 2, Chapter 3, and Chapter 7 in this work. Given matrix of data, $\mathbf{X}_{unscaled} \in \mathbb{R}^{n \times w}$, the SNV-transformed data can be found by subtracting the mean feature vector across all samples of training data, $\boldsymbol{\mu} \in \mathbb{R}^{1 \times w}$, and the feature-specific variance using the diagonal matrix defined as $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, ..., \sigma_w) \in \mathbb{R}^{w \times w}$ where $\sigma_1, ..., \sigma_w$ are the variances of each feature in the data [190]. This gives the scaled matrix, $\mathbf{X} \in \mathbb{R}^{n \times m}$.

Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ be a centering matrix that subtracts the mean sample from each sample of $\mathbf{X}$:

$$\begin{aligned} \mathbf{H} &= \mathbf{I}^{n \times n} - \frac{1}{n}\mathbf{1}^{\mathsf{T}}\mathbf{1} \\ \mu &= \frac{1}{n}\mathbf{1}\mathbf{X}_{unscaled} \end{aligned} \tag{A.1}$$

where $\mathbf{1} \in \mathbb{R}^{1 \times n}$ is a vector of 1's (i.e. $\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$). Let $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, ..., \sigma_w) \in \mathbb{R}^{w \times w}$ be a scaling matrix where $\sigma_1, ..., \sigma_w$ are the variances of each feature in the data [190].

$$\begin{aligned} \boldsymbol{\Sigma} &= \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_w \end{bmatrix} \\ \sigma_i &= \sqrt{\frac{1}{n-1}[\mathbf{X}_{unscaled}^{\mathsf{T}}\mathbf{X}_{unscaled}]_{ii}} \end{aligned} \tag{A.2}$$

The full SNV expression can be succinctly written as the above functions as

$$\mathbf{X} = \mathbf{H}\mathbf{X}_{unscaled}\mathbf{\Sigma}^{-1} \tag{A.3}$$

Where relevant, the SNV transformation is applied to all data but fit to only training data when used for quantification in this thesis.

APPENDIX B

PARTIAL LEAST SQUARES REGRESSION

Partial least squares regression (PLSR) is used to quantify spectra throughout this thesis, and so a section on PLSR is included here in the appendix. PLSR has also seen extensive use quantifying solutions and slurries typical of nuclear waste; the model has been well-reported in its ability to quantify complex mixture spectra with overlapping spectral bands, particularly when monitoring nuclear-waste solutions with vibrational spectroscopy. Recent work in nuclear-waste monitoring has applied PLSR to: locally linear regimes using piecewise PLSR with absorbance spectroscopy [25], multiple species in real Hanford waste using Raman spectroscopy [24], and sodium salt solutions with a limited training set using attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy and Raman spectroscopy [21].

Extensive literature exists that details and motivates the PLSR method [191, 192, 193, 102, 194]. The notation of Wold et al. is used here [102]. Assume that two data matrices, $\mathbf{X} \in \mathbb{R}^{n \times w}$ and $\mathbf{Y} \in \mathbb{R}^{n \times s}$ have been scaled via the SNV transform where there are n samples, w features (wavenumbers), and s chemical components[1].

The x-scores, $\mathbf{T}$, can be found from

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{\mathsf{T}} \tag{B.1}$$

where $\mathbf{W}$ are coefficients (weights). $\mathbf{X}$ is then modeled as a function of x-scores ($\mathbf{T}$) and loadings, $\mathbf{P}$, and some model error, $\mathbf{E}$:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathsf{T}} + \mathbf{E} \tag{B.2}$$

[1] In the literature, PLSR is often developed using complex (i.e. $\mathbb{C}^{n \times m}$) matrices and so the conjugate transpose, $\circ^*$, may be required in some spots where transpose ($\circ^{\mathsf{T}}$) is used here

$\mathbf{Y}$ is also modeled in a similar way, being decomposed into y-scores ($\mathbf{U}$), weights ($\mathbf{C}$), and model error ($\mathbf{G}$):

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^{\mathsf{T}} + \mathbf{G} \tag{B.3}$$

The notable advancement of PLSR is that the x-scores ($\mathbf{T}$) can be used to model $\mathbf{Y}$ with model error, $\mathbf{F}$:

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^{\mathsf{T}} + \mathbf{F} \tag{B.4}$$

Equation B.1 can be inserted into Equation B.4 to yield an equation where $\mathbf{Y}$ is an equation of $\mathbf{X}$ yielding a linear model:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{B} + \mathbf{F} \\ \mathbf{B} &= \mathbf{W}^{\mathsf{T}}\mathbf{C}^{\mathsf{T}} \end{aligned} \tag{B.5}$$

The parameters used in Equation B.5 can be found from the data (i.e. $\mathbf{X}$ and $\mathbf{Y}$) from existing algorithms, such as the nonlinear iterative partial least squares (NIPALS) algorithm [103, 195]. In this thesis, the scikit-learn version of PLSR is used with the NIPALS algorithm.

APPENDIX C

ANALYTICAL EXPRESSION FOR ERROR ESTIMATE

Results are presented in this section that show how an error estimate was analytically calculated for Chapter 3. In Chapter 3, $\mathbf{E_A}$ is approximated via a probabilistic approach by assuming that the elements of $\mathbf{E_A}$ (test data error matrix) and $\acute{\mathbf{E}}_\mathbf{A}$ (training data error matrix) are independent and identically distributed (i.i.d.):

$$\begin{aligned}(e_{i,1}, ..., e_{i,w} \text{ for } i = 1, ..., m) &\sim \mathcal{N}(\acute{\mu}, \acute{\sigma}^2) \\ (\acute{e}_{i,1}, ..., \acute{e}_{i,w} \text{ for } i = 1, ..., n) &\sim \mathcal{N}(\acute{\mu}, \acute{\sigma}^2)\end{aligned} \tag{C.1}$$

where $e_{i,j}$ is the j^{th} wavenumber of the i^{th} residual spectra of matrix $\mathbf{E}_A$ and $\acute{e}_{i,j}$ is the j^{th} wavenumber of the i^{th} residual spectra of matrix $\acute{\mathbf{E}}_\mathbf{A}$. The scalars $\acute{\mu}$ and $\acute{\sigma}$ are the matrix mean and matrix standard deviation, respectively, of the matrix of residuals found from the training data, $\acute{\mathbf{E}}_\mathbf{A}$:

$$\acute{\mathbf{E}}_\mathbf{A} = \acute{\mathbf{A}} - \acute{\mathbf{C}}\mathbf{K} \tag{C.2}$$

where $\acute{\mathbf{A}} \in \mathbb{R}^{n \times w}$ is a matrix of training spectra, $\acute{\mathbf{C}} \in \mathbb{R}^{r \times n}$ is a matrix of corresponding concentrations, and $\acute{\mathbf{E}}_\mathbf{A} \in \mathbb{R}^{n \times w}$ is a matrix of model training error. The i.i.d. assumption in this work allows a single Gaussian distribution to be fit to all elements of the residual matrix, $\acute{\mathbf{E}}_\mathbf{A}$ (including across different wavenumbers and different experiments). The assumption of i.i.d. and Gaussian error may be relaxed by using other distributions or a separate distribution for each wavenumber, in the case of a wavenumber-dependent error. After selecting an appropriate error distribution for the application, an estimated cumulative distribution function (CDF) is extracted from the selected distribution(s) and training data.

Using the CDF from training data, $\acute{\mathbf{E}}_A$, a calculated representative matrix, $\hat{\mathbf{E}}_\mathbf{A}$, is found

to bound expected model error, $\mathbf{E_A}$. Based on the i.i.d. assumption stated in Equation C.1, Equation C.3 approximates $\hat{\mathbf{E}}_\mathbf{A}$ as a matrix comprised of the scalar bound, b:

$$\hat{\mathbf{E}}_\mathbf{A} = \begin{bmatrix} b & \dots & b \\ \vdots & \ddots & \vdots \\ b & \dots & b \end{bmatrix} \in \mathbb{R}^{m \times w} \tag{C.3}$$

$$b = f(\acute{\mu}, \acute{\sigma}) \tag{C.4}$$

where b is a function of $\acute{\mu}$ and $\acute{\sigma}$. The value of b will be found from training data in this work. L is the probability that all elements of the residual matrix for a single residual spectrum ($\acute{e}_{i,1}$ through $\acute{e}_{i,w}$) are all less than or equal to a nonnegative upper bound on the residuals, b:

$$L = \mathcal{P}\Big(\bigcap_{j=1}^{w}(\acute{e}_{i,j} \leq b)\Big) = 0.5 \tag{C.5}$$

where $\bigcap$ represents the intersection of stochastic events. L is set to 0.5 so that, on average, 50% of residual spectra, $\mathbf{E_A}$, are over-estimated by $\hat{\mathbf{E}}_\mathbf{A}$ and 50% are under-estimated by $\hat{\mathbf{E}}_\mathbf{A}$.

From the assumption that the residuals at all wavenumbers are independently and identically distributed, $\acute{e}_{i,1}, ..., \acute{e}_{i,w} \sim \mathcal{N}(\acute{\mu}, \acute{\sigma}^2)$, the intersection of independent events simplifies Equation C.5 into the product of independent events since the residuals at every wavenumber all have equivalent probability of exceeding the scalar bound b under i.i.d. assumptions:

$$L = \mathcal{P}\Big(\bigcap_{j=1}^{w}(\acute{e}_{i,j} \leq b)\Big) = \prod_{j=1}^{w}\mathcal{P}(\acute{e}_{i,j} \leq b) = \mathcal{P}(\acute{e}_{i,j} \leq b)^w \tag{C.6}$$

$$L = \mathcal{P}(\acute{e}_{i,j} \leq b)^w \implies \mathcal{P}(\acute{e}_{i,j} \leq b) = \sqrt[w]{L} \tag{C.7}$$

Equation C.7 shows the probability of any individual residual, $\acute{e}_{i,j}$, being less than or equal to b. $\mathcal{P}(\acute{e}_{i,j} \leq b)$ can be simplified by using using training data to find the CDF of a Gaussian distribution, $F(b, \acute{\mu}, \acute{\sigma})$:

$$F(b, \acute{\mu}, \acute{\sigma}) = \frac{1}{2}\Big[1 + erf\Big(\frac{b - \acute{\mu}}{\acute{\sigma}\sqrt{2}}\Big)\Big] = P(e_{i,j} \leq b) \tag{C.8}$$

Combining Equation C.7 and Equation C.8 yield an expression for b:

$$b = \sqrt{2}\acute{\sigma} erf^{-1}\Big(2\sqrt[w]{L} - 1\Big) + \acute{\mu} \tag{C.9}$$

Equation C.9 gives b, the residual bound in $\hat{\mathbf{E}}_{\mathbf{A}}$ that was found by fitting $\acute{\mathbf{E}}_{\mathbf{A}}$ and is predicted to describe $\mathbf{E}_{\mathbf{A}}$ based on the assumption of i.i.d. measurement error between training and process experiments and uniformly distributed error between wavenumbers. For the computational experiments, b was roughly 2.7 standard deviations; there was experiment-to-experiment variation because different pseudorandom initializations had slightly different noise realizations. For experimental Raman spectroscopy data, b was approximated by 2.864 standard deviations. For experimental ATR-FTIR data, b was approximated by 2.424 standard deviations.

APPENDIX D
SUPPLEMENTAL RESULTS FOR COMPUTATIONAL NON-TARGET REMOVAL

The following results provide additional investigation into the studies of Chapter 3, including the computational efficiency of the methods presented as well as the experimental results analyzed in a "batch" rather than "real-time" manner.

D.1 Computational Efficiency and varied Training Data

For Computational Study 2, the different preprocessing methods were compared in terms of computation time of training and prediction (if there were separate training and prediction steps). This comparison may highlight which preprocessing methods were most suitable for real-time monitoring applications. It should be noted that different industries will have different requirements in terms of spectra acquisition times and quantification turnaround.

Figure D.1 shows the processing times with different amounts of non-target peak overlap. There was no significant trend in the computation time of these methods with increasing non-target peak overlap. This suggests that even in cases of a non-target peak that overlaps significantly with target peaks, the computation time of all methods did not significantly change. Figure D.2 shows the computation times of the preprocessing methods as the amount of measurement noise was varied. No appreciable trend was noted. Figure D.3 shows that the convolutional denoising autoencoder (CDAE) training time increases faster than linearly with time. Additionally, principal component analysis (PCA) training and nonnegatively constrained classical least squares (NCCLS) prediction increased faster than linearly with time, but was still on the order of fractions of a second for all cases studied in this work. Figure D.4 shows that blind source separation (BSS) independent component

analysis (ICA) increased faster than linearly with time. Additionally, NCCLS training and PCA prediction also increased faster than linearly with time.

These results suggested that CDAE cannot effectively train on large datasets in some real-time scenarios while BSS ICA cannot work with large batches of historical data in some real-time scenarios. Other than these two exceptions, all methods operated significantly faster than one second on the studied hardware for training and/or preprocessing, suggesting that they were all viable for most industrial real-time scenarios.

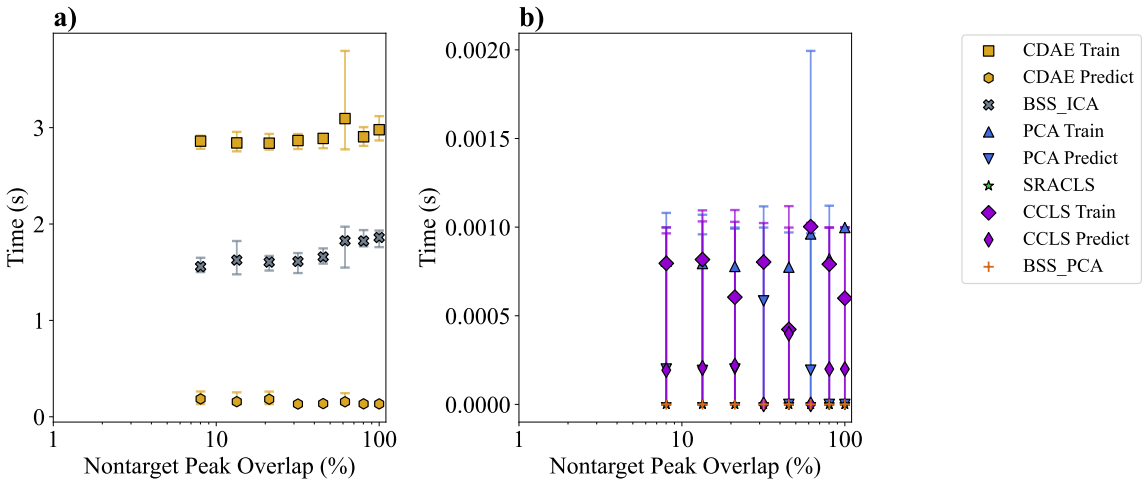


Figure D.1: Comparison of computation times for different preprocessing methods as a function of non-target peak overlap. Error bars correspond to full range of five replicates used for this study. The two plots are distinguished by different scaling where **a)** (larger computation times) contains quantification with no preprocessing and CDAE and **b)** (smaller computation times) contains PCA, spectral residual augmented classical least squares (SRACLS), BSS ICA, BSS PCA, and NCCLS.

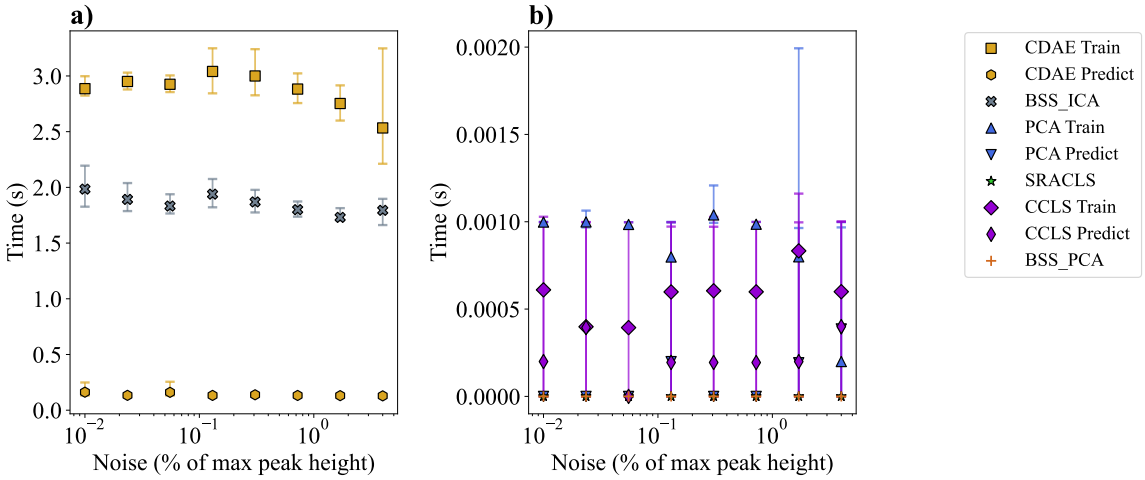


Figure D.2: Comparison of computation times for different preprocessing methods as a function of instrument (spectra) noise. Error bars correspond to full range of five replicates used for this study. The two plots are distinguished by different scaling where **a)** (larger computation times) contains quantification with no preprocessing and CDAE and **b)** (smaller computation times) contains PCA, SRACLS, BSS ICA, BSS PCA, and NCCLS.

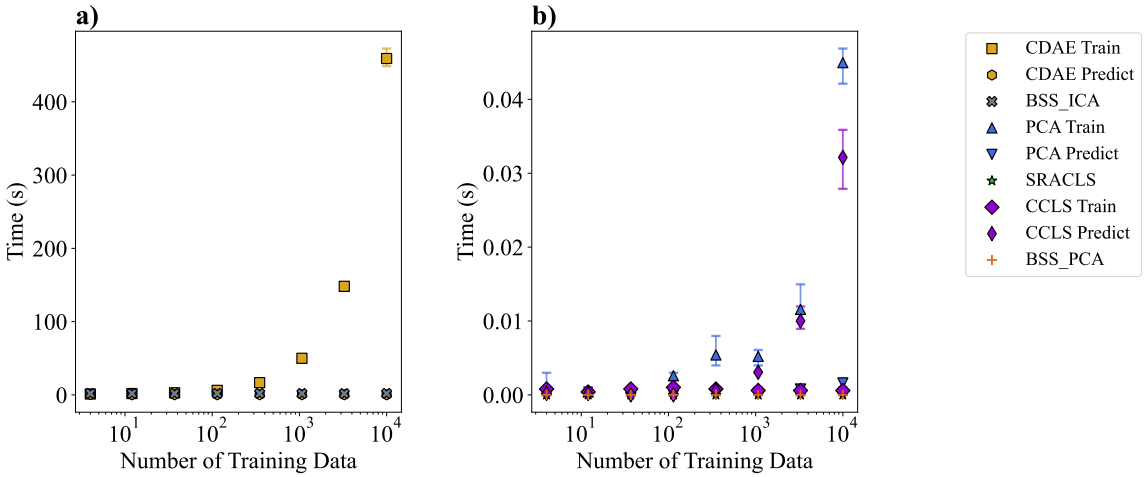


Figure D.3: Comparison of computation times for different preprocessing methods as a function of number of training data. Error bars correspond to full range of five replicates used for this study. The two plots are distinguished by different scaling where **a)** (larger RMSE values) contains quantification with no preprocessing and CDAE and **b)** (smaller root mean squared error (RMSE) values) contains PCA, SRACLS, BSS ICA, BSS PCA, and NCCLS.

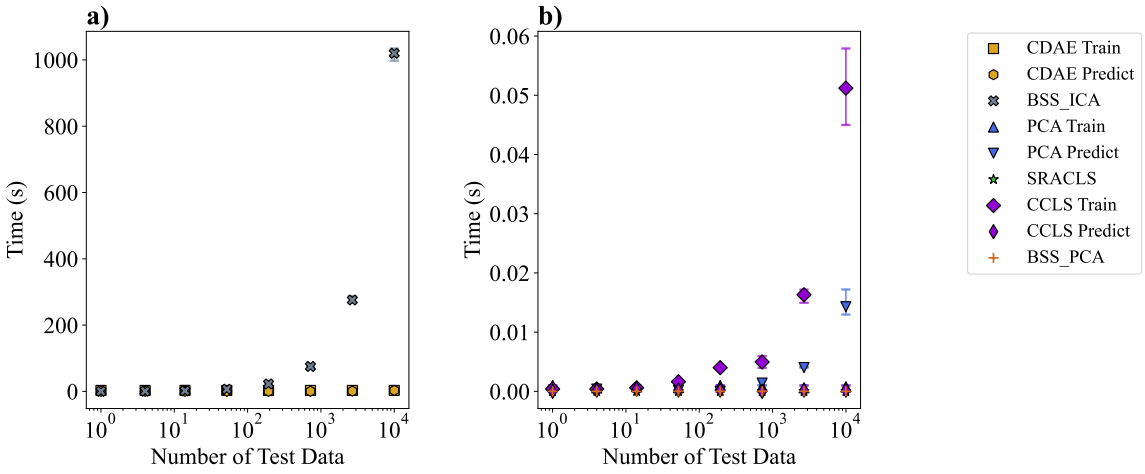


Figure D.4: Comparison of computation times for different preprocessing methods as a function of number of testing data. Error bars correspond to full range of five replicates used for this study. The two plots are distinguished by different scaling where **a)** (larger RMSE values) contains quantification with no preprocessing and CDAE and **b)** (smaller RMSE values) contains PCA, SRACLS, BSS ICA, BSS PCA, and NCCLS.

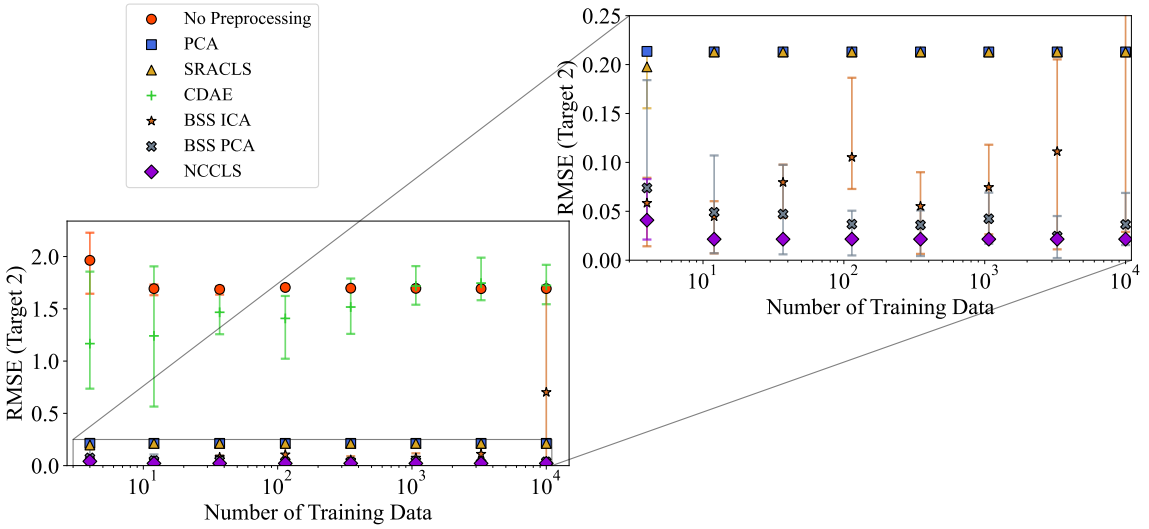


Figure D.5: Comparison of Target 2 RMSE for different preprocessing methods as a function of number of training data. Error bars correspond to full range of five replicates used for this study with pseudorandom spectra and model initializations. The inset plot is distinguished by a "zoomed in" scaling.

Figure D.5 shows the effect different amounts of training data (4–10,000) had on the preprocessing methods. In Figure D.5, all of the methods had slightly reduced performance and/or greater prediction variance at the lowest number of training data studied (four training data). This resulted from the quantification model being more vulnerable to noise given only four training data. Additionally, at training data exceeding 3,270 individual spectra, the performance of BSS ICA decreased substantially. This behavior was observed to correspond to the scikit-learn implementation of the fastICA algorithm being occasionally unreliable in finding a solution (notice the large variance in Figure D.5). Aside from the trends observed at low data and high data for BSS ICA specifically, all methods did not appear to be significantly affected by varying the amount of training data provided.

D.2 Experimental Results Analyzed in a Batch-Analysis Manner

The experimental data from Chapter 3 were also analyzed in a batch preprocessing (rather than real-time) analysis scenario and are shown in Figure D.6 and Figure D.7. PCA, BSS ICA, and BSS PCA improved performance in general when operating in a batch pre-

processing scenario. PCA was then the most accurate preprocessing method for the Raman spectra, while NCCLS remained the most accurate preprocessing method for the ATR-FTIR spectra.

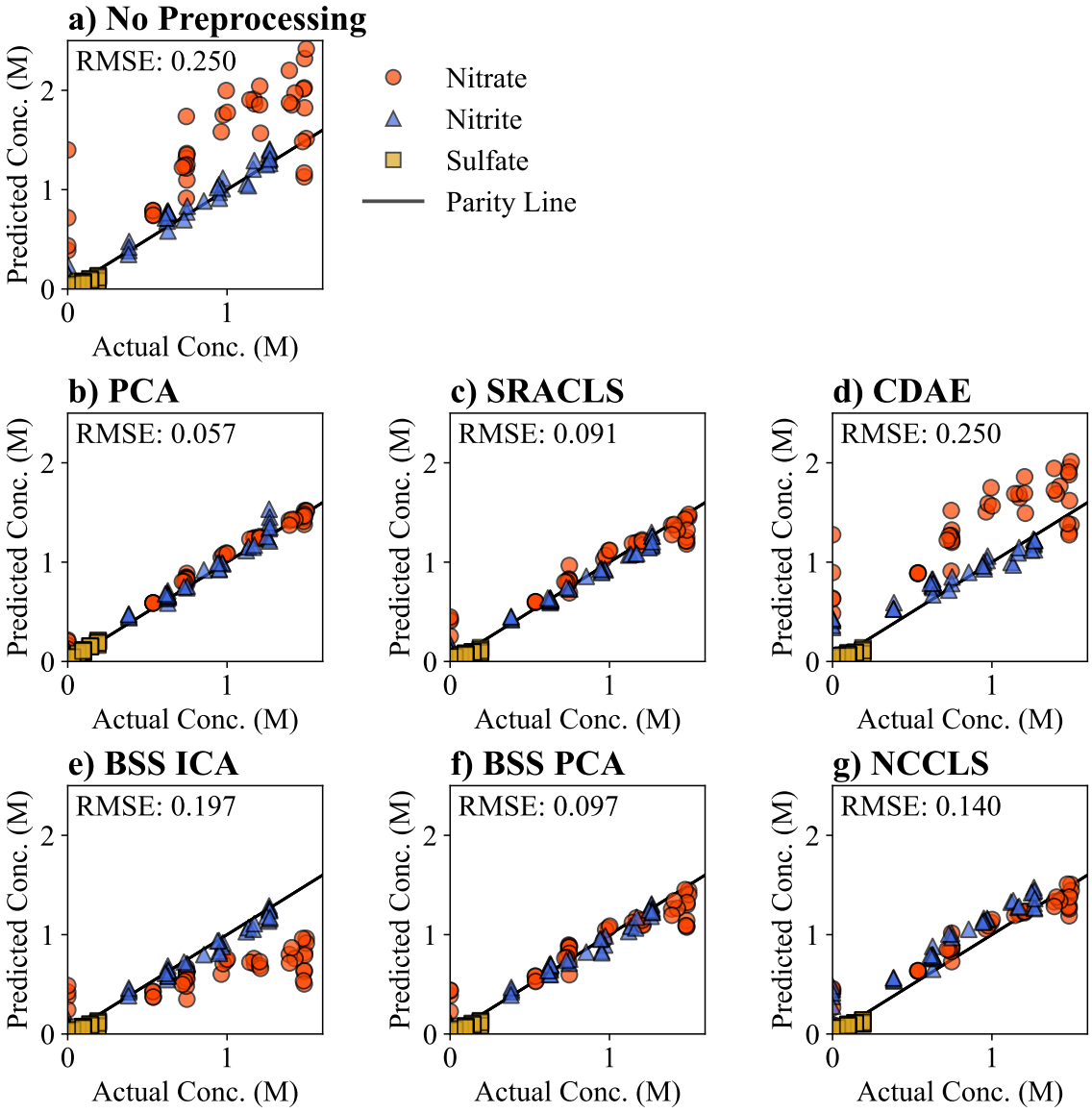


Figure D.6: Parity plots quantifying nitrate, nitrite, and sulfate from Raman spectra preprocessed using: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

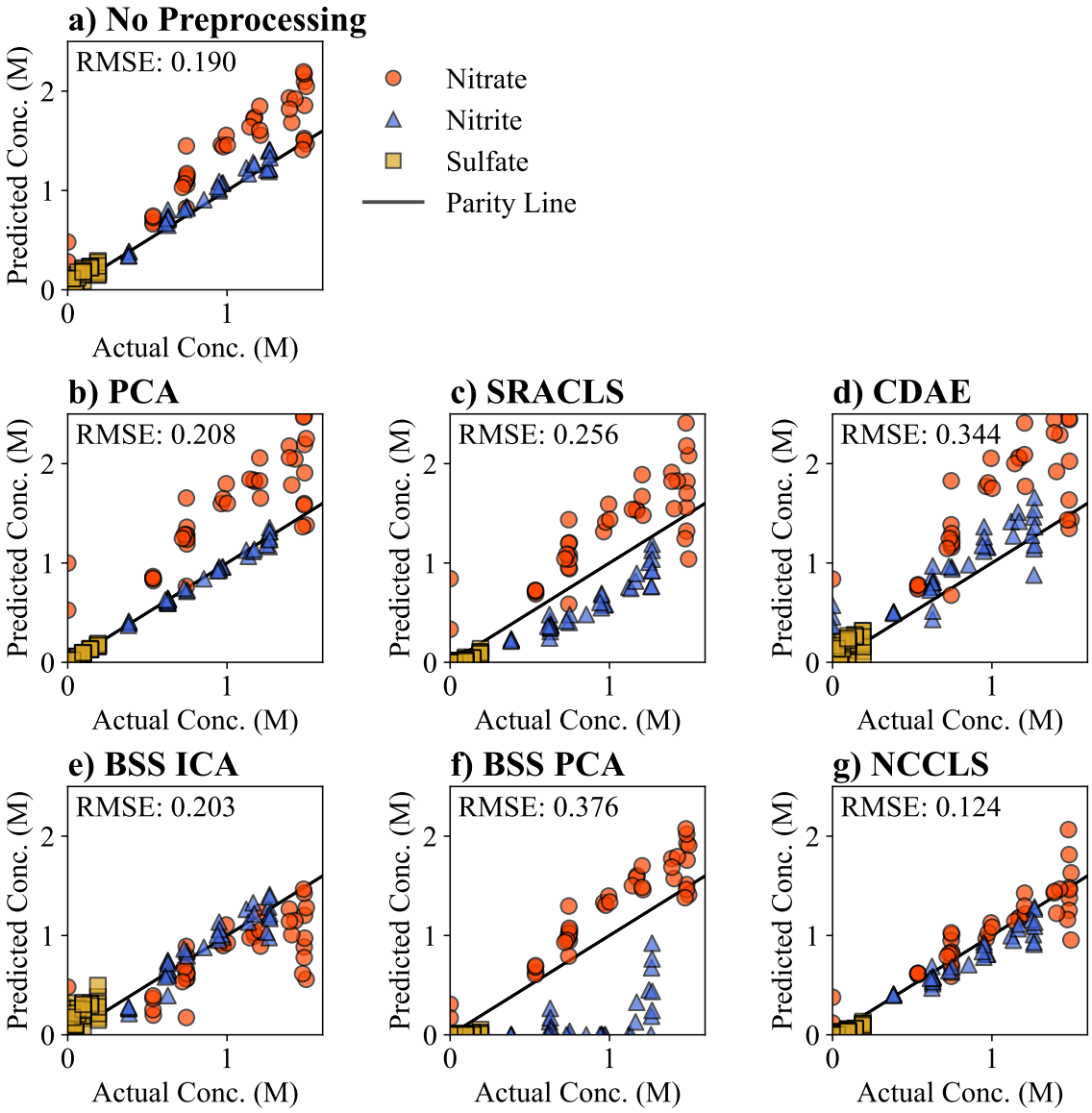


Figure D.7: Parity plots quantifying nitrate, nitrite, and sulfate from ATR-FTIR spectra preprocessed using: **a)** no preprocessing, **b)** PCA, **c)** SRACLS, **d)** CDAE, **e)** BSS ICA, **f)** BSS PCA, and **g)** NCCLS.

APPENDIX E

PEAK IDENTIFICATION FOR LOW-ACTIVITY WASTE SYSTEM

Reference spectra and peak identification was done in close collaboration with Dr. Rupanjali Prasad, and so the results have been removed from this thesis for not being independent work of the author. However, these results are relevant to the surrounding text and may be of interest to the reader. These results are highlighted here and presented in their entirety in a journal publication [19].

Soluble Species

The Fourier transform infrared (FTIR) and Raman reference spectra of soluble glass-forming chemical (GFC) components (vanadium pentoxide, boric acid, and lithium carbonate) in a 3 m NaOH solution (pH greater than 13) are shown in Figure E.1. All FTIR spectra in Figure E.1a have a water band at approximately 1640 cm^{-1} corresponding to the O-H bending band of water (ν_2 mode) [196]. Each of the spectra in Figure E.1 corresponds to a concentration of 1 m except for lithium carbonate (Li_2CO_3). The solubility of lithium carbonate is less than 1 m, hence the Raman and FTIR spectra correspond to a saturated solution of the salt (0.304 m, determined in this work) at 25 °C. The dissolution of boric acid ($B(OH)_3$) at high pH (greater than 13) is given by the following reaction:

$$B(OH)_3 + OH^- \rightleftharpoons B(OH)_4^- \tag{E.1}$$

At high pH, boric acid mainly dissociates into the borate ion [197, 198], $B(OH)_4^-$, whose presence can be further corroborated by examining the spectrum of boric acid solution in Figure E.1a (i). The FTIR spectrum exhibits spectral features at 1150 and 950 cm^{-1}. The broad peak at 1150 cm^{-1} corresponds to the B-O-H in-plane bending, while the peak

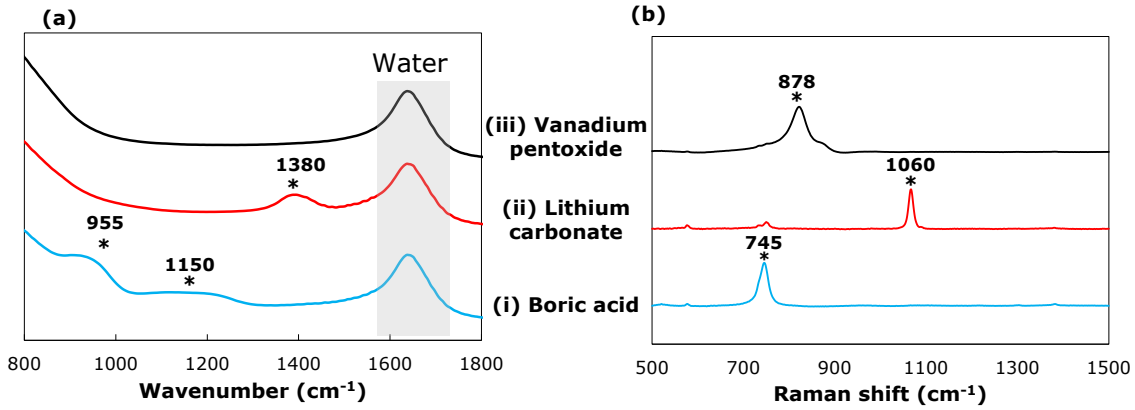


Figure E.1: **a)** FTIR and baseline-corrected **b)** Raman spectra of soluble GFC components in a basic 3 m NaOH solution: (i) boric acid ($B(OH)_3$), (ii) lithium carbonate ($Li_2(CO)_3$), and (iii) vanadium pentoxide (V_2O_5). The characteristic peaks have been marked with an '*' and their corresponding wavenumbers have been listed.

at 950 cm^{-1} is caused by the B-O asymmetric stretching vibrations in the borate ion [197, 198]. The corresponding Raman peak for the borate ion in Figure E.1b (i) is observed at 745 cm^{-1}, representing total symmetrical vibrations [199].

Lithium carbonate shows significant dissolution and dissociation (Equation E.2) at alkaline conditions into carbonate (CO_3^{2-}) ions.

$$Li_2CO_3 \rightarrow 2Li^+ + CO_3^{2-} \tag{E.2}$$

The FTIR and Raman spectra of lithium carbonate in 3 m NaOH solution are shown in Figure E.1a (ii) and Figure E.1b (ii), respectively. The FTIR spectrum exhibits a strong peak at 1380 cm^{-1} which originates due to the C–O asymmetrical in-plane stretch of the carbonate ion (CO_3^{2-}) [196]. The corresponding Raman spectrum also has a sharp peak around 1060 cm^{-1} arising due to the C–O symmetric stretching vibrations [200].

Vanadium pentoxide (V_2O_5) is an amphoteric oxide and is soluble in strong alkaline solutions to form metavanadate (VO^{3-}) and orthovanadate (VO_4^{3-}) ions:

$$V_2O_5 + 6OH^- \rightarrow 2VO_4^{3-} + 3H_2O \tag{E.3}$$

$$V_2O_5 + 2OH^- \rightarrow 2VO_3^- + H_2O \quad \text{(E.4)}$$

At pH greater than 13, the main ion present in the solution is the orthovanadate ion [201] (Equation E.3) which does not appear on the FTIR spectrum (Figure E.1a (iii)) but is Raman-active. The presence of the orthovanadate ion can be deduced by the spectral bands in the Raman spectrum (Figure E.1b (iii)). The Raman spectrum shows two bands at 878 and 825 cm^{-1}. The strongest band at 878 cm^{-1} is due to the symmetric stretching of the VO_3 units while the 825 cm^{-1} is associated with the symmetric stretching vibrations of the VO_2 units [202].

Insoluble Species

The FTIR and Raman references for the insoluble species were obtained by adding 5 g of a single type of solid (silica, kyanite, olivine, wollastonite, and zircon) to a 3 m NaOH solution (50 g/kg water) and stirring overnight. The silicates were observed to deposit on the ATR-FTIR probe window (Figure 4.2b). Therefore, FTIR reference spectra for silicates were obtained by pipetting around 5 mL of the solution containing solid particles, followed by centrifugation and filtration by passing the supernatant through a syringe filter (0.22 µm pore size) to ensure all solid particles were removed from the solution before measuring their FTIR spectra (Figure E.2a). The reference Raman spectra, on the other hand, were obtained by analyzing the slurry (containing solids suspended via mixing) with a Raman probe (Figure E.2b and Figure E.2c).

The FTIR spectra exhibit a single peak at 1640 cm^{-1} (Figure E.2a) which is attributed to the O–H bending vibrations of the water molecule [196]. Apart from the O-H bending vibrations from water at 1640 cm^{-1}, no other spectral features were observed in the FTIR spectra (Figure E.2a) for the five studied silicates. Similarly, no peaks were observed in the Raman spectra of the filtered solutions (data not shown). This again demonstrates negligible solubility that is shown by inductively coupled plasma (ICP) results [19].

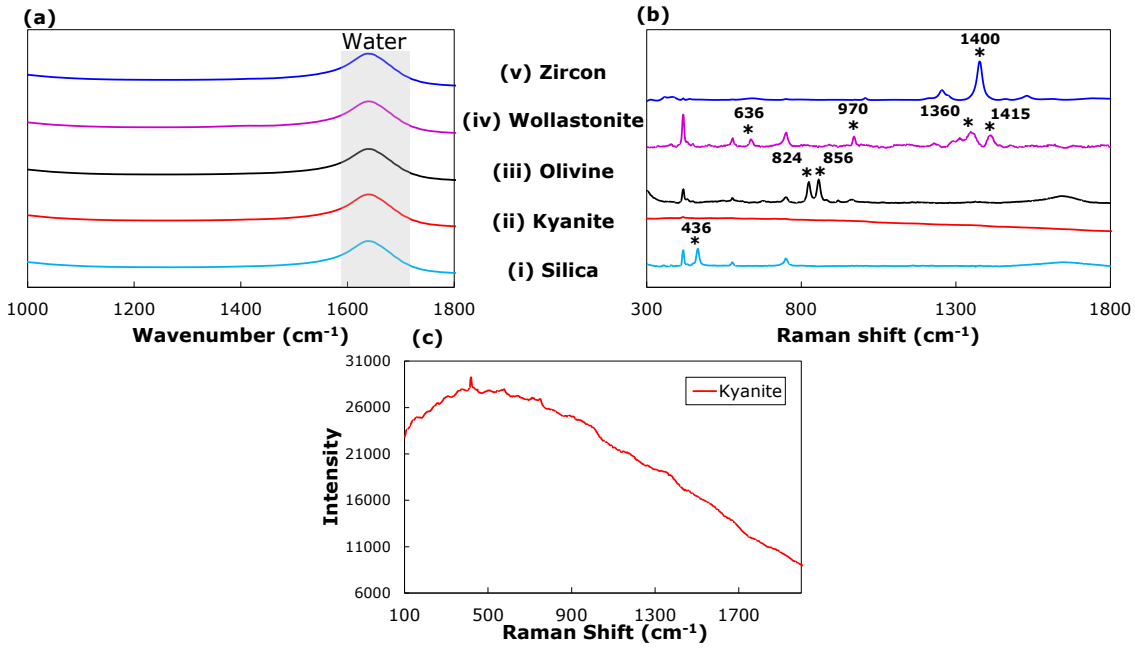


Figure E.2: Highest-peak normalized **a)** ATR-FTIR spectra of filtered silicate solution and baseline corrected **b)** Raman spectra of suspended silicates (50 g/kg water) in a basic 3 m NaOH solution: silica (SiO_2), kyanite (Al_2SiO_5), olivine (Mg_2SiO_4), wollastonite ($CaSiO_3$), and zircon ($ZrSiO_4$). The **c)** unnormalized kyanite spectrum showing broad fluorescence background. The Raman peaks corresponding to the solids have been marked with an '*' and their corresponding wavenumbers have been listed.

The reference Raman spectra of solids (silica and silicates) in 3 m NaOH solution are shown in Figure E.2b. Three peaks are present in every measurement located at: 419, 577, and 751 cm^{-1} due to the sapphire tip of the probe. The Raman spectrum of silica (SiO_2) in Figure E.2b (i) has a sharp peak at 436 cm^{-1} which arises due to the symmetric stretching vibrations of the Si–O–Si units [203, 204]. Similar Si–O–Si stretching interactions in olivine occur as a doublet at 824 and 856 cm^{-1} (Figure E.2b (iii)). Both features arise from coupled symmetric and asymmetric stretching vibrational modes of the SiO_4 tetrahedral units [205]. Raman spectra of wollastonite shows multiple features in the spectral region of 600 to 1500 cm^{-1} (Figure E.2b (iv)). The peak at 636 cm^{-1} is attributed to the Si–O–Si bending vibrations, whereas the bands at 970, 1360, and around 1415 cm^{-1} are due to Si–O stretching vibrations [206]. Zircon also exhibits a sharp peak at 1400 cm^{-1} arising from the Si–O stretching vibrations. The Raman spectrum of kyanite (Al_2SiO_5) does not

have strongly resolved peaks. However, the kyanite spectrum has a signature broad shape centered around 500 cm^{-1} (Figure E.2c) that is also observed in spectra of GFC mixtures.

Metal Oxides

The FTIR and Raman spectra of various oxides constituting the glass-forming chemicals (GFCs) are shown in Figure E.3a and Figure E.3b, respectively. The slurries probed using ATR-FTIR and Raman contain around 5 g of the solid compound dispersed in a 3 m NaOH solution (50 g/kg water). Except for the peak at 1640 cm^{-1}, which is attributable to the O–H bending vibrations in the water molecule, no spectral features are observed in the FTIR spectra (Figure E.3a). However, ICP data [19] show that some oxides, particularly zinc oxide (ZnO), are significantly soluble at high pH. Dissolution information combined with the observed spectra of Figure E.3a indicate that these oxides are not infrared active in the wavenumber range studied or that the sensitivity of the probe is not high enough to detect dissolved concentrations. The corresponding Raman spectra of the solid compounds exhibits several peaks for suspended particles of rutile (TiO_2, Figure E.3b (i)) and hematite (Fe_2O_3, Figure E.3b (iv)) that can be used for identification of the compounds in a mixture. Peaks at 143, 447, and 612 cm^{-1} are observed in the Raman spectra of rutile (Figure E.3b (i)) corresponding to the B_{1g}, E_g, and A_{1g} vibrational modes respectively. The Raman spectra of hematite has peaks at 227, 293, 418, and 610 cm^{-1} (Figure E.3b (iv)). While the peak at 227 cm^{-1} represents the A_{1g} phonon band mode, peaks at 293, 418, and 610 cm^{-1} are attributed to the E_g mode vibrations. These bands involve displacement of both iron and oxygen within $Fe(O)_6$ octahedral units [207, 208]. The other oxides, zinc oxide and tin oxide (SnO_2), do not have any visible spectral features in the measured Raman spectra except for background fluorescence (Figure E.3b (ii and iii)).

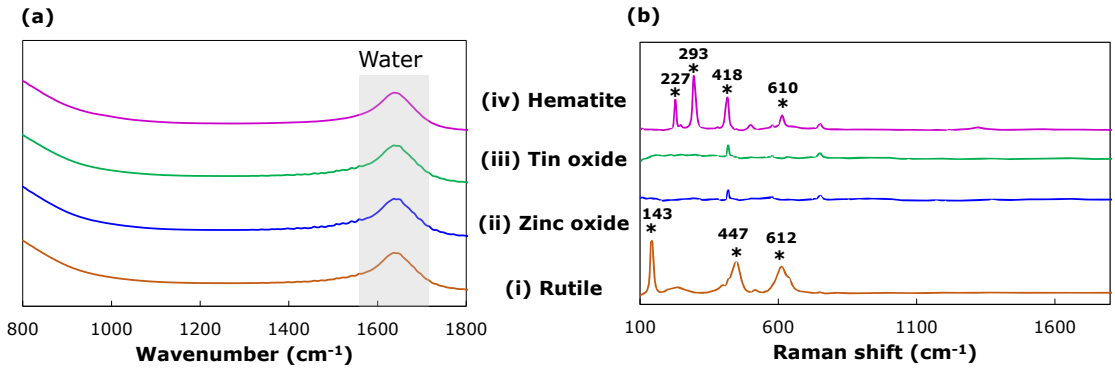


Figure E.3: **a)** FTIR and baseline corrected **b)** Raman spectra of metal oxides GFC components in a basic 3 m NaOH solution. (i) Rutile (TiO_2), (ii) zinc oxide (ZnO), (iii) tin oxide (SnO_2), and (iv) hematite (Fe_2O_3). The Raman peaks corresponding to the solids have been marked with an '*' and their corresponding wavenumbers have been listed.

APPENDIX F

INDEPENDENCE OF SLURRY SAMPLES FOR CHAPTER 4

Presented here are results to supplement Chapter 4 and verify that the presented quantification is from the spectra and not correlation between samples since samples were not i.i.d.. Without sufficient independence between predicted variables, the model may have accurate quantification without necessarily quantifying the spectra. Spurious correlation between variables becomes a greater challenge when working with a high dimensional input space (202 and 3101 features for ATR-FTIR and Raman spectra, respectively), high dimensional output space (5 quantified components for each spectroscopy), and a limited sample space (48 and 66 samples, respectively). An investigation of correlations between outputs (concentration) was performed. The coefficient of determination (the square of correlation, commonly seen as r^2) was used to account for large magnitude correlations. Figure F.1 shows the coefficient of determination of the components. An r^2 value of 1 is expected on the diagonal, because variables are perfectly correlated with themselves.

The heatmaps in Figure F.1 show the correlation between species for the dataset used in this work. To test the impact of data correlation, two PLSR models were constructed, one corresponding to quantified solids (kyanite, wollastonite, olivine, silica, and zircon) and the other corresponding to quantified solution species (nitrate, nitrite, carbonate, sulfate, and borate). Each PLSR model was constructed so that four of the respective quantified species were used as model inputs, while the remaining component was quantified. This is structurally similar to the quantification performed in Section 4.3.4 and Section 4.3.5 except that the input data is the concentrations of other species, rather than spectra. The two PLSR models of this section used four latent variables each, equal to the number maximum allowable with four input variables. This quantification isolates the effects of correlation between species: any prediction capability in Figure F.2 results from correlations

in the dataset. If the model is quantifying based on artifacts in the dataset, we expect the quantification accuracy in Figure F.2 to match the quantification accuracy in Figure 4.7 and Figure 4.10. Fig Figure F.2 shows the parity plots for both the soluble species and insoluble species studied.

From the results of the parity plots, it can be seen that a substantial amount of scatter exists for all the species when using other species as model inputs rather than spectra. This indicates that the PLSR models are quantifying the spectra and not making use of spurious correlations in the dataset to produce effective quantification.

The high correlation/prediction accuracy of borate in Figure F.2 warrants additional discussion. Borate produces the best quantification of all the species when using concentrations of other species for quantification. This likely results from the manner in which boric acid was added to solutions. Additions were done in two steps: a soluble species addition and a GFC addition. The soluble species addition consisted of nitrate, nitrite, carbonate, and sulfate (in addition to phosphate, acetate, and oxalate as minor species). The GFC addition, however, contained soluble boric acid and lithium carbonate. Because of this, boric acid was added in specific concentrations dictated by GFC composition rather than designed via pseudorandom uniform distribution like the other species. Despite this, the results in the manuscript with an R^2 value of 0.998 are substantially better than could be achieved through the correlation in Figure F.2.

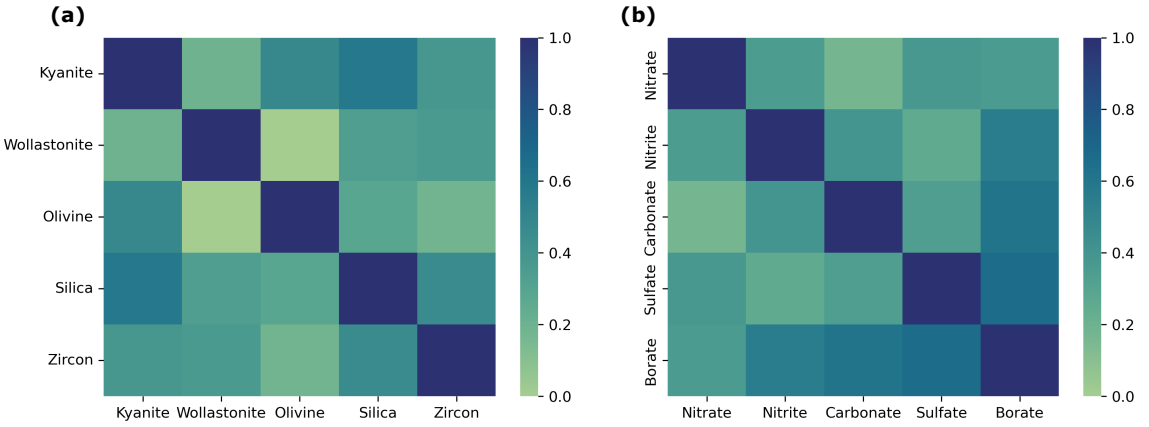


Figure F.1: Coefficient of determination plots for **a)** five quantified insoluble species and **b)** five quantified solution components.

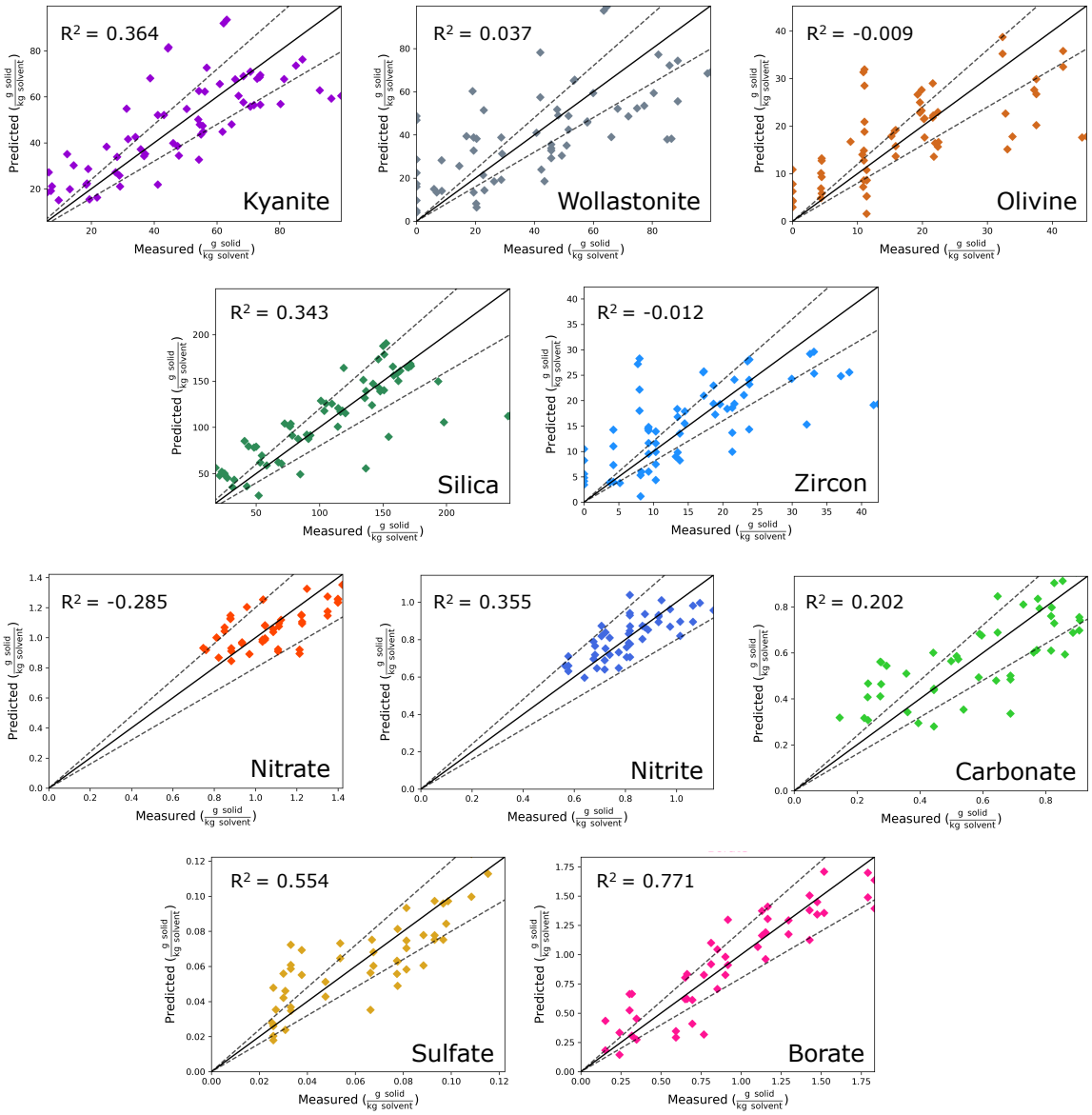


Figure F.2: Parity plots for both soluble and insoluble components when using the concentrations of the other species as inputs to a PLSR model. From these results, it can be concluded that the prediction accuracy achieved in Chapter 4 are not due to a correlated dataset.

APPENDIX G

STATE SPACE FORMULATION FOR MODEL-BASED FAULT DETECTION

The following equations show how the mass balance in Chapter 6 (i.e. Equation 6.8) is converted to state-space notation for Section 6.3. The below equations have their variables explained in Chapter 6.

$$\mathbf{m}_{k,t}^{\text{all}} = \begin{bmatrix} m_{k,t}^{1} \\ m_{k,t}^{2} \\ \vdots \\ m_{k,t}^{s-1} \\ m_{k,t}^{s} \end{bmatrix} \quad , \quad \mathbf{x}_t = \begin{bmatrix} \\ \mathbf{m}_{1,t}^{\text{all}} \\ \\ \\ \mathbf{m}_{2,t}^{\text{all}} \\ \\ \\ \vdots \\ \\ \\ \mathbf{m}_{v-1,t}^{\text{all}} \\ \\ \\ \mathbf{m}_{v,t}^{\text{all}} \\ \end{bmatrix} \tag{G.1}$$

$$\boldsymbol{\alpha}_{k,t}^{\text{all}} = \begin{bmatrix} \alpha_{k,t}^{1} \\ \alpha_{k,t}^{2} \\ \vdots \\ \alpha_{k,t}^{s-1} \\ \alpha_{k,t}^{s} \end{bmatrix} \quad , \quad \boldsymbol{\theta}_t = \begin{bmatrix} \vdots \\ \boldsymbol{\alpha}_{1,t}^{\text{all}} \\ \vdots \\ \boldsymbol{\alpha}_{2,t}^{\text{all}} \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{\alpha}_{v-1,t}^{\text{all}} \\ \vdots \\ \vdots \\ \boldsymbol{\alpha}_{v,t}^{\text{all}} \\ \vdots \end{bmatrix} \tag{G.2}$$

$$\mathbf{g}_{k,t}^{\text{all}} = \begin{bmatrix} g_{k,t}^{1} \\ g_{k,t}^{2} \\ \vdots \\ g_{k,t}^{s-1} \\ g_{k,t}^{s} \end{bmatrix} \quad , \quad \mathbf{u}_t = \begin{bmatrix} | \\ \mathbf{g}_{1,t}^{\text{all}} \\ | \\ | \\ \mathbf{g}_{2,t}^{\text{all}} \\ | \\ \vdots \\ | \\ \mathbf{g}_{v-1,t}^{\text{all}} \\ | \\ | \\ \mathbf{g}_{v,t}^{\text{all}} \\ | \end{bmatrix} \tag{G.3}$$

APPENDIX H

HIGH-LEVEL WASTE SLURRY DATA

Presented here are data used in Chapter 7 that are typical of multicomponent high-level waste (HLW) slurries with different practical faults tested.

H.1 Experimental Data

The starting solution (used for insoluble-phase training data and all validation and test data) was made in a using the composition: 88.1057 wt% H_2O, 10.573 wt% NaOH, and 1.322 wt% $Al(OH)_3$.

Table H.1: Solution compositions (in grams) for solution-phase training experiments.

Index	$NaNO_3$	$NaNO_2$	Na_2SO_4	H_3BO_3	H_2O	NaOH
0	7.08	4.76	0.85	0.90	101.09	11.99
1	10.62	5.59	1.86	2.70	101.09	11.99
2	6.37	4.21	0.00	3.58	101.01	12.00
3	8.51	7.52	1.55	11.64	101.01	12.00
4	7.79	6.42	0.43	6.26	101.34	12.09
5	11.32	8.06	1.13	12.51	101.34	12.09
6	9.21	3.92	0.28	5.37	100.36	12.01
7	16.99	6.13	0.99	7.16	100.36	12.01
8	12.04	7.24	1.98	0.00	100.79	12.14
9	16.30	6.69	1.43	8.05	100.97	12.01
10	13.46	6.97	2.13	4.46	100.33	12.14
11	9.93	5.30	1.70	13.43	100.68	12.05
12	12.76	5.87	0.72	10.74	100.32	12.10
13	15.59	5.03	0.57	8.94	100.61	12.10
14	14.17	4.47	1.29	1.79	101.05	12.04

Table H.2: Slurry compositions (in grams) for insoluble-phase training experiments.

Index	Fe_2O_3	Al_2SiO_5	SiO_2	$Al(OH)_3$	Starting Solution
0	4.66	5.70	4.96	4.65	113.72
1	11.47	10.14	10.40	2.72	113.72
2	3.31	2.74	0.34	8.72	113.52
3	7.37	20.44	8.88	12.20	113.52
4	6.04	14.55	2.65	12.79	113.68
5	6.81	23.40	6.54	9.51	113.68
6	8.73	1.28	1.85	11.43	113.66
7	23.71	13.06	5.74	2.74	113.66
8	14.19	19.00	11.20	3.28	113.45
9	22.37	16.06	8.11	15.52	113.48
10	16.92	17.51	11.96	10.05	113.75
11	10.10	8.67	9.64	23.64	113.56
12	15.56	11.60	4.21	19.61	113.65
13	20.98	7.18	3.44	16.86	113.55
14	19.61	21.93	1.07	18.21	113.94
15	18.26	4.22	7.30	6.01	113.74

Table H.3: Slurry compositions (in grams) for all validation and testing experiments. Symbols are used to designate the class data each experiment are ○ for validation (historical process) data and ⋆ for test data without faults, and ◇ for test data with faults. The adjusted heel fractions for Experiments 15 and 18 affect only subsequent batches, so these experiments have no fault associated with them.

Index	Batch	Tank	Fault	Starting Sol.	$NaNO_3$	$NaNO_2$	Na_2SO_4	H_3BO_3	Fe_2O_3	Al_2SiO_5	SiO_2	$Al(OH)_3$	Total Solids
○ 0	1	1	None	113.47	13.51	6.56	1.16	0.00	11.21	8.02	1.34	18.55	39.12
⋆ 1	1	2	None	113.61	13.23	6.63	0.93	14.38	11.46	18.61	12.01	17.65	59.73
○ 2	1	3	None	113.60	13.12	6.45	0.81	12.96	13.71	18.58	10.17	16.19	58.65
○ 3	2	1	None	113.56	14.94	4.82	0.50	0.00	17.36	6.82	1.38	9.13	34.69
○ 4	2	2	None	113.62	13.34	4.68	0.83	12.26	18.51	24.72	6.01	9.79	59.03
○ 5	2	3	None	113.56	12.70	5.16	0.74	12.27	17.09	22.38	6.22	11.52	57.21
⋆ 6	3	1	None	113.54	16.82	6.20	0.24	0.00	10.92	5.62	0.75	17.73	35.02
○ 7	3	2	None	113.57	16.47	5.82	0.47	14.97	9.97	21.49	5.00	14.97	51.43
◇ 8	3	3	Boric Acid pH Shift	113.63	14.81	5.80	0.77	13.25	11.27	21.57	4.94	14.11	51.89
◇ 9	4	1	Stirrer stopped for 5 min	113.58	15.37	5.98	1.08	0.00	17.29	8.13	1.37	14.42	41.21
◇ 10	4	2	FBRM not immersed	113.49	14.36	5.59	1.07	7.16	15.56	19.61	11.53	16.10	62.80
◇ 11	4	3	Raman not shrouded	113.65	13.32	5.76	1.23	7.26	15.45	17.42	11.58	15.31	59.76
◇ 12	5	1	Raman not immersed	113.96	15.10	7.78	2.01	0.00	16.42	2.24	1.50	12.48	32.64
○ 13	5	2	None	113.47	14.27	7.22	1.87	14.33	16.07	20.55	10.41	12.55	59.58
◇ 14	5	3	IR not immersed	113.62	13.31	6.57	1.75	12.67	16.46	18.17	10.27	11.67	56.57
⋆ 15	6	1	Heel fraction 0.73	113.52	10.39	4.96	1.79	0.00	9.96	1.66	0.98	6.00	18.60
◇ 16	6	2	Heel fraction 0.61	127.12	10.14	6.70	1.79	17.92	14.26	16.85	12.09	6.81	50.01
◇ 17	6	3	Heel fraction 0.69	112.72	8.64	6.18	1.42	14.52	15.10	16.73	9.35	6.49	47.67
⋆ 18	7	1	Heel fraction 0.92	113.33	12.34	7.92	0.25	0.00	19.43	4.72	1.91	19.68	45.74
◇ 19	7	2	Heel fraction 0.59	150.96	15.66	9.34	1.07	12.88	26.38	23.10	14.15	25.28	88.91
◇ 20	7	3	Heel fraction 0.76	116.26	12.01	6.68	1.02	10.19	17.47	15.40	9.92	15.92	58.71
⋆ 21	8	1	None	113.50	14.25	7.42	1.93	0.00	6.61	5.15	1.12	9.58	22.46
○ 22	8	2	None	113.50	12.90	7.12	1.94	9.82	8.66	14.08	4.81	10.18	37.73
⋆ 23	8	3	None	113.54	13.56	6.74	1.64	8.91	9.50	15.92	4.86	9.70	39.98
⋆ 24	9	1	None	113.60	7.86	5.25	2.09	0.00	19.09	3.07	0.45	18.12	40.73
⋆ 25	9	2	None	113.51	7.60	5.20	1.86	9.07	17.32	11.84	10.48	16.11	55.75
○ 26	9	3	None	113.50	7.39	5.64	1.88	9.36	16.33	13.66	10.65	15.99	56.63
◇ 27	10	1	Boric Acid	113.58	16.15	7.73	0.72	2.52	6.23	2.49	1.77	22.89	33.38
◇ 28	10	2	No kyanite	113.58	14.43	7.09	0.96	13.27	9.51	4.75	5.11	19.86	39.23
◇ 29	10	3	No kyanite	113.53	14.38	6.68	0.86	12.12	10.30	7.30	5.48	16.60	39.68
◇ 30	11	1	Carbonate 5 g	113.53	15.29	7.19	1.48	0.00	10.97	5.77	1.39	5.90	24.03
◇ 31	11	2	Carbonate 4 g	113.58	14.42	6.54	1.26	13.44	13.43	20.11	11.97	6.00	51.51
◇ 32	11	3	Carbonate 3.2 g	113.54	13.37	6.43	1.05	13.40	14.59	19.76	11.71	7.34	53.40
◇ 33	12	1	Zircon 5 g	113.53	12.40	6.72	0.45	0.00	17.72	2.25	1.65	12.09	33.71
◇ 34	12	2	Zircon 4 g	113.50	12.87	6.21	0.59	13.71	16.14	16.05	6.45	10.47	49.11
◇ 35	12	3	Zircon 3.2 g	113.51	12.17	5.85	0.65	12.09	17.33	17.42	7.50	9.31	51.56

APPENDIX I

FAULT DETECTION WITH DIFFERENT SUBSETS OF DATA

In this section, the multivariate statistical process monitoring (MSPM) (fault detection) from Chapter 7 is performed with Training Data used to generate the three-sigma region (Figure I.1). Additionally, MSPM was performed with just faults that produced abnormal sensor readings or composition — altered heel mass faults (Experiments 16, 17, 19, and 20), boric acid in the feed stream (Experiment 27), and incorrect GFC additions (Experiments 28 and 29).

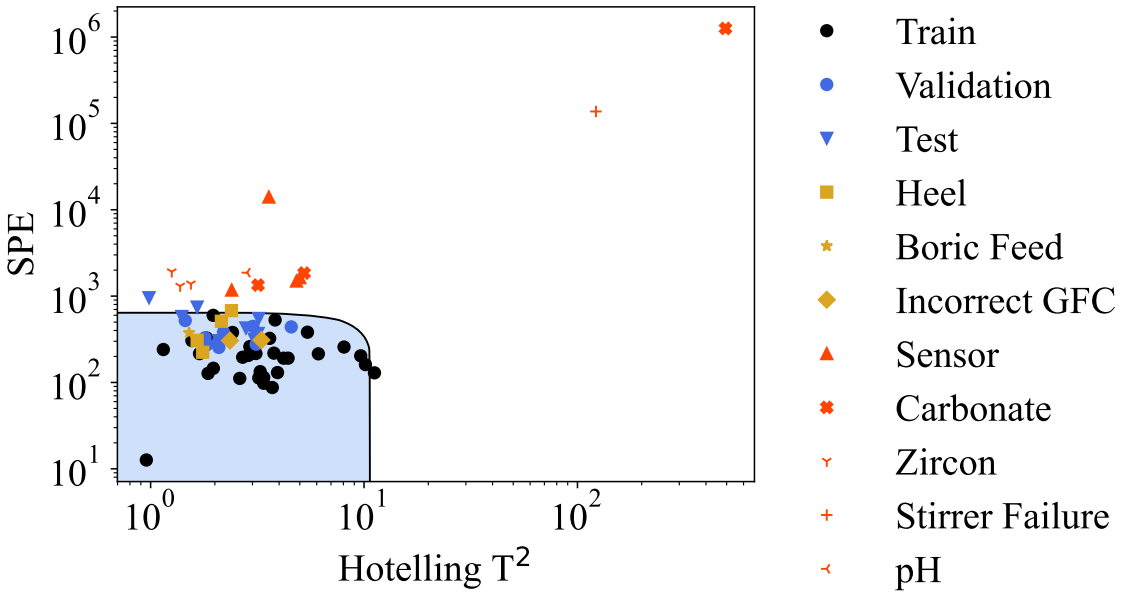


Figure I.1: Hotelling T^2 and squared prediction error (SPE) of all process data on logarithmically-scaled axes with a three-sigma confidence interval (ideally covering 98.9% of Gaussian data in two dimensions) fit to *training data*. Data without any known faults are displayed in blue. Data with process faults that had normal composition according to Table 7.1 and did not produce abnormal sensor measurements are shown in yellow. Data with the remaining process faults are shown in red. Training data are shown in black. The performance of the fault detection and conclusions drawn were unaffected by using training data to construct the three-sigma bounds.

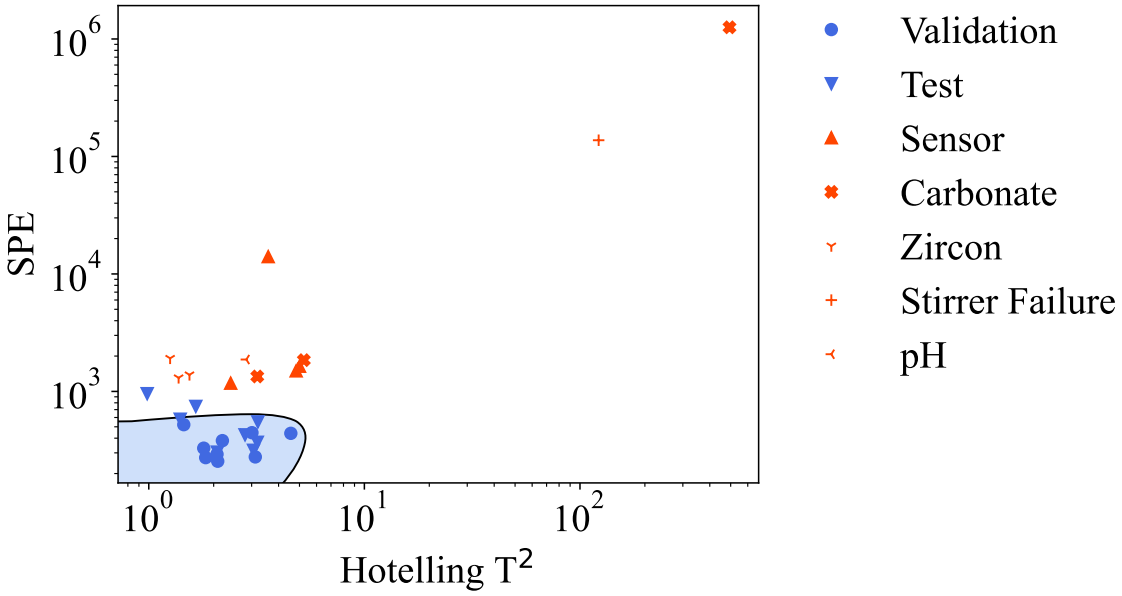


Figure I.2: Hotelling T^2 and reconstruction SPE on logarithmically-scaled axes with a three-sigma confidence interval (ideally covering 98.9% of data in two dimensions). Data without any known faults are displayed in blue. Data with process faults are shown in red.

Table I.1: Table of error metrics corresponding to Figure I.2.

Metric	**All Instruments**	**Raman**	**ATR-FTIR**	**FBRM**
Accuracy	**0.895**	0.737	0.632	0.579
Precision	0.857	0.818	0.857	**1.000**
Recall	**1.000**	0.750	0.500	0.333
F1-Score	**0.923**	0.783	0.632	0.500
Average	**0.919**	0.772	0.655	0.603

REFERENCES

[1] M. Benedict, T. H. Pigford, and H. W. Levi, *Nuclear Chemical Engineering*, 2nd. New York: McGaw-Hill, 1981, ISBN: 0070045313.

[2] R. A. Peterson, Ed., *Engineering Separations Unit Operations for Nuclear Processing*. Boca Raton: CRC Press, 2020, ISBN: 9781138605824.

[3] G. Yaghy et al., "Opportunities for process intensification technologies in nuclear effluent treatment: A review of precipitators, adsorbers and separators," *Chemical Engineering and Processing - Process Intensification*, vol. 191, no. March, p. 109 441, 2023.

[4] I. W. Donald, B. L. Metcalfe, and R. N. J. Taylor, "The immobilization of high level radioactive wastes using ceramics and glasses," *Journal of Materials Science*, vol. 32, pp. 5851–5887, 1997.

[5] R. A. Peterson et al., "Review of the Scientific Understanding of Radioactive Waste at the U.S. DOE Hanford Site," *Environmental Science and Technology*, vol. 52, no. 2, pp. 381–396, 2018.

[6] J. Marcial, B. J. Riley, A. A. Kruger, C. E. Lonergan, and J. D. Vienna, "Hanford low-activity waste vitrification: A review," *Journal of Hazardous Materials*, vol. 461, no. June 2023, 2024.

[7] M. E. Stone, C. C. Diprete, M. E. Farrar, A. M. Howe, F. R. Miera, and M. R. Poirier, "WTP Real-Time , In-Line Monitoring Program Task 2 : Determine the Technical Basis for Process Control and Task 5 : Process Control Challenges," Tech. Rep. December, 2017.

[8] M. E. Farrar, M. R. Poirier, C. C. DiPrete, and M. E. Stone, "WTP Real-Time In-Line Monitoring Program : Task 3 : LAW and EMF Flow Sheet Sampling Points," Savannah River National Laboratory, Tech. Rep. March, 2018.

[9] A. M. Lines et al., "Online, Real-Time Analysis of Highly Complex Processing Streams: Quantification of Analytes in Hanford Tank Sample," *Industrial and Engineering Chemistry Research*, vol. 58, no. 47, pp. 21 194–21 200, 2019.

[10] D. S. Kim, J. D. Vienna, and A. A. Kruger, "Preliminary ILAW Formulation Algorithm Description," Office of River Protection, Richland, Tech. Rep., 2012.

[11] G. F. Piepel, B. G. Amidan, A. Heredia-langner, D. R. Weier, and S. K. Cooley, "Statistical Methods and Results for WTP IHLW and ILAW Compliance," Tech. Rep. WTP-RPT-072, 2005.

[12] B. E. Wells et al., "Hanford waste physical and rheological properties : Data and gaps," Pacific Northwest National Laboratory, Tech. Rep., 2011.

[13] R. E. Gephart, "A short history of Hanford waste generation, storage, and release," Richland, WA, Tech. Rep., 2003.

[14] D. S. Kosson et al., "Hanford site-wide risk review project - final report," Consortium For Risk Evaluation with Stakeholder Participation, Tech. Rep., 2018.

[15] M. E. Stone, "Evaluation of a Material Balance Only and Material Balance with Real-Time In-line Monitoring Approaches for DFLAW Processing," Savannah River National Laboratory, Tech. Rep., 2019.

[16] A. M. Lines, J. M. Bello, C. Gasbarro, and S. A. Bryan, "Combined Raman and Turbidity Probe for Real-Time Analysis of Variable Turbidity Streams," *Analytical Chemistry*, vol. 94, no. 8, pp. 3652–3660, 2022.

[17] H. M. Felmy, N. T. Boily, A. S. Medina, S. A. Bryan, and A. M. Lines, "Raman Monitoring of Hanford tank simulants in high turbidity : qualitative analysis," Pacific Northwest National Laboratory, Richland, Tech. Rep., 2023.

[18] D. J. Griffin, M. A. Grover, Y. Kawajiri, and R. W. Rousseau, "Robust Multicomponent IR-to-Concentration Model Regression," *Chemical Engineering Science*, vol. 116, pp. 77–90, 2014.

[19] R. Prasad, S. H. Crouse, R. W. Rousseau, and M. A. Grover, "Quantifying dense multicomponent slurries with in-Line ATR-FTIR and Raman spectroscopies : A Hanford case study," *Industrial & Engineering Chemistry Research*, vol. 62, pp. 15 962–15 973, 2023.

[20] G. M. Maggioni, S. Kocevska, M. A. Grover, and R. W. Rousseau, "Analysis of Multicomponent Ionic Mixtures Using Blind Source Separation : A Processing Case Study," *Industrial & Engineering Chemistry Research*, vol. 58, pp. 22 640–22 651, 2019.

[21] S. Kocevska, G. M. Maggioni, R. W. Rousseau, and M. A. Grover, "Spectroscopic Quantification of Target Species in a Complex Mixture Using Blind Source Separation and Partial Least-Squares Regression: A Case Study on Hanford Waste," *Industrial & Engineering Chemistry Research*, vol. 60, no. 27, pp. 9885–9896, 2021.

[22] H. M. Felmy, H. E. Lackey, A. S. Medina, M. J. Minette, S. A. Bryan, and A. M. Lines, "Leveraging Multiple Raman Excitation Wavelength Systems for Process Monitoring of Nuclear Waste Streams," *ACS ES&T Water*, vol. 2, pp. 465–473, 2022.

[23] A. M. Lines et al., "Sensor Fusion: Comprehensive Real-Time, On-Line Monitoring for Process Control via Visible, Near-Infrared, and Raman Spectroscopy," *ACS Sensors*, vol. 5, no. 8, pp. 2467–2475, 2020.

[24] P. Tse, J. Shafer, S. A. Bryan, and A. M. Lines, "Quantification of Raman-Interfering Polyoxoanions for Process Analysis: Comparison of Different Chemometric Models and a Demonstration on Real Hanford Waste," *Environmental Science & Technology*, vol. 55, no. 19, pp. 12 943–12 950, 2021.

[25] R. Lascola, P. E. O'Rourke, and E. A. Kyser, "A Piecewise Local Partial Least Squares (PLS) Method for the Quantitative Analysis of Plutonium Nitrate Solutions," *Applied Spectroscopy*, vol. 71, no. 12, pp. 2579–2594, 2017.

[26] A. M. Lines et al., "Sensor fusion: Comprehensive real-time, on-line monitoring for process control via visible, near-infrared, and raman spectroscopy," *ACS Sensors*, vol. 5, pp. 2467–2475, 8 2020.

[27] R. J. Lascola, P. E. O'Rourke, and D. M. Immel, "Development of an L-Edge X-ray absorbance spectrometer for monitoring dissolver solutions in H-Canyon," Savannah River National Laboratory, Aiken, SC, Tech. Rep. May, 2024.

[28] S. H. Crouse, R. Prasad, M. A. Grover, and R. W. Rousseau, "Addressing Spectral Non-linearities when Quantifying Raman Spectra of Dense Slurries: Applications for Slurry Monitoring at Hanford," in *WM 2023 Conference*, Phoenix, AZ, 2023.

[29] S. H. Crouse, S. Kocevska, R. Prasad, M. A. Grover, and R. W. Rousseau, "Data Fusion of ATR-FTIR and Raman Spectroscopies for Component Quantification: Applications at Hanford. Proceedings of the Waste Management Symposium, Phoenix, AZ," in *WM2022 Conference*, 2022.

[30] S. H. Crouse, R. Prasad, R. W. Rousseau, and M. A. Grover, "Real-Time IR Spectroscopy Monitoring Coupled with Blind Source Separation for Applications in Nuclear Waste Processing," *Frontiers in Nuclear Engineering*, no. November, pp. 1–13, 2023.

[31] S. H. Crouse, R. W. Rousseau, and M. A. Grover, "A feature selection method for overlapping peaks in vibrational spectroscopy using nonnegatively constrained classical least squares," *Computers and Chemical Engineering*, vol. 189, no. July, p. 108 785, 2024.

[32] S. H. Crouse, R. Prasad, R. W. Rousseau, and M. A. Grover, "State and parameter estimation of non-ideal batch reactors with heel masses," in *American Control Conference*, IEEE, Jul. 2024, pp. 1186–1191, ISBN: 9798350382655.

[33] S. H. Crouse, R. Prasad, M. A. Grover, and R. W. Rousseau, "Detecting faults in nuclear waste slurry processing with in-line probes: A computational study," in *WM2024 Conference*, 2024.

[34] R. E. Gephart, "A short history of waste management at the Hanford Site," *Physics and Chemistry of the Earth*, vol. 35, no. 6-8, pp. 298–306, 2010.

[35] U.S. Department of Energy, "2019 Hanford Lifecycle Scope, Schedule and Cost Report," Tech. Rep., 2019.

[36] A. Goel, J. S. McCloy, R. Pokorny, and A. A. Kruger, "Challenges with vitrification of Hanford High-Level Waste (HLW) to borosilicate glass – An overview," *Journal of Non-Crystalline Solids: X*, vol. 4, no. August, p. 100 033, 2019.

[37] H. A. Colburn and R. A. Peterson, "A history of Hanford tank waste, implications for waste treatment, and disposal," *Environmental Progress and Sustainable Energy*, vol. 40, no. 1, 2021.

[38] S. E. Kelly, "A Joule-Heated Melter Technology for the Treatment and Immobilization of Low-Activity Waste," *Washington River Protection Solutions Report RPP-48935*, vol. 767, no. Rev. 0, 2011.

[39] J. D. Vienna and D. Kim, "Preliminary IHLW Formulation Algorithm Description," Office of River Protection, Richland, Tech. Rep., 2023.

[40] US Department of Energy, "2025 Hanford lifecycle scope, schedule, and cost report," Richland, WA, Tech. Rep., 2025.

[41] D. Chew, "Savannah River Site - Waste Tank Levels," Savannah River National Laboratory, Aiken, Tech. Rep. SRMC-LWP-2022-00001, Rev. 83, 2024.

[42] W. R. Wilmarth et al., "Review: Waste-Pretreatment Technologies for Remediation of Legacy Defense Nuclear Wastes," *Solvent Extraction and Ion Exchange*, vol. 29, no. 1, pp. 1–48, 2011.

[43] M. R. Norton, H. B. Shah, M. E. Stone, L. E. Johnson, and R. O. Driscoll, "Overview - Defense Waste Processing Facility Operating Experience," in *Waste Management Symposia*, Tucson, 2002, pp. 1–13.

[44] N. R. Council, *Advice on the Department of Energy's Cleanup Technology Roadmap: Gaps and Bridges*. Washington, DC, 2009, pp. 1–284.

[45] C. T. Randall, L. M. Papouchado, and S. L. Marra, "The Defense Waste Processing Facility, from Vision to Reality," 2000.

[46] S. H. Crouse et al., "Selected chemical engineering applications in nuclear-waste processing at the savannah river site," *Annual Review of Chemical and Biomolecular Engineering*, vol. 16, 2025.

[47] J. Ray, B. Culbertson, S. Marra, and M. Plodinec, "DWPF Glass Product Control Program," Tech. Rep., 2018.

[48] D. P. Lambert, M. S. Williams, C. H. Brandenburg, and J. D. Newell, "Sludge Batch 9 Simulant Runs Using the Nitric-Glycolic Acid Flowsheet," Savannah River National Laboratory, Tech. Rep. November, 2016.

[49] D. P. Lambert, A. D. Nikolov, D. T. Wasan, M. S. Williams, A. M. Howe, and W. H. Woodham, "A novel defoamer for processing nuclear waste: Testing and performance," *Environmental Progress and Sustainable Energy*, vol. 40, no. 4, pp. 1–13, 2021.

[50] D. P. Lambert and A. M. Howe, "Antifoam Development for Eliminating Flammability Hazards and Decreasing Cycle Time in the Defense Waste Processing Facility," Savannah River National Laboratory, Tech. Rep. May, 2021.

[51] W. H. Woodham, A. M. Howe, and M. J. Siegfried, "Sludge Batch 10 Flowsheet Testing with Non-radioactive Simulants," Savannah River National Laboratory, Tech. Rep. September, 2021.

[52] T. L. White, D. P. Lambert, J. R. Zamecnik, and W. T. Riley, "Ion Chromatography (IC) Analysis of Glycolate in Simulated Waste," Savannah River National Laboratory, Aiken, Tech. Rep., 2015.

[53] J. Cornel, C. Lindenberg, and M. Mazzotti, "Quantitative application of in situ ATR-FTIR and Raman Spectroscopy in crystallization processes," *Analytical Chemistry*, vol. 47, pp. 4870–4882, 2008.

[54] G. R. Naik and W. W. Editors, *Blind Source Separation: Advances in Theory, Algorithms and Applications*. Springer, 2014, pp. v–vi, ISBN: 9783642550157.

[55] J. V. Stone, "Independent component analysis: An introduction," *Trends in Cognitive Sciences*, vol. 6, no. 2, pp. 59–64, 2002.

[56] D. M. Haaland and D. K. Melgaard, "New Prediction-Augmented Classical Least-Squares (PACLS) Methods: Application to Unmodeled Interferents," *Applied Spectroscopy*, vol. 54, no. 9, pp. 1303–1312, 2000.

[57] D. M. Haaland and D. K. Melgaard, "New augmented classical least squares methods for improved quantitative spectral analyses," *Vibrational Spectroscopy*, vol. 29, no. 1-2, pp. 171–175, 2002.

[58] D. K. Melgaard, D. M. Haaland, and C. M. Wehlburg, "Concentration residual augmented classical least squares (CRACLS): A multivariate calibration method with advantages over partial least squares," *Applied Spectroscopy*, vol. 56, no. 5, pp. 615–624, 2002.

[59] C. M. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Scholkopf, Eds. Springer, 2006, ISBN: 978-0387-31073-2.

[60] S. L. Brunton and N. J. Kutz, *Data-Driven Science and Engineering*, 2nd. Cambridge: Cambridge University Press, 2022.

[61] R. Tauler, A. Izquierdo-Ridorsa, and E. Casassas, "Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 293–300, 1993.

[62] R. Tauler, "Multivariate curve resolution applied to second order data," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 133–146, 1995.

[63] J. Jaumot, A. de Juan, and R. Tauler, "MCR-ALS GUI 2.0: New features and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 140, pp. 1–12, 2015.

[64] A. de Juan and R. Tauler, "Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 3-4, pp. 163–176, 2006.

[65] A. De Juan, J. Jaumot, and R. Tauler, "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem," *Analytical Methods*, vol. 6, no. 14, pp. 4964–4976, Jul. 2014.

[66] C. H. Camp, "PyMCR: A Python Library for Multivariate Curve Resolution Analysis With Alternating Regression (MCR-AR)," *Journal of Research of the National Institute of Standards and Technology*, vol. 124, no. 124018, pp. 1–10, 2019.

[67] S. Kocevska, G. M. Maggioni, S. H. Crouse, R. Prasad, R. W. Rousseau, and M. A. Grover, "Effect of ion interactions on the Raman spectrum of NO3—: Toward monitoring of low-activity nuclear waste at Hanford," *Chemical Engineering Research and Design*, vol. 181, no. 1, pp. 173–194, 2022.

[68] J. Y. Yu, Y. Zhang, S. H. Tan, Y. Liu, and Y. H. Zhang, "Observation on the Ion Association Equilibria in NaNO3 Droplets using Micro-Raman Spectroscopy," *Journal of Physical Chemistry B*, vol. 116, no. 41, pp. 12 581–12 589, 2012.

[69] E. Serjeant and B. Dempsey, *Ionisation constants of organic acids in aqueous solution*. Oxford: Pergamon Press, 1979.

[70] M. das Graças Gomes, S. da S.S. Borges, L. G. Lopes, and D. W. Franco, "UV-visible spectrum of nitrous acid in solution: pKa determination and analytical applications," *Analytica Chimica Acta*, vol. 282, no. 1, pp. 81–85, 1993.

[71] S. H. Crouse, R. W. Rousseau, and M. A. Grover, "A feature selection method for overlapping peaks in vibrational spectroscopy using nonnegatively constrained classical least squares," *Computers and Chemical Engineering*, vol. 189, no. July, p. 108 785, 2024.

[72] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[73] D. M. Haaland and D. K. Melgaard, "New Classical Least-Squares/Partial Least-Squares Hybrid Algorithm for Spectral Analyses," *Applied Spectroscopy*, vol. 55, no. 1, pp. 1–8, 2001.

[74] M. K. Antoon, J. H. Koenig, and J. L. Koenig, "Least-Squares Curve-Fitting of Fourier Transform Infrared Spectra With Applications in Polymer Systems.," *Applied Spectroscopy*, vol. 31, no. 6, pp. 518–524, 1977.

[75] D. M. Haaland and R. G. Easterling, "Application of New Least-Squares Methods for the Quantitative Infrared Analysis of Multicomponent Samples.," *Applied Spectroscopy*, vol. 36, no. 6, pp. 665–673, 1982.

[76] X.-g. Fan, Y. Zeng, Y.-L. Zhi, T. Nie, Y.-j. Xu, and X. Wang, "Signal-to-noise ratio enhancement for Raman spectra based on optimized Raman spectrometer and convolutional denoising autoencoder," *Journal of Raman Spectroscopy*, vol. 52, pp. 890–900, 2021.

[77] J. Lv, Z. Chen, X. Luan, and F. Liu, "Denoising stacked autoencoders-based near-infrared quality monitoring method via robust samples evaluation," *Canadian Journal of Chemical Engineering*, vol. 101, no. 5, pp. 2693–2703, 2023.

[78] C. He et al., "Accurate Tumor Subtype Detection with Raman Spectroscopy via Variational Autoencoder and Machine Learning," *ACS Omega*, vol. 7, no. 12, pp. 10 458–10 468, 2022.

[79] J. Xu et al., "High-Speed Diagnosis of Bacterial Pathogens at the Single Cell Level by Raman Microspectroscopy with Machine Learning Filters and Denoising Autoencoders," *ACS Chemical Biology*, vol. 17, no. 2, pp. 376–385, 2022.

[80] S.-j. Baek, A. Park, A. Shen, and J. Hu, "A background elimination method based on linear programming for Raman spectra," *Journal of Raman Spectroscopy*, vol. 42, pp. 1987–1993, 2011.

[81] D. C. Heinz and C.-I. Chang, "Fully Constrained Least Squares Linear Spectral Mixture Analysis Method for Material Quantification in Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, 2001.

[82] L. Wang, D. Liu, and Q. Wang, "Geometric method of fully constrained least squares linear spectral mixture analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 6, pp. 3558–3566, 2013.

[83] K. Muteki et al., "Mixture component prediction using iterative optimization technology (Calibration-Free/Minimum Approach)," *Industrial and Engineering Chemistry Research*, vol. 52, no. 35, pp. 12 258–12 268, 2013.

[84] Z. Shi, J. Hermiller, and S. G. Muñoz, "Estimation of mass-based composition in powder mixtures using Extended Iterative Optimization Technology (EIOT)," *AIChE Journal*, vol. 65, no. 1, pp. 87–98, 2019.

[85] S. G. Muñoz and E. H. Torres, "Supervised Extended Iterative Optimization Technology for Estimation of Powder Compositions in Pharmaceutical Applications: Method and Lifecycle Management," *Industrial and Engineering Chemistry Research*, vol. 59, no. 21, pp. 10 072–10 081, 2020.

[86] W. Saeys, K. Beullens, J. Lammertyn, H. Ramon, and T. Naes, "Increasing Robustness Against Changes in the Interferent Structure by Incorporating Prior Information in the Augmented Classical Least-Squares Framework," *Analytical Chemistry*, vol. 80, no. 13, pp. 4951–4959, 2008.

[87] H. Martens and T. Naes, "Multivariate Calibration by Data Compression," in *Near-Infrared Technology in the Agricultural and Food Industries*, P. Williams and K. Norris, Eds., Minnesota: American Association of Cereal Chemists, 1987, ch. 4, pp. 57–87, ISBN: 0-913250-49-X.

[88] H. Martens and T. Naes, *Multivariate Calibration*. John Wiley & Sons, 1989, ISBN: 0 471 90979 3.

[89] P. J. Larkin, *IR and Raman Spectroscopy - Principles and Spectral Interpretation*. Elsevier Inc., 2011, ISBN: 978-0-12-386984-5.

[90] R. L. McCreery, *Raman Spectroscopy for Chemical Analysis*. 2000, vol. 157, ISBN: 0471252875.

[91] N. V. Sahinidis, "Optimization under uncertainty: State-of-the-art and opportunities," *Computers and Chemical Engineering*, vol. 28, no. 6-7, pp. 971–983, 2004.

[92] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[93] J. E. Jackson, *A User's Guide to Principal Components*. John Wiley & Sons, 1991, ISBN: 0-471-62267-2.

[94] U. Manmontri and P. A. Naylor, "A class of frobenius norm-based algorithms using penalty term and natural gradient for blind signal separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1181–1193, 2008.

[95] C. Ru, W. Wen, and Y. Zhong, "Raman spectroscopy for on-line monitoring of botanical extraction process using convolutional neural network with background subtraction," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 284, 2023.

[96] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. Buydens, and E. Marchiori, "Convolutional neural networks for vibrational spectroscopic data analysis," *Analytica Chimica Acta*, vol. 954, pp. 22–31, 2017.

[97] Y. Guo, W. Jin, W. Wang, Z. Guo, and Y. He, "Unsupervised convolutional variational autoencoder deep embedding clustering for Raman spectra," *Analytical Methods*, vol. 14, pp. 3898–3910, 2022.

[98] X. Yan, S. Zhang, H. Fu, and H. Qu, "Combining convolutional neural networks and on-line Raman spectroscopy for monitoring the Cornu Caprae Hircus hydrolysis process," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 226, 2020.

[99] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function," *Neural Networks*, vol. 6, no. 6, pp. 861–867, 1993.

[100] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[101] H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[102] S. Wold, M. Sjöström, and L. Eriksson, "PLS-Regression: a Basic Tool of Chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[103] H. WOLD, *Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments*. ACADEMIC PRESS, INC., 1973, pp. 383–407.

[104] IBM, *CPLEX Optimization Studio*, www.cplex.com.

[105] M. L. Bynum et al., *Pyomo – Optimization Modeling in Python*, 3rd. 2012, vol. 67, ISBN: 978-3-030-68927-8.

[106] D. M. Scott, "Recent advances in in-process characterization of suspensions and slurries," *Powder Technology*, vol. 399, p. 117 159, 2022.

[107] N. T. Boily, H. M. Felmy, A. S. Medina, J. M. Bello, S. A. Bryan, and A. M. Lines, "Development of an Attenuated Total Reflectance-Ultraviolet-Visible Probe for the Online Monitoring of Dark Solutions," *ACS Sensors*, vol. 10, no. 1, pp. 122–132, Jan. 2025.

[108] E. Simone, A. N. Saleemi, and Z. K. Nagy, "Application of quantitative Raman spectroscopy for the monitoring of polymorphic transformation in crystallization processes using a good calibration practice procedure," *Chemical Engineering Research and Design*, vol. 92, no. 4, pp. 594–611, 2014.

[109] Q. Wang, F. Li, X. Jiang, S. Wu, and M. Xu, "On-stream mineral identification of tailing slurries of tungsten: Via NIR and XRF data fusion measurement techniques," *Analytical Methods*, vol. 12, no. 25, pp. 3296–3307, 2020.

[110] J. D. Vienna, D. C. Skorski, D. S. Kim, and J. Matyas, "Glass Property Models and Constraints for Estimating the Glass to be Produced at Hanford by Implementing Current Advanced Glass Formulation Efforts," Pacific Northwest National Laboratory, Richland, WA, Tech. Rep. July, 2013.

[111] A. Westesen, B. Wells, and R. Peterson, "Identifying challenge sludges to process at the hanford site," *Particulate Science and Technology*, vol. 41, pp. 453–459, 4 2022.

[112] M. J. Siegfried and M. E. Stone, "LAW Simulant Recipes for Evaluation of Real-Time, In-Line Monitoring Instruments," Savannah River National Laboratory, Tech. Rep., 2020.

[113] G. Févotte, "In situ Raman spectroscopy for in-line control of pharmaceutical crystallization and solids elaboration processes: A review," *Chemical Engineering Research and Design*, vol. 85, no. 7 A, pp. 906–920, 2007.

[114] B. Wells, P. Gauglitz, L. Mahoney, and M. Fountain, "Technical Gaps in Hanford High-Level Waste Solids Settling Behavior and Settling Time Evaluation for Direct

Feed High-Level Waste (DFHLW) Operations," Pacific Northwest National Laboratory, Tech. Rep., 2020.

[115] H. J. Van Manen, B. Rob, and O. F. Van Den Brink, "Focal length determination of raman immersion ball probes in diverse media," *Applied Spectroscopy*, vol. 63, no. 3, pp. 378–380, 2009.

[116] B. J. Marquardt, T. Le, and L. W. Burgess, "Demonstration of a high precision optical probe for effective sampling of solids by Raman spectroscopy," *Proceedings of The International Society for Optics and Photonics*, vol. 4469, no. October 2001, pp. 62–69, 2001.

[117] R. L. Russell, P. P. Schonewill, and C. A. Burns, "Simulant development for LAWPS testing," Pacific Northwest National Laboratory, Tech. Rep., 2017.

[118] C. B. Minnich, L. Helmdach, J. Ulrich, and M. P. Feth, "Model-Based Recognition of Mid-Infrared Sensor Fouling in Paracetamol Crystallization," *Chemical Engineering and Technology*, vol. 38, no. 8, pp. 1303–1307, 2015.

[119] Å. Rinnan, "Pre-processing in vibrational spectroscopy-when, why and how," *Analytical Methods*, vol. 6, no. 18, pp. 7124–7129, 2014.

[120] C. J. Strachan, T. Rades, K. C. Gordon, and J. Rantanen, "Raman spectroscopy for quantitative analysis of pharmaceutical solids," *Journal of Pharmacy and Pharmacology*, vol. 59, pp. 179–192, 2007.

[121] B. Schrader, A. Hoffmann, and S. Keller, "Near-infrared Fourier transform Raman spectroscopy: Facing absorption and background," *Spectrochimica Acta Part A: Molecular Spectroscopy*, vol. 47, no. 9-10, pp. 1135–1148, 1991.

[122] G. L. Nelson, A. M. Lines, A. J. Casella, J. M. Bello, and S. A. Bryan, "Development and testing of a novel micro-Raman probe and application of calibration method for the quantitative analysis of microfluidic nitric acid streams," *Analyst*, vol. 143, no. 5, pp. 1188–1196, 2018.

[123] M. Lin, Y. Wu, and S. Rohani, "Simultaneous Measurement of Solution Concentration and Slurry Density by Raman Spectroscopy with Artificial Neural Network," *Crystal Growth and Design*, vol. 20, no. 3, pp. 1752–1759, 2020.

[124] S. W. Wong, C. Georgakis, G. D. Botsaris, K. Saranteas, and R. Bakale, "Online estimation and monitoring of diastereomeric resolution using FBRM, ATR-FTIR, and Raman spectroscopy," *Industrial and Engineering Chemistry Research*, vol. 47, no. 15, pp. 5576–5584, 2008.

[125] K. P. Burnham and D. R. Anderson, *Model Selection and Inference: A Practical Information-Theoretic Approach*, 1st. 1998, vol. 95, p. 341, ISBN: 9781475729191.

[126] M. Pelletier, "Quantitative Analysis Using Raman Spectrometry," *Focal Point*, vol. 57, no. 1, 2003.

[127] L. L. Simon et al., "Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review," *Organic Process Research and Development*, vol. 19, no. 1, pp. 3–62, 2015.

[128] M. R. Poirier, A. M. Howe, F. R. Miera, M. E. Stone, C. C. DiPrete, and M. E. Farrar, "WTP real-time in-line monitoring program tasks 4 and 6 : Data quality and management and preliminary analysis plan," Savannah River National Laboratory, Tech. Rep., 2017.

[129] D. C. Harris, *Quantitative Chemical Analysis*, Eighth. Clancy Marshall, 2010.

[130] R. J. Davies-Colley and D. G. Smith, "Turbidity, Suspended Sediment, and Water Clarity: A Review," *Journal of the American Water Resources Association*, vol. 37, no. 5, pp. 1085–1101, 2001.

[131] J. T. Kirk, *Light and Photosynthesis in Aquatic Ecosystems*. Cambridge University Press, 1994.

[132] E. T. Baker and J. W. Lavelle, "Effect of Particle Size on the Light Attenuation Coefficient of Natural Suspensions," *Journal of Geophysical Research*, vol. 89, no. C5, pp. 8197–8203, 1984.

[133] S. J. Julier and J. K. Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proceedings of the IEEE*, vol. 92, no. 12, p. 1958, 2004.

[134] E. A. Wan and A. T. Nelson, "Neural dual extended Kalman filtering: Applications in speech enhancement and monaural blind signal separation," *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, pp. 466–475, 1997.

[135] T. A. Wenzel, K. J. Burnham, M. V. Blundell, and R. A. Williams, "Dual extended Kalman filter for vehicle state and parameter estimation," *Vehicle System Dynamics*, vol. 44, no. 2, pp. 153–171, 2006.

[136] H. Khodadadi and H. Jazayeri-Rad, "Applying a dual extended Kalman filter for the nonlinear state and parameter estimations of a continuous stirred tank reactor," *Computers and Chemical Engineering*, vol. 35, no. 11, pp. 2426–2436, 2011.

[137] M. E. Farrar, M. R. Poirier, C. C. Diprete, and M. E. Stone, "WTP Real-Time In-Line Monitoring Program: Task 3: LAW and EMF Flow Sheet Sampling Points," Savannah River National Laboratory, Tech. Rep., 2018.

[138] E. A. Wan and A. T. Nelson, "Dual Extended Kalman Filter Methods," in *Kalman Filtering and Neural Networks*, S. Haykin, Ed., New York: John Wiley & Sons, 2001, ch. Five, pp. 123–173.

[139] M. He, Y. Wang, and E. Forssberg, "Slurry rheology in wet ultrafine grinding of industrial minerals: A review," *Powder Technology*, vol. 147, pp. 94–112, 1-3 Oct. 2004.

[140] D. Speed et al., "Physical, chemical, and in vitro toxicological characterization of nanoparticles in chemical mechanical planarization suspensions used in the semiconductor industry: Towards environmental health and safety assessments," *Environmental Science: Nano*, vol. 2, pp. 227–244, 3 Jun. 2015.

[141] D. Li, J. Liu, S. Wang, and J. Cheng, "Study on coal water slurries prepared from coal chemical wastewater and their industrial application," *Applied Energy*, vol. 268, Jun. 2020.

[142] A. Demirbas, G. Edris, and W. M. Alalayah, "Sludge production from municipal wastewater treatment in sewage treatment plant," *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*, vol. 39, pp. 999–1006, 10 May 2017.

[143] M. A. McDonald et al., "Reactive crystallization: A review," *Reaction Chemistry and Engineering*, vol. 6, pp. 364–400, 3 Mar. 2021.

[144] T. Mühler, C. M. Gomes, J. Heinrich, and J. Günster, "Slurry-based additive manufacturing of ceramics," *International Journal of Applied Ceramic Technology*, vol. 12, pp. 18–25, 1 Jan. 2015.

[145] Y. Liu, R. Zhang, J. Wang, and Y. Wang, "Current and future lithium-ion battery manufacturing," *ISCIENCE*, vol. 24, p. 102 332, 2021.

[146] J. J. Hastings, D. Rhodes, A. S. Fellerman, D. Mckendrick, and C. Dixon, "New approaches for sludge management in the nuclear industry," *Powder Technology*, vol. 174, pp. 18–24, 1-2 May 2007.

[147] J. Chun, T. Oh, M. Luna, and M. Schweiger, "Effect of particle size distribution on slurry rheology: Nuclear waste simulant slurries," *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 384, pp. 304–310, 1-3 Jul. 2011.

[148] P. A. Smith, D. R. Rector, and A. Shekarriz, "Microstructural and rheological characterization of colloidal aggregates of nuclear waste slurries," *Mineral Processing and Extractive Metallurgy Review*, vol. 20, pp. 311–324, 1 2000.

[149] J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, "Generalized contribution plots in multivariate statistical process monitoring," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, pp. 95–114, 2000.

[150] R. L. Postles and K. G. Brown, "The DWPF product composition control system at Savannah River: Statistical process control algorithm," in *Fifth Intrenational Symposium on Ceramics in Nuclear and Hazardous Waste Management*, Cincinnati, Ohio, 1991.

[151] K. G. Brown, "The product composition control system at Savannah River: Statistical process control algorithm," in *SAFEWASTE 93*, Avignon, France, 1993.

[152] J. L. Westcott, R. M. Jochen, and S. S. Prevette, "Using statistical process control to monitor radioactive waste characterization at a radioactive facility monitor radioactive waste characterization at a radioactive facility," in *Waste Management Symposium 2007*, Tuscon, AZ, 2007.

[153] S. J. Qin, "Statistical process monitoring: Basics and beyond," *Journal of Chemometrics*, vol. 17, no. 8-9, pp. 480–502, 2003.

[154] A. Ferrer, "Latent structures-based multivariate statistical process control: A paradigm shift," *Quality Engineering*, vol. 26, no. 1, pp. 72–91, 2014.

[155] J. S. Hunter, "The exponentially weighted moving average," *Journal of Quality Technology*, vol. 18, no. 4, pp. 203–210, 1986.

[156] C. Zhang, X. Gao, T. Xu, Y. Li, and Y. Pang, "Fault detection and diagnosis strategy based on a weighted and combined index in the residual subspace associated with PCA," *Journal of Chemometrics*, vol. 32, no. 11, pp. 1–17, 2018.

[157] M. S. Reis and G. Gins, "Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis," *Processes*, vol. 5, no. 3, 2017.

[158] I. Jul-Jørgensen, P. Facco, K. V. Gernaey, M. Barolo, and C. A. Hundahl, "Data fusion of Raman spectra in MSPC for fault detection and diagnosis in pharmaceutical manufacturing," *Computers and Chemical Engineering*, vol. 184, no. March, p. 108 647, 2024.

[159] R. Calvini, G. Foca, and A. Ulrici, "Data dimensionality reduction and data fusion for fast characterization of green coffee samples using hyperspectral sensors," *Analytical and Bioanalytical Chemistry*, vol. 408, no. 26, pp. 7351–7366, 2016.

[160] L. Strani, E. Mantovani, F. Bonacini, F. Marini, and M. Cocchi, "Fusing NIR and process sensors data for polymer production monitoring," *Frontiers in Chemistry*, vol. 9, no. October, pp. 1–9, 2021.

[161] J. Zhao, W. Li, H. Qu, G. Tian, and Y. Wei, "Real-time monitoring and fault detection of pulsed-spray fluid-bed granulation using near-infrared spectroscopy and multivariate process trajectories," *Particuology*, vol. 53, pp. 112–123, 2020.

[162] F. Tahir, M. T. Islam, J. Mack, J. Robertson, and D. Lovett, "Process monitoring and fault detection on a hot-melt extrusion process using in-line Raman spectroscopy and a hybrid soft sensor," *Computers and Chemical Engineering*, vol. 125, pp. 400–414, 2019.

[163] S. J. Hong and G. S. May, "Neural-network-based sensor fusion of optical emission and mass spectroscopy data for real-time fault detection in reactive ion etching," *IEEE Transactions on Industrial Electronics*, vol. 52, no. 4, pp. 1063–1072, 2005.

[164] E. Simone, A. N. Saleemi, and Z. K. Nagy, "In situ monitoring of polymorphic transformations using a composite sensor array of Raman, NIR, and ATR-UV/vis spectroscopy, FBRM, and PVM for an intelligent decision support system," *Organic Process Research and Development*, vol. 19, no. 1, pp. 167–177, 2015.

[165] T. R. Graham et al., "In situ 27Al NMR spectroscopy of aluminate in sodium hydroxide solutions above and below saturation with respect to gibbsite," *Inorganic Chemistry*, vol. 57, no. 19, pp. 11 864–11 873, 2018.

[166] A. Lopalco, A. A. Lopedota, V. Laquintana, N. Denora, and V. J. Stella, "Boric acid, a lewis acid with unique and unusual properties: Formulation implications," *Journal of Pharmaceutical Sciences*, vol. 109, pp. 2375–2386, 8 Aug. 2020.

[167] S. H. Crouse, L. F. De Jesus, R. W. Rousseau, and M. A. Grover, "Using data fusion to detect faults in high-level waste slurries with Raman spectroscopy, ATR-FTIR spectroscopy, and FBRM," in *WM 2025 Conference*, Phoenix, AZ, 2025.

[168] S. M. Azcarate, R. Ríos-Reina, J. M. Amigo, and H. C. Goicoechea, "Data handling in data fusion: Methodologies and applications," *TrAC - Trends in Analytical Chemistry*, vol. 143, 2021.

[169] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, pp. 403–414, 3 1995.

[170] J. M. Lee, C. K. Yoo, and I. B. Lee, "Statistical monitoring of dynamic processes based on dynamic independent component analysis," *Chemical Engineering Science*, vol. 59, no. 14, pp. 2995–3006, 2004.

[171] S. Wold, P. Geladi, K. I. M. Esbensen, and J. Ohman, "Multi-way principal components-and pls-analysis," vol. 1, no. July 1986, pp. 41–56, 1987.

[172] R. Dunia and S. J. Qin, "Subspace approach to multidimensional fault identification and reconstruction," *AIChE Journal*, vol. 44, no. 8, pp. 1813–1831, 1998.

[173] H.Hotelling, "Multivariate quality control illustrated by air testing of sample bombsights," *Techniques of statistical analysis*, p. 111, 1947.

[174] J. V. Kresta, J. F. Macgregor, and T. E. Marlin, "Multivariate statistical monitoring of process operating performance," *Canadian Journal of Chemical Engineering*, vol. 69, 1991.

[175] B. M. Wise and N. B. Gallagher, "The process chemometrics approach to process monitoring and fault detection," *Journal of Process Control*, vol. 6, no. 6, pp. 329–348, 1996.

[176] C. K. Kumara, W. J. Ng, A. Bandara, and R. Weerasooriya, "Nanogibbsite: Synthesis and characterization," *Journal of Colloid and Interface Science*, vol. 352, no. 2, pp. 252–258, 2010.

[177] A. U. Rajapaksha, M. Vithanage, R. Weerasooriya, and C. B. Dissanayake, "Surface complexation of nickel on iron and aluminum oxides: A comparative study with single and dual site clays," *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 405, pp. 79–87, 2012.

[178] J. Addai-Mensah, J. Dawe, R. Hayes, C. Prestidge, and J. Ralston, "The unusual colloid stability of gibbsite at high pH," *Journal of Colloid and Interface Science*, vol. 203, no. 1, pp. 115–121, 1998.

[179] D. D. Dunuwila, L. B. C. Ii, and K. A. Bergiund, "An investigation of the applicability of attenuated total reflection infrared spectroscopy for measurement of solubility and supersaturation of aqueous citric acid solutions," *Journal of Crystal Growth*, vol. 137, pp. 561–568, 1994.

[180] C. Su and D. L. Suarez, "Coordination of adsorbed boron: A FTIR spectroscopic study," *Environ. Sci. Technol*, vol. 29, pp. 302–311, 2 1995.

[181] D. Peak, G. W. Luther, and D. L. Sparks, "Atr-ftir spectroscopic studies of boric acid adsorption on hydrous ferric oxide," *Geochimica et Cosmochimica Acta*, vol. 67, pp. 2551–2560, 14 Jul. 2003.

[182] A. R. Heath, P. D. Fawell, P. A. Bahri, and J. D. Swift, "Estimating average particle size by focused beam reflectance measurement (FBRM)," *Particle and Particle Systems Characterization*, vol. 19, no. 2, pp. 84–95, 2002.

[183] A. Junghans et al., "Laser-induced breakdown spectroscopy (libs) process monitoring for the dflaw vitrification facility," in *WM2021 Conference*, Mar. 2021, ISBN: 978-0-9828171-8-6.

[184] J. Piper, E. Kurtz, E. Cockram, R. Lakis, and A. Junghans, "Enabling in-situ libs measurements of liquids and slurries enabling in-situ libs measurements of liquids and slurries," Los Alamos National Laboratory, Tech. Rep., Aug. 2024.

[185] J. A. Manrique-martinez et al., "Evaluation of multivariate analyses and data fusion between raman and laser-induced breakdown spectroscopy in binary mixtures and its potential for solar system exploration," *Journal of Raman Spectroscopy*, pp. 1–16, January 2019 2020.

[186] M. Z. Bell and A. Macfarlane, ""Fixing" the nuclear waste problem? The new political economy of spent fuel management in the United States," *Energy Research and Social Science*, vol. 91, no. June, p. 102 728, 2022.

[187] M. K. McBeth, M. Warnement Wrobel, and I. van Woerden, "Political ideology and nuclear energy: Perception, proximity, and trust," *Review of Policy Research*, vol. 40, no. 1, pp. 88–118, 2023.

[188] M. Gallucci, "A Glass Nightmare: Cleaning up the Cold War's Nuclear Legacy at Hanford," *IEEE Spectrum*, pp. 1–10, 2020.

[189] M. McCormick, S. L. Samuelson, D. Faulk, and J. Hedges, "Hanford Federal Facility Agreement and Consent Order: Hanford Public Involvement Plan," Tech. Rep., 2017, p. 38.

[190] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *American Statistician*, vol. 72, no. 4, pp. 309–314, 2018. arXiv: 1512.00809.

[191] R. D. Tobias et al., "An introduction to partial least squares regression," in *Proceedings of the twentieth annual SAS users group international conference*, SAS Institute Inc. Cary, NC, USA, vol. 20, 1995, pp. 1250–1257.

[192] P. Geladi and B. R. Kowalski, "Partial Least-Squares Regression - A Tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[193] D. M. Haaland and E. V. Thomas, "Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information," *Analytical chemistry*, vol. 60, no. 11, pp. 1193–1202, 1988.

[194] H. Abdi, "Partial least square regression (PLS regression)," *Encyclopedia for research methods for the social sciences*, vol. 6, no. 4, pp. 792–795, 2003.

[195] H. Wold, "Soft modelling by latent variables: The non-linear iterative partial least squares (nipals) approach," *Journal of Applied Probability*, vol. 12, pp. 117–142, S1 1975.

[196] Z. S. Nickolov, O. Ozcan, and J. D. Miller, "FTIR analysis of water structure and its significance in the flotation of sodium carbonate and sodium bicarbonate salts," *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 224, no. 1-3, pp. 231–239, 2003.

[197] D. Peak, G. W. Luther, and D. L. Sparks, "ATR-FTIR spectroscopic studies of boric acid adsorption on hydrous ferric oxide," *Geochimica et Cosmochimica Acta*, vol. 67, no. 14, pp. 2551–2560, Jul. 2003.

[198] C. Su and D. L. Suarez, "Coordination of Adsorbed Boron: A FTIR Spectroscopic Study," *Environ. Sci. Technol*, vol. 29, no. 2, pp. 302–311, 1995.

[199] J. R. Nielsen and N. E. Ward, "Raman spectrum and structure of the metaborate ion," *The Journal of Chemical Physics*, vol. 5, no. 3, p. 201, 1937.

[200] Y. Ma, W. Yan, Q. Sun, and X. Liu, "Raman and infrared spectroscopic quantification of the carbonate concentration in K2CO3 aqueous solutions with water as an internal standard," *Geoscience Frontiers*, vol. 12, no. 2, pp. 1018–1030, 2021.

[201] G. Deo and I. E. Wachs, "Predicting molecular structures of surface metal oxide species on oxide supports under ambient conditions," *Journal of Physical Chemistry*, vol. 95, no. 15, pp. 5889–5895, 1991.

[202] W. P. Griffith and T. D. Wickins, "Raman Studies on Species in Aqueous Solutions. Part 1. The Vanadates," *Journal of the Chemical Society A*, pp. 1087–1090, 1966.

[203] K. J. Kingma and R. J. Hemley, "Raman spectroscopic study of microcrystalline silica," *American Mineralogist*, vol. 79, no. 3-4, pp. 269–273, 1994.

[204] S. K. Sharma, J. F. Mammone, and M. F. Nicol, "Raman investigation of ring configurations in vitreous silica," *Nature*, vol. 292, no. 5819, pp. 140–141, 1981.

[205] K. E. Kuebler, B. L. Jolliff, A. Wang, and L. A. Haskin, "Extracting olivine (Fo-Fa) compositions from Raman spectral peak positions," *Geochimica et Cosmochimica Acta*, vol. 70, no. 24, pp. 6201–6222, 2006.

[206] Z. Yang, H. Peng, W. Wang, and T. Liu, "Crystallization behavior of poly(ϵ-caprolactone)/layered double hydroxide nanocomposites," *Journal of Applied Polymer Science*, vol. 116, no. 5, pp. 2658–2667, 2010.

[207] C. P. Marshall, W. J. Dufresne, and C. J. Rufledt, "Polarized Raman spectra of hematite and assignment of external modes," *Journal of Raman Spectroscopy*, vol. 51, pp. 1522–1529, 2020.

[208] I. Chamritski and G. Burns, "Infrared- And raman-active phonons of magnetite, maghemite, and hematite: A computer simulation and spectroscopic study," *Journal of Physical Chemistry B*, vol. 109, no. 11, pp. 4965–4968, 2005.

VITA

Steven Crouse was born on August 4^{th}, 1999, to parents Martin and Sandra Crouse in North Carolina. He graduate from the North Carolina School of Science and Mathematics in 2017 and from North Carolina State University in 2020 with a B.S. in chemical engineering and a minor in mathematics. Steven performed undergraduate research with Milad Abolhasani at NC State where he worked on a microfluidic platform for quantum dot synthesis and analysis. In 2020, he began a PhD at the Georgia Institute of Technology working with Dr. Martha Grover and Dr. Ronald Rousseau.

In his spare time, Steven enjoys playing guitar, reading, running, weightlifting, playing soccer, hiking, watching movies, watching football, and socializing with friends. His office space is characterized by an organized array of gifts given to him by friends during his PhD, and he can often be found there taking a twelve-minute nap to prepare his mind for "thinking."