DIFFRACTION SPECTROSCOPY OF METALLOPROTEINS

By

Darren A. Sherrell

A thesis submitted to the College of Graduate Studies and Research In partial fulfillment of the requirements for the Doctor of Philosophy In the Department of Geological Sciences University of Saskatchewan, Saskatoon, Canada

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the libraries of this university may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Department Head
Geological Sciences
University of Saskatchewan
114 Science Place Saskatoon
Saskatchewan, Canada
S7N 5E2

ABSTRACT

X-ray absorption is not only element specific, but atom specific: two atoms of the same element in different states or in different neighbourhoods will have slightly different absorption characteristics. These energy dependent atomic form factors are carried over to the diffraction intensities. The atomic form factors are sensitive not only to the the energy of the X-ray but also the diffraction criteria; providing individual local physical data at different ratios in various diffractions. This process is referred to as *site selectivity*, it is unique to Diffraction Spectroscopy, and is achieved only when the sample is in crystal form. Through this work, a technique has been devised to site-separate two atoms of iron from within a protein, that builds on prior small unit cell Diffraction Anomalous Fine Structure experiments and harnesses the collection and processing software commonly used in large unit cell crystallography. A technique (*dev* + PCA) has been developed to retrieve the small signals from individual atom-labels out of the large and noisy background of real diffraction taken across a spectrum. The intensity of the diffractions are calculated by integrating over multiple images, profiling spots, merging datasets, and scaling across the whole spectrum. This thesis explores how Diffraction Spectroscopy can be used effectively on large unit cells, namely those of proteins. Site-selective absorption experiments were conducted on large unit cell crystals at a 3rd generation synchrotron beamline, exclusively using existing equipment. The spectra generated were limited in scope but are an adequate proof of concept.

Remember, kids, the only difference between

screwing around and science

is writing it down.

-Adam Savage

ACKNOWLEDGEMENTS

In the end I would like to thank Dr. Jay Nix for his wise council, the use of his lab and for his insights into crystallography; also for being there, for always being there. I'd also like to thank professors Graham George and Ingrid Pickering for introducing me to the subject and keeping the faith over the many years this took to come to fruition. My thesis committee for their lively discussions and pertinent direction. Dr. Edwin Westbrook for teaching me about beamline physics and equipment and Dr. Stephan Friedrich for giving me my first chance and my second chance too. Dr. Gary Brudvig and Kari Young for their excellent small molecules. Dr. Julien Cotelesage for collecting data on those crystals. Aina Cohen for great advice and writing beamline software for our ever-changing methodology at Stanford Synchrotron Radiation Lightsource (SSRL) and Michel Fodje for implementing the same at the Canadian Light Source (CLS). Thomas Spatzal and Eva-Maria Roth for ferredoxin crystals. The ever-clear writings of Templeton and Templeton and those of Dr. Julie O. Cross. All those that did Diffraction Anomalous Fine Strucutre in the 1990's for forging the way and the DAFS bibliography at University of Chicago. To Dr. Mark Hackett for a single afternoons' conversation about Principle Component Analysis whose ideas have yet to be implemented. Dr. Brian Bewer for his insight into quantum mechanics. Jessica & Cooper Jasner and Wyatt Sherrell for keeping things on the lighter side. Mum, Dad and 32 Cordova for having my back. English is my first language but I'd like to give special thanks to those that made it read that way, my editors: Liz Glasgow, Chris Culpeper, Jon Bosworth, Dr. Jay Nix and Dr. LisaMarie Wands.

I thank Canadian Institutes of Health Research (CIHR) - Training grant in Health Research Using Synchrotron Techniques (CIHR-THRUST) for a Fellowship, and Natural Sciences and Engineering Research Council (NSERC) for additional funding. Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic

Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health, National Institute of General Medical Sciences (including P41GM103393) . The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH. A portion of the data in this thesis was performed at the CLS, which is funded by the Canada Foundation for Innovation, NSERC of Canada, the National Research Council Canada, the CIHR, the Government of Saskatchewan, Western Economic Diversification Canada, and the University of Saskatchewan. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

TABLE OF CONTENTS

Permission to Use……..i

Abstract……ii

Quote……iii

Acknowledgements……..iv

List of Figure……..x

List of Abbreviations…..……xiii

Table of Software …..……..……..….xiv

CHAPTER 1……..1

INTRODUCTION ..……..1

1.1 Motivation ..……..……..……1

1.2 Metalloproteins……..2

1.3 X-ray Absorption Spectroscopy (XAS) …..……..4

1.4 Macromolecular Crystallography (MX)……..5

1.5 Diffraction Spectroscopy (DS) ……..6

1.6 The Experiment……6

1.7 Analysis……..……7

1.8 Ways and Means……8

CHAPTER 2……10

X-RAY ABSORPTION SPECTROSCOPY THEORY ..……..…10

2.1 Classical ……..10

2.2 Quantum ..……..…11

CHAPTER 3……16

DIFFRACTION THEORY……..16

3.1 Scattering Light - Classical……..……..…16

3.2 Laue Equations……..……..…17

3.3 Intensity of a diffracted spot……19
3.3.1 Temperature Factor and Occupancy ……20
3.3.2 Lorentz Factor……21
3.3.3 Polarization……21
3.3.4 Self-Absorption……22
3.3.5 Detector……23
3.4 Approach to Calculating Intensities ……23
3.5 Hamiltonian of Diffraction……25
3.5.1 Sign Convention ……28
3.6 Comparison with Absorption……29

CHAPTER 4 ……33
DIFFRACTION SPECTROSCOPY THEORY……33
4.1 Atomic Specificity……33
4.2 Structure Factor Calculation ……35
4.3 The Example ……38
4.3.1 Expansion of the Target Atoms……40
4.3.2 Variation in Intensity……44
4.3.3 Opposite Bias ……45
4.3.4 Realism ……45
4.4 Cylinder Projection……47

CHAPTER 5 ……49
ANALYSIS OF DIFFRACTION SPECTROSCOPY ……49
5.1 Data Extraction……49
5.1.1 Outlier Rejection, Dixon Q-test and Feature Scaling ……51
5.2 Calculating Deviation *(dev)* ……52
5.3 Extracting Signal ……54
5.3.1 Separating Sets ……54
5.3.2 Principal Component Analysis ……55

CHAPTER 6 ……58
SIMULATED DIFFRACTION……58
6.1 Raw Ingredients to Simulated Diffraction ……58
6.2 Simulated Ferredoxin ……60
6.3 Results of Simulated Experiment ……62
6.4 DeskTools ……66
6.5 Simulated Ferredoxin Summary ……67

CHAPTER 7 ……73
MYOGLOBIN ……73
7.1 The Protein ……73
7.2 The Experiment ……75
7.3 Collection History ……78
7.4 The Dimension 1 Anomaly ……79
7.5 Results of Myoglobin PCA ……82
7.6 Fourier Transforming of Dimension 2……84
7.7 Myoglobin Conclusion ……87

CHAPTER 8 ……88
FERREDOXIN……88
8.1 The Protein ……90
8.2 The Comparison Spectra ……91
8.3 The Data……95
8.3.1 Collection……95
8.3.2 Processing ……95
8.4 Results……99
8.5 BacksubRot.py Fitting Algorithm ……102
8.6 Analysis of the Components of the Diffraction after Fitting……107
8.7 Conclusion……108

CHAPTER 9 ……111

DISCUSSION 111

9.1 Short Term Upgrades.....113

9.1.1 PCA Investigation 113

9.1.2 Simultaneous XAS 113

9.1.3 Continuous Collection 114

9.1.4 Multiple Target Atoms.....114

9.1.5 Data Mining the PDB 115

9.1.6 Temperature and Normal Polarization.....115

9.1.7 Cluster DS.....116

9.1.8 High Contrast Collection.....116

9.1.9 Relationship to the Rees Method of Separation.....116

9.1.10 Software Upgrade.....117

9.2 Middle Distance Upgrades.....117

9.2.1 Limits.....117

9.2.2 Diffraction Anomalous Fine Structure.....117

9.2.3 Bond Polarization.....118

9.2.4 Phasing.....118

9.3 Long Term.....119

9.4 Outro.....120

Appendix I: Scattering Theory.....121

Appendix II: Simulated Ferredoxin XAS from PDB structures 126

Appendix III: Fast Fourier Transform of Kramers-Kronig in Python.....131

Appendix IV: PDB File Conversion of FEFF.inp in Python 134

Bibliography 139

LIST OF FIGURES

CHAPTER 2.....**10**

2-1 K-edge Absorption Spectra of Various Heavy Elements.....15

CHAPTER 3.....**16**

3-1 Diffraction Condition/Direction as Integer Multiples of 2π.....18

3-2 Kramers-Kronig Anomalous Dispersion Relation.....30

3-3 Feynman Diagrams of Absorption vs Diffraction.....31

3-4 Feynman Diagrams vs Atomic Form Factor 32

CHAPTER 4 .....**33**

4-1 Anomalous Dispersion of Similar Irons.....34

4-2 Ferredoxin Unit Cell 36

4-3 Anomalous Dispersion of Elements within Ferredoxin.....37

4-4 Argand Diagram of Every Atom in Ferredoxin.....39

4-5 Argand Diagram of the Complex Contribution from a Single Iron.....41

4-6 Total Structure Factor with Single Iron Anomalous Contribution.....43

4-7 Total Structure Factor for two Diffractions with Opposite Iron Emphasis.....46

4-8 Cylinder Spectra of Anomalous Dispersion for Two Different Irons 48

CHAPTER 5 .....**49**

5-1 Testing the PCA Module.....57

CHAPTER 6 .....**58**

6-1 Iron Sulphur Active Site of Ferredoxin 61

6-2 Simulated Ferredoxin Diffractions in HKL-space 63

6-3 Results of Principle Component Analysis on Simulated Diffraction.....65

6-4 Cylinder Spectra of the PCA Module working on *dev*(Fe1:0.95) and *dev*(Fe2:0.95) for Simulated Ferredoxin ……..…..……..…69
6-5 Cylinder Projection of the Rotated and Scaled PCA for Simulated Ferredoxin ……70
6-6 Reduced Iron Spectrum vs Rotated and Scaled from PCA of *dev*(Fe1). ……..…71
6-7 Oxidized Iron Spectrum vs Rotated and Scaled from PCA of *dev*(Fe2)……72

CHAPTER 7 ……..73
7-1 Myoglobin Active Site……..…74
7-2 Screen Output from DeskTools running Myoglobin ……..……77
7-3 The Dimension 1 Anomaly vs Energy and Time……80
7-4 PCA Results of Myoglobin ..……..81
7-5 Fluorescence Spectra vs Dimension 3……83
7-6 Kramers-Kronig of Dimension 2 of Myoglobin DS……..…85
7-7 Fluorescence Spectra vs Kramers-Kronig of Dimension 2……..…86

CHAPTER 8 ……88
8-1 Images of Ferredoxin Crystal I2 Collection Experiment ……..……89
8-2 Kramers-Kronig Transform of Ferredoxin Reduced Iron ……..……..92
8-3 Kramers-Kronig Transform of Ferredoxin Oxidized Iron ……..……..93
8-4 Cylinder Spectra of the Dispersion Relation from EXAFS Absorption Spectra of Reduced and Oxidized Ferredoxin Irons ……..……..94
8-5 Master_XDS.inp File for Ferredoxin I2……97
8-6 XSCALE.inp File for Ferredoxin I2……..98
8-7 Abbreviated Screen Output From DeskTools ……100
8-8 Results of the PCA Module Working on Ferredoxin ……101
8-9 Background Subtraction of the Fe2 PCA Components 2 and 3 ……103
8-10 Background Subtraction of the Fe1 PCA Components 2 and 3……104
8-11 Final Comparison of DS PCA Rotated and Scaled Ferredoxin Inner Iron with Fe3+ XAS Spectra ……105

8-12 Final Comparison of DS PCA Rotated and Scaled Ferredoxin Outer Iron with Fe2+XAS Spectra 106
8-13 Savitzky-Golay Fit of Ferredoxin DS PCA 110

LIST OF ABBREVIATIONS

AFA	Abstract Factor Analysis
ALS	Advanced Light Source
CCD	Charge Coupled Device
CLS	Canadian Light Source
CL	Cromer Lieberman
CMOS	Complementary Metal Oxide Semiconductor
DAFS	Diffraction Anomalous Fine Structure
DANES	Diffraction Anomalous Near Edge Spectroscopy
dev	Sample Standard Deviation of Intensities
DS	Diffraction Spectroscopy
d*TREK	Processing Software
EXAFS	Extended X-ray Absorption Spectroscopy
FEFF	Ab Initio Multiple Scattering Software
FFT	Fast Fourier Transform
FORTRAN	Formula Translating System
KK	Kramers-Kronig
MAD	Multi-wavelength Anomalous Dispersion
MX	Macromolecular Crystallography
PCA	Principle Component Analysis
PDB	Protein Data Bank
SSRL	Stanford Synchrotron Radiation Laboratory
SVD	Single Valued Decomposition
XAS	X-Ray Absorption Spectroscopy
XDS	X-ray Diffraction Software

TABLE OF SOFTWARE

Program Name or Major Subroutine	Function
process_w_xds.py	Process each energy wedge with *xds_par* using the same master XDS.INP file.
make_multipleXSCALEinp.py	Create XSCALE.inp file for scaling multiple runs automatically from processed XDS_ASCII.HKL files.
DeskTools.py (Top Level Program Name)	Take input PDB file, hkl file(s) and keyword arguments and calculate *dev* dictionary. Execute PCA subroutine and plot. Make single and multiple synthetic files. Execute DAFS for single atom (Myoglobin) or multiple (Ferredoxin).
DeskTools.BriefCase()	Extract relevant values from PDB file. Create {[hkl]: Intensity} dictionary from collected data. Get Cromer-Mann coefficients and Anomalous {[energy][label]: [fp, fpp]} dictionary from Cromer-Liberman or input spectra
DeskTools.PaperPad()	Fill out unit cell from asymmetric sub-unit data and symmetry operations. Create dictionary of intensities based on theory (not data), for use in calculating *dev.*
DeskTools.MathematicalMethods()	Calculate reciprocal metric tensor, d-space, PCA, cubic mean std, row normalization, Savitzky-Golay
DeskTools.WipeBoard()	Create single synthetic file. Calculate for use in *dev.* Make *dev* dictionary. Dixon-Q test and outlier rejection.
rescale_rotate_save_and_plot.py	Take output PCA components from DeskTools and scale, rotate and plot them against comparison spectra.
BackSubRot.py	Background Subtraction and rotation of the cylindrical projected components
plot_3D.py & view_data.py	Various plotting methods: 2D, 3D and HKL-space.
fftkk.py	Kramers-Kronig Transform f using Cromer (fortran) values.
pdb2feff.py	Convert *filename.pdb* to *feff.inp,* execute FEFF8 and plot.

All software written by author.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Due to their wavelength, X-rays have been used since 1912 [1] to determine the nature and arrangement of systems at the atomic scale. The absorption wavelength of elements within some of the most interesting systems, molybdenum cofactors in nitrate reductase or the iron-sulphur clusters of thioredoxin-like ferredoxins for example, coincides with the wavelength of X-rays that diffract at experimentally useful angles. This energy-coincidence phenomena between diffraction and absorption has been adopted for what is called Multi-wavelength Anomalous Dispersion (MAD) phasing [2] of macromolecular crystals. Newer macromolecular crystallography (MX) beamlines designed and built to exploit this phenomenon.

The Protein Data Bank (PDB) was formed as a warehouse of the 3-dimensional structures of large macromolecules. In 1995, there were only 15 synchrotron beamlines dedicated to MX worldwide, and since then an average of 6 beamlines per year have been built and commissioned [3] to service the ever-growing structural biology community. Approximately 120 synchrotron MX beamlines exist at this time and account for 88% of the total structures deposited in the PDB. A few of these beamlines have already been decommissioned while many more are under construction. MAD phasing in crystallography monitors the change of intensity of the diffracted light at specific energies. The intensity of diffracted X-rays is half of the information supplying unit cell electron density maps; the other half is the phase. The diffraction intensity can be evaluated and used to help phase structures by tuning the energy of the X-rays to favourable absorption frequencies of target atoms within the cell. This causes well-

characterized intensity changes, which can be used to lock in the phase that is lost when the intensity is recorded. MAD-phasing typically only records the intensity at 3 energies. Diffraction Spectroscopy [4, 5] (DS), in contrast, follows the modulations of intensity over a wide range of energies, giving a spectrum. Since 1983, DS has only been conducted on very small molecules, well-ordered solids with small unit cells or highly symmetric systems [4] and always on a handful of carefully chosen diffraction spots [5]. The purpose of this thesis is to demonstrate that DS can be applied to much larger molecular systems using a standard MAD-capable beamline. A discussion of a mathematical framework for extracting DS results from such a complex system and its implications going forward is also included. Beyond the theoretical framework that has been developed, proof of single and multi-iron containing protein experiments are conducted. The results obtained in these simple and complex systems indicate that the techniques developed herein could help answer many interesting questions in biochemistry and biology.

1.2 Metalloproteins

It has been estimated that transition metal ions are in approximately 11% of all proteins. Two basic roles are fulfilled by these essential transition metals: (1) transfer, such as in nitrate reductase; and (2) catalytic, such as in metalloenzymes. Metalloenzymes are enzymes which contain a metal as an integral part of their active sites. These metals are responsible either directly or indirectly for a considerable portion of the interesting chemistry. Metalloenzymes catalyze a wide range of biochemical reactions. Cytochrome oxidase, the ultimate consumer of the O_2 that we breathe, contains a binuclear Cu-Fe active site. Photosystem II has a photosynthetic O_2 evolving complex, with a Mn_4Ca cluster at its active site. Nitrogenase, a key enzyme in the global nitrogen cycle contains a complex molybdenum-iron-sulfur cluster. These examples of metalloenzymes are remarkable not only for their chemistry or structure but also in their unmatched efficiency; the mechanisms of these processes are complex and need detailed study. An accurate understanding of the transition metals physical and

electronic structures is essential for understanding the metal complex's role within the larger protein structures.

The ultimate goal of focus of this research is to gain a better understanding of these structures and to apply DS to that end. An essential prerequisite for this is to develop the technique of DS so that it can be applied to complex systems, and for that purpose two-iron plant ferredoxin was selected. Ferredoxin is a much studied metalloprotein, the structure and biochemistry of which are well documented in the literature [7, 8]. Functionally, ferredoxin is a one electron capacitor: retaining a charge, and when appropriate, transferring that charge.

The two techniques involved in DS are X-ray Macromolecular Crystallography and X-ray Absorption Spectroscopy. The former provides a crystallographic atomic resolution model of a protein (beta sheets, alpha helices, locations of the metal complexes) and serves to explain the structure-function relationship of a macromolecule [9]. However, sometimes understanding the shape is just a starting point; X-ray Absorption Spectroscopy is employed to gain greater understanding of an active-sites' interatomic distances [10] if there is a metal present. XAS is usually performed separately from X-ray Macromolecular Crystallography, under different conditions, on a different beamline, and usually with a non-crystalline sample. XAS gives an order of magnitude improvement, than macromolecular crystallography, in bond length determination in the immediate cluster surrounding the target atom, and is comparable to small molecule crystallography. XAS is limited in that it sees all of a particular metal in a sample where DS is able to hone in on a single absorber (target atom). By combining the two techniques, Diffraction Spectroscopy can supply complementary information to crystallography by elucidating, spectroscopically, details of the an individual metal sites' environment.

1.3 X-ray Absorption Spectroscopy (XAS)

XAS has a relatively short history, with its first appearing in 1971 in a landmark paper by Sayers, Stern and Lytle [11] concerning amorphous and crystalline Germanium. Lytle et. al. showed using Fourier analysis of the oscillations on the high end of absorption can be represented as a sum of normalized Gaussians and that the Gaussians could model the radial distribution of the absorber's nearest neighbours. Lytle's XAS is a probe that gives very accurate information of the target atom coordination and its environment. XAS is also not constrained by crystallinity and can be collected in a variety of modalities such as absorption, fluorescence and electron yield. XAS can give the number of neighbours, the neighbouring atom types (to a lesser degree) as well as very accurate bond distances. The technique has been further developed to exploit the polarized synchrotron beam [12] by aligning the polarized beam with molecule orientation. It has also been refined to highlight magnetic effects as well as ligand field splitting [13].

XAS can normally measured by transmission on a dilute sample or fluorescence [14] on a concentrated (even solid) sample or a combination of both. It may also be used in imaging elemental distributions throughout a much larger (cm x cm) sample [15]. As X-rays pass through the sample, they are absorbed in very specific amounts at different energies depending on the element of the target atom and its environment. One of the drawbacks of XAS is that it excites all atoms of a given element in the sample, whether or not they are the target atom. As well as keeping the experiment free from contaminating elements of the same type, if there are more than one of the same element of interest in the sample itself the complexity increases in the interpretation of the data. Modern XAS analysis places the target atom's orientation, neighbouring atoms and location within a sample and anchors it to its surroundings using inference from chemistry, density functional theory or by visual inspection of the crystallographic structure. A more detailed discussion of XAS is presented in Chapter 2.

1.4 Macromolecular Crystallography (MX)

The use of X-ray crystallography in the elucidation of the three-dimensional structure of molecules has a history going back almost 100 years when Bragg [16] discovered the structure of sodium chloride. Physical chemists and biologists make common use of X-ray crystallography because the information gained is profoundly important to the understanding of molecules and macromolecules and in the design of future experiments. It is not an understatement to say that MX was a revolution to biology by throwing previously surmised atomic arrangements into sharp relief.

Almost all biological structures are now discovered using crystallography beamlines at synchrotrons. And newer, so called, 'third generation' synchrotron MX beamlines operate in the 5 - 18 KeV (2.47 - 0.67Å) range. Within this energy range lie the K-edge absorption edges, which are good for phasing, of elements from vanadium to zirconium as well as many L3 edges of elements such as tungsten or tantalum. Biologically relevant elements that fall within this spectrum include Fe, Mn, Se, Zn, and Cu. While a powerful technique, solving the three-dimensional structure of a macromolecule comes at a cost. For the protein's three-dimensional structure to be calculated, it must first be *crystallized*. This is the predominant rate-limiting step in X-ray crystallography. Crystals, and consequently their structures of challenging macromolecules can be decades in their discovery [17]; however crystallography has been successfully used to determine more than 85,500 structures [18].

The experimental data in X-ray crystallography are measurements of electron density of the macromolecule; atomic positions are located at the centres of the areas of highest density. A firm knowledge of the amino acid sequence and chemically realistic configurations allow the crystallographer to assign the correct atoms into the electron density map. The Fourier transform of the unit cell's density in real space is a set of structure factors (F) in reciprocal space. Diffraction occurs every time a reciprocal lattice point passes through the Ewald sphere [19]. Diffraction is then satisfied by rotating a (uniform single) crystal and selecting an appropriate sphere radius. The

location of a reciprocal lattice point is due to crystal symmetry. The intensity of a diffraction spot is from the contents of the unit cell, and the Ewald sphere's radius is inversely proportional to the X-ray's wavelength.

1.5 Diffraction Spectroscopy (DS)

The potential absorption of X-rays by a target atom, by causality [21], effects the diffraction in a well understood way. Atoms with different locations within the unit cell contribute at different levels to each diffraction by its energy dependent atomic form factor. DS can provide metal-ion local physical data and also distinguish between metals of the same element within the same sample. This process is referred to as site selectivity, is unique to DS, and is achieved only when the sample is in its crystal form. The unique signature from each target atom of the same element (with a different coordination or surrounding environments) allows diffractions that prefer that atom to be mined for its signature anomalous dispersion spectrum. The location of a diffracted X-ray is by dint of the crystal form, but the intensity of the diffracted spot depends on the contents of the unit cell.

This thesis explores how DS can be used effectively on much larger unit cells, namely those of proteins. Presented is a technique for extracting XAS-style data from elementally identical but not site-equivalent heavy (metal) atoms within a protein crystal. This is appropriate for metalloprotein crystals with transition metal elements at different locations or with different electronic configurations within the same macromolecule.

1.6 The Experiment

One of the driving forces behind these experiments is to discover whether DS can be done on very large macromolecules. Theoretically, absorption information is being utilized by MAD-phasing, but at the outset of this study it was unknown whether a beamline and macromolecular crystals could provide sufficient quality data for extraction of XAS-style spectra. An equally intriguing facet of the research is whether it could be

performed with existing equipment on a standard third generation beamline, allowing all beamlines of this and future generations the capacity to perform this experiment where and when necessary. The data collection strategy is straight forward. The beamline is tuned to an energy slightly below the absorption edge of the target element; diffractions are collected over a thin wedge of approximately 10° in 1° rotations. This is similar to a normal data collection except over a thinner wedge; during normal data collection it is not unusual to collect 60°-360° in 1° rotations. The beamline energy is then stepped up, and the exact same 10° are collected again. This is repeated at 50-70 energy points across the absorption edge of the target atom and, if the crystal is hardy, the whole experiment can be repeated again for better statistics. This collection method produces a few thousand diffractions over the range that the target atoms experience the absorption edge. When the diffraction intensity is combined with a high resolution structure from the same crystal, the diffractions can be categorized and analyzed for the anomalous dispersion spectra of the target atoms. Chapters 6, 7 and 8 discuss the experiments.

1.7 Analysis

Through this work, a technique has been devised to site-separate the atoms of interest that builds on small molecule DS utilizing collection and processing software commonly used in large unit cell crystallography. The intensity of the diffractions are calculated by integrating over multiple images, profiling spots, merging datasets, and scaling across the whole spectrum.

Diffraction theory describes that the very small signals from the individual target atoms will mix within a single diffracted spot at different ratios depending on the Miller indices of the diffraction and the locations of the atoms within the unit cell. The target atoms would normally be dwarfed by all the other atoms; however, there is an observable variation in intensity across the absorption spectrum of the element. DS theory allows for identification of which diffractions contribute most strongly from each of the separate target atoms in a solved crystal and for interpretation of the resulting

spectra. Two advances were made in this work in analyzing the data. First, the ability was realized to rationally assign the target atoms to mutually exclusive sets of diffractions and mining the data more effectively. Second, removing outliers and applying a Principle Component Analysis (PCA) subroutine to each set allows for extraction of eigenvectors and eigenvalues directly related to the anomalous signal. Working to separate signal from noise consumes much of the analysis and is the reason that PCA was implemented instead of simply observing a separate spectrum from each diffraction. The large unit cells and area detectors contribute to noise but they also provide the opportunity to observe many thousands of diffractions, occurring under the same laboratory environment, simultaneously. Utilizing the number of diffractions instead of the quality of any individual diffraction, it is possible to take advantage of the advances in computing to mine for faint signals in noisy data. New computing is also used for calculating the theoretical values of each diffraction, which is crucial to separating the diffraction into those favourable to one atom over the other. A detailed discussion of analysis of the data and a mathematical framework for biasing diffractions is given in Chapter 5.

1.8 Ways and Means

This work combines two well-established areas of applied physics as they relate to biological systems, XAS and MX. The systems that are included below have large unit cells, ideal simulated diffraction, non-ideal real diffraction from real crystals and mixed redox target atoms. The methodology applied to disentangle these systems has benefits and drawbacks. Dissecting the diffractions that bias one atom, in particular, over another and the implementation of PCA is broad insofar as what is swept up into the analysis. This broad scope obscures some attributes, and in complexity lies detail: Chapter 9 discusses the inclusion of a more nuanced approach and the benefits/ drawbacks of riding roughshod over aspects of diffraction and absorption.

Colours assigned to elements in thesis are consistent throughout and are identical to the colours used by molecular graphics program PyMol [76] except in one

important instance: for the outer iron of ferredoxin, it does not use the normal ‘sorbus orange’ colour, it has the colour assignment of ‘dodger blue’, usually associated with uranium. The images contained in this thesis are 600dpi and can be increased to approximately 400% without loss of quality.

CHAPTER 2

X-RAY ABSORPTION SPECTROSCOPY THEORY

An atom's absorption spectrum is directly related to its electronic configuration, which in turn depends on its local atomic neighbourhood. The location, structure and function of these absorbers is the goal of this research and the absorption spectrum has a direct relationship with the diffracted spectrum via the energy dependent correction to the atomic form factor, $f_1(E)+if_2(E)$, where absorption and f_2 are intimately dependent. In order to calculate diffraction intensities at a variety of energies it is essential to be familiar with the origins of absorption.

2.1 Classical

As light passes through a material, a portion is absorbed while the rest is transmitted. The thicker the material, the more absorption occurs. If the incident intensity of the light is I_0, then the attenuated intensity, I, of the transmitted light is given by Beer-Lambert's Law [21, 22]:

$$I = I_0 e^{-\mu t} \tag{2.1}$$

Where t is the thickness of the material and μ is the absorption coefficient. Values for μ can be measured in the lab using a variety of techniques. It is instructional to understand how each type of atom behaves when it interacts with light, and the sum of all the interactions is given by μ. When the energy of the light is in the vicinity of the energy required to promote an electron to a higher energy state of a given atom, the material becomes dramatically more opaque as the photon is absorbed. As the energy of the light, E, is increased the absorption occurs suddenly, has a step-like function (Figure 2-1) and is referred to as the 'absorption edge'. The portion of the spectrum

near to the absorption edge is complex and has transitions to bound states superimposed upon absorption due to excitation to the continuum. The absorption bands that occur in the spectrum have locations that are highly specific to each element. The precise positions and shapes are shifted and redrawn by the oxidation state of the atom and the neighbourhood in which it is located [12]. This chemical specificity is utilized by noting that across a limited spectrum of an edge, the other elements in the system have a uniform absorption coefficient (Figure 4-3). These bulk atom absorption profiles are not linear, but the change is smooth (approximately E^{-3}).

2.2 Quantum

Closely related to the absorption coefficient is the absorption cross section [24], σ, $\mu=\sigma\rho$, where ρ is the density. This classical absorption cross section, σ, is described quantum mechanically as the transition rate, $T_{i\to f}$, multiplied by the energy absorbed per transition, $\hbar\omega$, divided by the flux [25], ξc, where ξ is the energy density (Equation 2.4) and c, is the speed of light.

$$\sigma = \frac{T_{i\to f}}{\xi c}\cdot\hbar\omega \tag{2.2}$$

In this thesis it is assumed that the target atom, after absorption, has enough time to relax before another photon is incident: no one target atom sees two photons within a few femtoseconds of each other. The transition rate, $T_{i\to f}$, is the probability of absorption per unit time and the angular frequency, ω, is associated with the incident photon. This simplified expression will be revisited once it has been considered in parts [25]. The standard normalized wavefunction, A, for monochromatic plane waves is a sum of the annihilation (denoted with an apostrophe, *) and creation vectors also known as ladder operators [25, 26]:

$$\mathbf{A}(\mathbf{r},t) = \frac{1}{\sqrt{V}}\left(A\hat{e}e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + A^{*}\hat{e}^{*}e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\right) \tag{2.3}$$

Where t, is time, $\boldsymbol{k}$ is the wavevector and $\boldsymbol{r}$ is the direction of travel. The amplitude of $\boldsymbol{A}$, is knowable and normalizable; however, just like the volume term, V, it cancels out in the end. The denominator of 2.2, energy density, is calculated using [25]:

$$\xi = \frac{1}{4\pi}\left|\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}\right|_{avg}^{2} \tag{2.4}$$

Giving the flux:

$$\xi c = \omega^2 |A|^2 / 2\pi c V \tag{2.5}$$

The numerator in 2.2 is dominated by the probability of a transition of the system from the initial state to some final state [28]. The initial state is one photon and an unexcited atom whereas the final state is one excited atom alone. The prior statement is important, it defines absorption as a single photon and a single atom, then at a later time only a single atom: the photon has been annihilated[1].

The Hamiltonian ($\boldsymbol{H}$) describing the whole space is composed of those describing the photons (radiation)($\boldsymbol{H}_{rad}$), the target atom ($\boldsymbol{H}_{atom}$)and the interaction of photons with the target atom ($\boldsymbol{H}_{int}$):

$$\mathbf{H} = \mathbf{H}_{rad} + \mathbf{H}_{atom} + \mathbf{H}_{\text{int}} \tag{2.6}$$

$$\mathbf{H} = \mathbf{H}_{rad} + \mathbf{H}_{atom} + \frac{1}{2m}\left[\mathbf{p} - \frac{e}{c}\mathbf{A}\right]^2 \tag{2.7}$$

Where m is the mass of the absorbing electron and $\boldsymbol{p}$ is the momentum operator. Only the interaction Hamiltonian is valid for absorption and how it evolves in time. Strictly speaking, the interaction Hamiltonian is the sum of all interactions; however, it is assumed that only one photon and one electron are involved in the period of this event. This construction will also ignore the proton and spin magnetic moment as their contributions are orders of magnitude smaller [28]. By expanding the brackets in Equation 2.7 while allowing $\boldsymbol{A}$ and $\boldsymbol{p}$ to commute and note that the momentum operator squared term goes to zero:

[1] In Chapter 3, for scattering, there will be annihilation and creation.

$$\mathbf{H}_{\text{int}} = \frac{1}{2m}\left[\left(\frac{e}{c}\mathbf{A}(\mathbf{r},t)\right)^2 - 2\frac{e}{c}\mathbf{A}(\mathbf{r},t)\cdot\mathbf{p}\right] \tag{2.8}$$

For absorption, the first term must also be eliminated as second order in A requires one of the following to occur: two annihilations, two creations, or an annihilation and creation. None of these combinations is possible for a single photon absorption. However, second order in A will be revived when considering scattering in the next chapter:

$$\mathbf{H}_{\text{int}} = \frac{-e}{mc}\left[\mathbf{A}(\mathbf{r},t)\cdot\mathbf{p}\right] \tag{2.9}$$

The interaction Hamiltonian acts on the wavefunction, $\Psi(t)$, which evolves in time by the recursive expression for time dependent perturbation [21] and can be extended by repeatedly inserting $\Psi(t)$ into $\Psi(t')$:

$$\left|\Psi(t);\mathbf{k},\alpha\right\rangle = \left|\Psi(0);\mathbf{k},\alpha\right\rangle + \frac{1}{i\hbar}\int_0^t \mathbf{H}_{\text{int}}(t')\left|\Psi(t');\mathbf{k},\alpha\right\rangle dt' \tag{2.10}$$

For absorption it is sufficient to consider only the first order perturbation and then allow $\Psi(t)=\Psi(0)$, taking the time dependence out of the wavefunction explicitly. This is achieved using separation of variables, $\Psi(t,x)=g(t)f(x)$ [26], and the time dependent Schrödinger equation, a solution of which is:

$$\left|\Psi(x,t)\right\rangle = e^{-iE_0t/\hbar}\left|\Psi(0)\right\rangle \tag{2.11}$$

Once the time dependence is used once, for the first perturbation everything is set back and the time *in*dependent Hamiltonian interaction (denoted with an apostrophe) is:

$$\mathbf{H}'_{\text{int}} = \frac{-e}{mc}\frac{|A|}{\sqrt{V}}\left[\hat{\mathbf{e}}^{\alpha}\cdot\hat{\mathbf{p}}e^{i\mathbf{k}\cdot\mathbf{r}}\right] \tag{2.12}$$

The probability function is calculated in the usual manner. For first order perturbation for only a single absorber:

$$\left\langle\Psi_f(t);\mathbf{k},\alpha\middle|\Psi_i(t);\mathbf{k}',\alpha'\right\rangle = \frac{1}{i\hbar}\frac{-e}{mc}\frac{|A|}{\sqrt{V}}\left\langle f\middle|\hat{\mathbf{e}}^{\alpha}\cdot\hat{\mathbf{p}}e^{i\mathbf{k}\cdot\mathbf{r}}\middle|i\right\rangle\int_0^t e^{i\left(E_f-E_i-\hbar\omega\right)t'/\hbar}\,dt' \tag{2.13}$$

using [24]:

$$\lim_{t\to\infty}\int_0^t \frac{e^{i\omega t'}}{2\pi}dt' = \delta(\omega) \tag{2.14}$$

where $\delta(\omega)$ is a Dirac delta function:

$$\left\langle \Psi_f(t);\mathbf{k},\alpha \middle| \Psi_i(t);\mathbf{k}',\alpha' \right\rangle = \frac{2\pi}{\hbar}\left(\frac{-1}{i}\right)\left(\frac{e}{mc}\frac{|A|}{\sqrt{V}}\right)\langle f|\hat{\mathbf{e}}^{\alpha}\cdot\hat{\mathbf{p}}e^{i\mathbf{k}\cdot\mathbf{r}}|i\rangle\delta\left(E_f - E_i - \hbar\omega\right) \tag{2.15}$$

By inspection, this compares with Fermi's Golden Rule for transition rates [29, 30] such that:

$$T_{i\to f} = \frac{d}{dt}P_{i\to f} = \frac{2\pi}{\hbar}\left|\mathbf{M}_{fi}\right|^2\delta\left(E_{fi}\right) \tag{2.16}$$

The matrix element, for first order perturbation for a single final state is then:

$$M_1 = \frac{e}{mc}\frac{|A|}{\sqrt{V}}\langle f|\hat{\mathbf{e}}^{\alpha}\cdot\hat{\mathbf{p}}e^{i\mathbf{k}\cdot\mathbf{r}}|i\rangle \tag{2.17}$$

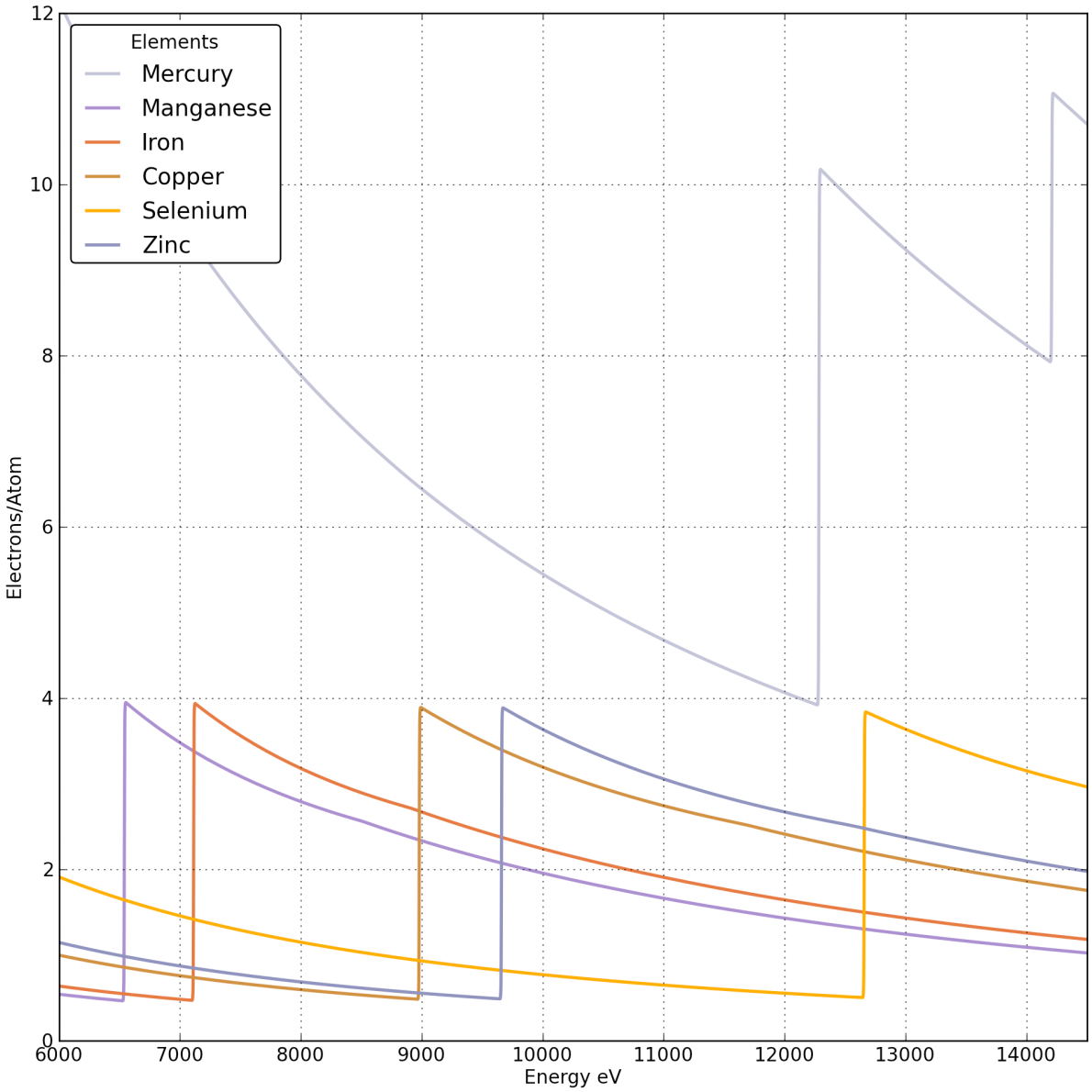
By inserting (2.17) into (2.16) and utilizing (2.5) so that the scattering cross section from (2.2) for a single target atom is summed over all possible final states, the familiar equation for plane wave absorption [32] is revealed:

$$\sigma = \frac{4\pi^2e^2}{\omega m^2 c}\sum_f\left|\langle f|\hat{\mathbf{e}}^{\alpha}\cdot\hat{\mathbf{p}}e^{i\mathbf{k}\cdot\mathbf{r}}|i\rangle\right|^2\delta\left(E_f - E_i - \hbar\omega\right) \tag{2.18}$$

In the X-ray region for target atoms that are transition metals, Eq. 2.18 is a core hole effect, so it is element specific. The total absorption for a material is the sum of the individual absorptions given by Eq 2.18. In the experiments conducted here the vast majority of atoms will have a small featureless contribution. However, the target atoms will have an abrupt step-like contribution (Figure 2-1) in the same region with detailed features that betray its local environment (Figure 4-1). This absorption profile directly influences the diffraction profile across the same spectral region, the intimate relationship between the two is unveiled in the following chapter.

Figure 2-1

K-edge Absorption Spectra of Various Heavy Elements



Six lines are given for various elements within the spectrum available within energies of 6000eV to 14500eV. Manganese, Iron, Copper, Zinc and Selenium K_1 edge and Mercury L_2 and L_3 edges. The edges are based on theoretical calculations by Cromer and Liberman [32, 33, 34]. These calculations do not include Near Edge effects, Atomic scattering factors or Extended X-ray Absorption Fine Structure.

CHAPTER 3

DIFFRACTION THEORY

There are two main parts to diffraction theory covered in this chapter, one that relates to the phenomena of diffraction itself and one that encapsulates how the energy dependence affects the atomic form factor. This second, energy dependent, part is an extension of the quantum mechanical treatment used in the previous chapter, and where the derivation of the relationship between absorption and diffraction originates. These formulations can be found scattered throughout many texts and a desire to have a self-consistent treatment for both (absorption and diffraction) in one place was the motivation for discussing them here.

3.1 Scattering Light - Classical

When an electromagnetic wave impinges on a crystal and is scattered from a location centred at ρ_j the position of the scatterer within a single unit cell of the crystal at that location is described as a vector in two parts: one vector *to* the unit cell origin $\boldsymbol{\rho}_{mnp}$ by the three integers[1] mnp of the primitive vectors of the crystal lattice and a second vector *from* the unit cell origin *to* the scattering centre, $\boldsymbol{\rho}_j(xyz)$. Both vectors use the primitive vectors of the crystal lattice, $\boldsymbol{abc}$:

$$\boldsymbol{\rho} = \boldsymbol{\rho}_{mnp} + \boldsymbol{\rho}_j \tag{3.1}$$

$$\boldsymbol{\rho} = \left(m\mathbf{a} + n\mathbf{b} + p\mathbf{c}\right) + \left(x_j\mathbf{a} + y_j\mathbf{b} + z_j\mathbf{c}\right) \tag{3.2}$$

Using the standard equation for a wave propagating in a direction $\boldsymbol{r}$ with amplitude E_0:

$$E(\mathbf{r}) = E_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{3.3}$$

[1] mnp are necessary integers as they fill the space with a regular array of unit cells, which is the definition of a crystal.

The resulting scattered wave has the form [35, 36]:

$$E_{sc} = \frac{C}{r} E_0 e^{i\mathbf{k}\cdot\boldsymbol{\rho}} e^{i(\mathbf{k}'\cdot\mathbf{r}-\omega' t)} \tag{3.4}$$

By grouping the spatial terms, using $\boldsymbol{R}=\boldsymbol{r}+\boldsymbol{\rho}$, where $\boldsymbol{R}$ is from the crystal origin to the detector and $\boldsymbol{r}$ is from the scattering centre to the detector. If the detector is at a sufficiently large distance, the scatterer $\boldsymbol{k}'$ and $\boldsymbol{R}$ are in the same direction. The resulting scattered wave is then the same as the equation for a single wave, a $1/r$ term that conserves energy, a constant of proportionality, C, and a phase factor that depends on the change in direction of the incoming and outgoing wave vectors $(\boldsymbol{k}'-\boldsymbol{k})$:

$$E_{sc} = \left(CE(\mathbf{R})/r\right) e^{-i\boldsymbol{\rho}\cdot(\mathbf{k}'-\mathbf{k})} \tag{3.5}$$

The total scattering will be a sum over all the unit cells, mnp, multiplied by the scattering centres, j, within a single unit cell. By employing Equation 3.1 and defining the change in wave vector as $\Delta\boldsymbol{k}=\boldsymbol{k}'-\boldsymbol{k}$, the total scattering amplitude is given by $S_{\Delta\boldsymbol{k}}$:

$$S_{\Delta\mathbf{k}} \underset{def}{\equiv} \sum_{mnp} e^{-i\boldsymbol{\rho}_{mnp}\cdot\Delta\mathbf{k}} \cdot \sum_{j} e^{-i\boldsymbol{\rho}_j\cdot\Delta\mathbf{k}} \tag{3.6}$$

3.2 Laue Equations

Diffraction occurs in highly localized directions due to the first part of the scattering amplitude and Figure 3-1:

$$\sum_{mnp} e^{-i\boldsymbol{\rho}_{mnp}\cdot\Delta\mathbf{k}} = \sum_{m} e^{-im(\mathbf{a}\cdot\Delta\mathbf{k})} \sum_{n} e^{-in(\mathbf{b}\cdot\Delta\mathbf{k})} \sum_{p} e^{-ip(\mathbf{c}\cdot\Delta\mathbf{k})} \tag{3.7}$$

As can be seen in Figure 3-1 the sum is only non-zero when ϕ is an integer multiple of 2π; hkl are the integers:

$$\begin{aligned} \mathbf{a}\cdot\Delta\mathbf{k} &= 2\pi h \\ \mathbf{b}\cdot\Delta\mathbf{k} &= 2\pi k \\ \mathbf{c}\cdot\Delta\mathbf{k} &= 2\pi l \end{aligned} \tag{3.8}$$

Which are the Laue equations of diffraction. Note mnp and hkl are all integers which force the exponent in (3.7) to be 1 and reduces the sum for a parallelepiped of length M, to:

Figure 3-1

Diffraction Condition/Direction as Integer Multiples of 2 π

Six lines are given for various primes up to and including $N=31$ because it is visually instructive. As N becomes larger, the closer it represents a delta function anchored at integer multiples of 2π. When N is in the range of 5,000 to 50,000, as in a real crystal, then diffraction would only occur in these directions and be negligible everywhere else.

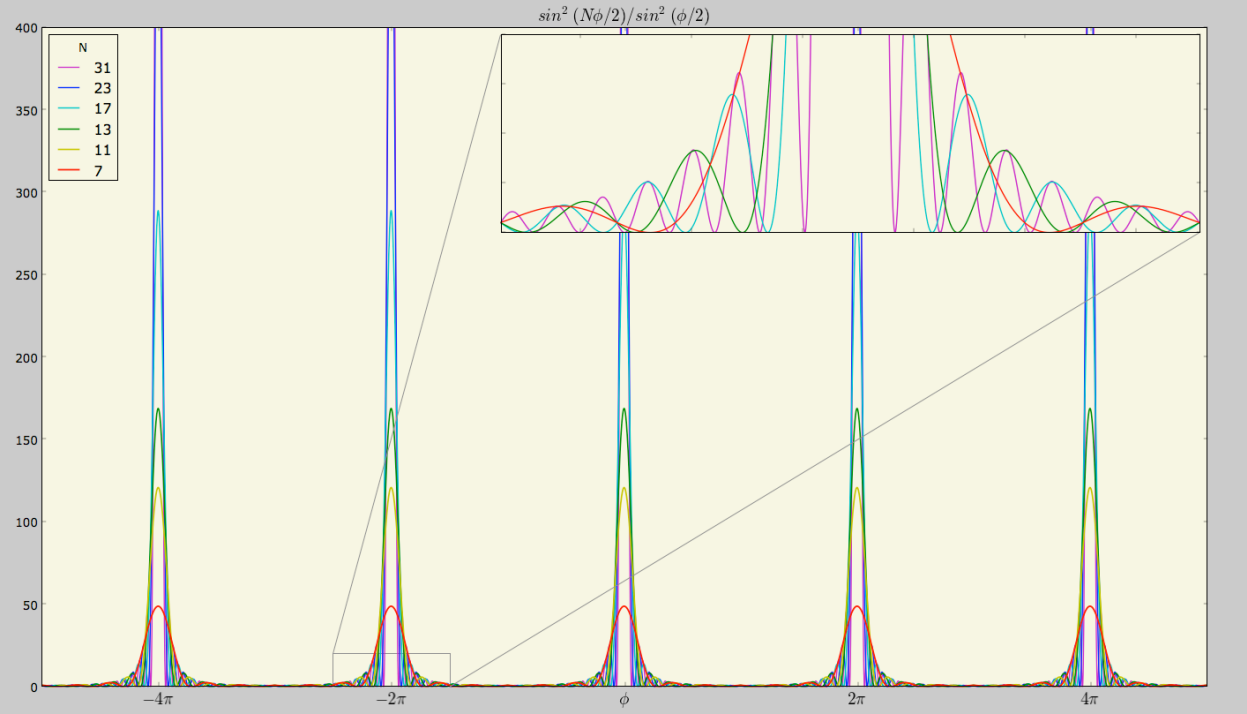


Figure 3-1 shows the equation:

$$\frac{\sin^2\left(N\phi/2\right)}{\sin^2\left(N/2\right)}$$

which is a sum of the geometric series:

$$\sum_m e^{-im(\mathbf{a}\cdot\Delta\mathbf{k})} = \sum_{m=0}^{N} r^m = \frac{1-r^{N+1}}{1-r}$$

for which the intensity of diffraction is the squared modulus. For large N, as in a crystal, $N\approx N+1$ such that:

$$\left|\frac{1-r^{N+1}}{1-r}\right|^2 = \frac{\sin^2\left(N\phi/2\right)}{\sin^2\left(N/2\right)}$$

$$r = e^{-i\phi} \quad ; \quad \phi = \Delta\mathbf{k}\cdot\mathbf{a}$$

$$\sum_{mnp} e^{-i\boldsymbol{\rho}_{mnp}\cdot\Delta\mathbf{k}} = \sum_{mnp} e^{-i2\pi(mh+nk+pl)} = M^3 \tag{3.9}$$

which is just another scale factor. The definition of the total scattering amplitude can further be reduced as $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are fundamental vectors of the reciprocal lattice with a 2π normalization factor:

$$\begin{aligned} S_{\Delta\mathbf{k}} &= M^3 \sum_j e^{-i\left(x_j\mathbf{a}+y_j\mathbf{b}+z_j\mathbf{c}\right)\cdot\left(h\mathbf{A}+k\mathbf{B}+l\mathbf{C}\right)} \\ &= M^3 \sum_j e^{-i2\pi\left(hx_j+ky_j+lz_j\right)} \\ &= M^3 \sum_j e^{-i2\pi\left(\mathbf{h}\cdot\mathbf{r}_j\right)} = M^3 \sum_j e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j} \end{aligned} \tag{3.10}$$

The subscript $\Delta\boldsymbol{k}$ for $S_{\Delta\boldsymbol{k}}$ can be described as the reciprocal lattice vector $\boldsymbol{h}=(hkl)$ for which the Laue equations hold true (are non zero). Without any loss in generality the sum, j, can be expanded, representing a single scattering event located at xyz within the unit cell with the scattering *centred at* j of an atom with **atomic form factor**, f_j [37] (Appendix II). This form factor can differ not only between the elements, but also amongst atoms of the same elemental type but in different configurations. The form factor is dimensionless, but frequently given values in electrons or electrons per atom. The sum over the form factors for a particular reciprocal lattice vector, hkl, is called the **structure factor**, $F(hkl)$, and the intensity of its diffraction, $I(hkl)$, is proportional to the squared modulus:

$$F\left(hkl\right) = \sum_j f_j e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j} \tag{3.11}$$

$$I\left(hkl\right) \propto \left|F\left(hkl\right)\right|^2 \tag{3.12}$$

3.3 Intensity of a Diffracted Spot

In the full kinematic version the atomic form factor is an integral of the electron concentration function [35] and in practice is calculated using tabulated values by

Cromer-Mann. There is a lot hidden in the proportionality symbol of equation 3.12, diffraction using X-rays on a real crystal has many factors. In order to use the equality symbol the temperature, occupancy, Lorentz factor, polarization, self absorption, detector efficiency and handful of constants [28, 31, 38, 39, 56] must be included:

$$I(\Delta k,E) = a_1\left[L(\Delta k,E)\cdot P(\Delta k)\cdot A(\Delta k,E,t)\cdot D(E)\cdot \left|F(\Delta k,E)\right|^2 \right] + a_2 + a_3 E \qquad (3.13)$$

For convenience the substitution of $\Delta\boldsymbol{k}$ for hkl is made. In the next Chapter and for the rest of this thesis, the energy dependence is sought and therefore E will be shown explicitly here[1].

3.3.1 Temperature Factor and Occupancy

There are four parameters given for every atom in the Protein Data Bank (PDB), the world's largest repository of protein structures: the position, the element, its occupancy and temperature factor. The temperature factor (B-factor)[2] and the Occupancy (O) are subsumed into the structure factor sum as they are atom specific:

$$F(\Delta k,E) = \sum_j f_j e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j}\cdot DWF_j\cdot O_j \qquad (3.14)$$

The B-factor is important to diffraction as it is a measure of the uncertainty in the position of the atom and has the effect of decreasing the intensity. The most common implementation for proteins is the B-factor, a single number related to the mean squared displacement of the atom that is often used as a short hand to describe a structure's orderliness. If the B-factor is given isotropically [40]:

$$DWF_j = e^{-\frac{B_j}{4}\left(\frac{1}{d}\right)^2} \qquad (3.15)$$

Where d is the distance between scattering planes and is a function of $\Delta\boldsymbol{k}$.

Atoms within a structure are constrained by bonding and as such are more likely to move in some directions than others. If this level of detail is available then

[1] In many texts wavelength, λ, is used instead of energy, E.

[2] This is the crystallographic Debye-Waller Factor which is different from the XAS DWF.

anisotropic temperature factors can also be used and employ a tensor form of the more general B-factor.

Occupancy is how often an atom appears to occupy a position; it usually has a value of 1, however sometimes an ion will only be bound to a few molecules or a large ligand may be found in a few different conformations. Therefore it is not uncommon to see values of 1, 0.5. Values of 0.6 and 0.4 are also prevalent in order to discern between two distinct configuration in a single PDB file.

3.3.2 Lorentz Factor

Also known as the kinematical factor, the Lorentz factor "is proportional to the time of reflection permitted to each reflection, or inversely proportional to the velocity with which the plane passes through the condition of reflection" [40]. It is most often described as being a scale factor, however it is energy dependent and for DS should be calculated as such [41, 42]. Θ_B is the Bragg angle of the diffraction:

$$L(\Delta k, E) = \left(\frac{1}{E}\right)^3 \frac{1}{\sin\theta_B(\Delta k)} \tag{3.16}$$

3.3.3 Polarization

Polarization comes in two forms, both phenomena originate from research conducted by Leonid Azaroff at the Illinois Institute of Technology [43, 44]. The first form is the relationship between the direction of the impinging E-vector and the diffraction plane, herein called, 'normal polarization'. The second form is the relationship between the impinging E-vector and the directions of the bonds of the target atom, herein called, 'bond polarization'.

The intensity of the diffracted ray is dependent on the polarization of the incident beam and the directions it scatters. A small adjustment for use of a double crystal

monochromator must also be made when applying Kahn's equations [45 46][1]. The normal polarization is then:

$$E''^2 = \kappa^4 \kappa'^2 \left[E_\sigma{}^2 \alpha \left(\cos^2\phi \cdot \cos^2 2\theta_B + \sin^2\phi \right) + E_\pi{}^2 \left(\sin^2\phi \cdot \cos^2 2\theta_B + \cos^2\phi \right) \right] \quad (3.17)$$

Where E''^2 is equal to the intensity after passing the monochromator and the crystal, Θ_B is the Bragg angle, ϕ is the polar angle on the face of the detector and $\alpha = cos^2 2\Theta_{mono}$ where Θ_{mono} is the monochromator Bragg angle. The κ and κ' contain in all the factors that are independent of the angle of reflection [43]. At a synchrotron source where the light is almost 100% plane (π) polarized the second part of 3.17 dominates.

Bond polarization effects are due to the angle of the E-vector with respect to the bond. When the impinging E-vector is parallel to a bond it is strongest and when it is perpendicular it is diminished [44]. Templeton and Templeton [47, 48, 49] formalized the mathematics with a tensor description of the structure factor whilst maintaining the polarization indexes, π and σ, from the normal polarization calculations. In the following chapter the relationship between absorption and anomalous signal in a diffracted ray is shown to be related and hence polarization can be considered when calculating diffracted intensities. A treatment of the effect of polarization on XAS with regards to dipole and quadruple allowed transitions in a crystal of cupric chloride dihydrate is given by Pickering and George [50].

3.3.4 Self-Absorption

Self-absorption is the process where the intensity of the diffraction is attenuated by regular absorption as it passes through the crystal and is a function of the thickness of the sample and the absorption coefficient, see equation 2.1. This effect is expected to be small, but also very difficult to calculate as the exact dimensions of each crystal are different and slowly rotating. It is difficult to assign crystal thickness versus angle and also calculate the likelihood that self-absorption will take place in the direction of the

[1] Kahn notes a typographic error in Azaroff's original 1955 publication which is critically important.

diffracted ray. The buffer and cryoprotectant, in the crystal loop, will also have an effect on the self-absorption. Protein crystals have a very low density of target atoms, there are typically 2-40 per unit cell which can contain tens of thousands of bulk atoms. There is the complication that at some diffraction angles the *escape route* from the crystal for a diffraction may coincide with dense rows of absorbers, empty rows or a plum pudding distribution. Self absorption will have detrimental effects on analysis of the spectrum of diffracted rays as self-absorption will come from all the types of target atoms within the crystal (buffer/cryo-protectant) and not only from the target atom we are attempting to separate. Self-absorption of the diffraction ray could add unwanted structural changes in the spectrum and cause confusion.

3.3.5 Detector

Many types of detectors are used to record the intensity and position of a diffracted ray. There is no one equation to be calculated for detector variations. Each detector has its own sensitivities: irregularities in the phosphor, lens or tapered fibres, temperature variations amongst others[1] are compensated for by using dark images, flood fields and noting dead or bright pixels [51]. Synchrotron beamlines for MX regularly account for variations in detector sensitivity as well as calculating detector efficiency. It is the vast improvements of area detectors and charge coupled devices (CCD) that have made large scale collections of diffractions routine. The sensitivity and speed of collection is closing in on parity, one count on the detector for each photon, these improvements are one of the main catalysts for this research project.

3.4 Approach to Calculating Intensities

The driving force behind the work presented in this thesis is an attempt to separate the anomalous dispersion spectra from two target atoms in different oxidation states and in different conformations within a very large unit cell in which the target atoms are vastly outnumbered by bulk atoms. The competing (or concurrent) desire to

[1] Zingers for instance.

fully understand all the mechanisms that effect the spectra is ever present however due the underlying complexity these experimentalists have returned to a more holistic approach. The method chosen here of investigating the spectra originates in the techniques first applied to absorption spectroscopy. Before the advent of good theoretical absorption software such as FEFF [52] model compounds and a library of previous experimental spectra were relied upon to confer information about spectra that were being measured for the first time. As these are the first experiments to be conducted with this new methodology it is important to extract the spectra whilst also looking for underlying explanations. Occupancy and isotropic temperature factors are included in the simulated diffractions calculated in this work. Detector variations, Lorentz factor and normal Polarization are handled by the internals of the detector and the processing software, XDS [53]. The detector, Lorentz and normal polarization are accounted for in laboratory data collection on the beamline but are not accounted for in the theoretical or simulated diffractions as they do not effect the size of a target atoms contribution to the intensity with respect to the overall intensity. Up to this point polarization has been avoided in DS by limiting diffraction to those perpendicular to the polarization of the impinging E-vector. The crystals were orientated so that the diffraction occurred in this orientation. This is impossible with an area detector however normal polarization diminishes only the scale of the anomalous phenomena but not its shape so it can be largely ignored. The more complex bond polarization effect from bond directions as they relate to the E-vector has the attribute of increasing or decreasing *parts* of the phenomena and it should be regarded with more care. DS as applied to large macromolecular crystals could well avoid these effects if the total oscillation angle taken is wide and the orientations of the target atoms do not lie in too high a symmetry. Which is to say that this effect needs more research as it could accentuate or depress parts of the anomalous dispersion spectrum either for good or ill. In complexity lies greater detail and we look forward to accounting for this effect in future experiments. Self-absorption is expected to be small and near-impossible to compensate for, or in this experimentalists view so ungainly that it deserves an entire dissertation all of its own.

3.5 Hamiltonian of Diffraction

In the vicinity of the absorption edge the atomic form factor, f_j, becomes complex [54]; this 'anomalous dispersion' contribution stems from the finite period of time in which the photon might have been absorbed by the atom, yet continued down the diffraction path. This period is governed by Heisenberg's Uncertainty Principle, within which the newly created photo-electron can be promoted to a higher state or probe its surrounding environment before reabsorbing and returning the electron to its original state[1]. Therefore, this more subtle form factor, Equation 3.18, is sensitive not only to the scattering vector, $\Delta\boldsymbol{k}$, but the energy, E, of the photons and the intermediate states of the photo-electron wavefunction, $|c\rangle$. These wavefunctions govern an atom's susceptibly to absorption and the nature of the intervening period:

$$f(\Delta\mathbf{k},E) = f_0(\Delta\mathbf{k}) + f_1(E) + if_2(E) \tag{3.18}$$

For this more complete description of scattering, non-relativistic quantum mechanical effects must be included. Photon-atom interactions were given in Chapter 2 for absorption at different energies by a single photon and atom. In order to demonstrate the scattering process the same form of the interaction Hamiltonian is used. However, only instances where there is single annihilation *and* a single creation of the wavevector is considered. This is accomplished by using both parts of the interaction Hamiltonian from Equation 2.8, given again here:

$$\mathbf{H}_{\text{int}} = \frac{1}{2m}\left[\left(\frac{e}{c}\mathbf{A}(\mathbf{r},t)\right)^2 - 2\frac{e}{c}\mathbf{A}(\mathbf{r},t)\cdot\mathbf{p}\right] \tag{3.19}$$

This interaction Hamiltonian is combined with the time dependent perturbation theory and evaluated separately for the $\boldsymbol{A}^2$ term to first order and the momentum ($\boldsymbol{Ap}$) term to second order, see Figure 3-3. Restricting the calculation to one absorbing atom and ignoring the spin and magnetic moments, the first order in $\boldsymbol{A}^2$ is as follows:

$$\left\langle \Psi_f(t);\mathbf{k},\alpha \middle| \Psi_i(t);\mathbf{k}',\alpha' \right\rangle_{(1)} = \frac{1}{i\hbar}\int_0^t \left\langle \Psi_f(t) \middle| \frac{1}{2m}\left(\frac{e}{c}\mathbf{A}(\mathbf{r},t)\right)^2 \middle| \Psi_i(t');\mathbf{k},\alpha \right\rangle dt' \tag{3.20}$$

[1] This can happen before the photon is absorbed, but the effect is orders of magnitude smaller.

Separating out the constants while using the classical radius of an electron, $r_e = e^2/mc^2$, the integral is evaluated and then divided by the flux (Equation 2.5). The ingoing and outgoing photons must be set to the same wavelength, $\omega=\omega'$, which defines elastic scattering, but allows for a change in direction, $k \neq k'$, which allows for diffraction:

$$\frac{\left\langle \Psi_f(t);\mathbf{k},\alpha \middle| \Psi_i(t);\mathbf{k},\alpha \right\rangle_{(1)}}{t} = -\frac{2\pi c r_0}{\omega^2} \left\langle i \middle| \hat{e} \cdot e^{-i(k'-k)r} \middle| i \right\rangle \delta\left(E_f - E_i \pm \hbar\omega\right) \tag{3.21}$$

This is the fundamental, f_0, non-anomalous contribution to the diffraction, $\Delta\boldsymbol{k}$, from a scatterer located at $r_j=(x_j\boldsymbol{a}+y_j\boldsymbol{b}+z_j\boldsymbol{c})$:

$$f_{0,j}(\Delta k) e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j} \underset{def}{\equiv} \left\langle i \middle| \hat{e} \cdot e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j} \middle| i \right\rangle \tag{3.22}$$

Summing over all the atomic form factors within a unit cell returns the definition of the structure factor to quantum form:

$$F(\Delta k) = \sum_j -f_{0,j}(\Delta k) e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j} = \sum_j -\left\langle i \middle| \hat{e} \cdot e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j} \middle| i \right\rangle \tag{3.23}$$

To calculate the anomalous dispersion correction, the energy dependent parts (real and imaginary) of the interaction Hamiltonian must be evaluated. The second order contributions from the momentum operator are:

$$\left\langle \Psi_f(t);\mathbf{k}',\alpha' \middle| \Psi_i(t);\mathbf{k},\alpha \right\rangle_{(2)} = \left(\frac{1}{i\hbar}\right)^2 \int_0^t \int_0^{t'} \left\langle \Psi_f(t) \middle| \left(\frac{1}{2m} \frac{-2e}{c} \mathbf{A}(\mathbf{r},t) \cdot \mathbf{p} \right)^2 \middle| \Psi_i(t') \right\rangle dt'' dt' \tag{3.24}$$

The matrix element must be summed over all possible intermediate states, c, of the target atom. This is a similar situation to absorption, Figure 3-3. The final state of annihilation, $\langle f|$, for absorption, is replaced with the annihilation/creation of intermediate states, $|c\rangle\langle c|$, and the final state is similar to the initial state as this is diffraction:

$$\left\langle \Psi_f(t);\mathbf{k}',\alpha' \middle| \Psi_i(t);\mathbf{k},\alpha \right\rangle_{(2)} = \left(\frac{1}{i\hbar}\right)^2 \frac{r_0}{m} \int_0^{t'} \int_0^t \sum_c \left\langle f \middle| \mathbf{p}\cdot\mathbf{A} \middle| c \right\rangle e^{i\left(E_f - E_c\right)t'/\hbar} \left\langle c \middle| \mathbf{p}\cdot\mathbf{A} \middle| i \right\rangle e^{i(E_c - E_i)t''/\hbar} dt'' dt' \tag{3.25}$$

The electromagnetic wave has its usual factors, evaluation of each integral brings down a $\hbar/i$:

$$\left\langle \Psi_f(t);\mathbf{k},\alpha \middle| \Psi_i(t);\mathbf{k},\alpha \right\rangle_{(2)} = \left(\frac{1}{i\hbar}\right)^2 \frac{r_0}{m}\frac{|A|^2}{V}\left(\frac{\hbar}{i}\right)^2 \sum_c \left[\frac{\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}'^* e^{-i\mathbf{k}'\cdot\mathbf{r}}|c\rangle\langle c|\mathbf{p}\cdot\hat{\mathbf{e}}e^{+i\mathbf{k}\cdot\mathbf{r}}|i\rangle}{E_f - E_c - \hbar\omega} + \right.$$
$$\left. + \frac{\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{+i\mathbf{k}\cdot\mathbf{r}}|c\rangle\langle c|\mathbf{p}\cdot\hat{\mathbf{e}}'^* e^{-i\mathbf{k}'\cdot\mathbf{r}}|i\rangle}{E_c - E_i + \hbar\omega'} \right] \delta\left(E_f - E_i - \hbar\omega + \hbar\omega'\right) \tag{3.26}$$

The stipulation is that the system returns to its original configuration. By dividing through by the ingoing and outgoing flux, the second order contribution is evaluated in the *forward scattering limit:* $\omega=\omega'$, $\boldsymbol{k}=\boldsymbol{k}'$ and $\boldsymbol{e}=\hat{\boldsymbol{e}}'=\hat{\boldsymbol{e}}$. The dependence of the anomalous dispersion has been removed from the direction of scatter by equating $\boldsymbol{k}$ and $\boldsymbol{k}'$. Although, the angular dependence has been shown to be either small or non-existent 'More work is needed' [24]. One way to view this assumption is by noting that most diffraction is a glancing blow to the electron cloud whereas absorption is a core-electron effect:

$$\frac{\left\langle \Psi_f(t);\mathbf{k},\alpha \middle| \Psi_i(t);\mathbf{k},\alpha \right\rangle_{(2)}}{t} = \left(\frac{2\pi c r_0}{\omega^2}\right)\frac{1}{m}\sum_c \left[\frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}}|c\rangle\right|^2}{E_f - E_c - \hbar\omega} + \frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}}|c\rangle\right|^2}{E_c - E_i + \hbar\omega} \right] \delta\left(E_f - E_i \pm \hbar\omega\right) \tag{3.27}$$

The total amplitude for scattering is then the addition of the two matrix elements, summed over all atoms, j, within the unit cell as well as summing all intermediate states, c, for susceptible target atoms. A phenomenological damping term, $i\eta$, is also included that prevents the denominator going to zero:

$$\left|F(hkl,E)\right|^2 \propto \left| \sum_j \left[-\langle i|\hat{e}\cdot e^{-i\Delta\mathbf{k}\cdot\mathbf{r}_j}|i\rangle + \frac{1}{m}\sum_c \left(\frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}_j}|c\rangle\right|^2}{E_f - E_c - \hbar\omega - i\eta} + \frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}_j}|c\rangle\right|^2}{E_c - E_i + \hbar\omega - i\eta} \right) \right] \right|^2 \tag{3.28}$$

All that is left is to equate this with Equation 3.18 by separating the anomalous contribution into its complex form using:

$$\lim_{\eta\to 0^+} \frac{1}{u + i\eta} = P\frac{1}{u} - i\pi\delta(u) \tag{3.29}$$

Where P is the Cauchy Principle Value:

$$\left|F(hkl,E)\right|^2 \propto \left|\sum_j \left[-\langle i|\hat{e}\cdot e^{i\Delta\mathbf{k}\cdot\mathbf{r}_j}|i\rangle + \frac{1}{m}\sum_c \left(P\left[\frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}_j}|c\rangle\right|^2}{E_f - E_c - \hbar\omega} + \frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}_j}|c\rangle\right|^2}{E_f - E_c + \hbar\omega} \right] + i\pi\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}_j}|c\rangle\right|^2 \delta\left(E_f - E_c - \hbar\omega\right) \right) \right] \right|^2$$

(3.30)

Comparing this to Equation 3.18:

$$\begin{aligned} f_0(\Delta\mathbf{k}) &= \langle i|\hat{e}\cdot e^{-i\Delta\mathbf{k}\cdot\mathbf{r}}|i\rangle \\ f_1(E) &= \frac{1}{m}\sum_c P\left[\frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}}|c\rangle\right|^2}{E_f - E_c - \hbar\omega} + \frac{\left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}}|c\rangle\right|^2}{E_f - E_c + \hbar\omega} \right] \\ f_2(E) &= \frac{\pi}{m}\sum_c \left|\langle f|\mathbf{p}\cdot\hat{\mathbf{e}}e^{i\mathbf{k}\cdot\mathbf{r}}|c\rangle\right|^2 \delta\left(E_f - E_c - \hbar\omega\right) \end{aligned}$$

(3.31)

Where f_0, elastic scattering, is independent of energy and equates with the kinematic diffraction. And $f_1(E) + if_2(E) = \Delta f(E)$ is the energy dependent anomalous contribution in the forward scattering limit where the real and imaginary components are related to each other by the Kramers-Kronig dispersion relation:

$$\begin{aligned} f_1(\omega) &= +\frac{2}{\pi}P\int_0^\infty \frac{\omega'}{\omega'^2 - \omega^2} f_2(\omega')d\omega' \\ f_2(\omega) &= -\frac{2}{\pi}P\int_0^\infty \frac{\omega}{\omega'^2 - \omega^2} f_1(\omega')d\omega' \end{aligned}$$

(3.32)

f_1 is a positive cusp shaped symmetric function in phase with the elastic scattering and f_2 is a positive antisymmetric step function 90° out of phase: the so called *imaginary* contribution (Figure 3-2).

3.5.1 Sign Convention

There are historical and convenient forms in which the the sum of f_0, f_1 and f_2 are given the signs that precede them and show how their Kramers-Kronig transforms relate to each other. Many of them are misleading. These sign conventions extend

throughout the diffraction and absorption literature with the exception of one: clarity was brought to the topic by J. O. Cross's thesis [31]. The Thomson scattering, f_0, is negative, which does not matter when it is being squared for intensities, however, authors were forced to surreptitiously assign negative values to f_1 in MAD crystallography, DAFS and DANES (DS) as the effect was in the opposing direction when the anomalous dispersion was included; by swapping ω' for ω in the denominator of the Kramers-Kronig transform. It is uncomfortable to consider structure factors, f_0, with a negative value and f_1 as positive however this is how it must be unless we change the sign of Beer-Lambert's attenuation coefficient (Equation 2.1), which is highly unlikely.

3.6 Comparison with Absorption

There is a simple relationship between the imaginary part of diffraction and absorption by equating f_2 from Equation. 3.31 with σ in Equation 2.18:

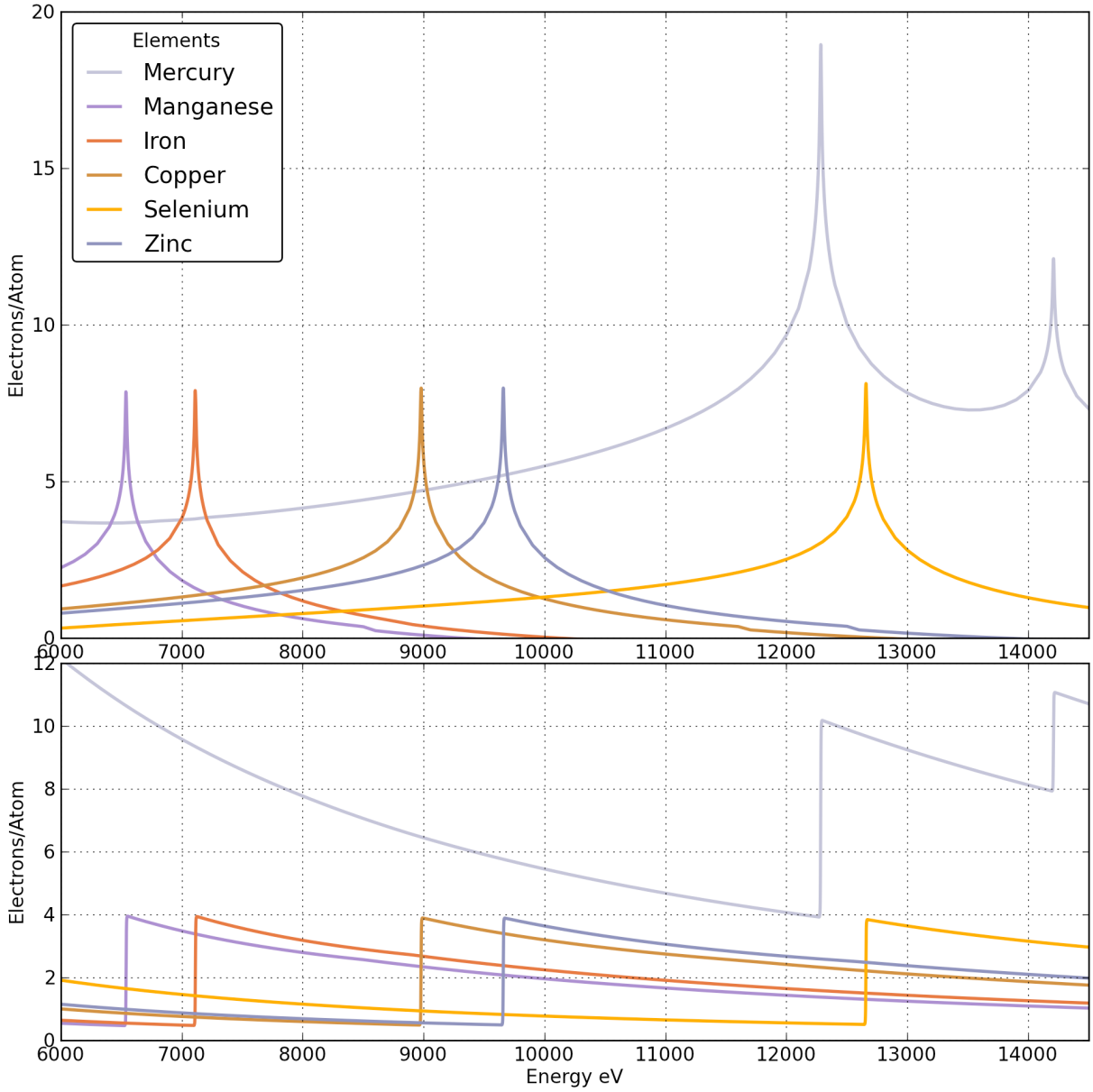
$$f_2(E) = \frac{\omega}{4\pi c r_e}\sigma(E) \tag{3.33}$$

It is here that the two different experimental methods (absorption and diffraction) are fused, each betraying he other. By experimentally determining an atom's absorption profile, $\sigma(E)$, it is possible to calculate its anomalous contribution to diffraction [5, 24, 26, 31, 55][1] using the Kramer-Kronig dispersion relation. The imaginary part of the complex structure factor, f_2, is proportional to the product or the energy, E, and the absorption, σ, of the same atom. Long before the energy of a photon is capable of promoting a core electron the diffraction signal is increasingly effected due to the dispersion relation in which f_2 is related to f_1. The effect on the real part can be over twice as large as that of the imaginary part (See Figure 3-2).

[1] Each reference gives an equivalent equation with some simple algebra and physics.

Figure 3-2

Kramers-Kronig Anomalous Dispersion Relation

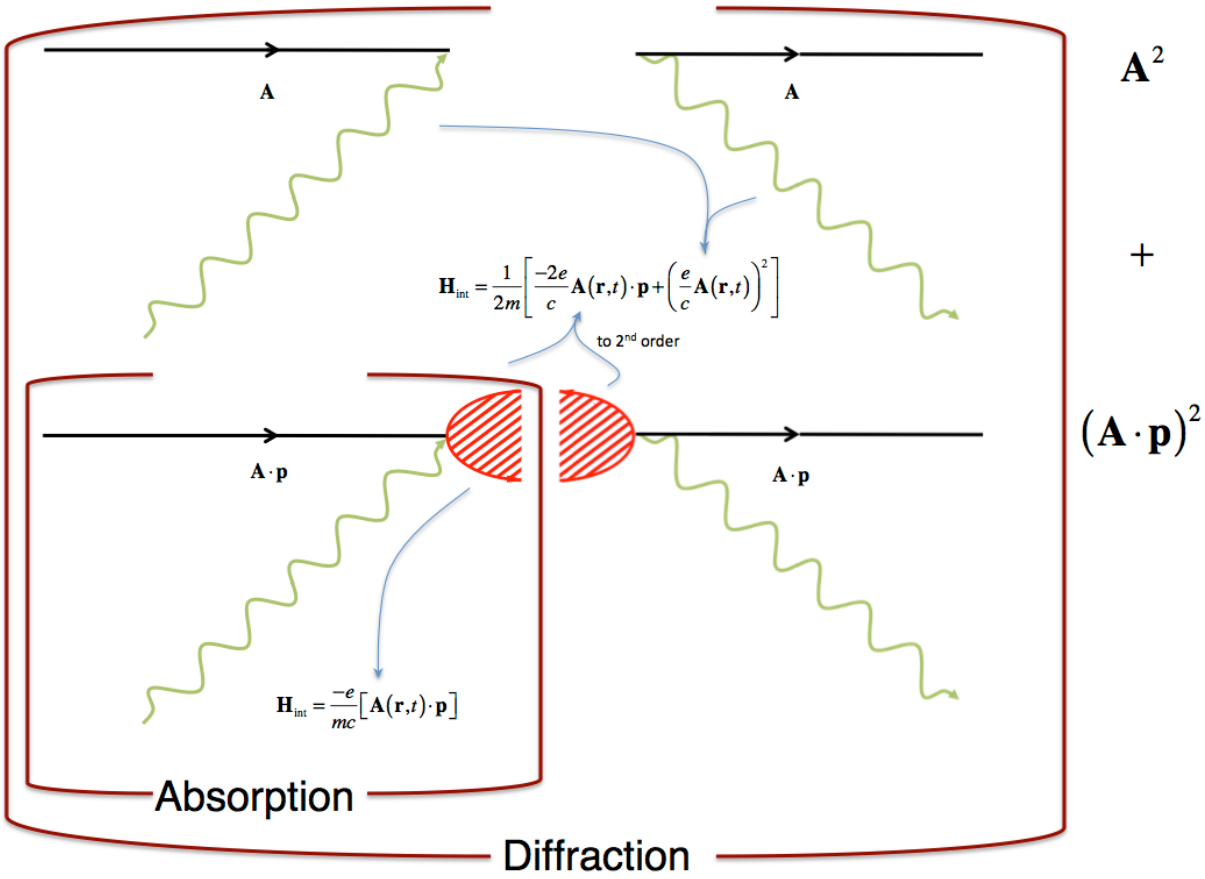


Six lines are given for various elements: Manganese, Iron, Copper, Zinc and Selenium K edge and Mercury L-II and L-III edges and their Kramers-Kronig transform. f_1 has a positive cusp and f_2 a positive step. The edges are based on theoretical numbers by Cromer and Liberman [32, 33].

Figue 3-3

Feynman Diagrams for the Interaction Hamiltonian

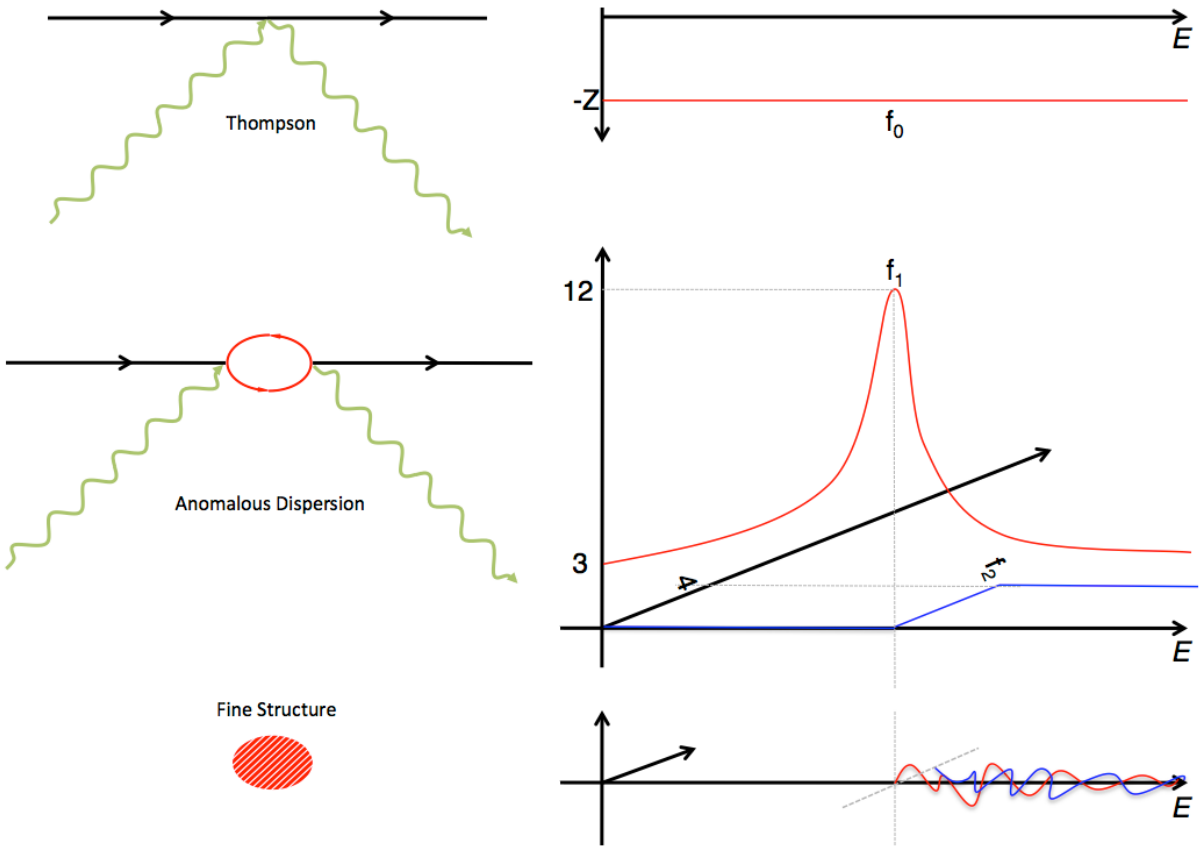
Absorption vs. Diffraction



Comparison of Feynman diagrams for absorption and diffraction. Absorption requires annihilation where as diffraction needs second order in A and pA: one annihilation and one creation.

Figure 3-4

Feynman Diagrams vs Atomic Form Factor



The decomposition of the Feynman diagrams into its contribution to the atomic form factor. f_0 is energy independent and governed by the Z number. Anomalous dispersion (f_1, f_2) are related to bare-atom absorption and the hash marks are for the fine structure. All numbers are approximate.

CHAPTER 4

DIFFRACTION SPECTROSCOPY THEORY

4.1 Atomic Specificity

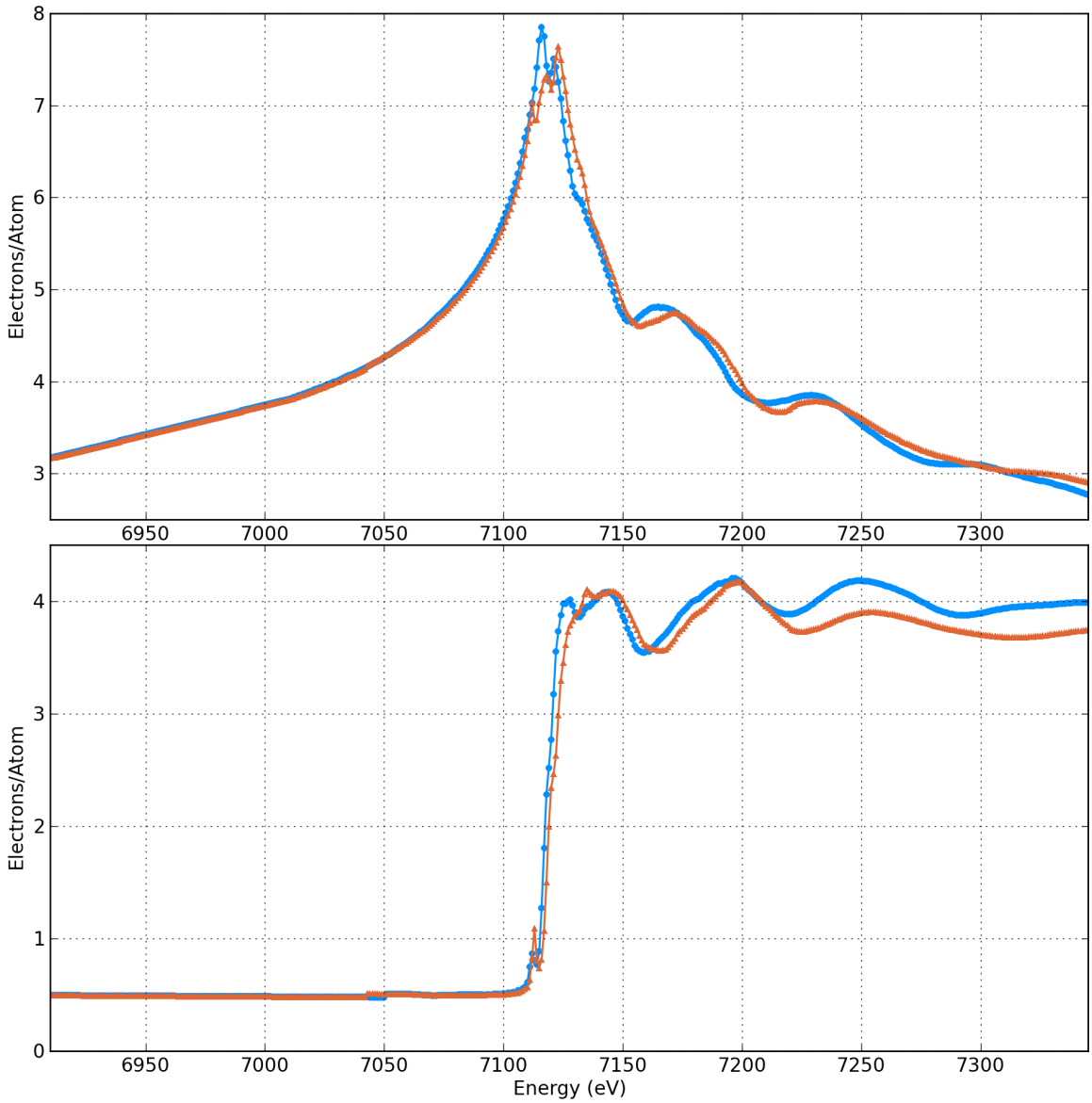
Absorption is not only element specific, but atom specific: two atoms of the same element in different states or in different neighbourhoods will have slightly different absorption profiles (Figure 4-1). These profiles are carried over to the diffraction intensities by the arguments given above. In real experiments, the absorption profiles of f_2 are measured from an experiment, then its Kramers-Kronig mate, f_1, is calculated, and the two are then used to predict anomalous dispersion effects[1]. In order to investigate the spectra of diffraction, appropriate absorption spectra are required.

The spectra featured in Figure 4-1 is a combination of empirical and theoretical observations. The absorption profiles of each of the two irons are constructed using near-edge data from a real experiment on reduced and oxidized iron-sulphur clusters. The Extended X-ray Absorption Fine Structure (EXAFS) oscillations are generated from software package FEFF using real atomic positions from a ferredoxin crystal taken at very high resolution (Appendix II). These two absorption spectra were spliced together and scaled to fit the theoretical iron absorption K-edge from Cromer and Liberman. The theoretical edges were adjusted to coincide with the observed edge. The f_1 spectra were then generated using software that was written specifically for this task, fftkk.py (Appendix III). The software utilizes a technique developed by Templeton and Templeton [60] and the subroutines were transliterated from a FORTRAN program written by Graham George and Ingrid Pickering: fftkk.f. These types of similar-but

[1] For use in MAD phasing experiments

Figure 4-1

Anomalous Dispersion of Similar Irons



Partially simulated iron dispersion spectra, reduced (dodger blue) and oxidized (sorbus orange). The piecewise continuously differentiable absorption function (*bottom*) was created from three subsections: the pre-edge and scale from Cromer-Liberman, the near edge from a reduced and oxidized rubredoxin experiment (ps-rd), and the XAS oscillations using FEFF. The FEFF calculations were generated from a high resolution crystal (1CZP.pdb) structure of Ferredoxin [61]. The corresponding real part (*top*), f_1, was generated using a computer subroutine: fftkk.py (Appendix III).

different spectra are what DS is trying to deconvolute from experiments. When the two spectra are combined within a single diffraction, great care must be taken in choosing just the right set of diffractions that preferentially express one over the other

4.2 Structure Factor Calculations

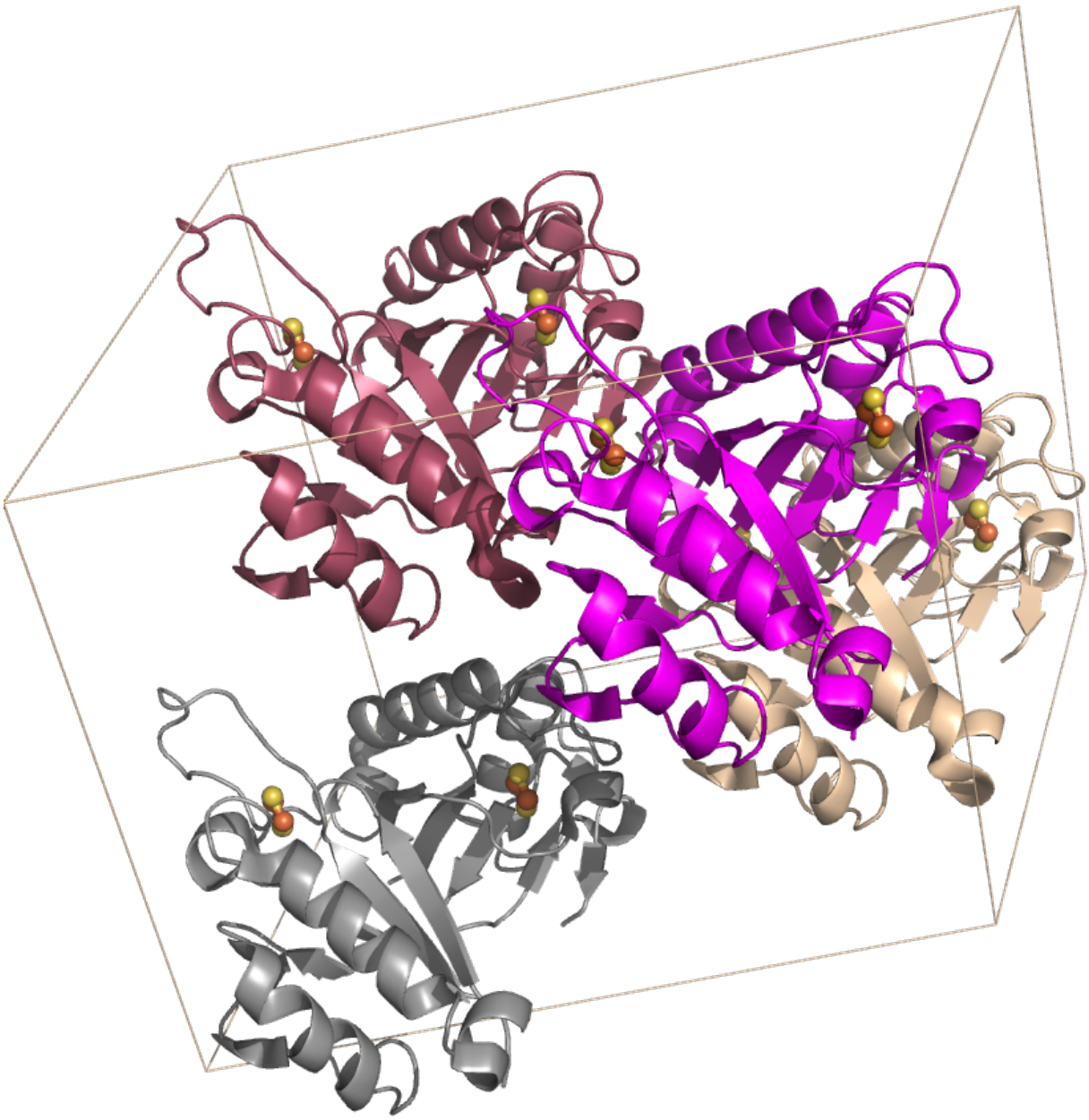
DS theory relies on the energy dependence of individual diffractions. This can be labeled more clearly by substituting $2\pi\boldsymbol{h}$ for $\Delta\boldsymbol{k}$ in the structure factor equation, where $\boldsymbol{h}$ is the vector normal to the diffraction plane and indicates the Miller indices hkl. The structure factor, F, and the atomic form factor, f_j, can then be neatly represented in a compact form:

$$F(\mathbf{h},E)=\sum_j f_j(\mathbf{h},E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_j)} \tag{4.1}$$

In order to investigate the contributions that every atom supplies to an individual diffraction, it is useful to consider a real example. The protein that we are using for this example will be the ferredoxin from our experiments. This protein was solved during a solution run and the locations ($\boldsymbol{r}$) of the atoms within the unit cell will be used to calculate the phase part of each atomic form factor. The underlying energy *in*dependent Thomson scattering, f_0, is calculated using Cromer-Mann coefficients and the d-spacing of the diffraction plane, which is a function of $\boldsymbol{h}$. Anomalous dispersion was calculated using Cromer-Liberman values for the so-called 'bulk' atoms (C, N, O, S, Zn) at 7117eV and the semi-empirical values shown in Figure 4-1 for the reduced and oxidized target atoms of iron. This example will analyze a single diffraction: $\boldsymbol{h} = hkl = (-9, 11, 12)$. In the unit cell of this ferredoxin there are 4040 Carbons, 1080 Nitrogens, 1524 Oxygens, 88 Sulphurs, 8 Zincs and 16 Irons. There are only 2 irons per protein, but there are 8 proteins in each unit cell. The analysis is conducted for a single diffraction over the absorption K-edge of iron and therefore no significant energy dependent anomalous dispersion contributions from any of the bulk atoms is expected (Figure 4.2). The variation is so small across this spectrum that the anomalous dispersion from each of

Figure 4-2

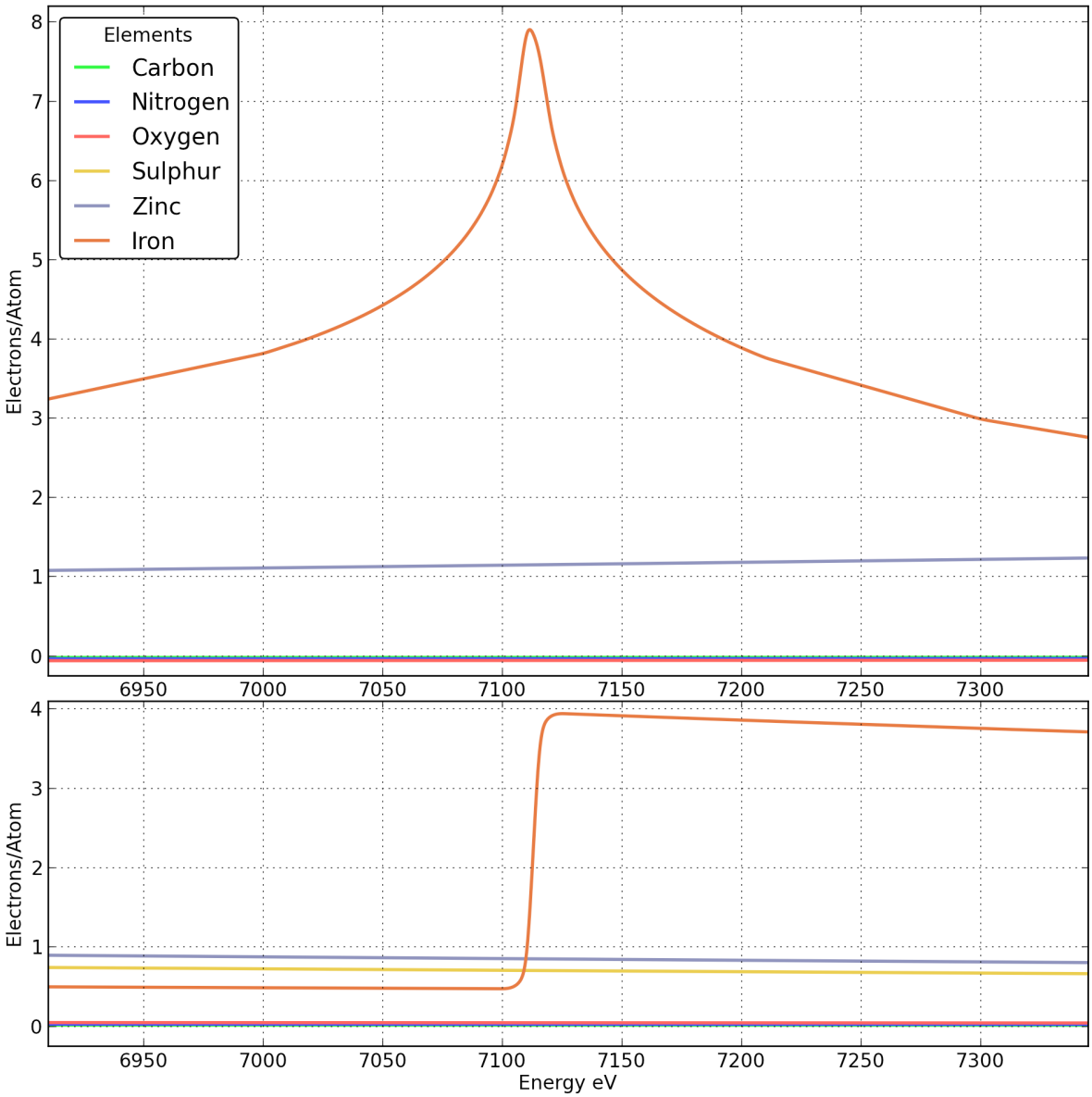
Ferredoxin Unit Cell



The ferredoxin unit cell has 4 asymmetric sub-units (ASU), each ASU has two ferredoxin proteins and each protein 2 irons. The eight 2Fe2S molecule are shown in ball-an-stick and the rest of the protein as a cartoon.

Figure 4-3

Anomalous Dispersion of Elements within Ferredoxin



The real and imaginary contributions of f_1 and f_2 (*top* and *bottom* respectively) in the spectral region of the Iron K-edge. Contributions from the bulk atoms are small and smooth The iron contribution shown here is that of a lone iron calculated using Cromer-Liberman, broadened by convolution with a Voigt function.

the bulk atoms can be considered not a *function* of energy. For instance: Zinc, which has the largest atomic form factor and anomalous contribution of the bulk atoms has a change of less than 0.5%, with all other bulk atoms being an order of magnitude smaller.

4.3 The Example

It will be shown that our example is highly biased toward one of the two irons in the protein: this will neatly demonstrate the site-selectivity of DS. By using equation 4.1 it is possible to separate the contributions that come from atoms that have an energy dependence in iron K-edge region from those that do not. In order to visualize this better, the calculation is broken into to a sum of sums from a sum over all atoms in the unit cell, where each sum is dedicated to a particular element:

$$\begin{aligned} F(\mathbf{h},E) = &\sum_{C}^{4040} f_C(\mathbf{h}) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_C)} + \sum_{N}^{1080} f_N(\mathbf{h}) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_N)} + \sum_{O}^{1524} f_O(\mathbf{h}) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_O)} + \cdots \\ &\cdots + \sum_{S}^{88} f_S(\mathbf{h}) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_S)} + \sum_{Zn}^{8} f_{Zn}(\mathbf{h}) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Zn})} + \sum_{Fe}^{16} f_{Fe}(\mathbf{h},E) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe})} \end{aligned} \tag{4.2}$$

With the structure factor broken up into its atomic constituents, one will notice that a very large part of the sum is not a function of energy and will have the same value across the entire spectrum:

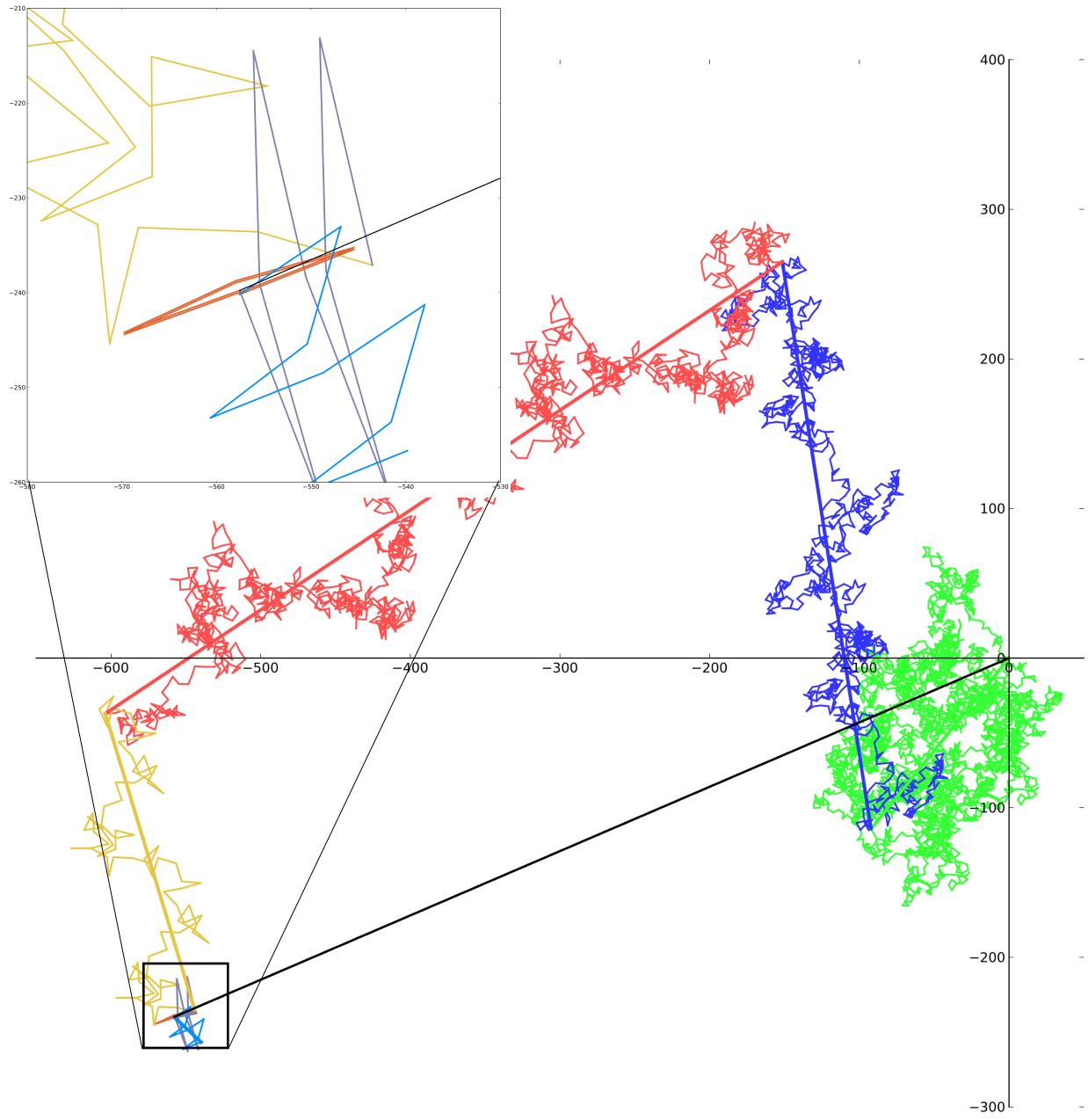
$$F(\mathbf{h},E) = F_C(\mathbf{h}) + F_N(\mathbf{h}) + F_O(\mathbf{h}) + F_S(\mathbf{h}) + F_{Zn}(\mathbf{h}) + \sum_{Fe}^{16} f_{Fe}(\mathbf{h},E) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe})} \tag{4.3}$$

These element specific sub-factors can be further simplified. The progression of this sum of sums can be seen in Figure 4.3 where the thousands of individual elements that are not a function of energy are first soaked up into element specific vectors and then represented by a single black vector, $F_{Z\text{-}Fe}(\mathbf{h})$:

$$F(\mathbf{h},E) = F_{Z-Fe}(\mathbf{h}) + \sum_{Fe}^{16} f_{Fe}(\mathbf{h},E) e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe})} \tag{4.4}$$

Figure 4-4

Argand Diagram of Every Atom within Ferredoxin



The sum of the structure factors from a ferredoxin unit cell with total structure factor for each element for ***h***=*(-9, 11, 12)*. The solid black line is the total structure factor from the bulk atoms- it extends from the origin to the start of the two iron structure factors. The individual atoms can also be made out along with their sub-total structure factors, colour coded thus: green near origin(C), dark blue(N), light red(O), yellow(S), lavender(Zn), orange(Fe1) and blue(Fe2) (see Figure 4.2).

$F_{Z\text{-}Fe}(\mathbf{h})$ could just as well be labeled 'bulk' in the equation (or have no label at all) as the lack of energy in the parentheses infers its position within the sum. Figure 4.3 demonstrates that the highly complex contributions from within a crystal can quickly be simplified.

4.3.1 Expansion of the Target Atoms

There are two target atoms: Fe1 and Fe2, one of each in the protein and eight of each in a unit cell. Their individual totals for this *hkl* are very different, which is why this particular diffraction, $\mathbf{h}=(-9, 11, 12)$, was chosen. An examination of Figure 4-4 reveals that Fe1 (orange sorbus) looks like a very flat rhombus. In fact it is two rhombi on top of each other, each side of the rhombus is an atom-vector. What this effectively means is that the sum of the eight atoms of Fe1 do not go very far. Put another way, Fe1 does not significantly contribute to the total structure factor. The eight atoms of Fe2 (dodger blue) make a significant contribution to the location of the end of this vector sum[1]. This is illuminated in the sum by separating the two iron labels into sub-factors of their own:

$$F(\mathbf{h},E) = F_{Z-Fe}(\mathbf{h}) + \sum_{Fe1}^{8} f_{Fe1}(\mathbf{h},E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe1})} + \sum_{Fe2}^{8} f_{Fe2}(\mathbf{h},E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe2})} \tag{4.5}$$

And substitute in the energy dependent structure factor from Equation 3.18:

$$\begin{aligned} F(\mathbf{h},E) = F_{Z-Fe}(\mathbf{h}) + \sum_{Fe1}^{8}\left[f_{0,Fe1}(\mathbf{h}) + f_{1,Fe1}(E) + if_{2,Fe1}(E)\right]e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe1})} + \cdots \\ \cdots + \sum_{Fe2}^{8}\left[f_{0,Fe2}(\mathbf{h}) + f_{1,Fe2}(E) + if_{2,Fe2}(E)\right]e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe2})} \end{aligned} \tag{4.6}$$

Figure 4.4 illustrates how small the total contribution of Fe1 is to the structure factor as well as how large the three components (f_0, f_1 and f_2) of each of the Fe2 atoms are. Further simplification can be made by noting that the Thomson, f_0, scattering for all 16 iron atoms is identical *and* energy independent:

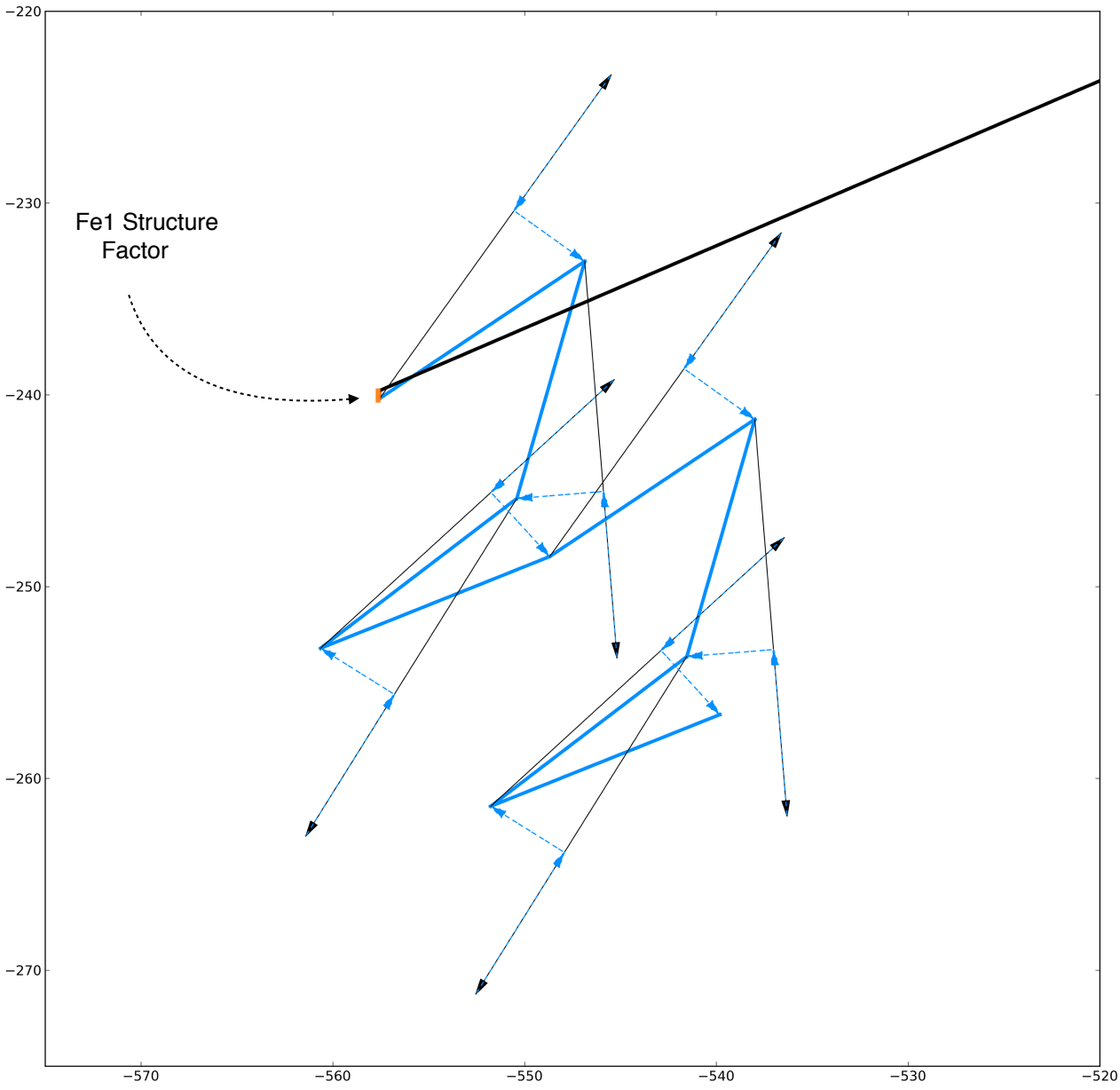
$$f_{0,Fe1}(\mathbf{h}) \equiv f_{0,Fe2}(\mathbf{h}) \tag{4.7}$$

---

[1] Inspiration for these types of diagrams came from Feynman's book on QED.

Figure 4-5

Argand Diagram of the Complex Contribution from a Single Iron



A close up of the total bulk structure factor (long black line), the total Fe1 structure factor (orange), and a breakout of the 8 individual contributions from each Fe2 atom in the unit cell for $\boldsymbol{h}$=(-9, 11, 12) with their anomalous contributions. f_0 Thomson (thin black arrow), f_1 with opposite phase to Thomson (blue) and f_2 perpendicular to Thomson and f_1, (blue dashed). The subtotals of $f_{Fe2} = f_0 + f_1 + if_2$ are the thick blue lines seen in Figure 4-4.

Allowing the bulk atoms structure factor, $F_{Z\text{-}Fe}(\mathbf{h})$, to also absorb the Thomson scattering of the irons to become the *feature-free* or *background* structure factor, F_Z, as it contains all the atom types but none of the absorption edges and is energy independent in the range of this spectrum:

$$F(\mathbf{h},E) = F_Z(\mathbf{h}) + \sum_{Fe1}^{8}\left[f_{1,Fe1}(E) + if_{2,Fe1}(E)\right]e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe1})} + \sum_{Fe2}^{8}\left[f_{1,Fe2}(E) + if_{2,Fe2}(E)\right]e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe2})} \tag{4.8}$$

It is important to grasp the step between equation 4.6 and 4.8. The structure factor has become a sum in three parts: all the energy *in*dependent Thompson scattering, the energy dependent part of atoms labeled Fe1, and atoms labeled Fe2. This would be represented in Figure 4-5 by adding each individual f_0 (smaller black arrows of Fe2) to the large bulk atom arrow (long black arrow).

The sum of the energy dependent part of Fe1 is tiny for this diffraction (the orange dot). By consolidating that part of the structure factor as well as separating the sum of the real and imaginary parts[1] of Fe2, leads to:

$$F(\mathbf{h},E) = F_Z(\mathbf{h}) + \left[\sum_{Fe1}^{8} f_{1,Fe1}(E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe1})} + i\sum_{Fe1}^{8} f_{2,Fe1}(E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe1})}\right] + \left[\sum_{Fe2}^{8} f_{1,Fe2}(E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe2})} + i\sum_{Fe2}^{8} f_{2,Fe2}(E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe2})}\right] \tag{4.9}$$

$$F(\mathbf{h},E) = F_Z(\mathbf{h}) + \left[F_{1,Fe1}(E) + iF_{2,Fe1}(E)\right] + \left[F_{1,Fe2}(E) + iF_{2,Fe2}(E)\right] \tag{4.10}$$

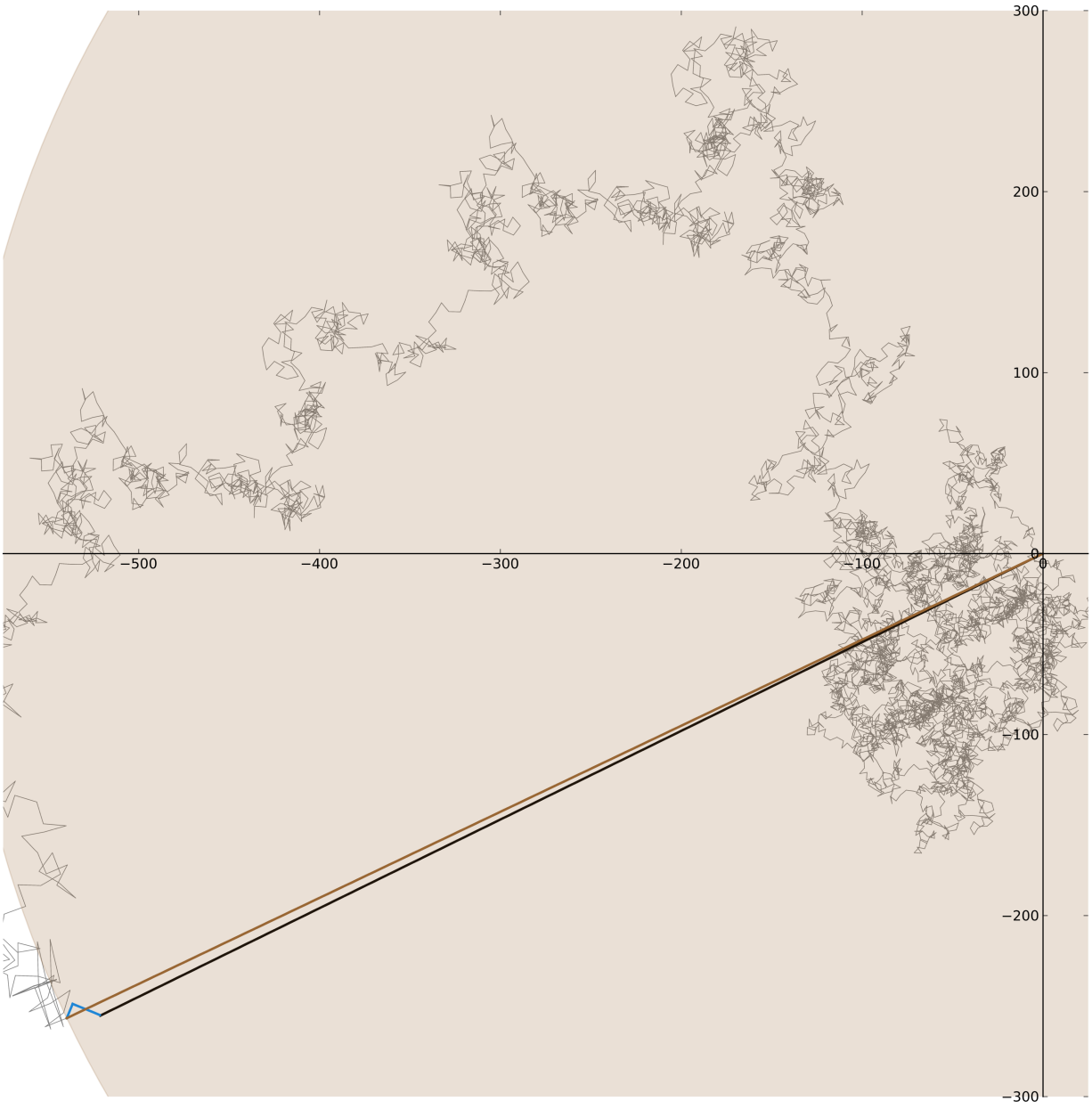
$$F(\mathbf{h},E) = F_Z(\mathbf{h}) + \Delta_{Fe1}(E) + \left[F_{1,Fe2}(E) + iF_{2,Fe2}(E)\right] \tag{4.11}$$

This equation is represented in Figure 4-6 by the solid line for $F_Z(\mathbf{h})$: the minuscule contribution by the energy dependent part of the Fe1 is barely visible in between the black background structure factor and the 'real' part of the blue Fe2 anomalous contribution. As a point of interest, one may also stare at the eight atoms that constitute Fe2 in Figure 4-5 and note the two-fold symmetry from the two asymmetric sub-units, each sub-unit having four atoms apiece.

[1] The Greek capital delta (Δ) is used instead of the lowercase delta (δ) because different diffractions will have larger or smaller values.

Figure 4-6

Total Structure Factor with Single Iron Anomalous Contribution



The total structure factor for the all atoms in the unit cell (brown vector) as a sum of the feature-free structure factor (Thomson, black vector) and the total real and imaginary parts of Fe2's 8 atoms (blue). The transparent circle inscribed by the total structure factor indicates the size of the intensity (when multiplied by π).

4.3.2 Variation in Intensity

The intensity of a diffraction is proportional to the squared modulus of the structure factor as pointed out in Equation 3.12, adapted for the change in notation here:

$$I(\mathbf{h},E) \propto \left|F(\mathbf{h},E)\right|^2 \tag{4.12}$$

This is analogous to the area of a circle inscribed by vector $F(\boldsymbol{h},E)$ with a scale factor of π (light brown shaded area of Figure 4-6). This is how anomalous signal looks for a particular diffraction at a particular energy; the variation in intensity of the diffracted spot is identical to the variation of the area of the circle (divided by π). Through this whole example, only a single energy point has been considered: 7117eV. At this energy, there is a large anomalous contribution to both irons (see Table 4-1). Figure 4-6 shows that the anomalous contribution from Fe2 to the area of the circle will not be purely real or purely imaginary, but some mixture of both, by noting that the angle of the blue Fe2 vectors are oblique with respect to $F_Z(\boldsymbol{h})$.

Table 4-1

Anomalous Contributions by Atom-label

Element	Δf at 7117eV
C	-0.02 + 0.01i
N	-0.04 + 0.02i
O	-0.06 + 0.04i
S	-0.37 + 0.70i
Zn	+1.14 + 0.85i
Fe1	+9.07 + 4.19i
Fe2	+8.71 + 4.54i

When analysis is performed on diffractions from a complex unit cell like that of a macromolecule, it is easy to realize that, with so many thousands of atoms and so many thousand of diffractions, two different target atoms can have an almost smooth set of

contributions by ratio: Fe1 from some maxima to minima and correspondingly Fe2 from minima to maxima. The intensity between diffractions from a macromolecule crystal vary by orders of magnitude and the contribution from just a handful of atoms is relatively small. In the proceeding chapter we formulate how to standardize the bias between two target atoms. The example given for $\boldsymbol{h}=(-9, 11, 12)$ demonstrates that a single diffraction can have a significant contribution from the anomalous dispersion of a single atom type in a large macromolecular unit cell while suppressing an atom of a very similar absorption profile.

4.3.3 Opposite bias

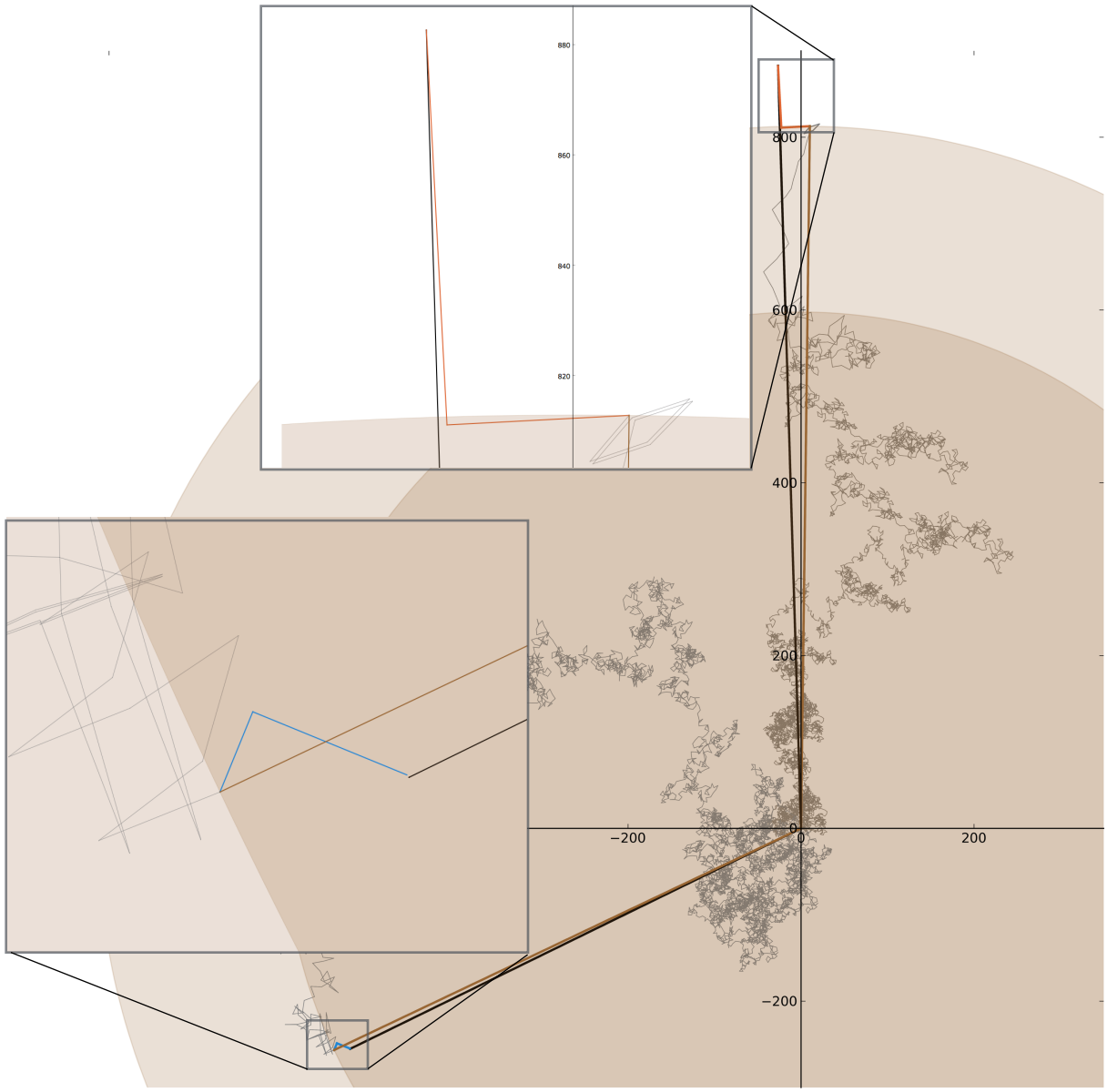
To demonstrate some of the variation available, Figure 4-8 shows the example above with $\boldsymbol{h}_1=(-9, 11, 12)$, but also with $\boldsymbol{h}_2=(10, 26, 0)$. In the case of $\boldsymbol{h}_2$ the first iron, Fe1 (orange), is strung out and the second iron (blue) is all wrapped up on itself. This is opposite to the previous example. What is probably most remarkable about $\boldsymbol{h}_2$ is that the phase of the real part of the anomalous signal is almost parallel with the phase of the background structure factor (the black vector and long orange vector are almost collinear). This means that any variation in intensity of this diffraction across the absorption spectrum of iron will be almost one hundred percent real, f_1, in nature. The imaginary f_2 part from atom-label Fe1 is ninety degrees out of phase with f_1, and therefore its projection, at a tangent to the circle, will effect the overall phase of the structure factor but have very little bearing on the resulting intensity (area of the circle).

4.3.4 Realism

DS can be used to calculate and analyze individual diffractions. The focus of this project, however was to see if it was possible to extract *spectroscopic* data from protein crystals at a normal third generation MX beamline. In order to do that, it was necessary to ignore individual diffractions and take a more holistic approach to recovering the information. The above

Figure 4-7

Total Structure Factor for two Diffractions with Opposite Iron Emphasis



An argand diagram of two total structure factors $\boldsymbol{h}_1=(-9, 11, 12)$ and $\boldsymbol{h}_2=(10, 26, 0)$. The total structure factors (brown lines) are composed of individual factors from each atom (light grey). Black is the total feature-free scattering from all elements and attached to those are the major anomalous contributions f_1 and f_2. For $\boldsymbol{h}_1=(-9, 11, 12)$ the only significant anomalous contribution is Fe2 (blue) and for $\boldsymbol{h}_2=(10, 26, 0)$ its Fe1 (orange).

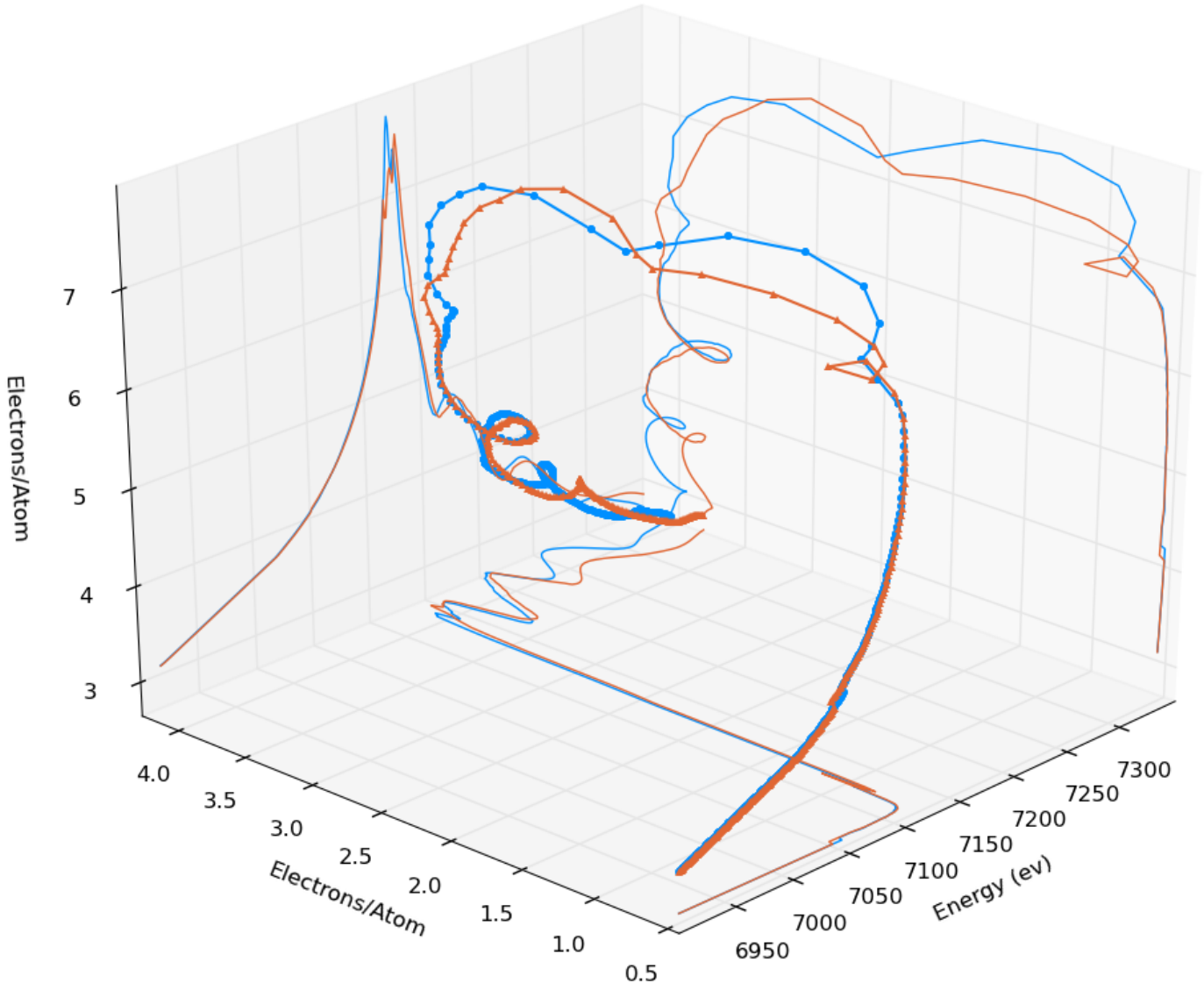
examples are idealized: it assumes that the exact position in space of every element is known with only empty space in between and free from all errors. As one can imagine, this is hardly the case with a real crystal. Due to the size of the iron's diffraction signal, there is a high confidence level on where the iron's are in space with respect to the unit cell. The phase of the anomalous signal with respect to the background structure factor could be very different, and noise will always obfuscate the underlying information. If there is a lot of noise in the signal, it can be two or three times the size of the signal of the anomalous dispersion. In light of these drawbacks, there is one big advantage of collecting large unit cell diffraction: thousands of diffractions all collected at the same time and under identical laboratory environmental conditions.

4.4 Cylinder Projection

A good visual interpretation of the anomalous dispersion, introduced here, is currently called the 'cylinder projection'. The Kramers-Kronig transform demonstrates that the real and imaginary parts of dispersion are actually two views of the same phenomena. It is an identical situation to the sine and cosine function being two perpendicular projections of a single helix. It is useful to remember that the contribution from anomalous dispersion to the intensity of a diffraction is a projection of a cylinder toward the origin. The cylinder is generated from the real and imaginary parts of the anomalous dispersion, and the phase of the cylinder is governed by the phases of the individual structure factors (from that atom-label). With two atom-labels, the effect on the total structure factor is the sum of the two cylinder projections, one cylinder for each atom-label.

Figure 4-8

Cylinder Projection of Anomalous Dispersion for Two Different Irons



The cylinder projection of two different iron atoms: oxidized ferric iron (orange) and reduced ferrous iron (blue). When the cylinder is projected onto two perpendicular surfaces (bottom and left vertical), the Kramers-Kronig dispersion relation can be seen quite clearly. The rear right vertical wall is the Argand projection (occasionally seen in the literature [57]).

CHAPTER 5

ANALYSIS OF DIFFRACTION SPECTROSCOPY

5.1 Data Extraction

A technique has been developed to retrieve the small signals from individual atom-labels out of the large and noisy background of real diffraction. In previous sections, the theoretical production of a diffraction intensity and how it varies with energy was explained. The act of observing the diffraction intensity with a detector obscures the phase information. The absorption profile of the atom-labels of interest are mixed in a variety quantities in different diffraction spots, each with a different phase. It is instructive to understand how intensity variations are extracted from a real dataset in light that the phase part is absent.

The first step is to take a solution dataset of the crystal so that the locations of the atoms in the relevant unit cell can be used instead of a similar or hypothetical structure. The solution set is also used for unit cell parameters and crystal orientation in the DS experiment. This allows the crystal processing software to lock in the orientation and force the Miller indices (hkl) to be the same at subsequent energies. Quantity is almost as important as quality for the method chosen, so selecting a rotation wedge in which large numbers of reflections occur can also be done at this stage. The fundamental methodology of a DS run at a beamline is to shoot a 10° wedge in 1° rotations starting with the X-ray energy well below the core hole absorption of the element of interest and then step the energy across the edge. Repeating the wedge at 50-70 energy steps (a wedge for each energy) over the iron absorption edge will produce a narrow spectrum of 400-500eV. The spectrum may be narrow, but it is sufficiently detailed across the edge by executing larger energy steps at the beginning

and end of the spectrum and gradually tapering the step size to 1eV steps across the edge section. This is similar to energy steps taken while measuring a standard X-ray absorption spectrum in fluorescence; the energy step file is sometimes called a region file by absorption spectroscopists. The crystal is then translated so that a fresh section can be exposed an identical DS run is collected. By laterally translating the crystal, its orientation with respect to diffraction is unchanged and, therefore, another solution set is not required at this point. The crystal is translated two times, giving three DS datasets, assuming the crystal is large enough and/or the beam is small enough. These thin 10° wedges are processed using crystallography software [53] incorporating prior knowledge of the crystal orientation and cell parameters taken during the solution set. The end product of processing supplies a list of *hkl*s and their intensities at the associated energy. If the crystal can be translated, the three wedges from the same energy point can be scaled and merged for better statistics. Each wedge contains approximately 4000-9000 reflections depending on unit cell size, crystal orientation, crystal quality, symmetry, detector size and X-ray energy. The result of the processed data is a 2D matrix of the intensities of each diffraction spot (labelled $\boldsymbol{h}_1$ to $\boldsymbol{h}_n$) as a function of energy from lowest to highest (labelled E_{lo} to E_{hi}), Equation 5.1. In this matrix, each line represents a discrete diffraction spot. This is a semi-sparse matrix and, depending on crystal parameters, energy and processing software, only 60-80% of reflections will be successfully recorded at every energy. Reflections that are not recorded at every energy, for whatever reason, are rejected. The habit of automatically rejecting diffractions that do not reflect at every energy prevents partial diffractions from inconsistently weighting the results and discourages hole-filling, by the experimentalist, in the matrix which would artificially reduce noise:

$$\mathbf{M} = \begin{pmatrix} I_{\mathbf{h}_1, E_{lo}} & \cdots & I_{\mathbf{h}_1, E_{hi}} \\ \vdots & \ddots & \vdots \\ I_{\mathbf{h}_n, E_{lo}} & \cdots & I_{\mathbf{h}_n, E_{hi}} \end{pmatrix} \tag{5.1}$$

The full/dense matrix of reflections is called $\boldsymbol{M}$ and may contain somewhere in the region of 200,000 - 400,000 individual reflections. A number of operations on $\boldsymbol{M}$ are

conducted, such as feature scaling, outlier rejection, low intensity rejection and overall scaling, so that each diffraction, which can vary by orders of magnitude between reflections, is part of a consistent set.

5.1.1 Outlier Rejection, Modified Dixon Q-test and Feature Scaling

When real data is collected (as opposed to simulated data) there are two criteria for which the data can be rejected. Failure to meet the criteria does not reject an individual intensity, because $\boldsymbol{M}$ must be full, a judgement is made to outright reject all intensities for that hkl at every energy. The first outlier rejection criteria uses a mean subtracted diffraction and reject the hkl if one of its values falls outside of an interval. For the myoglobin an interval of [0.5, 1.5] was used and for ferredoxin a slightly more inclusive interval of [0, 2] was applied. The second condition of rejection applied here is a modified Dixon's Q-test [58] in which the data has *not* been rearranged in order of increasing value (which occurs in the normal Dixon Q-tests). Diffractions are rejected if they have values greater that 0.6 for myoglobin and 0.7 for ferredoxin. This test acts as a discontinuity discriminator and, at a very broad level, it allows for lots of noise in the signal but removes egregious diffractions. Once hkls have survived the rejection criteria they are feature scaled. Feature scaling is a process in which a diffraction is standardized across the spectrum:

$$I_{scaled} = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \tag{5.2}$$

The list of reflections that survive are recorded and the feature scaled data of those reflection moves onto overall scaling.

Overall scaling is the value associated with a_1 in Equation 3.13. The overall mean of matrix $\boldsymbol{M}$ is scaled with the overall mean of the simulated matrix generated by software *DeskTools.py (Appendix V). DeskTools* was written in house and has subroutines for theoretical diffraction that include B-factors and occupation. Lorentz factors, normal polarization and detector variations are adjusted for in the data

processing software. Bond-polarization and self-absorption are not calculated as their values are expected to small and demonstrably time consuming. Values for a_2 and a_3 are also not accounted for as the simulated data generated is still so dissimilar on average from values collected on real crystals, the noise levels of which are so high that to apply values to a_2 and a_3 would be disingenuous.

These prior steps are required before analysis can start; reduced the experiment down to a single matrix of information, one must slice the matrix up into three parts before continuing to do any analysis. The three sections are *hkl*s favouring the Fe1 atom, *hkl*s favouring the Fe2 atom and the rest of the reflections. In order to trisect the matrix $\boldsymbol{M}$ in this way, a method to discriminate reflections based on contributions from the target atoms has been developed: called *dev*.

5.2 Calculating Deviation (dev)

Within the narrow spectrum, the sum over all atoms j can be broken up into atoms that do not have dependence on energy and those that do, bulk atoms and target atoms respectively. The method of deviation calculations that follows can be extended for more than two targets without loss in generality[1]. The total structure factor for each diffraction at each energy can be symbolized as follows, from Equation 3.11:

$$\left[F\right]_{\mathbf{h},E} = \left[F_Z + \Delta_{Fe1} + \Delta_{Fe2}\right]_{\mathbf{h},E} \tag{5.3}$$

and, by extension, the intensity:

$$I_{\mathbf{h},E} = \left|\left[F_Z + \Delta_{Fe1} + \Delta_{Fe2}\right]_{\mathbf{h},E}\right|^2 \tag{5.4}$$

Ranking intensity it is a simple matter of calculating each diffraction with:

1) No anomalous contribution (i.e., no energy dependence):

$$I^0_{\mathbf{h}} = \left|\left[F_Z\right]_{\mathbf{h}}\right|^2 \tag{5.5}$$

2) Contributions from the non-anomalous atoms and all the Fe1 atoms:

[1] See the discussion section.

$$I_{\mathbf{h},E}^{Fe1} = \left| \left[F_Z + \Delta_{Fe1} \right]_{\mathbf{h},E} \right|^2 \tag{5.6}$$

3) Contributions from non-anomalous and all the Fe2 atoms:

$$I_{\mathbf{h},E}^{Fe2} = \left| \left[F_Z + \Delta_{Fe2} \right]_{\mathbf{h},E} \right|^2 \tag{5.7}$$

Note that these intensities are used only to calculate the effect that each anomalous scatterer has on a diffraction and are not used in calculating actual expected intensities. To determine each target atom's contribution to the intensity, *sample standard deviation* is used [59], summing across the spectrum, where the energy steps go from $E_1=E_{lo}$ to $E_N=E_{hi}$:

$$\sigma_{\mathbf{h}}{}^{Fe1} = \sqrt{\frac{\sum_{E=1}^{N} \left(I_{\mathbf{h},E}{}^{Fe1} - I_{\mathbf{h}}{}^{0} \right)^2}{N-1}} \quad \sigma_{\mathbf{h}}{}^{Fe2} = \sqrt{\frac{\sum_{E=1}^{N} \left(I_{\mathbf{h},E}{}^{Fe2} - I_{\mathbf{h}}{}^{0} \right)^2}{N-1}} \tag{5.8}$$

These deviations need to be normalized in order to rank whether a diffraction, *hkl*, should be assigned to the Fe1 subset for analysis, the Fe2 subset or whether it should be rejected. This is done by defining the anomalous deviation or *dev:*

$$dev(Fe1)_{\mathbf{h}} \overset{def}{\equiv} \left(\frac{\sigma_{\mathbf{h}}^{Fe1}}{\sigma_{\mathbf{h}}^{Fe1} + \sigma_{\mathbf{h}}^{Fe2}} \right) \quad dev(Fe2)_{\mathbf{h}} \overset{def}{\equiv} \left(\frac{\sigma_{\mathbf{h}}^{Fe2}}{\sigma_{\mathbf{h}}^{Fe1} + \sigma_{\mathbf{h}}^{Fe2}} \right) \tag{5.9}$$

It is important to note that each *dev* has a range from 0 to 1 and, more importantly, that $dev(Fe1)_h+dev(Fe2)_h=1$, allowing reflections to be ranked by their relative contributions. It is also pertinent to mention that all manner of calculus may go into ranking diffractions: signal versus intensity, signal versus angle (resolution), X-ray E-vector versus crystal orientation, or a combination of all and: each of which could include background noise. The method devised above is not the obvious calculation for assigning how much a single atom contributes to a single diffraction. It is however a perfect little tool for assigning intensities to the group in that if it does have a contribution it will be from that particular atom. As seen in the example provided in chapter 4, $\boldsymbol{h}=(-9, 11, 12)$, if the Lorentz factor, B-factor, normal polarization and self-absorption are included for every atom in the crystal, Fe2 would still dominate over Fe1 for that reflection. On the

downside, the method does not include bond-polarization which could diminish Fe2's overall effect.

5.3 Extracting Signal

5.3.1 Separating the sets

Each of the two sets of 8 irons in the unit cell has a phase that is a product of the atomic positions and the reflection, hkl, in the exponent. The anomalous contribution from each iron has two parts: f_1 is parallel and in phase with the Thompson scattering, and f_2 is perpendicular. The overall effect of one iron type (atom-label) on the intensity is the sum of the 8 symmetry-equivalent irons coinciding to make a large contribution and how the magnitude and direction of those irons relate to the *feature-free* structure factor. Principal Component Analysis (PCA) was chosen to analyze the data at it relies on f_1 and f_2 being orthogonal to each other. The aforementioned definition of dev is used to determine each reflection's propensity toward one iron over the other and place them in the matrix $\boldsymbol{M}$:

$$\mathbf{M} = \left(\begin{array}{c|ccc|c|c} \mathbf{h}_1 & I_{\mathbf{h}_1, E_{lo}} & \cdots & I_{\mathbf{h}_1, E_{hi}} & dev(Fe1)_{\mathbf{h}_1} & dev(Fe2)_{\mathbf{h}_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{h}_n & I_{\mathbf{h}_n, E_{lo}} & \cdots & I_{\mathbf{h}_n, E_{hi}} & dev(Fe1)_{\mathbf{h}_n} & dev(Fe2)_{\mathbf{h}_n} \end{array} \right) \tag{5.10}$$

Applied to $\boldsymbol{M}$ is a threshold on the dev columns and only reflections with deviations that overcome this value are kept; the $\boldsymbol{M}$ matrix is sliced into three parts: favourable to Fe1, favourable to Fe2, and the rejected middle. For example, in the simulated diffraction experiment in the next chapter, a threshold of 0.95 was chosen. This means that reflections, that have a dev of 0.95 or greater from Fe1 are put into one matrix and those with a dev over 0.95 from Fe2 are put in the other. These are mutually exclusive sets of data (anything with a dev of 0.95 in Fe1 has a dev of 0.05 in Fe2, and vice versa): all other reflections are discarded. The final two matrices are only diffraction intensities: reflections ($\boldsymbol{h}$) down one axis and energy (E) along the other:

$$\mathbf{M}_{dev(Fe1:0.95)} = \begin{pmatrix} I_{\mathbf{h}_i,E_{lo}} & \cdots & I_{\mathbf{h}_i,E_{hi}} \\ \vdots & \ddots & \vdots \\ I_{\mathbf{h}_n,E_{lo}} & \cdots & I_{\mathbf{h}_n,E_{hi}} \end{pmatrix} \quad \mathbf{M}_{dev(Fe2:0.95)} = \begin{pmatrix} I_{\mathbf{h}_j,E_{lo}} & \cdots & I_{\mathbf{h}_j,E_{hi}} \\ \vdots & \ddots & \vdots \\ I_{\mathbf{h}_n,E_{lo}} & \cdots & I_{\mathbf{h}_n,E_{hi}} \end{pmatrix} \tag{5.11}$$

In laboratory experiments, the entries in the matrix are processed intensities using XDS software on real crystal diffraction. For the simulated experiments, they are generated by *DeskTools*. These subset matrices greatly favour one of the irons over the other, and it is simply a matter of listing the reflections from each dev matrix and extracting those from the dataset of the real data.

5.3.2 Principal Component Analysis

Principal component analysis attempts to find patterns in data by calculating eigenvectors and eigenvalues for a multi-dimensional covariant matrix [62]. For one of the matrices $\boldsymbol{M}$ above an entry in the covariant matrix, C, would be:

$$\operatorname{cov}\left(\mathbf{h}_i,\mathbf{h}_j\right) = \sum_{n=E_{lo}}^{E_{hi}} \frac{\left(\mathbf{h}_{i,n} - \bar{\mathbf{h}}_i\right)\left(\mathbf{h}_{j,n} - \bar{\mathbf{h}}_i\right)}{\left(n-1\right)} \tag{5.12}$$

where the bar denotes the average and covariance is closely related to variance which is a more general form of sample standard deviation:

$$C = \begin{pmatrix} \operatorname{cov}\left(\mathbf{h}_1,\mathbf{h}_1\right) & \cdots & \operatorname{cov}\left(\mathbf{h}_1,\mathbf{h}_n\right) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}\left(\mathbf{h}_1,\mathbf{h}_n\right) & \cdots & \operatorname{cov}\left(\mathbf{h}_n,\mathbf{h}_n\right) \end{pmatrix} \tag{5.13}$$

The python subroutine [63] does all the heavy lifting by calculating the eigenvectors and -values V and D of the covariance matrix:

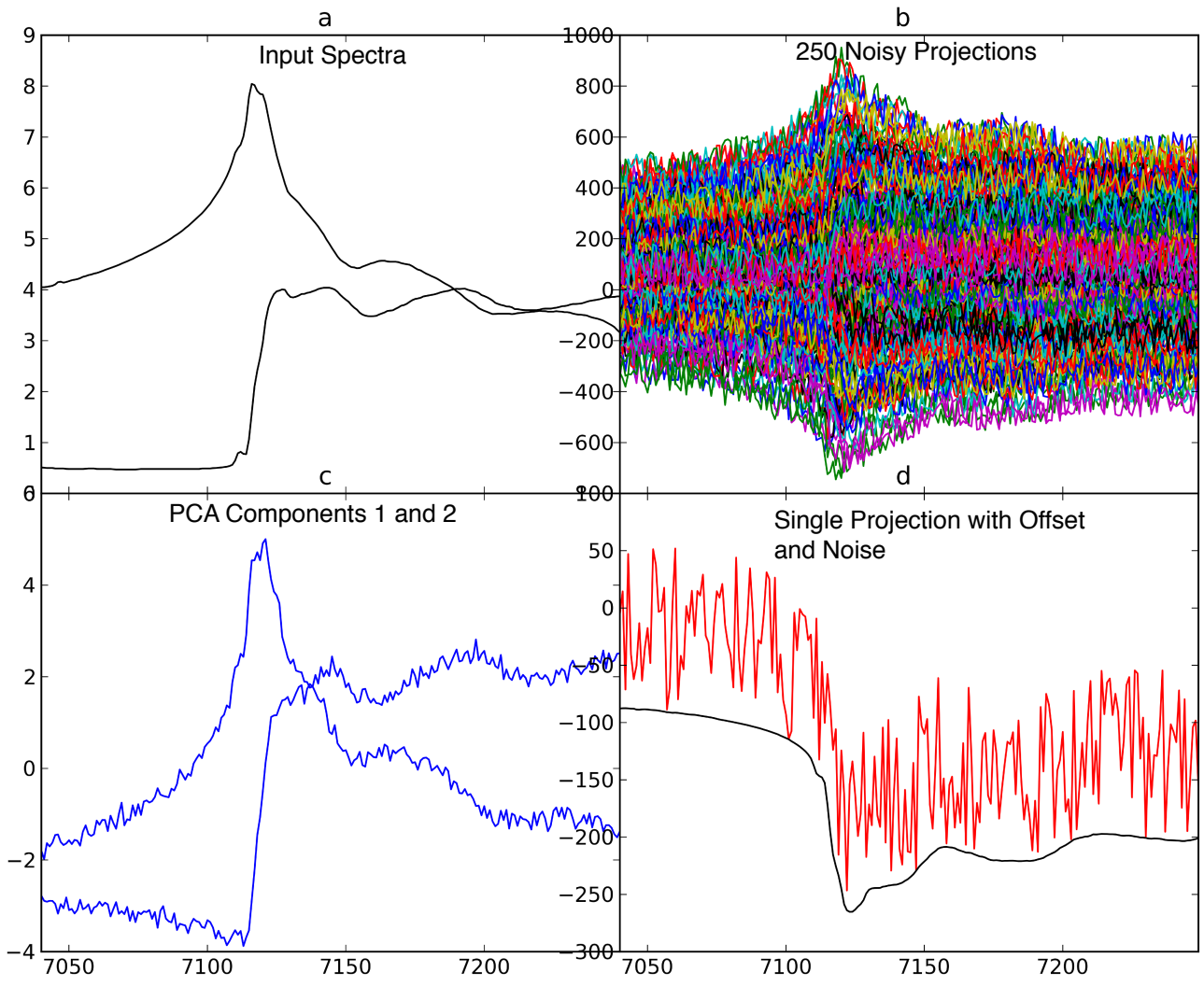
$$V^{-1}CV = D \tag{5.14}$$

During the testing phase of the PCA module, a randomly rotated helix projected onto two orthogonal planes was used with various random levels of offsets and noise. The results were so convincing that the tests were repeated using a raw anomalous

dispersion spectrum. With 250 exceptionally noisy input data, the software was able to pull out two orthogonal spectra that are very similar in shape and structure to the inputs, (Figure 5-1). It is clear that combined with scaling and a slight rotation that the original data could be retrieved using this technique, if the signal to noise was not too extreme. PCA with a rotation is also known as abstract factor analysis [62]. Three drawbacks of PCA are: 1) the scale of the results can only be approximated by multiplying by the square-root of the eigenvalue; 2) the offset with regards to the baseline is lost; and 3) the sign of the calculated spectra is, for all intents and purposes, assigned randomly every time. Despite these drawbacks, the inclusiveness of this method is appropriate for the large-scale noisy data that is created when diffracting from a large macromolecule unit cell crystal which only has a handful of small signals embedded in it. The underlying spectra can be regarded either as mixtures of the two Kramers-Kronig dispersions, each taking a sine and a cosine of the phase, or as a single cylinder projected toward the origin with the PCA results also being paired off and forced into a counterclockwise cylinder (see Figure 6-4). It is advantageous to the input and output cylinders to release them of their coordinate systems and compare features. In near edge and fine structure , it is not uncommon to scale and fit splines as well as unmoor the exact position of absorption. PCA is an excellent tool for supplying the underlying *shape* of a signal, but it is impossible to assign values to that shape in a meaningful way unless it is scaled to a known spectra. The act of changing signs, shifting and scaling as well as combining perpendicular components with an assumed handedness should give pause for thought In the following chapters, a discussion on how to scale with regards to real data is given along with a look at the benefits of a slight rotation if the quality of results is high.

Figure 5-1

Testing the Principal Component Analysis Module



a) A pair of Kramers-Kronig dispersion relations used as the input spectra. b) The pair are joined rotated and projected 250 times with a random noise multiplier as well as an offset. d) A closer look at single projection with and without noise. c) the results of the PCA python module working on the 250 noisy spectra/data. The two top components with the highest scores are shown. No fitting algorithm applied to the results however they have been scaled by one over the square root of their score.

CHAPTER 6

SIMULATED DIFFRACTION

6.1 Raw Ingredients to Simulated Diffraction

Chapters 7 and 8 cover two laboratory experiments. In chapter 7 a single iron containing macromolecule diffraction spectrum from myoglobin is compared to its fluorescence spectrum; and in chapter 8, the site separated diffraction spectra from two inequivalent irons in a 2Fe-2S cluster from the large macromolecule ferredoxin is given. These two experiments require a detailed knowledge of the atomic positions as well as insight into which diffractions collected during the experiment are valuable for analysis. The software *DeskTools* was written to evaluate the results of the one- and two-metal proteins. In order to evaluate the dev of relevant hkl intensities, *DeskTools* was written to simulate diffraction from a PDB file. With these new programs it became possible to calculate and analyze an ideal simulated dataset.

Prior to experiments on actual crystals, simulations were conducted whereby a small sphere of reciprocal lattice vectors (hkl), publicly available structures from the PDB as well as arbitrary spectra were applied to the two different target atoms. Initially a crude 5 point spectrum was used, then a real absorption spectrum was used to simulate the anomalous dispersion. *DeskTools* takes these inputs and generates $I(\mathbf{h},E)$ for diffractions of the reciprocal lattice points at energies on the spectrum. These simulations were successful and proved invaluable with ferredoxin in particular, demonstrating the strength of the technique which will be discussed in detail below.

Simulated diffraction was created in *DeskTools* using atomic positions, energy data points, a list of diffractions (hkl) and the B-factors; from an experiment on a real

crystal *I2*. The atomic form factor Thomson scattering values, $f_{0,j}$, were calculated using Cromer-Mann coefficients. The dispersion values for f_1 and f_2 of *non*-target bulk atoms are taken from Cromer-Liberman. The anomalous dispersion values for target atoms uses the partially simulated spectra from Figure 4-1 at energy points from the region file that cover iron K-edge absorption.

dev was created to deconvolute diffractions by contributions from site separate atoms. Simulated diffraction was created to simulate diffraction but not necessarily to simulate reality. As described in Section 3.3 in order to simulate real diffraction many factors need to be included. B-factors and occupancy do not significantly effect any of the calculations with regards to calculating *dev*. In order to calculate an atom's contribution to *dev*, temperature is not included, however when simulated diffractions are generated, temperature and occupancy are included. Simulated diffraction intensities are calculated using the squared modulus of the following structure factor:

$$F(\mathbf{h},E)=\sum_j f_j(\mathbf{h},E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_j)}e^{-B_j/4d_{\mathbf{h}}^2}O_j \tag{6.1}$$

Where $\boldsymbol{r}_j$ is the location of the atom j, B_j is the temperature factor, O_j is the occupancy and d_h is the d-spacing for the reflection which can be solved using Bragg's Law [17] where $n=1$:

$$2d_{\mathbf{h}}\sin\theta = n\lambda \tag{6.2}$$

The simulated diffraction is not a perfect simulation of real data as it does not include radiation damage, air scatter, noise, self-absorption, bond polarization, or systematic errors. Nothing obscures the shape of the anomalous scattering factors (in simulated diffraction) associated with each iron, except the intrinsic complication of diffraction itself (and temperature). Therefore *dev* is only attempting to deconvolute and site separate atoms from the complications associated with the loss of phase from diffraction.

A simulated run of both the myoglobin single iron and the ferredoxin two-iron was conducted, however, no report has been made on the single iron because the output

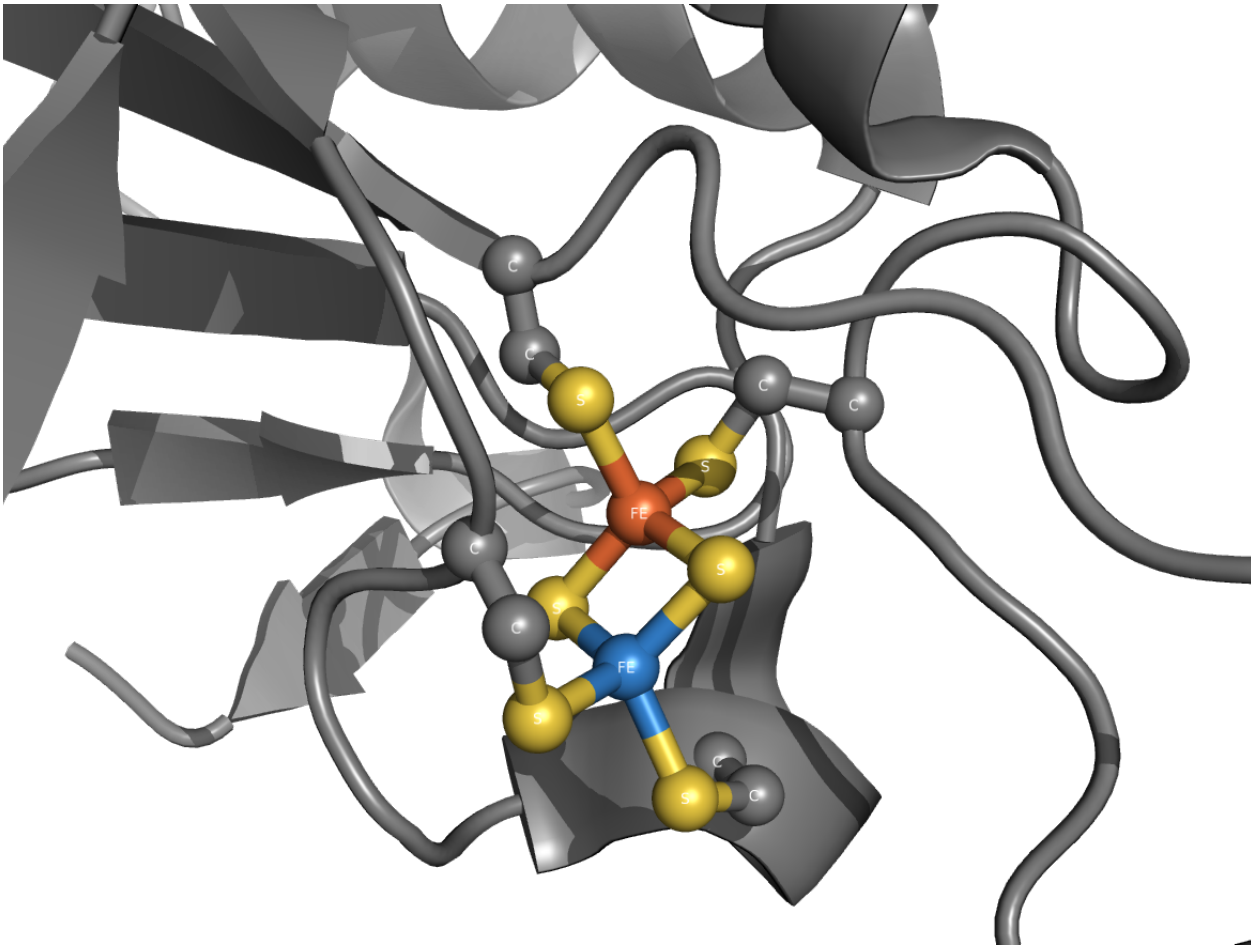
spectral results were indistinguishable from the input. With the multi-iron ferredoxin experiment, the nature of crystallography and lack of phase information starts to interfere with the separation of the two iron signals in the diffraction and its analysis. Extensive use of experimentally acquired initial values are exploited and then extrapolated using the equation for intensity, Equation 5.4. The initial simulated experiments were successful in site separating anomalous dispersion signals and once real structures and data had been acquired, the experiments were repeated using simulations to probe the limits of the technique. Repeating the simulated experiments using the reciprocal lattice vectors (hkl) obtained in a laboratory DS run with atomic positions from a solution derived at the the same time demonstrated the strength of PCA in separating and retrieving the underlying spectra from simulated data. The results described below clearly show the ultimate goal of DS: decoupling the site specific spectra.

6.2 Simulated Ferredoxin

The two irons of ferredoxin are in different oxidation states as the surface iron interacts with molecules outside of the protein, it has been experimentally determined [9, 61, 62] to be the reduced iron. The FEFF suite of programs was used to calculate the EXAFS spectrum of the surface iron and to compare it with that of the iron deeper in the body of the protein. All high resolution (<2Å) 2Fe-2S cluster containing ferredoxins in the PDB are included in a detailed analysis of FEFF spectra from PDB structures in Appendix II. The Cyanobacterium Anabaena PCC7119 ferredoxin (1CZP.pdb) was used as the model for the EXAFS oscillations in this chapter as the structure has very high resolution (1.17Å) models for both oxidized and reduced states [61]. The EXAFS oscillations were calculated to $k=6$ with a maximum of 4 multiple scattering legs. Each iron is tetrahedrally bound to 4 sulphurs, sharing 2 bridging

Figure 6-1

Iron-Sulphur Active Site of Ferredoxin 1CZP



The 2Fe-2S active site of Ferredoxin. The two tetrahedrally bonded irons are coloured orange and blue. The orange, Fe^{3+}, ferric iron, Fe2, is deeper in the protein with a stable oxidation state, it is anchored by cysteine 9 and 22. The blue, reduced, ferrous Fe^{2+}, surface iron Fe1, is believed to be the operative part of the 'active site'. It is anchored by cysteine 55 and 59.

sulphurs, and all atoms within 5Å of the the target atom were included in the calculations, see Figure 6-1. The EXAFS oscillations were then combined with near edge spectra from an unrelated absorption experiments of both an oxidized and reduced iron in order to have theoretical spectra that cover the region of interest, 6910eV - 7345eV. These semi-theoretical oxidized and reduced iron absorption spectra were then put through the fftkk.py software to calculate the imaginary, f_2, part of the anomalous signal. The original spectra can be seen in Figure 3-1, the FEFF calculations are documented in Appendix II and the Kramers-Kronig Fourier transform software is available in Appendix III.

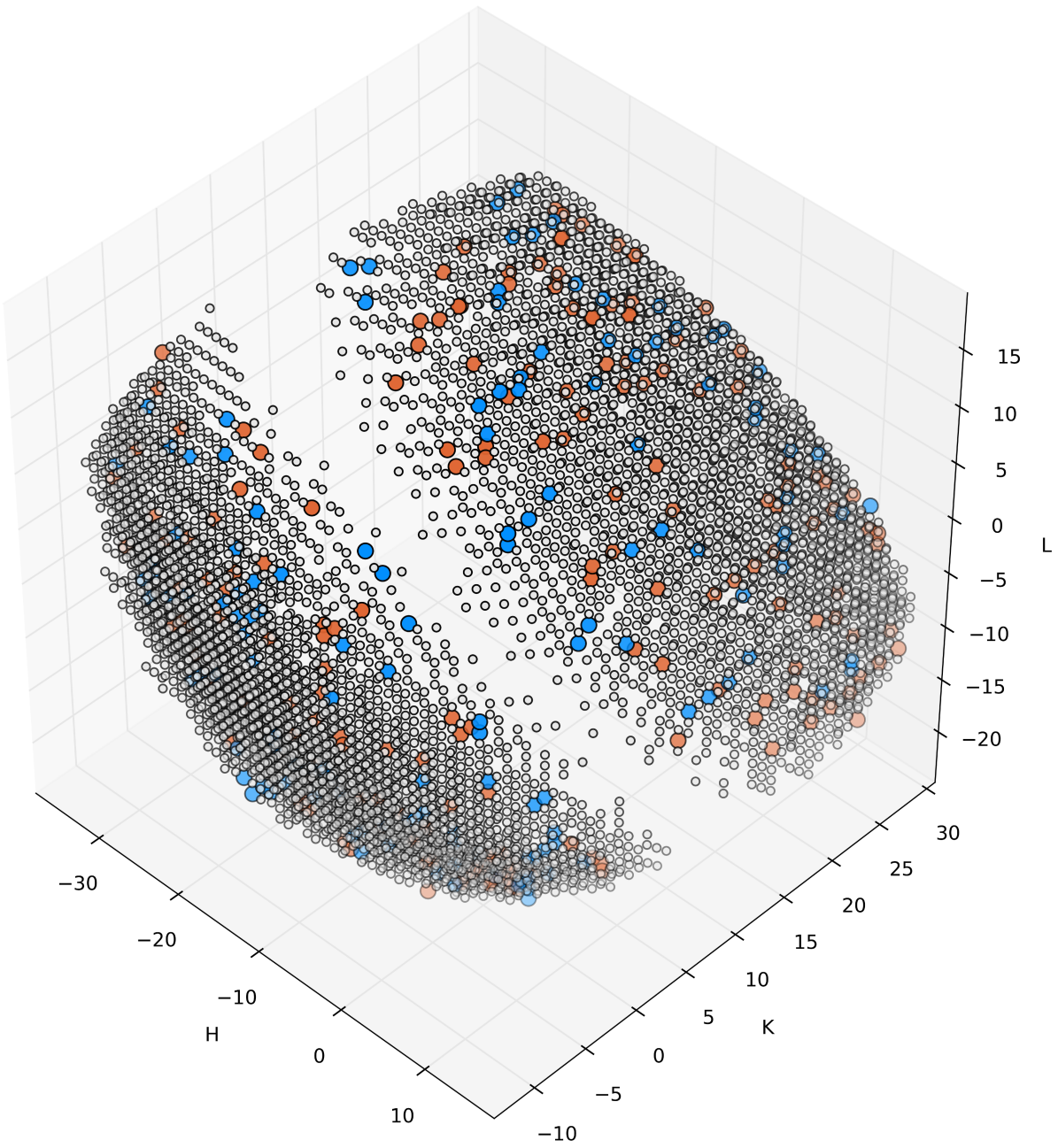
Initial values for a list of intensities is generated by *DeskTools.* Each atom in the unit cell is designated with an element Z, a position $\boldsymbol{r}_j=(x, y, z)$, and a temperature B-factor, from the PDB file and *DeskTools* calculates all the symmetry related atoms within the unit cell. The list of reflections were taken from experiments[1] with crystal *I2*. The Thomson scattering uses the d-spacing of the reflection plane which is calculated using the unit cell dimensions ($a, b, c, \alpha, \beta, \gamma$) also from the PDB file. Different oxidation states were then assigned to the two irons by giving each one the anomalous spectra theorized above and a the list of intensities is recalculated for every energy in the spectrum. This is the simulated data which no longer contains any phase information as the absolute value of the total structure factor has been squared, see Equation 5.4.

6.3 Results of Simulated experiment

With the list of reflections that dev calculated for this matrix of intensities, the intensity is free from external noise as the data is simulated. A very high threshold can be used to separate the matrix into its three parts: favouring the first iron, Fe1, the second, Fe2, and rejecting all the reflections whose mixture does not favour one over the other. A threshold of 0.95 was chosen and by doing so reduced the 5428 reflections into two sub-matrices with 179 reflections favouring Fe1 and 141 favouring Fe2.

[1] Details of how this list of reflections is acquired are given in the experimental ferredoxin chapter (8).

Figure 6-2



A plot of the 5428 reflections in hkl-space. Reflections favouring the Fe2, oxidized (orange), reflections favouring Fe1 are reduced (blue) are shown for a threshold of 0.95. The discarded reflections are white. The blank section through the middle is associated with the rotation axis

$$\mathbf{M}_{dev(Fe1:0.95)} = \begin{pmatrix} I_{\mathbf{h}_1,E_{6910}} & \cdots & I_{\mathbf{h}_1,E_{7345}} \\ \vdots & \ddots & \vdots \\ I_{\mathbf{h}_{173},E_{6910}} & \cdots & I_{\mathbf{h}_{173},E_{7345}} \end{pmatrix} \quad \mathbf{M}_{dev(Fe2:0.95)} = \begin{pmatrix} I_{\mathbf{h}_1,E_{6910}} & \cdots & I_{\mathbf{h}_1,E_{7345}} \\ \vdots & \ddots & \vdots \\ I_{\mathbf{h}_{140},E_{6910}} & \cdots & I_{\mathbf{h}_{140},E_{7345}} \end{pmatrix} \tag{6.3}$$

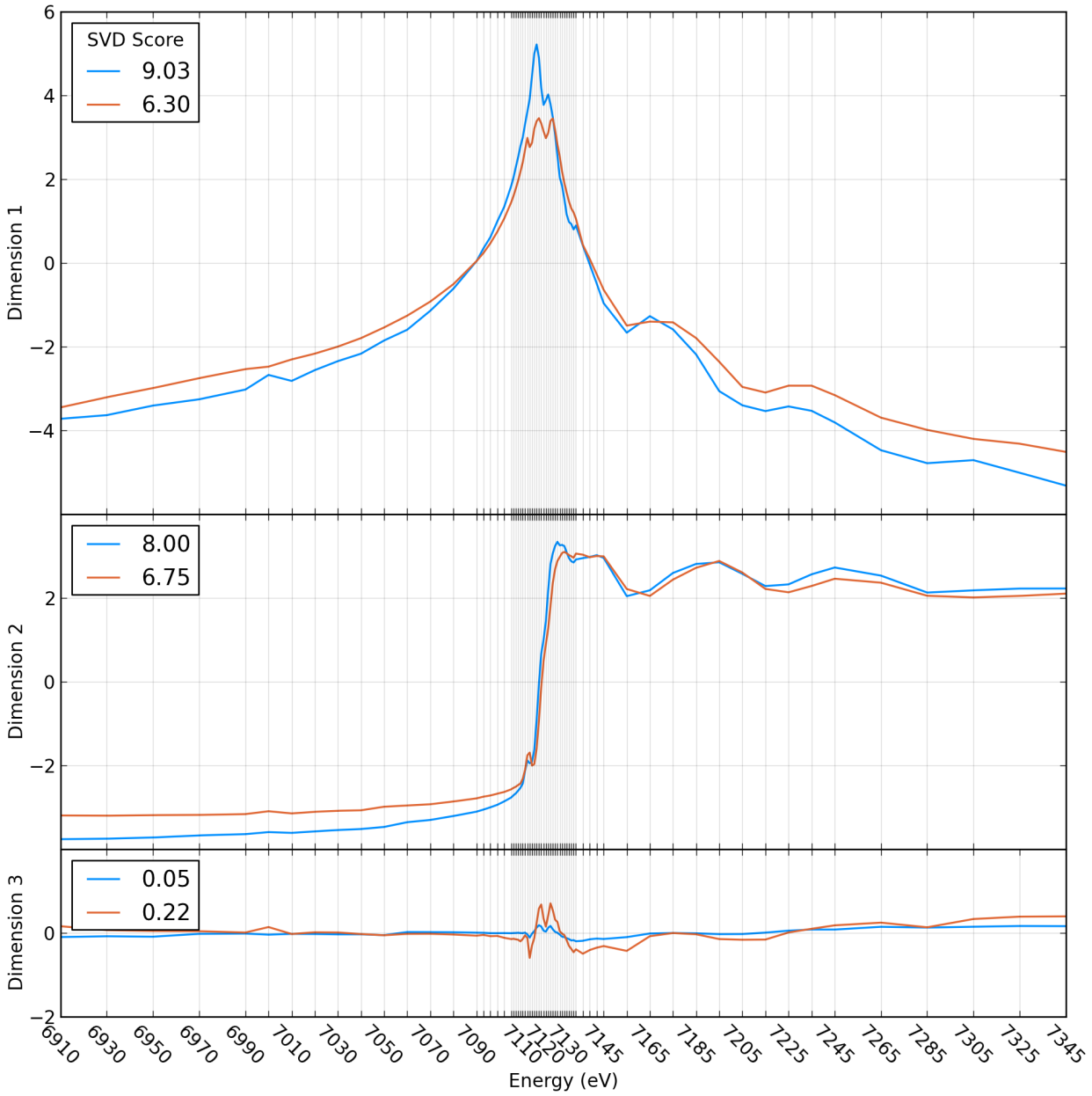
Each intensity contains all elements in the unit cell, however, one matrix has been calculated where all the Fe2 structure factors are wrapped up so their total contribution is low and the other where Fe1 is equally low. The contribution from the atom-label of interest is the sum of two perpendicular vectors, f_1 and f_2, from the dispersion relation and can vary in magnitude from near zero to close to 8 times the the anomalous dispersion[1]. Because all phase information is lost with intensities two signals are mixed in at unknown levels in each reflection. There are 179 and 141 reflections in each of the Fe1 and Fe2 matrices, respectively. The small, but significant 0.05 *dev* from the opposing atom-label will act like noise in the signal. To each of these matrices the PCA module is applied, which uses eigenvalue decomposition to separate linearly uncorrelated variances. The anomalous signals in the reflections are related through the Kramers-Kronig transformation, but this is opaque to PCA which only attempts to maximize variance in successive dimensions. However Kramers-Kronig pairs are orthogonal to each other as are the resulting dimensions from PCA. In theory only $n+1$ channels are required to decompose n variables; with channels numbering 179 and 141 the results presented are sufficient, but not perfect. The small amount of the other atom's signal in the opposing matrix is strongly correlated: the two cylinders are very similar, and therefore disruptive to isolating one from the other.

PCA returns as many components as the number of reflections and each has an associated eigenvalue, also referred to as, 'the score'. The results require two signals from each matrix, however, it is useful to look at the top three as the third component

[1] There are 8 of each type of iron; as structure factors are a sum, eight times the anomalous dispersion contribution is the limit for this crystal.

Figure 6-3

Results of Principle Component Analysis on Simulated Diffraction



The top three dimensions (components) from a PCA are shown for two separate matrices of diffraction from a simulated DS run conducted on two-iron ferredoxin along with their scores. One matrix favouring the reduced iron (blue) and the other favouring the oxidized (orange) iron.

gives an accurate illustration for how quickly the score drops off. PCA does not return experimental values for the ordinate as the input values are mean-subtracted before undergoing decomposition, however, they are an indication of the size of the variance. Figure 6-3 shows the top three results of PCA from the two matrices; $dev(Fe1{:}0.95)$ and $dev(Fe2{:}0.95)$ are blue and orange respectively.

6.4 DeskTools

It takes approximately twenty minutes to run each of the two time intensive parts of a simulated run on a standard desktop computer. The first part strips the *hkl*s from the processed XDS reflection files and generates a matching simulated reflection file using Equation 6.1. The second part takes the *hkl*s from the simulated reflection files and generates a dense matrix and then calculates the deviation values which returns the list of reflections that are to be used by the PCA module. For this simulated run the dense matrix of 5428 reflections has a resolution range of 35Å - 1.77Å. The period of time it took to evaluate the PCA slows exponentially depending on the size of the $\boldsymbol{M}_{dev(atom\text{-}label:threshold)}$ matrix due to the exponential nature of calculating the covariance matrix. For lists under 200 reflections, as in this experiment, it takes about one minute. DeskTools saves each step in a binary file so when adjustments are made to parameters it may utilize those parts that have previously been calculated, but are not effected. The output dimensions/spectra are displayed in Figure 6-3.

Atom-label Fe1 in 1CZP.pdb is assigned with the reduced iron spectrum and Fe2 with the oxidized before diffraction is simulated and these spectra are returned by using dev and the PCA module. The easiest feature to recognize is the triple peaked crown of the oxidized iron f_1: compare the orange line of dimension 1 in Figure 6-3 to oxidized iron in Figure 3-1. Four features are immediately clear that differentiate the results from the input spectra: 1) The lack of a perfectly flat baseline in dimension 2 before the inflection. 2) The main inflection point in dimension 2 is almost shared. 3) The score size of dimension 1 vs dimension 2 of the orange result is opposite to that of the blue. 4) All

three peaks of orange dimension 1 are level, whereas in the input spectra they are ascending in value.

The next step, when the results are returned, is to compare the input spectra with the PCA results as the values from PCA are free floating unsigned lines. In Section 5.3.2, it is noted that PCA is a subsection of Abstract Factor Analysis (AFA) which includes a rotation. Therefore, in order to properly scale input spectra to the results f_1, f_2, dimension 1 and dimension 2 are normalized so their lowest and peak values are in the range of -1 to +1 (feature scaling). Next, the cylinder forms are projected and the PCA results are rotated around the mutual origin/axis at *(E, 0, 0)* until there is a least residual fit with the input spectra. It is possible to then reverse the normalization factor of the input spectra and apply it to the newly rotated projections of the PCA and compare them (Figure 6-6 and 6-7). Dimensions 1 and 2 for Fe1 are rotated -0.6° and Fe2 is rotated -1.7°; when the cylinder is reprojected they are rescaled by normalization factors of 4.6 and 4.0 for f_1 and f_2 respectively, plus a shift to bring their centroids into alignment.

6.5 Simulated Ferredoxin Summary

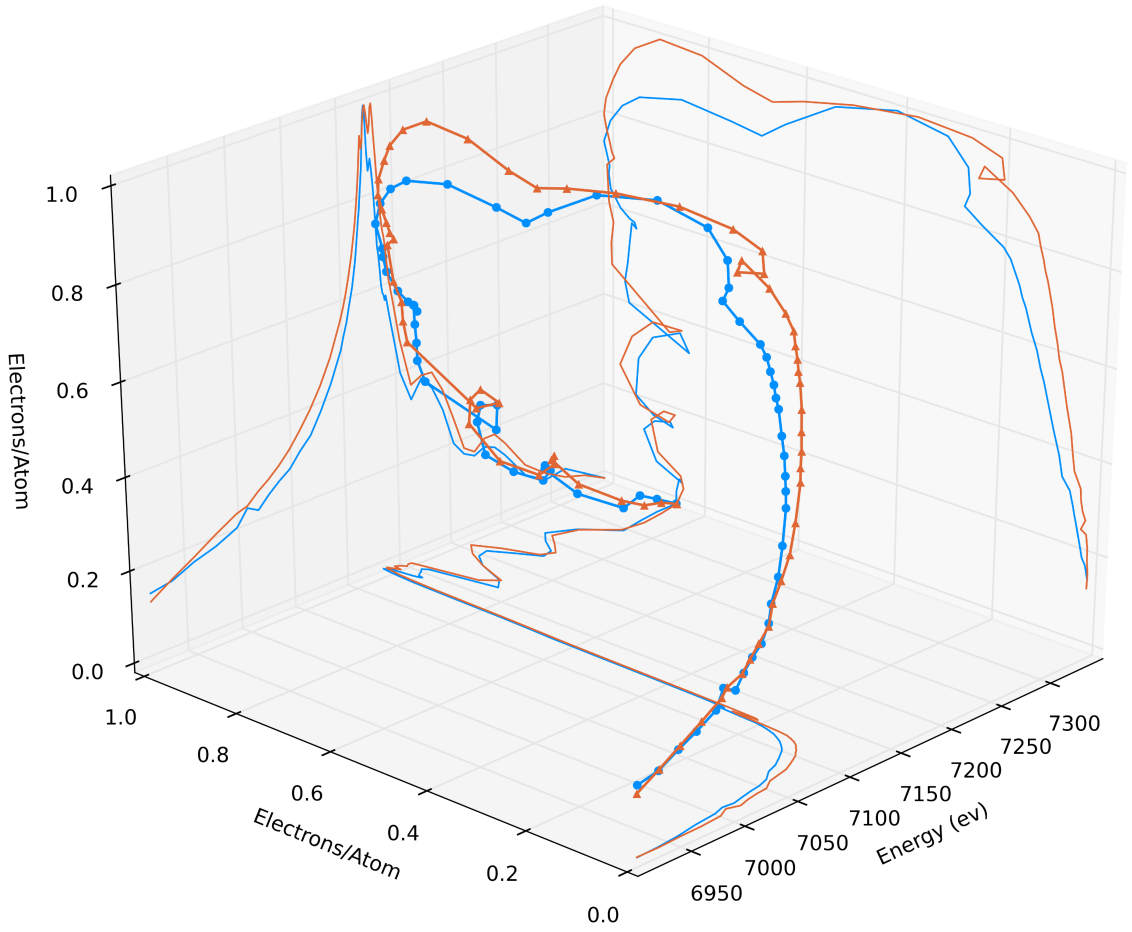
In summary, the 8 atoms labeled Fe1 are associated with an oxidized iron spectra, the 8 atoms labeled Fe2 are associated with a reduced iron spectra. When their simulated intensities are calculated, collated and separated by a deviation threshold they return the spectra assigned to them when operated on by AFA. The results in Figure 6-4 and 6-5 are so closely matched to the input with almost all major features transferred. One feature that persists throughout all attempts of getting the output and the input as sensibly rotated and scaled as possible is the slight misaligned inflection point of the Fe1 output. It is shifted down in energy by a fraction, but it is toward Fe2; making their inflection points almost identical. It is possible that Dimension 3 has the residual information in it (Figure 6-3), but there is no clear way to apply it without bias on the part of the experimentalist. Future development of this fitting technique may yield better results. It is possible to automate the handedness of the

cylinder, pick a centroid axis (based on end points[1]) about which to rotate and write a least squares algorithm for the best fit. Part of the error occurring from PCA is that each dimension is orthogonal by maximum variance, which is not the same as the three dimensions used by the input (E, f_1, f_2). This is not unexpected, but it makes the arguments that these *are* spectra, or spectral equivalents, less intuitive and it is unclear how to rotate (through a fourth dimension) to include the signal left out in dimension 3. These are small errors, Figures 6-6 and 6-7 clearly demonstrate the validity of this technique. Simulated experiments such as these will continue to be performed alongside real crystallographic experiments to track results and expectations.

[1] Ideally, it would be very distant points generated by a spline.

Figure 6-4

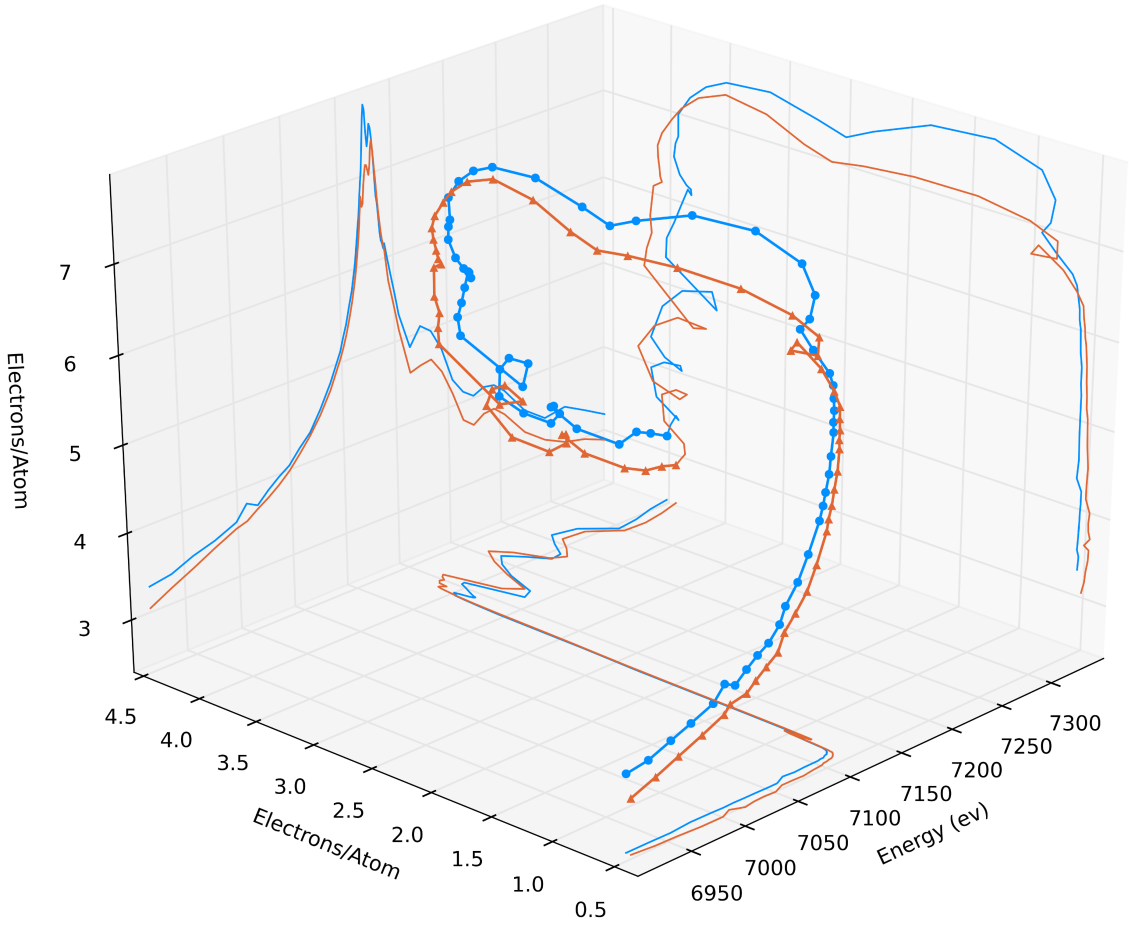
Cylinder Projection of the PCA Module Working on *dev*(Fe1:0.95) and *dev*(Fe2:0.95) for Simulated Ferredoxin



The results of PCA module working on the separated matrices of simulated ferredoxin. The cylinder projection of Dimensions 1 and 2. Those associated with the reduced iron (blue) and the oxidized iron (orange).

Figure 6-5

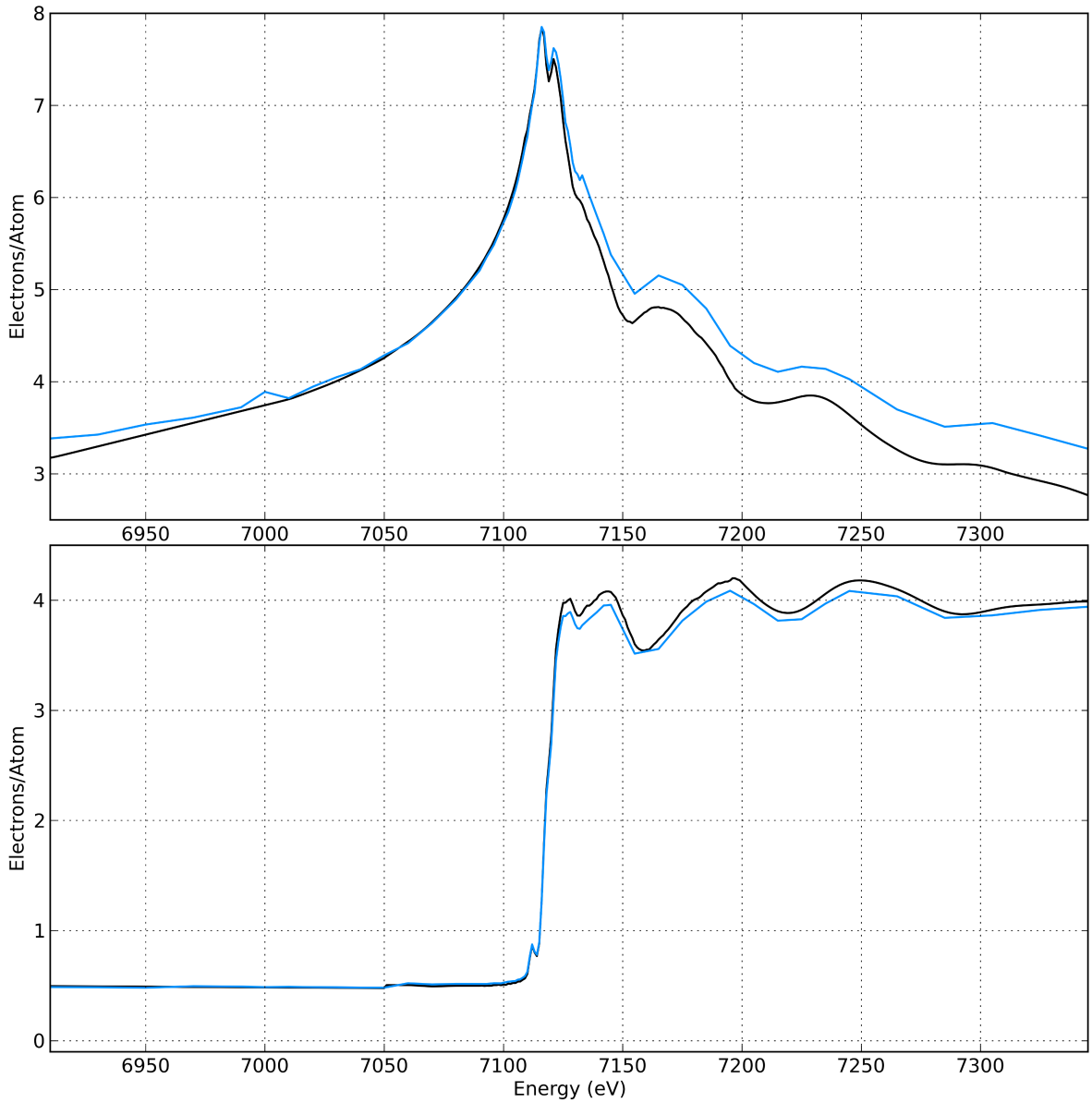
Cylinder Projection of the Rotated and Scaled PCA for Simulated Ferredoxin



The results of PCA module working on the separated matrices of simulated ferredoxin. The cylinder projection of Dimensions 1 and 2. Those associated with the reduced iron (blue) and the oxidized iron (orange). The cylinder has been rotated and scaled to the value to fit the values of the input spectra.

Figure 6-6

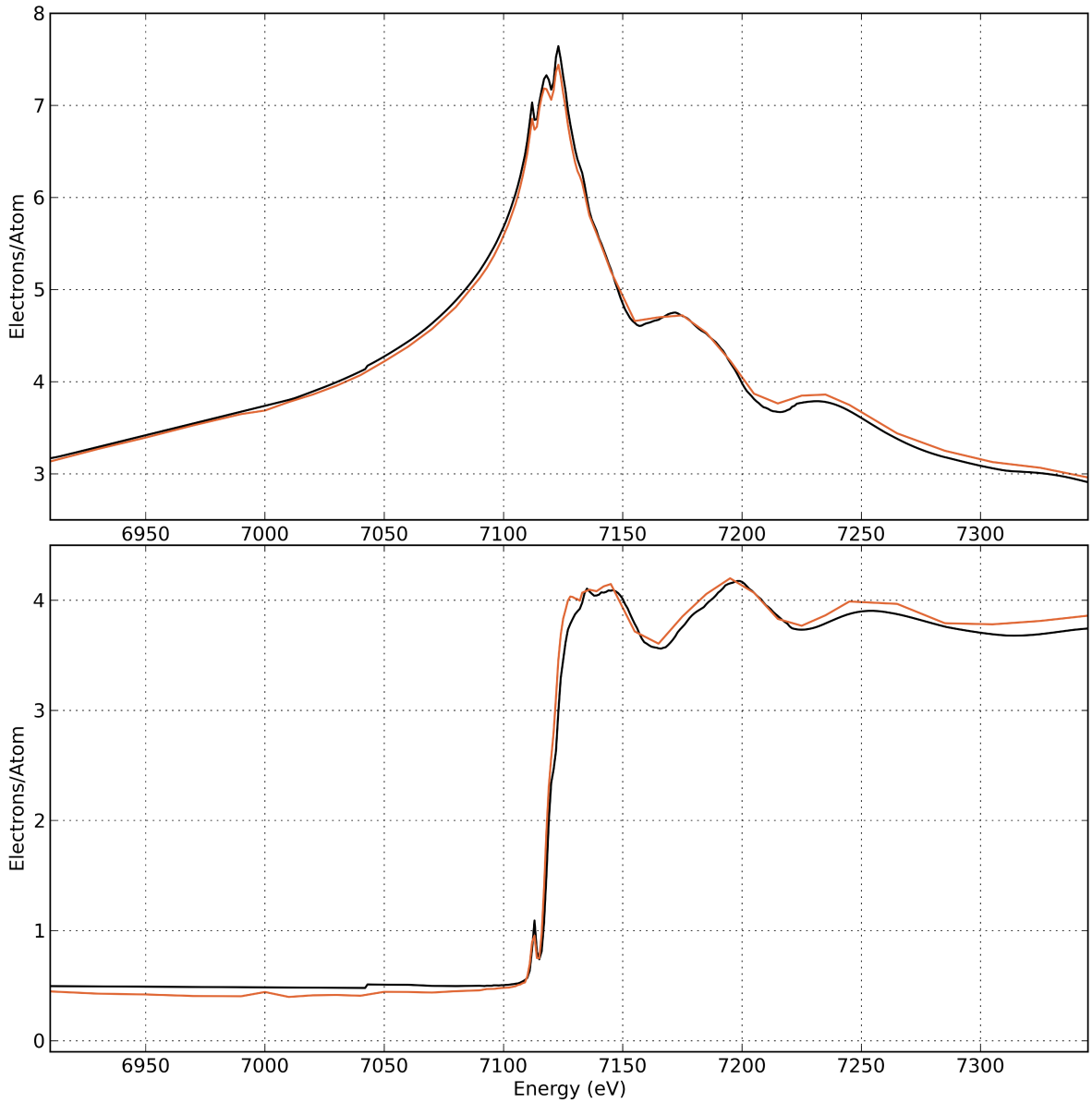
Reduced Iron Spectrum vs Rotated and Scaled from PCA of *dev*(Fe1)



The comparison of the input spectra for the reduced iron and the simulated diffraction results. This is a subset of the simulated diffractions that favour atom Fe1 outer iron that is associated with this spectrum. The real and imaginary parts of the anomalous dispersion are combined into a cylinder, feature scaled and then dimensions 1 and 2 are also combined, scaled and rotated until a minimum residual is returned. The resulting projections were then scaled back to the original values of the dispersion relation.

Figure 6-7

Oxidized Iron Spectrum vs Rotated and Scaled from PCA of *dev*(Fe2)



The comparison of the input spectra for the oxidized iron and the simulated diffraction results. This is a subset of the simulated diffractions that favour atom Fe2 outer iron that is associated with this spectrum. The real and imaginary parts of the anomalous dispersion are combined into a cylinder, feature scaled and then dimensions 1 and 2 are also combined, scaled and rotated until a minimum residual is returned. The resulting projections were then scaled back to the original values of the dispersion relation.

CHAPTER 7

MYOGLOBIN

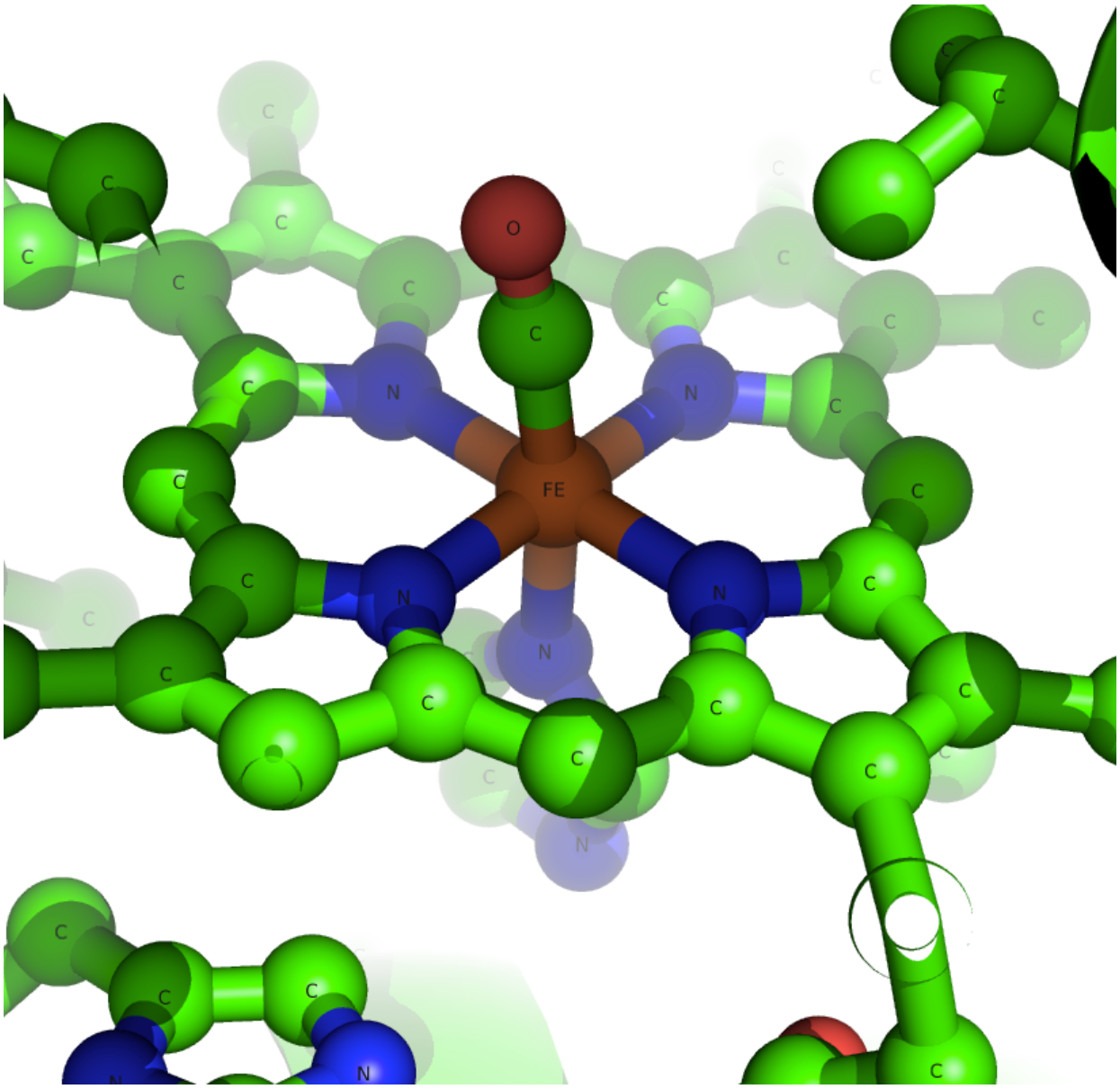
Myoglobin contains a single iron and as such will have a single simple absorption spectrum. The absorption will translate into the diffraction via the anomalous dispersion correction. The correction factor is a complex number characterized by f_1, the real part, and f_2, the imaginary part, where f_2 is closely related to the absorption by Equation 3.33 and is visually similar. With myoglobin the objective is to separate these two orthogonal signals by applying component analysis (PCA) to all the diffractions whilst excluding diffractions that are excessively noisy and do not meet inclusion criteria based on a modified Dixon-Q test.

7.1 The Protein

Myoglobin (Mb) is a small red-tinged, iron-rich protein that is an important part of the respiratory cycle. Myoglobin was selected because it is an easily acquired, iron-containing single metal protein. Sperm whale myoglobin is very easily obtained from Sigma-Aldrich chemical company. There was relatively little legwork involved in securing crystals of very good quality as they come already crystallized in a granular form not too dissimilar to pink sugar. Recombinant P6 Mb solution from Sigma (M-7527) was concentrated to 20 mg/ml and used without further purification. Crystals of Mb were grown at 295K using the hanging-drop vapour-diffusion method [77]. PDB entry 1jw8 [69], it crystallizes with one protein per asymmetric sub unit, six proteins per unit cell with P6 symmetry, therefore within the unit cell there are only six irons out of nearly ten thousand bulk atoms.

Figure 7-1

Myoglobin Active Site6



Active site of met-myoglobin is tucked into a hydrophobic pocket within the protein and has CO on the upper half of the porphyrin ring and a histidine below. The Iron atom is Fe^{3+} oxidation state.

$$F(\mathbf{h},E)=\sum_{C}^{5118} f_C(\mathbf{h})e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_C)}+\sum_{N}^{1374} f_N(\mathbf{h})e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_N)}+\sum_{O}^{3018} f_O(\mathbf{h})e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_O)}+\cdots$$
$$\cdots+\sum_{S}^{36} f_S(\mathbf{h})e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_S)}+\sum_{Fe}^{6} f_{Fe}(\mathbf{h},E)e^{-i2\pi(\mathbf{h}\cdot\mathbf{r}_{Fe})} \quad 7.1$$

These are met-myoglobin proteins and are all in the same ferric state where each iron is a stable Fe^{3+}. The immediate surroundings of the iron are quite spectacular. The iron sits in the centre of a porphyrin ring, two of the iron's d-orbitals are within the plane of the ring and the remaining is perpendicular. The CO binds above the plane of the ring, below the plane the iron binds with histidine 94.

7.2 The Experiment

The experiments detailed in this thesis were conducted remotely [72] at Stanford Synchrotron Radiation Light source using beamline 9-2 and the Canadian Light Source CMCF2. 9-2 is a wiggler beamline with a Rhodium coated flat mirror, a toroidal focusing mirror, a Si-111 double crystal monochromator and a Mar325 detector. To all intents and purpose they are spectroscopic beamlines and some, such as CMCF2 at the CLS, have been built with spectroscopy in mind. Developments in the efficiency of area detector sensitivity means that modern detectors are almost photon counters. The Mar325 has a Detector Quantum Efficiency (DQE) of 0.8 between 8-12KeV, which is to say that they average 10 counts on the CCD for every 12 photons, approximately.

This myoglobin proof-of-principle experiment for a single metal crystal deviates from the simulated (Chapter 6) and ferredoxin (Chapter 8) experiments insomuch as there is only one anomalous atom per protein (the iron). All reflections, excluding outliers, are included in the analysis, therefore the use of *dev* is not required. This allows for the inclusion of many more reflections as all diffractions are in some way biased toward a single iron. The matrix of intensities for this atom still contains two different signals (f_1 and f_2), both from the same iron. Due to the phases of each atom changing with each reflection, the f_1 and f_2 signals have multitudinous orientations with

respect to the bulk atom structure factor. The data is normalized by feature scaling the intensities and then applying abstract factor analysis to recover the shape of the anomalous dispersion part of the atomic scattering factor.

The crystal has a unit cell of just over 320,000 $\mathring{A}^3$ with dimensions 90.38Å, 90.38Å, 45.34Å and angles 90°, 90°, 120°. A total of 9552 atoms, not including Hydrogens. Data were taken at 59 energies from 6910eV to 7300eV. The myoglobin crystals diffracted well, the crystal orientation was easy to obtain and maps were easily generated. For the DS run, eight 1.5° oscillations were collected using a specialized script written by Jinhu Song and Aina Cohen[1] at SSRL that took advantage of blu-ice/ DCSS scripting engine [70, 71]. At the time of the experiment the ability to easily process the data took a much more important role than the total intensity of each reflection, which does not take into account the signal-to-noise ratio that would later consume much of the analysis. The average wedge generated 9030 reflections, however very high-resolution spots do not survive across the entire spectrum. There were a total of 10865 recorded reflections across the entire spectrum and 7890 of them were rejected for either falling out of range over the spectrum or from not having values assigned to them across all energies by the processing software.

Only 2975 reflections were contiguous over the spectrum with resolutions from 39Å to 2.13Å, despite the data collecting successfully out to 1.6Å. Of the 2975 reflections, 175 were rejected for having a Dixon Q-test value greater than 0.6 and another 1412 were rejected as outliers. Outliers, in this instance, were determined as reflections with values outside the interval [0.5, 1.5] of their mean subtracted intensity. This left 1388 reflections that proceeded to the PCA module. The matrix $\boldsymbol{M}$ of 1388 reflections were processed with the PCA module. Although this number is quite large, the results from the analysis shows relatively high levels of noise, which is the consequence of not doing multiple runs.

[1] Staff Scientist/Leader, operations and development group for structural molecular biology at SSRL.

Figure 7-2

Screen Output from *DeskTools* running Myoglobin

```
DS One Target Atom Run
UC Vol from PDB: 320749
[90.38, 90.38, 45.34, 90.0, 90.0, 120.0]
[[ 0.011064  0.006388  0.      ]
 [ 0.        0.012776  0.      ]
 [ 0.        0.        0.022056]]
Symmetry recovered from REMARKS 290 of 1jw8.pdb
6 symmetry operations
['X', 'Y', 'Z']
['-Y', 'X-Y', 'Z']
['-X+Y', '-X', 'Z']
['-X', '-Y', 'Z']
['Y', '-X+Y', 'Z']
['X-Y', 'X', 'Z']
ASU from a PDB file: 1jw8.pdb
[853, 1, 229, 503, 6] 1592
['C', 'N', 'O', 'S', 'Fe']
[853, 229, 503, 6, 1]
Target Atom Positions
[[ 23.921   9.18   -5.724]
 [ 26.626  30.565  -2.691]
 [ 31.178  18.341  28.294]
 [ 31.807  16.869   0.967]
 [ 32.956  13.049  22.584]
 [ 45.19   26.09   -0.627]]
[[ 34.783   4.14   10.249]]
5
[[ 8169. -4084.     0.]
 [-4084.  8169.     0.]
 [    0.     0.  2056.]]
Calculated Unit Cell Volume 320742
Spectrum
|    |    |    |    |    | | | | | | ||||||||||||||||||||||||||||||||||||||||| | | |  |    |    |     |       |       |
59 Total:  6910.0 ---> 7300.0
Retrieving pickle file data_dictionary: mb_1-data-dict.pickle
bibliography/ownwork/2013_paper_data/Myoglobin/1b/
Resolution Range 39 --> 1.6
Number of HKLs is 10865
Number of hkls in data_dict                    : 10865
DATA hkls rejected for outside bracket 0.5 1.5 : 1412
DATA hkls rejected for dixonQ too high is 0.6  : 175
DATA hkls kept                                 : 1388
Calculating and creating the Priniple Componants file
Myoglobin/1b/ pca1_0.6_0.5_1.5.pickle
```

This is the verbose output from DeskTools.py running a myoglobin DS simulated experiment, wherein a number of checks are conducted. The output shown here is for the successful run discussed in this chapter.

7.3 Collection History

Over the period of a year, a dozen DS runs were conducted on myoglobin. The myoglobin experiments started in mid-2010 at SSRL in Stanford, they were the first datasets taken and at the time multiple runs on separate sections of the crystals were not being conducted. The prevailing theory at the time was that using multi-crystal runs would yield more reflection and better quality data. A broader oscillation of 1.5Å was also used, as was an inferior region file for the energies.

The concept of applying PCA had yet to enter the experimentalists' lexicon and an equal amount of pre-edge spectra was taken as post edge. Edge position was also considered the most valuable part of the experiment and as such the post-edge oscillations barely cover classic Near-Edge (XANES) spectral regions. Though multiple florescence spectra were taken, multiples of a single position were not conducted, EXAFS was also not conducted. Six complete datasets were viable for processing, some were taken prior to the introduction of ‘top-off’ injection into the synchrotron ring and were rejected for having massive discontinuities due to data collection over a ‘fill cycle'. A few collections failed due to software glitches, filename mishandling, and normal errors such as pin slipping. These common beamline problems were expertly addressed and handled by the collaborator at SSRL (Aina Cohen). By late 2011, data was successfully collected using the original regime where multiple runs on the same wedge were *not* conducted.

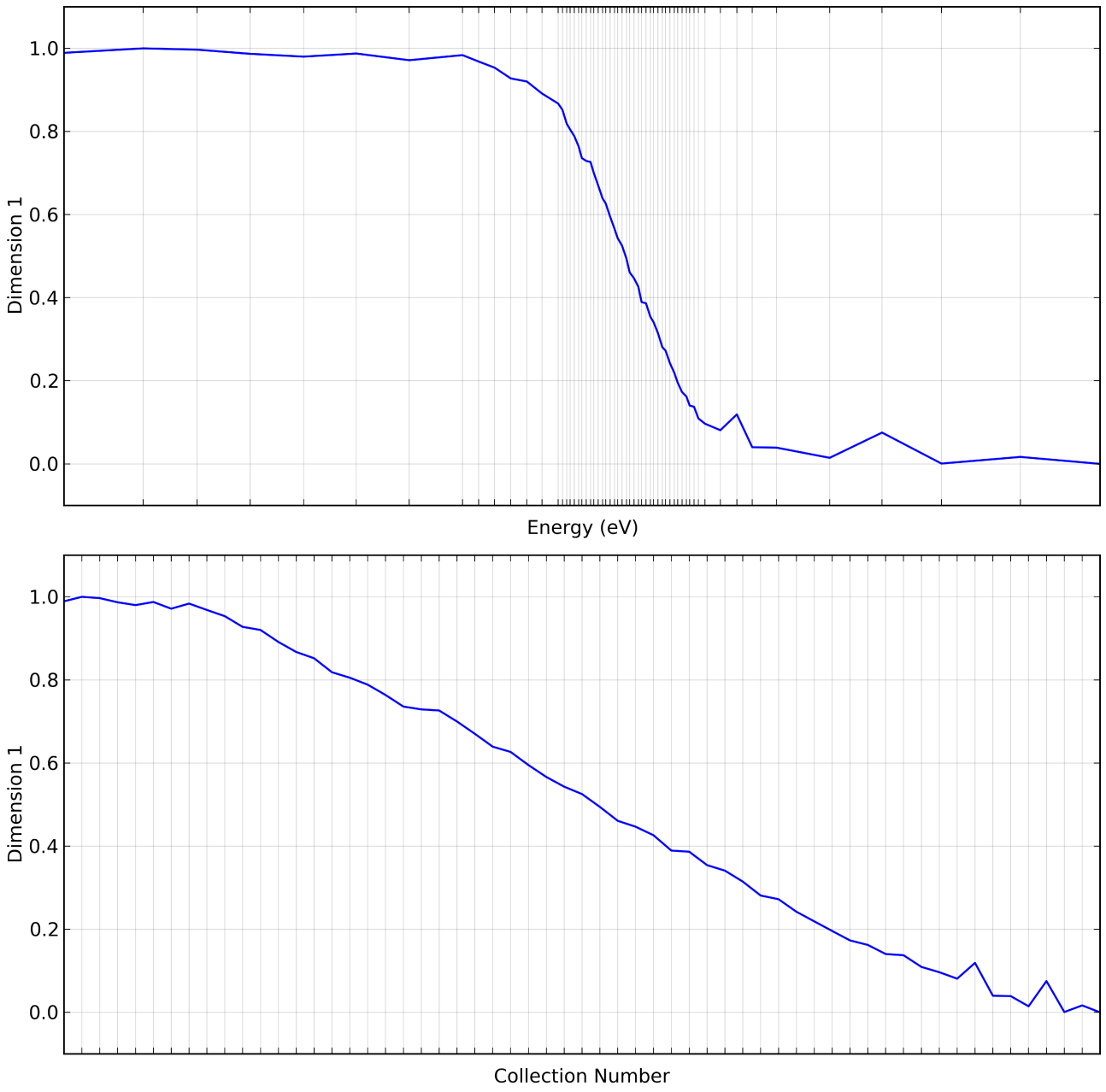
A reliable method of processing the data was developed to handle each of the 59 energies in an identical manner. Online backup was located at beamline 4.2.2 at the Advanced Light Source in Berkeley California. This space allowed for the production of scripts to automate the processing software, d*TREK [65]. The results for the myoglobin stem from early scripts in which the data was processed by holding the crystal orientation and allowing the detector distance to vary. This simulates the energy change, an adaptation of processing that has its origins in Multi-wavelength Anomalous Dispersion (MAD) crystallography.

7.4 The Dimension 1 Anomaly

When PCA is performed on matrix $\boldsymbol{M}$ in the manner addressed in Chapter 5, the largest component by far is a smooth inflection of the first dimension shown in Figure 7-3 (top). Dimension 1 was considered a major contribution to the anomalous dispersion. However, this high scoring contribution was identical and pervasive throughout all data collection, independent of all other factors. It occurs across all subsets of matrix $\boldsymbol{M}$ including subsets that have low affinity to the heavy atoms. This led to the conclusion that dimension 1 is independent of the dispersion. Due to the absence of signs in PCA and the lack of a theoretical framework, the component was arbitrarily biased with a negative sign (since this phenomena is independent of heavy atoms, a detrimental effect is more probable). Figure 7-3 plots dimension 1 against energy (top) as well as linearly in time (bottom). Plotting the effect negatively and against the number of exposures (time equivalent), it appears that the crystal holds itself together before being overcome by the effect. As their is no correct ordinate value, it is impossible to tell how steeply this effects the diffraction. It is supposed that dimension 1 is recording the decay of the crystal's ability to diffract through radiation damage or another process such as an increase in overall temperature buildup. Experiments on crystals without heavy atoms and at differently spaced timing may well shed light on this mystery, however time was not afforded to this. It will need to be fully characterized in the future if DS is going to be used regularly by third party experimenters, however at this point it is a known unknown and analysis moved to discovering the anomalous dispersion features of PCA components from matrices that maximize that part of the signal. It is also worthy to note that dimension 1 does not appear to be susceptible to the effects of the anomalous dispersion, which means that the PCA module is able to separate variances stemming from different effects.

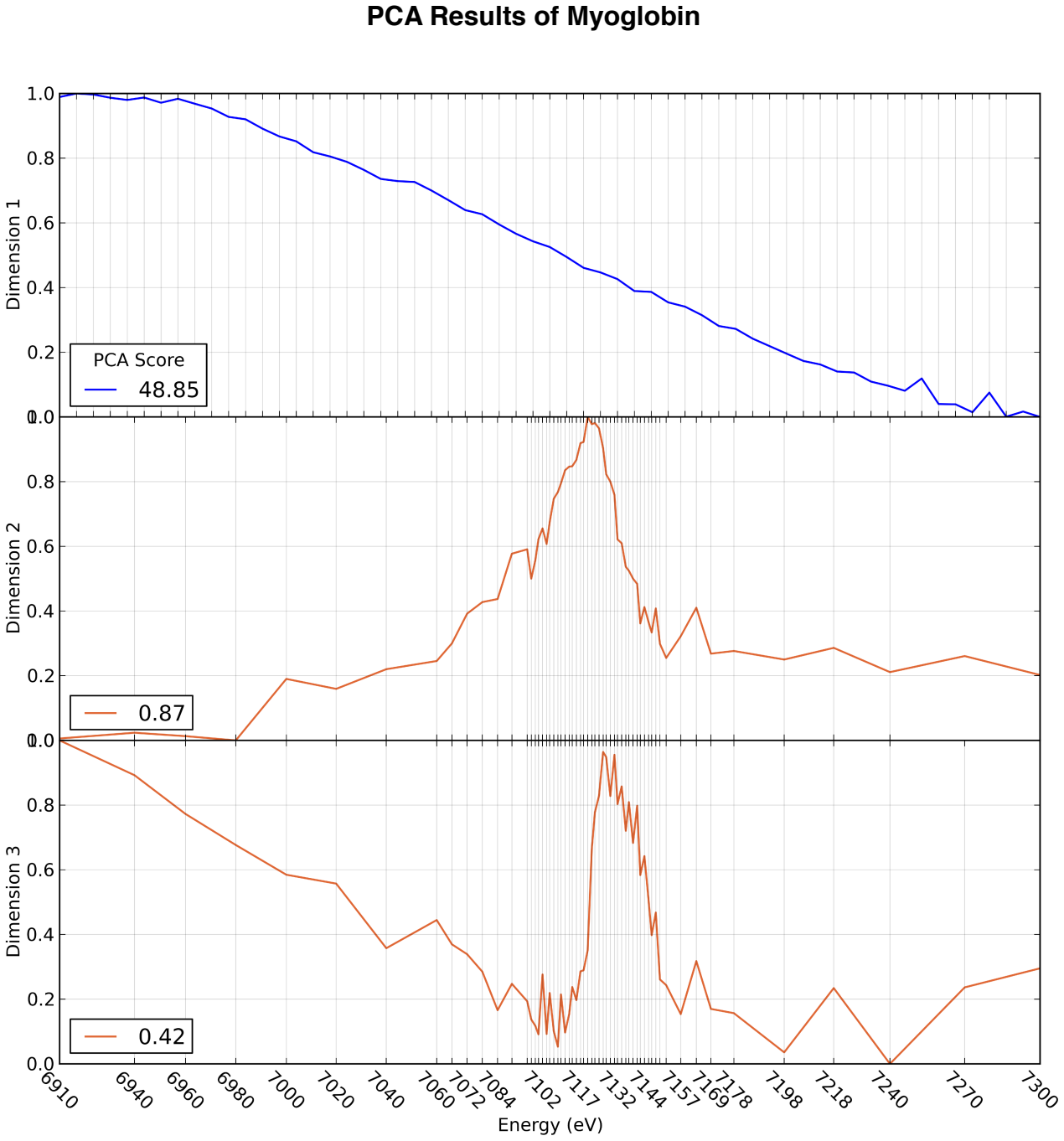
Figure 7-3

The Dimension 1 Anomaly vs Energy and Time



The first dimension output from PCA of diffractions from myoglobin. *(Top)* A plot of the component vs the energy at which it was collected. *(Bottom)* Plot of the same component given against the collection number (time equivalent).

Figure 7-4



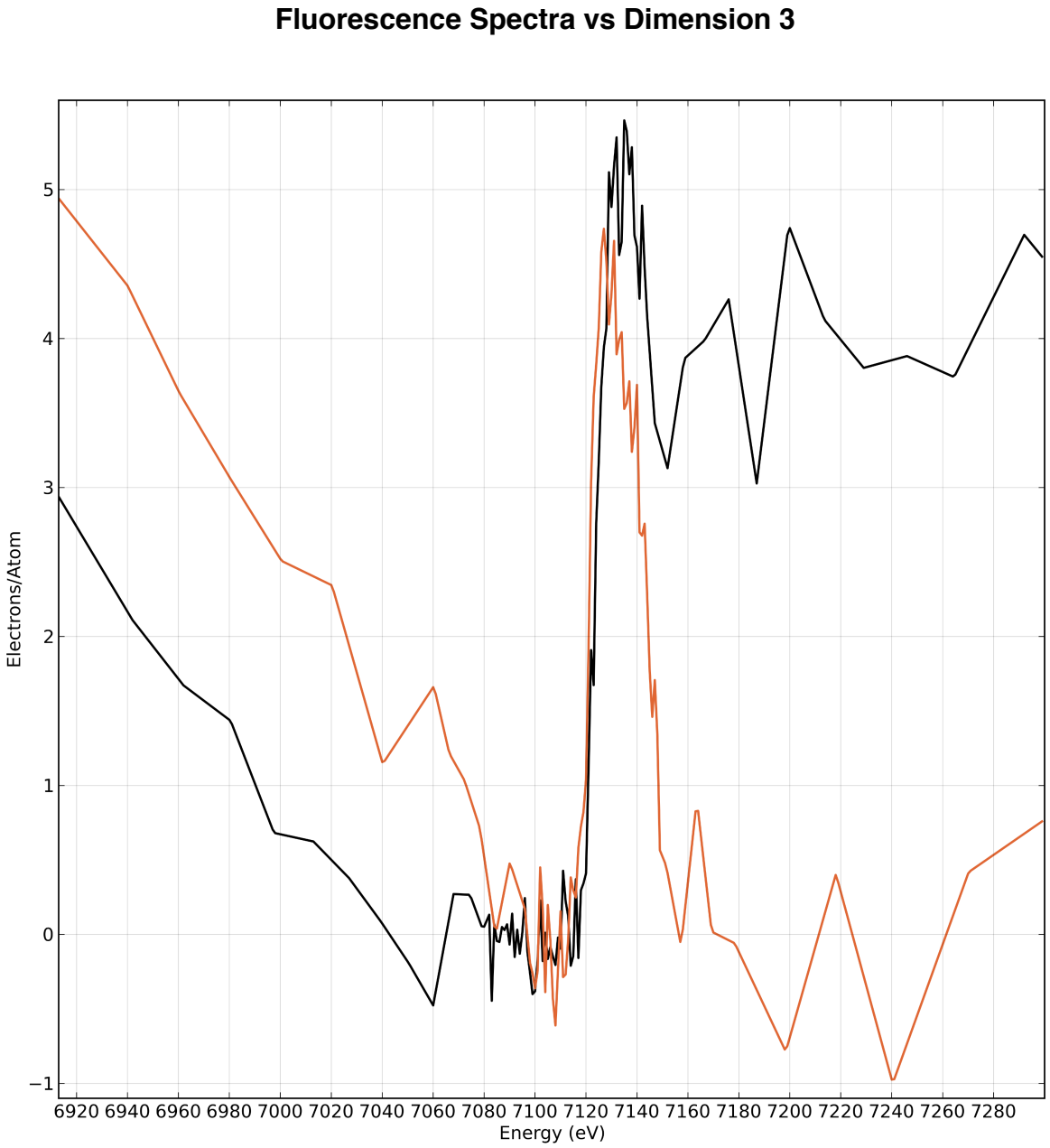
The top three principle components from a myoglobin DS run. Dimension 1 is plotted linearly with time of exposure and dimension 2 and 3 is plotted against energy. Each has been normalized to the to have a maximum value of 1 and a minimum of 0. Dimension 1 is given a negative slope to indicate a decline in diffraction strength, Dimension 2 is a positive cusp and Dimension 3 is a positive step to reflect the real and imaginary parts of the anomalous dispersion.

7.5 Results of Myoglobin PCA.

The myoglobin DS experiments are the first that have been conducted as far as the literature on the subject goes. The dimension 1 anomaly demonstrates that this is uncharted territory. You can see from Figure 7-4 that the next two components detected by the PCA module have a striking similarity to a cusp and step in shape. This is highly reminiscent of the dispersion relation. The 2:1 ratio of their relative scores is as expected from the size of the anomalous dispersion relation where the real part, f_1 has values approximately double those of f_2. The noise in dimension 2 is also visibly less than that of dimension 3, and even though the successive dimensions have only slightly lower scores, they rapidly devolve into shapeless noise. With a mixed two-iron or multi-iron protein, the successive dimension after dimension 3 may include signal from other absorbers in the iron region, however this is not the case with myoglobin. PCA looks to maximize variance in the covariance matrix and the intensities for reflections has been feature scaled, which introduces a skewing of the signal giving the appearance of a much steeper background function. Absorption spectra, that are related to f_2 by Equation 3.33, require background subtraction to flatten the pre- and post-edge due to the approximate E^{-2} decrease in absorption, this slight decrease over this region should be dramatically enhanced by feature scaling, which is done here. It is hypothesized that the feature scaling is what gives dimension 3 more of a symmetric triangle wave background rather than an antisymmetric square-wave, step-like function.

The object of retrieving the absorption profile form the anomalous signal of diffraction leads naturally to comparing dimension 3 with an absorption spectra from the same sample. No direct absorption spectroscopy was performed on myoglobin at the time of these experiments, nor are they available for this paper through other means. Fluorescence spectra were taken at the time of the experiment, these are much cruder than those that are taken is a normal XAS experiment. Although not ideal, they are an excellent indication of the absorption profile. Figure 7-5 shows the comparison of dimension 3 with the fluorescence spectra. The fluorescence, in black, has a nice clear

Figure 7-5



The comparison of a single fluorescence spectra (black) and dimension 3 (orange) from PCA in the DS regime performed on the same myoglobin crystal. The fluorescence spectra is simply the raw counts from the fluorescence spectra scaled to the signal counts in the beamline. Dimension 3 has been scaled for comparison.

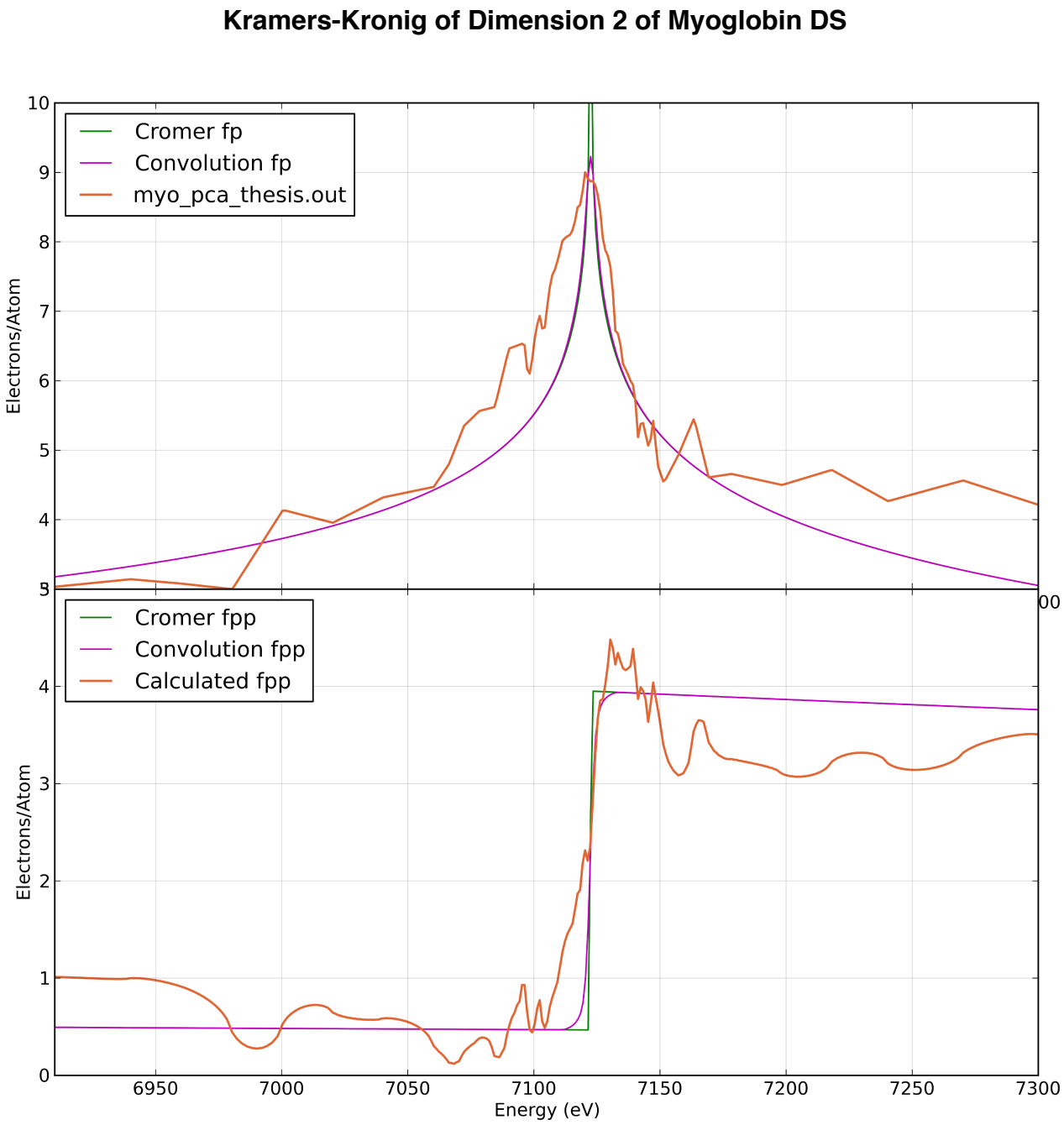
step-like behaviour, which dimension 3 (orange) does not. Ignoring the high level of noise in both, the lack of structure is evident in the pre-edge and there is a strong jump in both values in the near-edge. The exact position of the edge is shifted in dimension 3 from the diffraction. These sorts of dissimilarities suggest that dimension 2 could be a better contender for characterizing the anomalous dispersion.

7.6 Fourier Transforming of Dimension 2

Although the comparison of dimension 3 gives a close approximation to the absorption spectra, a more rigorous estimation should come from the Kramers-Kronig transform of dimension 2 because it originates from f_1 which should have approximately twice the signal size as f_2. The dimension 2 component has a higher score and visibly less noise. In order to calculate the Kramer-Kronig transform of dimension 2 the fftkk.py module from Appendix III was executed. Theoretical Cromer-Liberman values (black) are generated using a FORTRAN subroutine written by Prof. Graham George, these were convolved with a Voigt function which simulates broadening (magenta) in the energy due to the Darwin width of the monochromator crystal and lifetime broadening of the X-rays. The input values for f_1 is the scaled PCA component from dimension 2. The resulting output, the f_2 approximation, is much less noisy than the first approximation above using dimension 3.

Figure 7-7 shows the final results of the myoglobin DS experiment. PCA applied to a subset of diffracted x-rays from a large myoglobin macromolecule crystal, repeatedly exposed over the spectral range of an iron absorption edge, can detect the small variations. Although collection techniques have improved since the first year of these types of experiments, it is enlightening to see that the small, anomalous dispersion signal buried deep under noisy reflections is retrieved despite the phase of the anomalous dispersion taking a myriad of values.

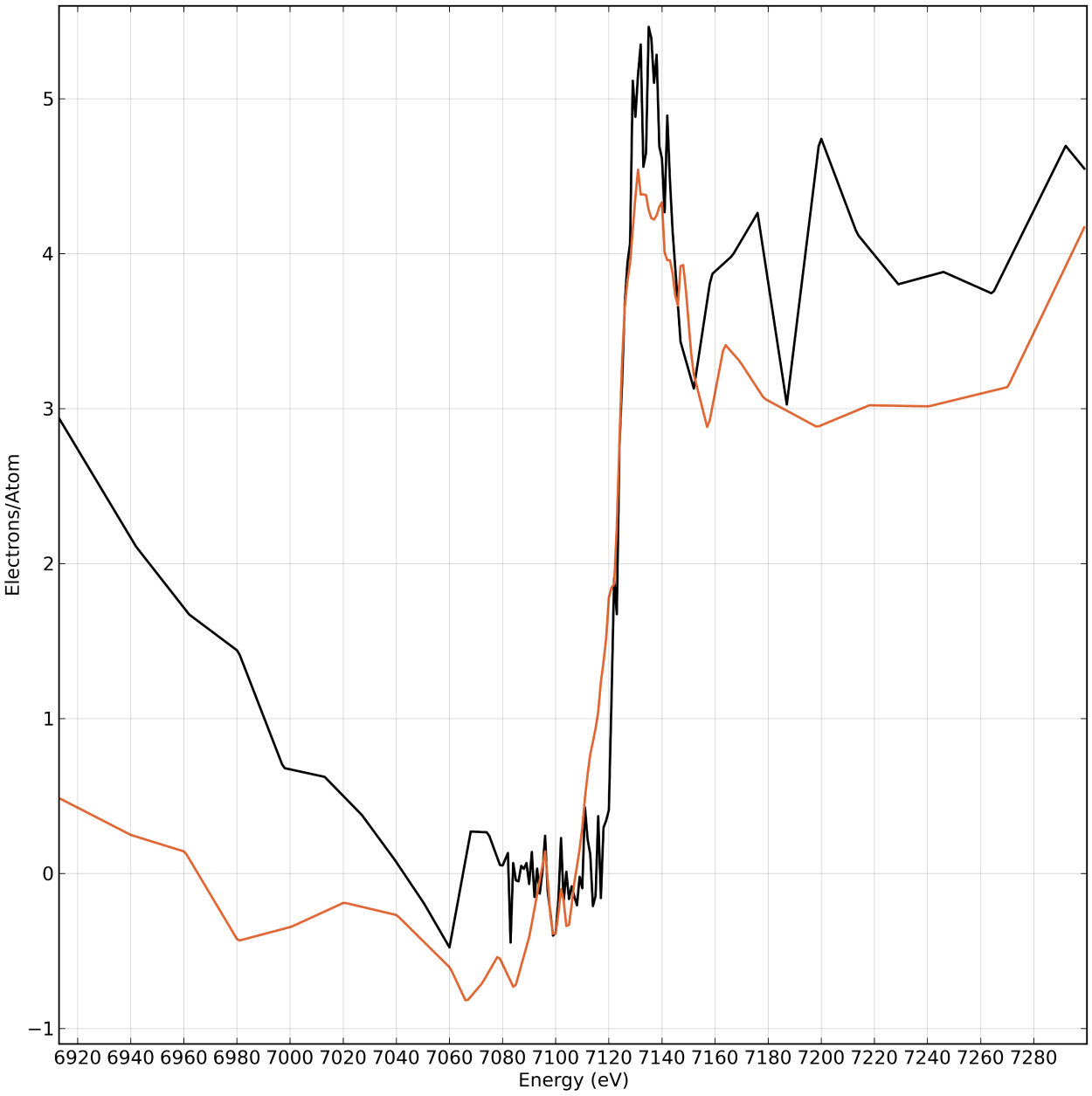
Figure 7-6



The underlying Kramers-Kronig relationship is seen by plotting the Cromer-Mann values along with the results of conducting a Fast Fourier Transform (FFT) on input spectra. Dimension 2 of the myoglobin PCA is scaled and fitted to the f_1 and the calculated f_2 after using the fftkk.py module (Appendix III).

Figure 7-7

Fluorescence Spectra vs Kramers-Kronig of Dimension 2



The comparison of a single fluorescence spectra (black) and the Kramers-Kronig of dimension 2 (orange) from PCA in the DS regime, both performed on the same myoglobin crystal. The fluorescence spectra is simply the raw counts from the fluorescence spectra scaled to the signal counts in the beamline. The KK of dimension 2 has been scaled for comparison.

7.7 Myoglobin Conclusion

This is a vastly improved result over the comparison with dimension 3 (Figure 7-5) with respect to edge position and step-likeness. As fluorescence is a better understood area of science, the features of its spectrum are also better understood. The combination of Diffraction, PCA and Kramers-Kronig returns something very akin to absorption. The spectra share enough traits to be compared to one another and are different enough to warrant skepticism. The single iron myoglobin DS experiment brings the research one step closer to its ultimate goal: retrieving separate anomalous dispersion spectra from a multi-metal macromolecular crystal.

The myoglobin experiments were not conducted with an ideal methodology. Currently there is no ideal collection strategy, however myoglobin does demonstrate that the PCA of noisy diffraction can return lower-noise signals that are directly related to the absorption properties of the target atoms within the crystal. This prior collection strategy, with the lack of a well characterized absorption profile, do not detract from the affirmation of the underlying theory: that DS can be performed on a 3rd generation MX beamline with a few software tools (to run the experiment, automate the processing, analyze the data, and perform a transform).

CHAPTER 8

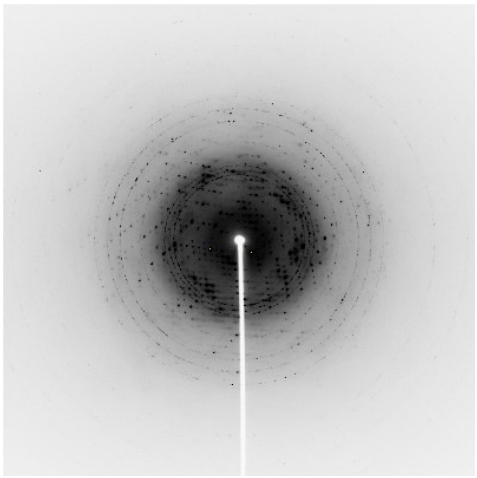
FERREDOXIN

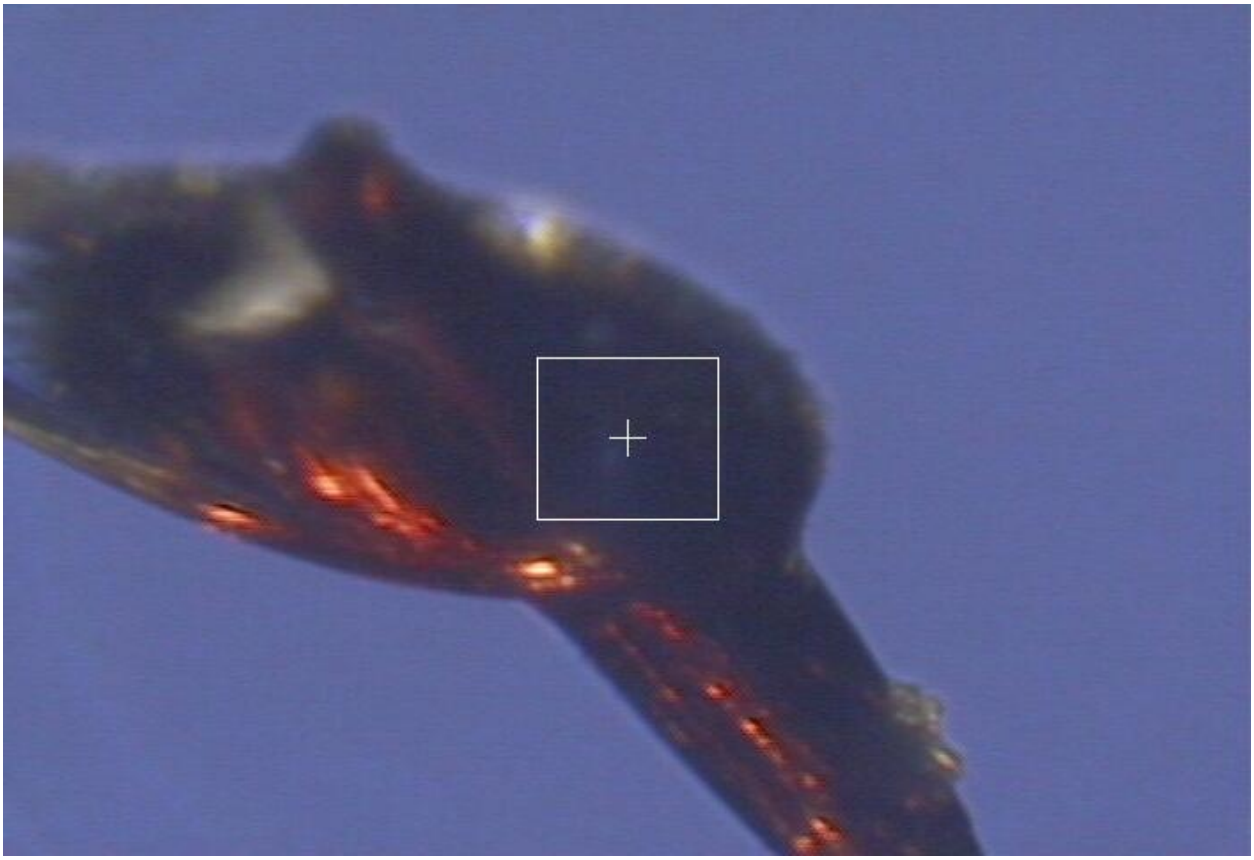
The objective with the ferredoxin is similar to that of myoglobin; extract the absorption-like spectra from diffraction. However, ferredoxin contains two irons that are in different electronic and physical conformations which will manifest as marginally different profiles in the spectral region under which the experiment is performed. The goal is to identify diffractions that prefer one iron over the other (and vice versa) and treat each of these subsets of diffractions separately for analysis. In analysing a set of diffractions that bias a single irons atomic location an attempt is made to identify its oxidation state by comparison to model spectra for irons known to be in a similar configuration.

Ferredoxin was investigated in two ways in this paper to demonstrate that these experiments can be conducted at modern beamlines. First, an entirely simulated experiment was conducted in which the oxidation states and neighbourhood conformations for each iron were calculated using energy dependent atomic form factors. The resulting diffractions were analyzed to see whether under perfect conditions; *dev* and PCA could deconvolute and, therefore, separate the anomalous dispersion for those selected atoms. Secondly, an identical experiment with a (similar) real crystal and resulting diffraction was conducted and is presented here. The experiment in this chapter gives evidence in the form of more detailed absorption profiles from site separated atoms in large unit cells, demonstrating that DS can be executed on many other structures. The additional goal of this thesis is to supply a new tool to crystallographers that relies on years of experience and experiments previously conducted by absorption spectroscopists.

Figure 8-1

Images of Ferredoxin Crystal I2 Collection Experiment





A screen shot of a single diffraction image and a photograph of the ferredoxin crystal I2 with the in-line microscope at beamline 9-2.

8.1 The Protein

The ferredoxin protein [73], PDB entry *1m2a*, is more complicated than the myoglobin inasmuch as it contains two target atoms in different oxidation states, conformations and locations. Many types of ferredoxin have been solved in the PDB, and an investigation of them shows similarities in the conformations of the inner and outer irons (see Appendix II). In this section, the diffraction spectra are separated by atom-label; the inner, ferric Fe1 from outer, ferrous, Fe2. The two irons, bridged by two sulphurs (see Figure 6-1), are 2.7Å from each other; each iron has 4 sulphurs coordinated with it. The crystal has a unit cell of just over 175,000Å^3 with dimensions 67.40Å, 59.02Å, 46.61Å and angles 90°, 110°, 90°. A total of 6756 atoms, not including Hydrogens. Data were taken at 67 energies from 6910eV to 7300eV in 1° oscillations.

From earlier experiments on myoglobin, it became clear that the signal to noise would need to be improved because the PCA module was still finding the energy-dependent signal from individual diffractions exceptionally noisy. By comparing the simulated ferredoxin data quality (less than two hundred spectra) with that of the myoglobin (over one thousand), it was decided that multiple runs on the same wedge would be needed. Simply exposing the crystal the same wedge for longer periods of time is limited by the dynamic range of the detector, lower resolution diffractions become overloaded and interfere with processing the crystal. Additionally, repeating the same wedge at the same energy also increased deterioration of the diffraction as wedge is already being exposed 67 times (once at each energy). The crystals were grown by Eva-Maria Roth under the supervision of Dr. Thomas Spatzal[1]. This is when crystal translation and multiple runs became the new collection regime. As can be seen in Figure 8-1 (*bottom),* the crystals grew as long needles which greatly simplified translating and re-exposing the crystal at a fresh location. A reduction of the oscillation angle from 1.5° to 1.0° also helped the software lock in the locations of the reflections with respect to each other. This is important as the crystal orientation stays the same

[1] Einsle Lab, University of Freiburg

as the diffractions move slightly, in the images, due to the change in energy. 'Fine-phi slicing' (taking sub 1° oscillations) would continue to improve processing, profile fitting and signal to noise; however, the experiment is time consuming so a decision was made to stop at 1°.

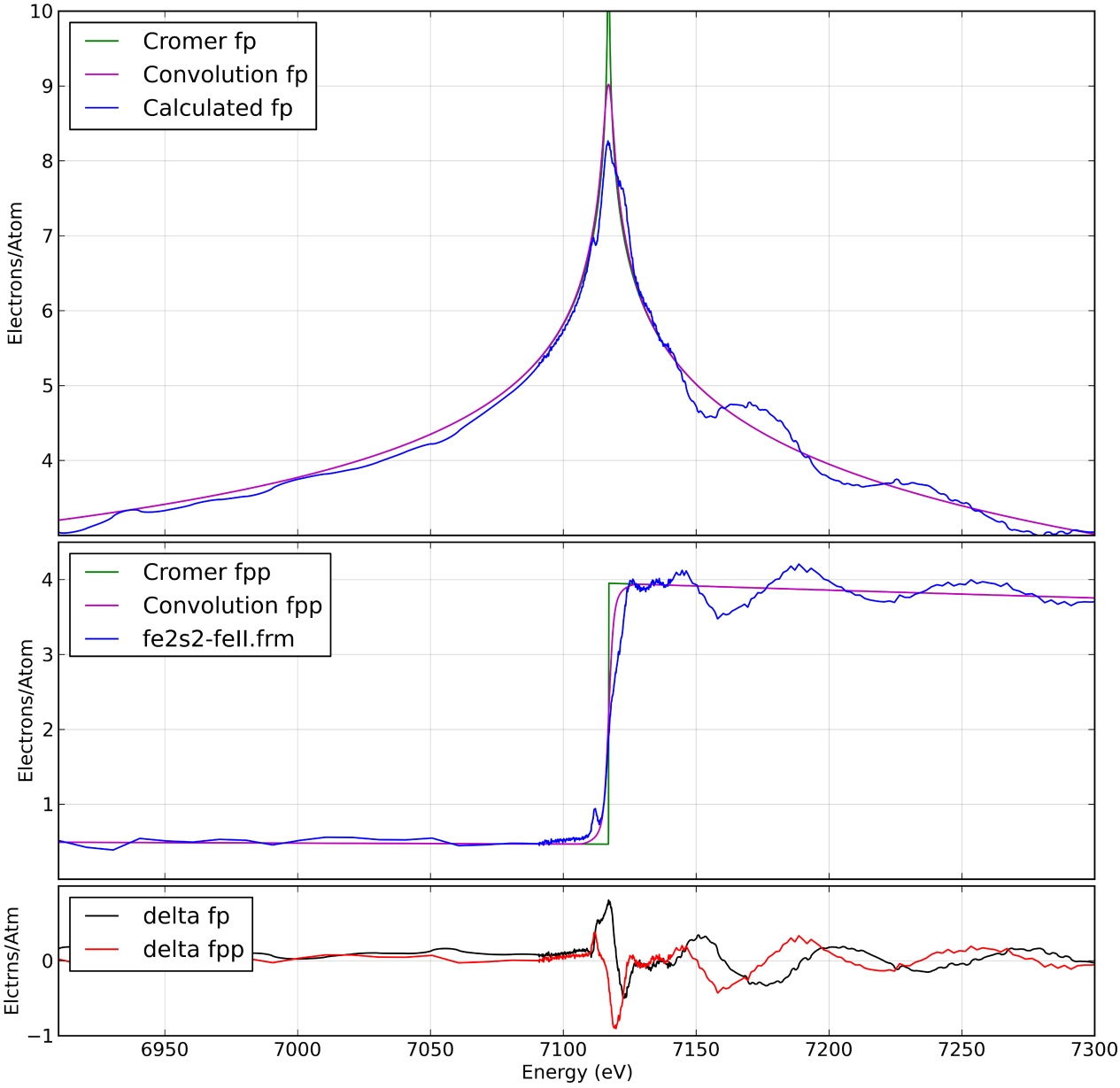
8.2 The Comparison Spectra

In the simulated ferredoxin experiment, the anomalous dispersion was generated using FEFF for the absorption spectra of one crystal with the near-edge stitched on from an unrelated iron absorption experiment. With the real ferredoxin, absorption spectra had already been measured by the George/Pickering group on an oxidized and reduced Aneabaena, and the analysis of the spectra created in this experiment will be compared to that. Absorption spectra are taken on the sample as a whole, therefore, the 'reduced' aneabaena start as 50/50 mix of reduced and oxidized iron. In order to separate the reduced irons half of Fe^{3+} is subtracted from the mixed spectra to generate a Fe^{2+} spectra. This new spectra for Fe^{2+} was determined this way using the EXAFSPAK [66] *backsub* module. The quality of the Fe^{2+} spectra suffers a little from this, but it is still an improvement over creating the spectra from PDB files (Appendix II). These absorption spectra are also used in creating theoretical spectra utilized by *dev* to see which sets of intensities bias each iron.

In order to calculate the structure factors for these irons, the absorption spectra were processed with the fftkk.py software so that the real, f_1, part of the anomalous dispersion can be calculated. The results for the Kramers-Kronig are given in Figures 8.-2 and 8-3. A comparison of the two anomalous dispersions, using the cylinder projection, are in Figure 8.4. The dispersion relations can be used as the energy dependent part of the structure factors for simulated diffraction calculations in *dev* as well as comparing them to the site separated spectra principle components from the DS experiment.

Figure 8-2

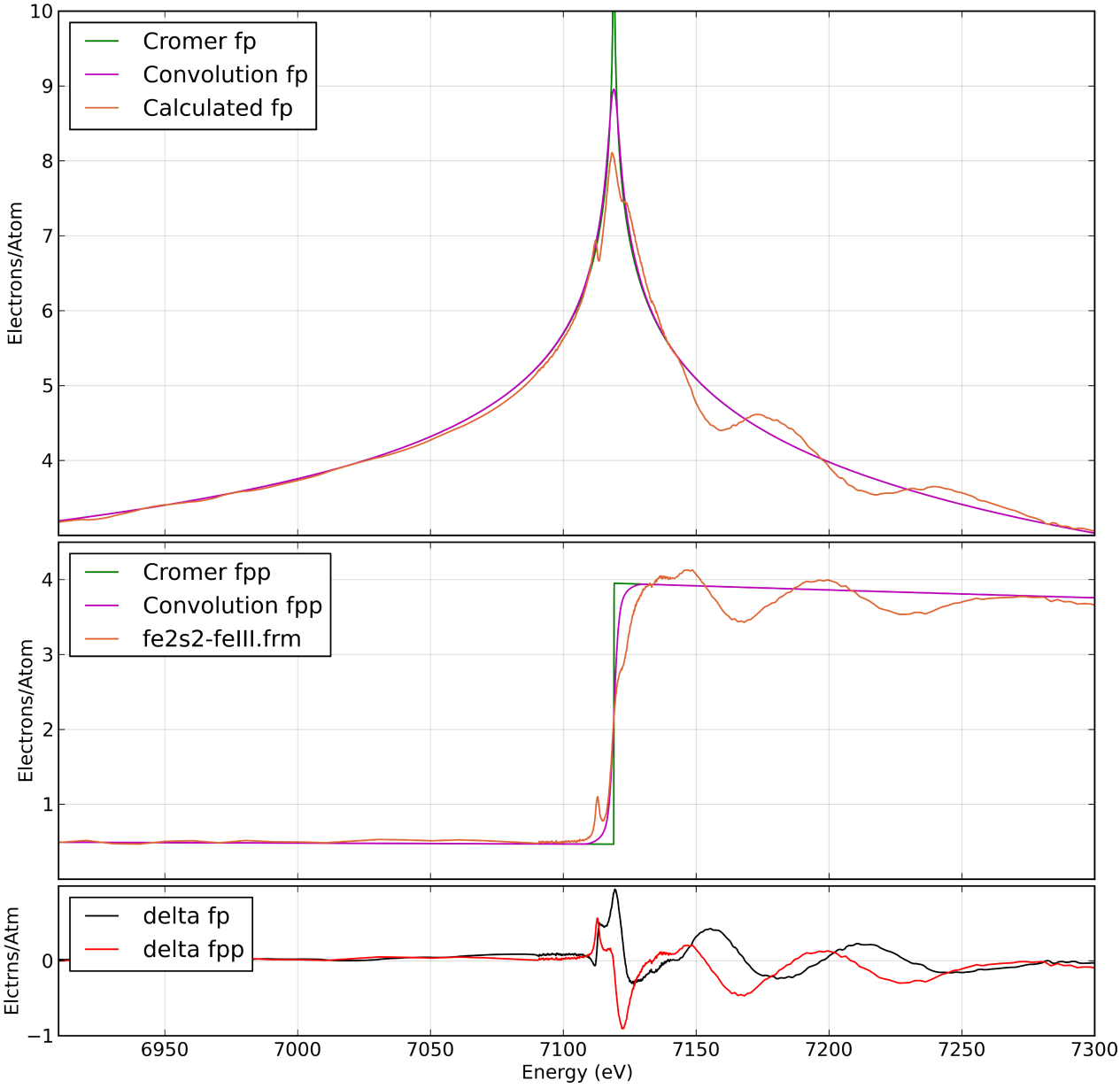
Kramers-Kronig Transform of Ferredoxin Reduced Iron From Absorption Experiment



Green: Cromer-Liberman (CL) values for reduced Iron shifted to coincide with the inflection point of the absorption. Magenta: CL values convoluted with a Voigt to approximate X-ray lifetime broadening and monochromator Darwin width. *(Top)* Kramers-Kronig of the absorption spectrum from ferredoxin reduced iron.: f_1. *(Middle)* The absorption spectrum: f_2. *(Bottom)* The difference between CL-Voigt and f_1 or f_2.

Figure 8-3

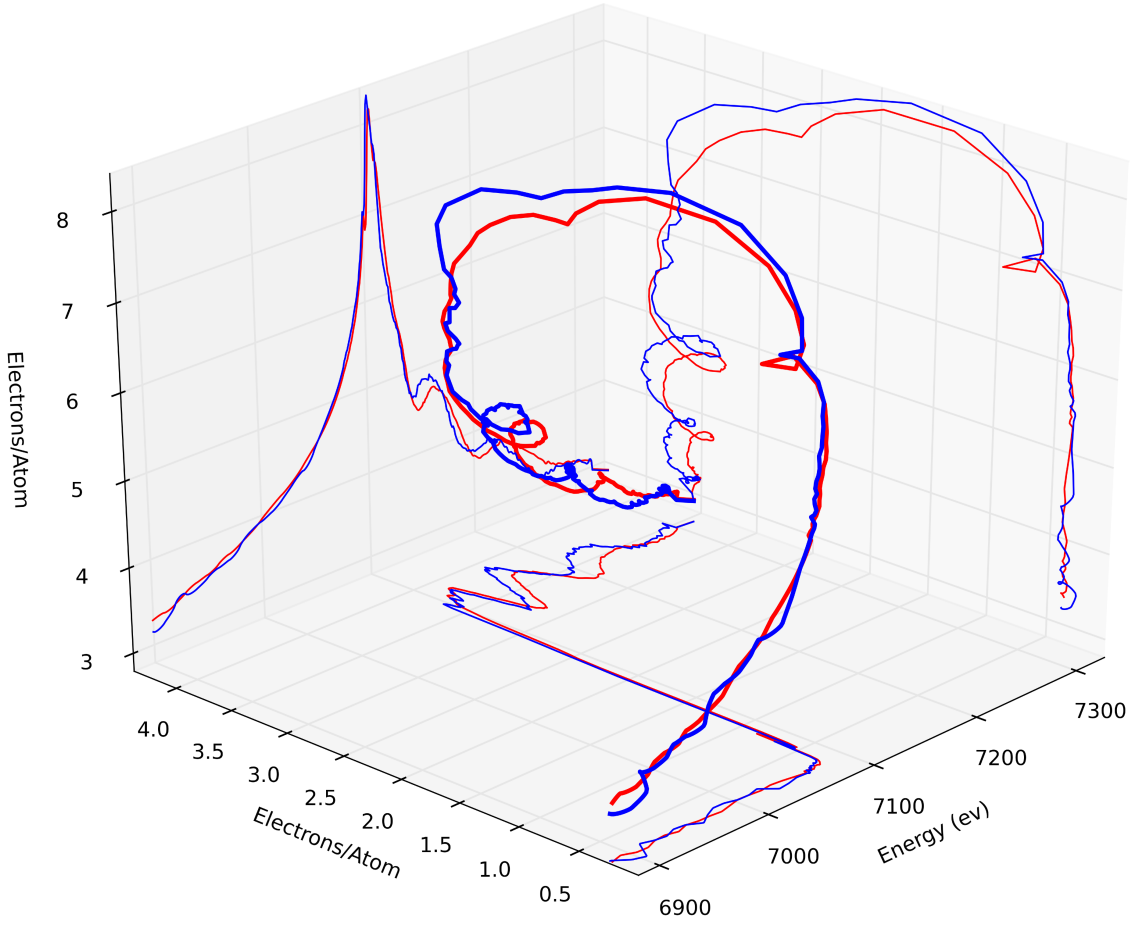
Kramers-Kronig Transform of Ferredoxin Oxidized Iron From Absorption Experiment



Green: Cromer-Liberman (CL) values for oxidized Iron shifted to coincide with the inflection point of the absorption. Magenta: CL values convoluted with a Voigt to approximate X-ray lifetime broadening and monochromator Darwin width. *(Top)* Kramers-Kronig of the absorption spectrum from ferredoxin reduced iron.: f_1. *(Middle)* The absorption spectrum: f_1. *(Bottom)* The difference between CL-Voigt and f_1 or f_2.

Figure 8-4

Cylinder Spectra of the Dispersion Relation from EXAFS Absorption Experiment: Spectra of Reduced and Oxidized Ferredoxin Irons.



(Centre) Cylinder spectra from Figure 8-2 and Figure 8-3. Reduced (blue) and Oxidized (red). *(Ground)* Absorption: f_2. *(Back Left)* Kramers-Kronig of the absorption: f_1. *(Back Right)* Argand projection.

8.3 The Data

8.3.1 Collection

There were only a handful of crystals that were viable. Due to the low quality of the diffraction from these crystals; three sweeps[1] were performed on different sections along the crystal. This was achieved by laterally translating the crystal (parallel to the rotation axis) to expose a fresh section The newer collection methodology improved the results dramatically despite the quality with which these crystals diffracted. The ferredoxin data used in the final study comes from crystal *I2,* and it diffracted[2] to 1.95Å. The beam was a 50x170µm[2] rectangle, and dose mode was switched off. Data were taken in 1° oscillations, and 13° wedges at 67 energy points: a total of 871 (13*67) images per sweep. Each image was exposed for 1 second; a total time of 93 minutes elapsed to collect all three sweeps. Within each 31-minute sweep, the crystal was exposed for 14.5 minutes, which is a 17% improvement in time over myoglobin. The gain in speed of collection is attributed to streamlining software at SSRL: resetting the goniometer concurrently while image read-out was being performed. These incremental improvements add up, saving beamtime for other uses, making more experiments of the same type possible as well as making these experiments more attractive to third parties.

8.3.2 Processing

A total 201 (3*67) wedges were processed, 67 wedges per sweep. A python module was written (*process-w-xds.py*) to automatically process each wedge at each energy using the crystal orientation from the solution dataset that was solved for the protein structure. A 'master' processing file (xds.inp) was created by processing the crystal at a variety of energies across the spectra using the solution orientation and unit

---

[1] A sweep is a term used to describe collection across the spectrum. In the context this DS run it is a collection of a wedge at each energy from 6910-7345eV.

[2] At 7345eV the resolution was 1.86Å but all gains in resolution are lost when calculating the dense matrix.

cell dimension from the solved structure run (see Figure 8-5 for the master file). The program also assigned the number of processing cores to be used in parallel as well as setting the correct wavelength and filenames. It took approximately 2 hours to process each sweep on a 2.4Ghz Intel Core i5 MacBook Pro (2010). Two keywords options were used that differ from normal processing: 1) `FRIEDEL'S_LAW=FALSE` was set so no compensation for anomalous signal was performed, and 2) `SPACE_GROUP_NUMBER=1` was set to force all data to be processed in P1 symmetry so that no compensation was made for symmetry-related diffractions.

Another program (*multiple_xscale.py*) was written to automate the scaling between the 3 datasets at the same energy. This program created the XSCALE.inp filethat is run by the XDS software (Figure 8-6) to merge the three datasets at each energy and also used the `FRIEDEL'S_LAW=FALSE` keyword option.

The result of collecting, processing and scaling was 67 *hkl* files, one for each energy. Each *hkl* file contains a list of *hkl*s, their observed intensities and sigma values, which are an estimate of the standard deviation of the profile fitting. Once the processed data were merged and scaled by XSCALE.inp, the resulting *hkl* files were ready for analysis by the *DeskTools* program. A large amount of information is supplied to the investigator as the *DeskTools* performs its operations (Figure 8-7). The readout from the program gives a number of useful pieces of information: the name of the solved crystal, which atoms it is using for target atoms and their assigned oxidation states, the number and types of bulk atoms, the resolution, the Cromer-Mann coefficients, unit cell volume according to the PDB file versus the one calculated, and, lastly, the component scores and quantity of *hkl*s that the threshold produces.

Figure 8-5

Master_XDS.inp File for Ferredoxin I2

```
!-------------------xds.inp-----------------
JOB= ALL
MAXIMUM_NUMBER_OF_PROCESSORS= 3
MAXIMUM_NUMBER_OF_JOBS= 2
!-------------------Dataset parameters--------
X-RAY_WAVELENGTH= 1.77
DETECTOR_DISTANCE= 124.40
STARTING_ANGLE=0.0
OSCILLATION_RANGE= 1.0000
SPACE_GROUP_NUMBER=1
UNIT_CELL_CONSTANTS=67.200 59.800 47.200 90.00 110.30 90.00
UNIT_CELL_A-AXIS=     8.932   -62.806    22.513
UNIT_CELL_B-AXIS=    11.441   -18.064   -54.883
UNIT_CELL_C-AXIS=    40.519    22.889     0.944
NAME_TEMPLATE_OF_DATA_FRAMES=/Volumes/DATA/SSRL/data/Jan14_2012/I2/
I2_fero_DAFS_1_6.9100_?????.mccd
!NAME_TEMPLATE_OF_DATA_FRAMES=/Volumes/DATA/SSRL/data/Jan14_2012/I2/
I2_fero_DAFS_1_6.9100_?????.mccd DIRECT TIFF
DATA_RANGE= 1 13
SPOT_RANGE= 1 13
!-------------------Beamline parameters-------
NX=4096    NY=4096
QX= 0.079346 QY= 0.079346
ORGX= 2048.0 ORGY= 2048.0
    DETECTOR=MARCCD
    MINIMUM_VALID_PIXEL_VALUE=1
    !STRONG_PIXEL=3.0
    OVERLOAD=65535
    !MINIMUM_ZETA=0.05
    TRUSTED_REGION=0.00 1.05
    !TEST_RESOLUTION_RANGE=10.0 3.0
    !TOTAL_SPINDLE_ROTATION_RANGES=10 180 10
    !STARTING_ANGLES_OF_SPINDLE_ROTATION=0 180 5
    !VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS=6000 30000
    INCLUDE_RESOLUTION_RANGE=40 0
    ROTATION_AXIS=1.0 0.0 0.0
    INCIDENT_BEAM_DIRECTION=0.0 0.0 1.0
    FRACTION_OF_POLARIZATION=0.9
    POLARIZATION_PLANE_NORMAL=0.0 1.0 0.0
    DIRECTION_OF_DETECTOR_X-AXIS=1.000 0.000 0.000
    DIRECTION_OF_DETECTOR_Y-AXIS=0.000 1.000 0.000
    !MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT=6
    FRIEDEL'S_LAW=FALSE
```

The master input file for automatic processing for this ferredoxin. Most of the master file is kept between energies. Filename, X-ray wavelength, and the number of jobs and processors are the only variables that change in successive processing.

Figure 8-6

XSCALE.inp File for Ferredoxin I2 (abbreviated)

```
OUTPUT_FILE=6.9100_XDS_ascii.hkl
   INPUT_FILE=../I2_fero_DAFS_1/6.9100_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_2/6.9100_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_3/6.9100_XDS_ASCII.HKL
   MERGE=TRUE
   FRIEDEL'S_LAW=FALSE

OUTPUT_FILE=6.9300_XDS_ascii.hkl
   INPUT_FILE=../I2_fero_DAFS_1/6.9300_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_2/6.9300_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_3/6.9300_XDS_ASCII.HKL
   MERGE=TRUE
   FRIEDEL'S_LAW=FALSE

OUTPUT_FILE=6.9500_XDS_ascii.hkl
   INPUT_FILE=../I2_fero_DAFS_1/6.9500_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_2/6.9500_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_3/6.9500_XDS_ASCII.HKL
   MERGE=TRUE
   FRIEDEL'S_LAW=FALSE

OUTPUT_FILE=6.9700_XDS_ascii.hkl
   INPUT_FILE=../I2_fero_DAFS_1/6.9700_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_2/6.9700_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_3/6.9700_XDS_ASCII.HKL
   MERGE=TRUE
   FRIEDEL'S_LAW=FALSE

OUTPUT_FILE=6.9900_XDS_ascii.hkl
   INPUT_FILE=../I2_fero_DAFS_1/6.9900_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_2/6.9900_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_3/6.9900_XDS_ASCII.HKL
   MERGE=TRUE
   FRIEDEL'S_LAW=FALSE

OUTPUT_FILE=7.0000_XDS_ascii.hkl
   INPUT_FILE=../I2_fero_DAFS_1/7.0000_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_2/7.0000_XDS_ASCII.HKL
   INPUT_FILE=../I2_fero_DAFS_3/7.0000_XDS_ASCII.HKL
   MERGE=TRUE
   FRIEDEL'S_LAW=FALSE
            etc etc (for 67 sets of data)
```

The first six energies of the ferredoxin's three sweeps are shown. There are 67 in the full XSCALE.inp file. Three ASCII.HKL files per energy are merged and scaled using this file. The resulting output files use the lowercase ascii.hkl suffix.

8.4 Results

The average of the 3 sweeps produced 6690 *hkl*s in the lowest energy (6910eV) wedge and 8011 *hkl*s in the last (7345eV); 5428 *hkl*s were contiguous over the 67 energy points and 3 sweeps. Merging and scaling this data greatly improved the signal-to-noise ratio of the diffractions; 1874 diffractions were rejected as outliers or having too large a Dixon Q-test value. Fully 3554 diffractions were retained which is a greater percentage than when a single sweep was collected. A *dev* threshold of 0.80 was chosen as having a high degree of preference for one iron over the other, while maintaining the maximum number of reflections for PCA. There were 427 reflections retained for atom-label Fe1 and 352 for Fe2. The atom-labels for this ferredoxin, Fe1 and Fe2, are reversed as compared with 1CZP ferredoxin in the simulated chapter. To recap: Fe1 is the inner iron (previously assigned as the oxidized Fe^{3+} ferric iron), and Fe2 is the outer iron (previously assigned as the reduced Fe2+ ferrous iron). The diffractions were then feature-scaled and put through the PCA module, the scores for which are in the legends of Figure 8-8. As with the myoglobin data, the first component from analysis of both datasets has a near-linear decline with respect to collection time, this 'dimension 1 anomaly' is discussed in Section 7.4. The scores for dimension 2 are both approximately twice as large as those for dimension 3, as expected. Dimension 2's pre-edge region is much less noisy that myoglobin.

As shown in Figure 8-8, the scores for the inner iron (orange) are all higher than those of the outer iron (blue); the ratio is 90% in agreement with the ratio of the number of reflections collected:

$$\frac{(6.98+2.68+1.30)}{(5.25+2.11+1.06)} \approx \frac{427}{352} \tag{8.1}$$

The effect on the inner iron, by radiation damage, should be less with respect to reduction and have a more consistent spectra than the outer iron (Appendix II); that is evidenced with the PCA scores and by visual inspection. Both results share a maximum value at 7120eV in dimension 2 and a corresponding inflection at the same

Figure 8-7

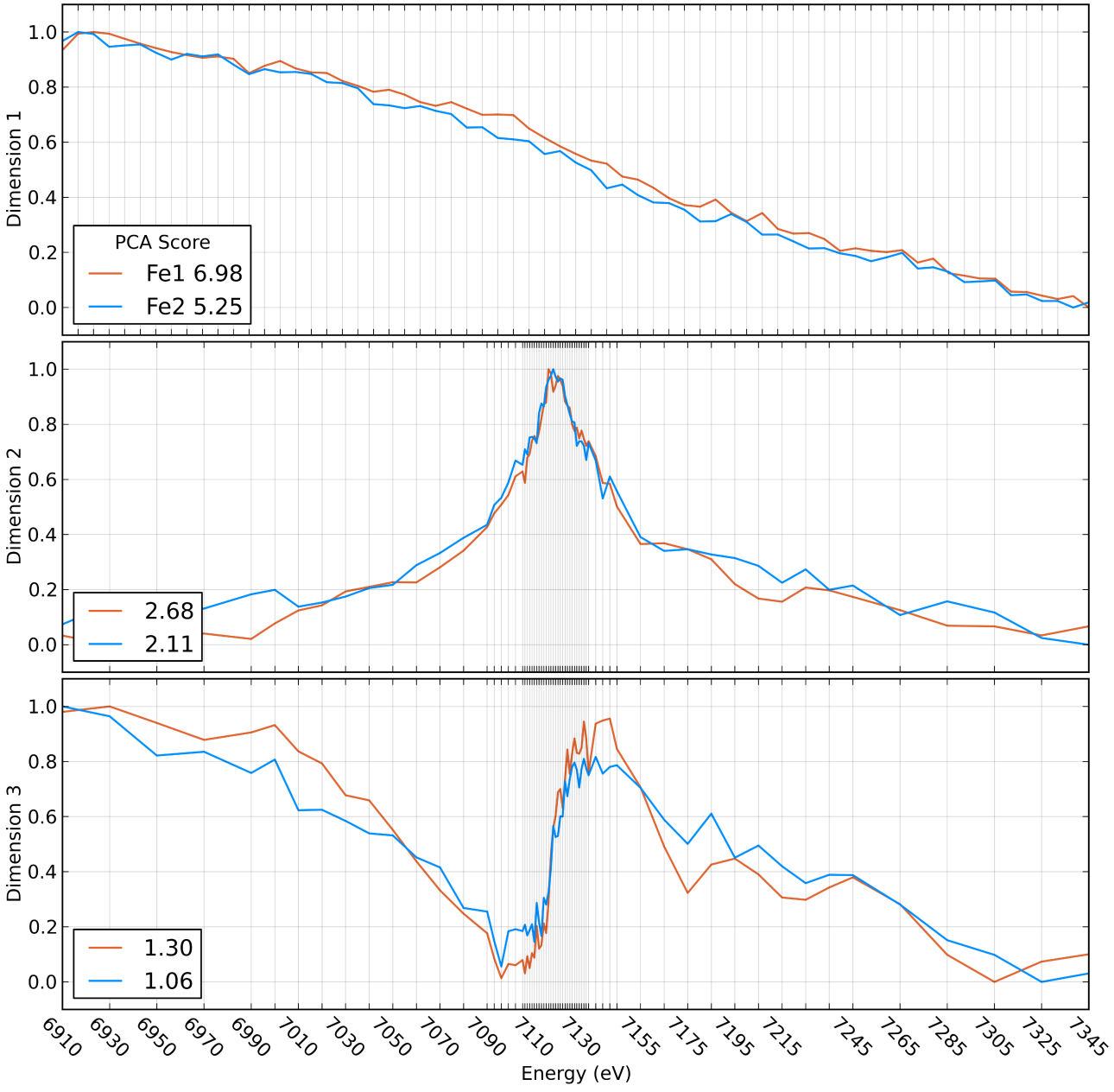
Screen Output From DeskTools (abbreviated)

```
xtal_info for pdb file
UC Vol: 174595
[67.4, 59.02, 46.61, 90.0, 109.67, 90.0]
[[ 0.014837  0.        0.005303]
 [ 0.        0.016943  0.      ]
 [ 0.        0.        0.022784]]
PDB - Symmetry recovered from SYM directory file c2.sym
4 symmetry operations
['X', ' Y', ' Z']
['-X', ' Y', ' -Z']
['X+1./2.', ' Y+1./2.', ' Z']
['-X+1./2.', ' Y+1./2.', ' -Z']
ASU from a PDB file: Ferrodoxin_I2middle5_JN.pdb
HETATM 1593 FE1  FES A 201      54.182  15.302  55.348  1.00 17.73          FE
HETATM 1594 FE2  FES A 201      53.600  15.593  52.740  1.00 43.67          FE
HETATM 1602 FE1  FES B 202      74.292  15.381  56.164  1.00 14.14          FE
HETATM 1603 FE2  FES B 202      74.904  15.777  59.042  1.00 36.20          FE
[1010, 2, 2, 270, 381, 22, 2] 1689
['C', 'FE1', 'FE2', 'N', 'O', 'S', 'ZN']
Fe1 ->      Fe3+
Fe2 ->      Fe2+
 [1010, 270, 381, 22, 2, 2, 2]
['C', 'N', 'O', 'S', 'Zn', 'Fe3+', 'Fe2+']
['C', 'N', 'O', 'S', 'Zn', 'Fe1', 'Fe2']
[[ 4543.     0. -1057.]
 [    0.  3483.     0.]
 [-1057.     0.  2172.]]
Unit Cell Volume 174592
['6.9100_XDS_scaled.hkl'…'7.3450_XDS_scaled.hkl']
Spectrum
67 Total:  6910.0 ---> 7345.0
Creating pickle file data_dictionary: I2-FeroMerge123-data-dict.pickle
Resolution Range 35 --> 1.77
Number of HKLs is 5428
Cromer-Mann
{'Fe2+': 'Reduced_Hybrid.out', 'Fe3+': 'Oxidized_Hybrid.out'}
Example of anomalous dictionary entry at energy: 7119.0
Zn    +1.14 +0.85j
Fe2+  +7.26 +2.52j
C     -0.02 +0.01j
O     -0.06 +0.04j
Fe3+  +7.28 +2.00j
S     -0.37 +0.70j
N     -0.04 +0.02j
Creating pickle file theo_dict: I2-FeroMerge123-theo-dict.pickle
Creating pickle file theo_dict: 0_I2-FeroMerge123-theo-dict.pickle
Creating pickle file theo_dict: 1_I2-FeroMerge123-theo-dict.pickle
Creating pickle file theo_dict: 2_I2-FeroMerge123-theo-dict.pickle
Number of hkls in data_dict: 5428
Usable HKL list is 3554 long
Calculating and creating the Priniple Componants file
evals1: [ 6.97846047  2.68154944  1.29617549  0.54269905  0.51239883]
evals2: [ 5.24699352  2.11381552  1.06442259  0.50942268  0.44623708]
Threshold 0.8:    427 352
```

An edited version of the screen-out from running *DeskTools* on ferredoxin crystal I2 used in this section.

Figure 8-8

Results of the PCA Module Working on Ferredoxin



The top three components (dimensions) from PCA module working on the separated matrices of diffractions from a DS run on ferredoxin, feature scaled, with scores. Atom-label Fe1, the inner iron, associated with oxidized iron (orange) and Fe2, the surface reduced iron (blue)

point in dimension 3. Ideally, the two inflection points of dimension 3 would be slightly separated with the outer iron shifted to lower energy, as we expect this to be the reduced iron. The components are less noisy than myoglobin and also contain non-negligible differences. The dimension 1 anomaly persists with the ferredoxin crystal and if the designation of dimension 1 as negative (a detrimental effect) is correct, it indicates that the hkls associated with the outer iron succumb to the effect more quickly than the oxidized iron. Similarly, dimension 2 of the outer iron started to be effected at lower energies more than than the inner iron. The components of dimension 3 were the noisiest, as expected, but also had significant overall difference in shape.

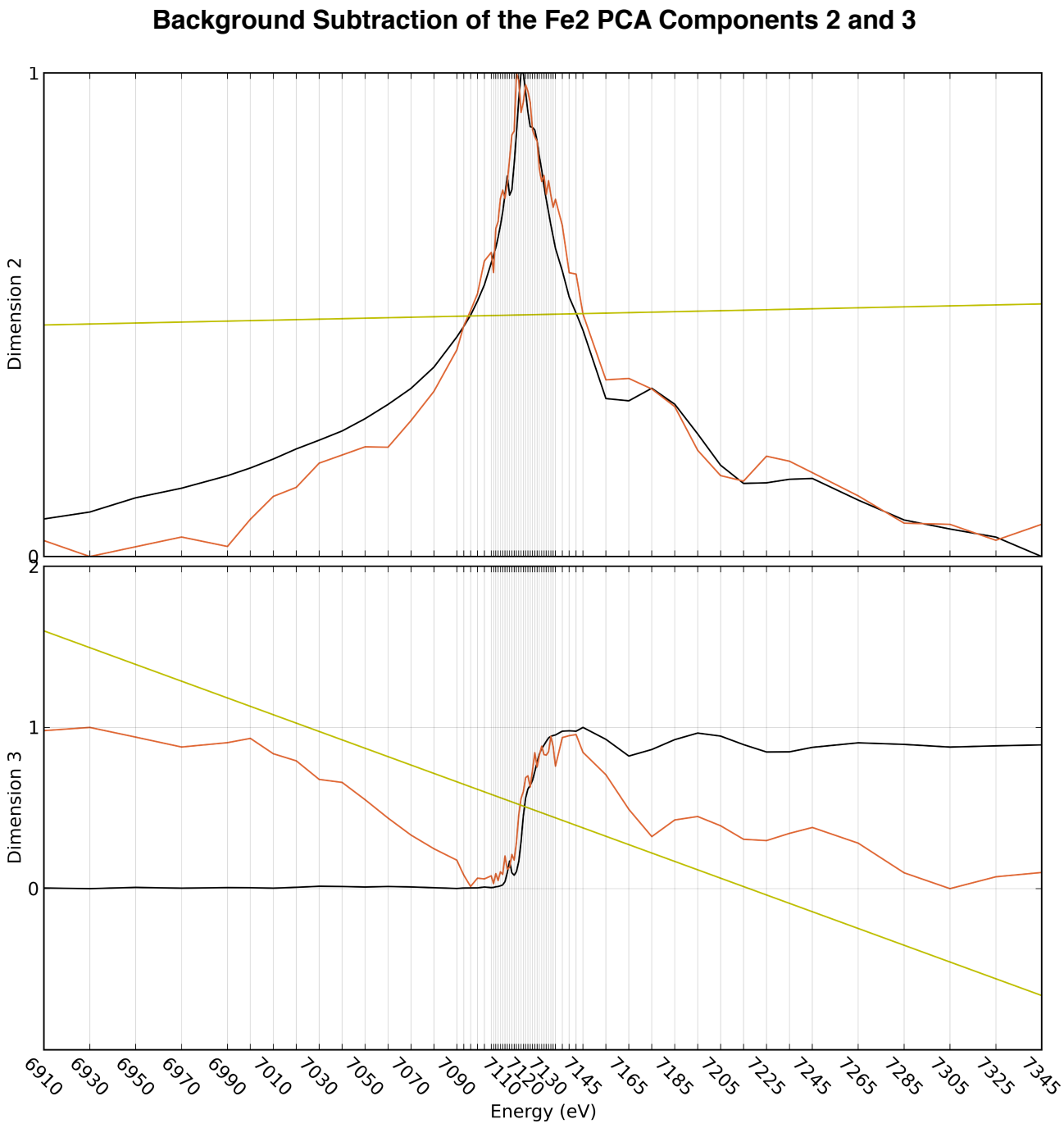
8.5 BacksubRot.py Fitting Algorithm

After the components are calculated from the separate atom-labels, a fitting algorithm is applied to compare them to the Kramers-Kronig from XAS spectra (Figures 8-2 and 8-3). The method implemented by *BacksubRot* feature-scaled both the components and the comparison spectra so that they both have maxima and minima values of 1 and 0, respectively. A linear (with respect to energy) background is calculated by rotating a line about the mid-point on the energy spectrum and the ordinate value. The background is then subtracted from the component, and a least-squares sum is calculated between the result and the comparison spectrum. The *Backsub,* which produces the lowest number from least squares, is ultimately subtracted from the component. The results of the *Backsub* algorithm (Figures 8-9 and 8-10) were then ready to be acted upon by the rotating portion (*Rot*) of the program. The resulting component pair from *Backsub* was projected into the cylinder form and then rotated through the range $\phi=[0,1]$ using:

$$\dim 2_{new} + \dim 3_{new} = \left(\dim 2_{orig} + \dim 3_{orig}\right) e^{i 2\pi\phi} \tag{8.2}$$

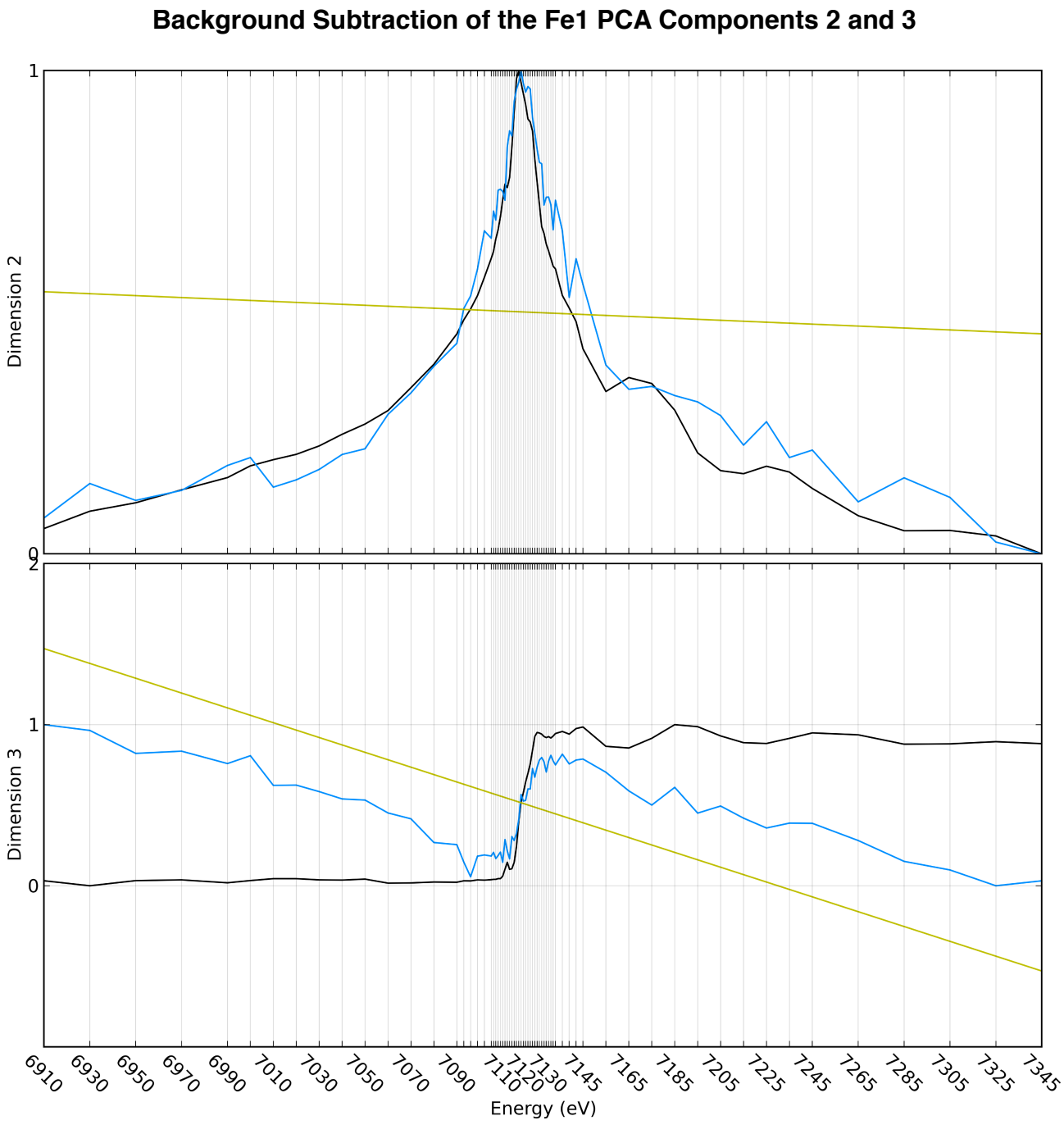
Just as with the background subtraction operation, the least-squares are calculated using the new dimensions 2 and 3 and the comparison spectra. When a minima is found, the new dimensions are retained. The background-subtracted and

Figure 8-9



Mid way through the background subtraction module of *BacksubRot.py* program. Components 2 and 3 (orange) from the inner iron are compared to the feature-scaled Kramers-Kronig pair from XAS spectrum Fe2+ (black). The calculated minimum backgrounds are also shown (yellow).

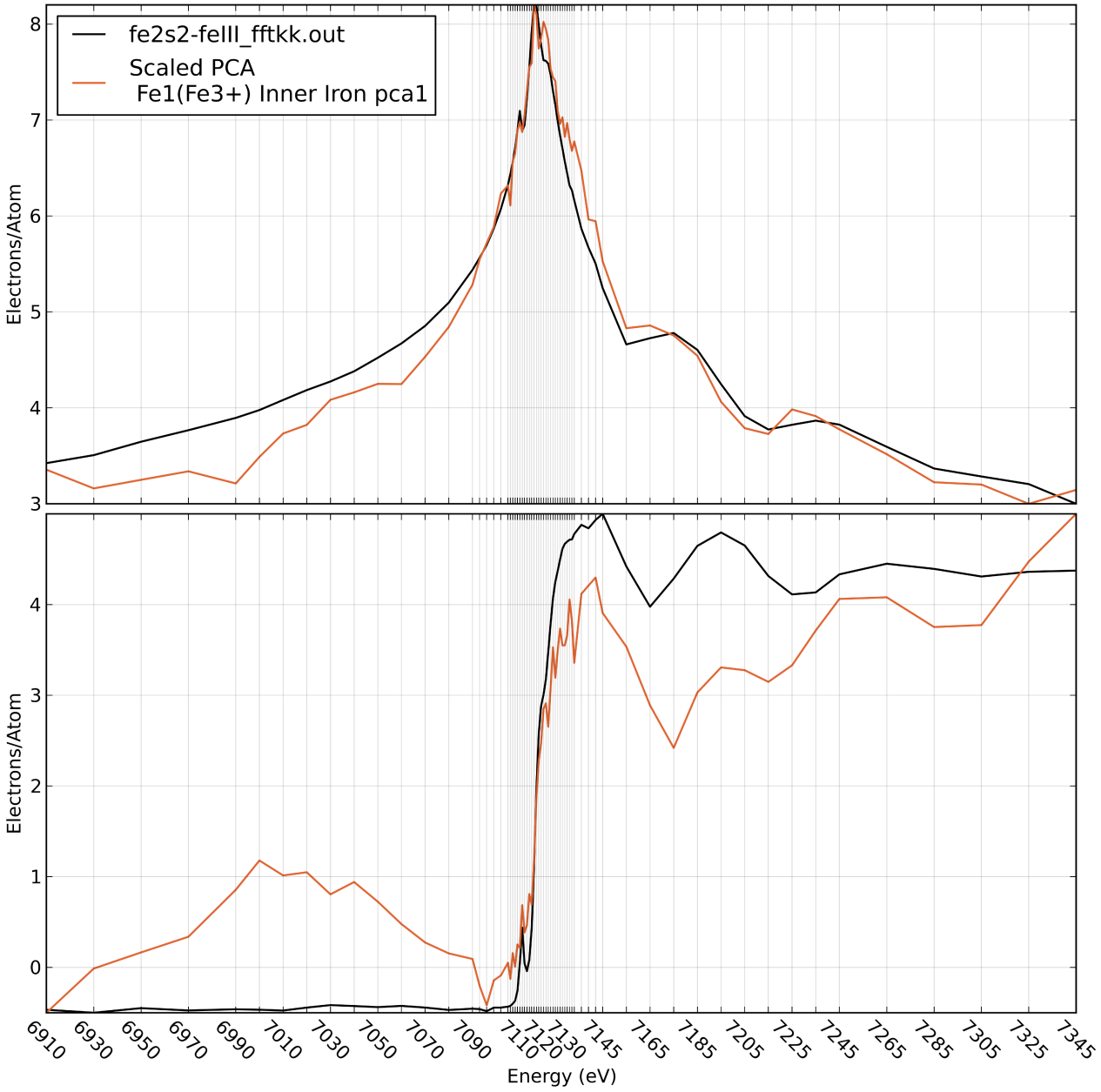
Figure 8-10



Mid way through the background subtraction module of *BacksubRot.py* program. Components 2 and 3 (blue) from the outer iron are compared to the feature-scaled Kramers-Kronig pair from XAS spectrum Fe3+ (black). The calculated minimum backgrounds are also shown (yellow).

Figure 8-11

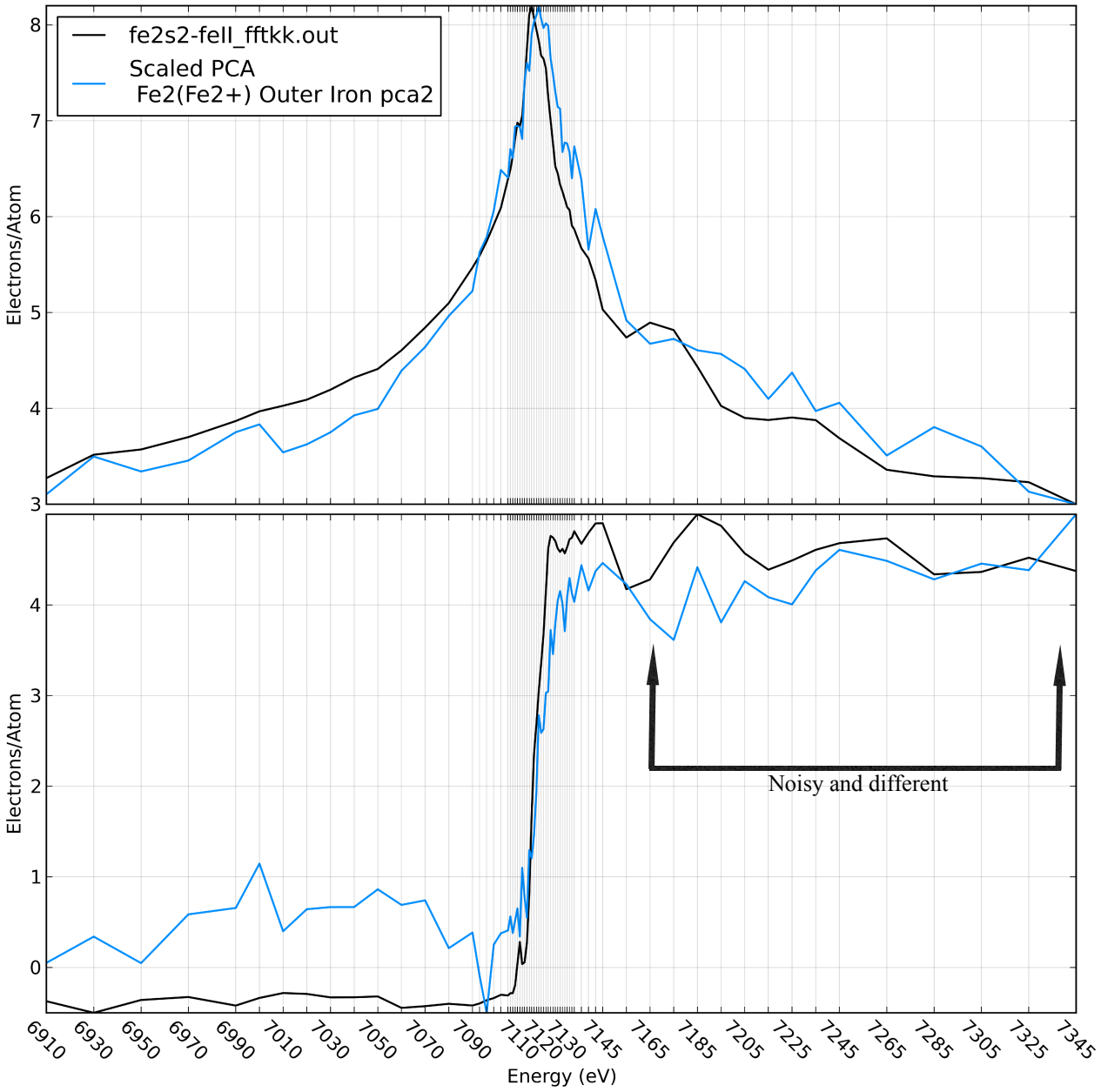
Final Comparison of DS PCA Rotated and Scaled Ferredoxin Inner Iron with Fe3+ XAS Spectra



Orange: the scaled and rotated 2nd and 3rd principle components of a subset of diffraction from ferredoxin that are biased toward atom-label Fe1 and suppressed in regards to Fe2. Black: the absorption spectrum (bottom) of reduced iron from ferredoxin and its Kramers-Kronig mate (top)

Figure 8-12

Final Comparison of DS PCA Rotated and Scaled Ferredoxin Outer Iron with Fe2+ XAS Spectra



Blue: the scaled and rotated 2nd and 3rd principle components of a subset of diffraction from ferredoxin that are biased toward atom-label Fe2 and suppressed in regard to Fe1. Black: the absorption spectrum (bottom) of oxidized iron from ferredoxin and its Kramers-Kronig mate (top)

rotated dimension 2 and dimension 3 are then scaled so that they can be compared to the original XAS spectra. This process is conducted on each set of components, from the outer iron and the inner iron.

8.6 Analysis of the Components of the Diffraction after Fitting

The initial conjecture that site separated absorption profiles from subsets of diffractions can be attained is clearly seen here and proved by the simulated diffraction. That each profile is biased toward one atomic position has also been confirmed however the proof that these profiles are of sufficient quality to make a definitive statement about each atom's environment is an over-stretch. The inner iron has excellent agreement and stands out as the best data taken to date, the outer iron has a more complicated interpretation. Separate spectra from iron atoms in the 2Fe-2S of ferredoxin has never been achieved so it is not obvious that our spectrum is the spectrum we seek. The data is too limited and noisy for EXAFS analysis and a threshold of 0.8 is not as definitive as the threshold of 0.95 used in simulated diffraction. The objective of the experiment was to deconvolute the spectra from each atom which has been achieved. This holistic method is reminiscent of the early years of absorption where there was a need for a library of prior examples from which to compare. As these are the first of this type of spectra there exists no such library, though running Kramers-Kronig on absorption spectra seems to be a very close second. FEFF does have the ability to calculate DANES spectra but calculating the Near-Edge part of the spectrum has still not been perfected either for normal absorption or for DANES.

Discussing the spectrum of the real part, f_1, of anomalous dispersion from diffractions is novel in and of itself but with improved data having both pairs of spectra from f_1 and f_2 could be used as a self consistent check and a way to tackle noise. Currently dimension 3 is so noisy that when combined and rotated with dimension 2 that it increases noise in dimension 2 significantly: which further complicates analysis.

The results of the background subtracted, rotated and scaled components of the PCAs are shown in Figures 8-11 and 8-12. There are three features that the component-spectra contain that are possible significant indicators.

1) The overall cusp of the outer iron is shifted down in energy when compared to the inner iron even as it shares a maxima. This can best be seen with a smoothing function in Figure 8-13 *(top)*. The method of collecting these spectra starts on the low energy side so this effect is less damaged by radiation.

2) The peaks right after the inflection are slightly muted similar to that of the reduced spectrum from absorption. Figure 8.13 *(bottom)*.

3) The oscillations after the edge of the outer iron are significantly different from those of the inner iron. They are not different from the oxidized absorption spectrum *and* similar to the reduced … just different (Figure 8-12).

Figure 8-13 shows a comparison of the components after a Savitzky-Golay fitting function has been executed which helps visualize the spectra with less noise however as the spacings in energy are not even the fitting function smooths out the details in the oscillations above the absorption edge.

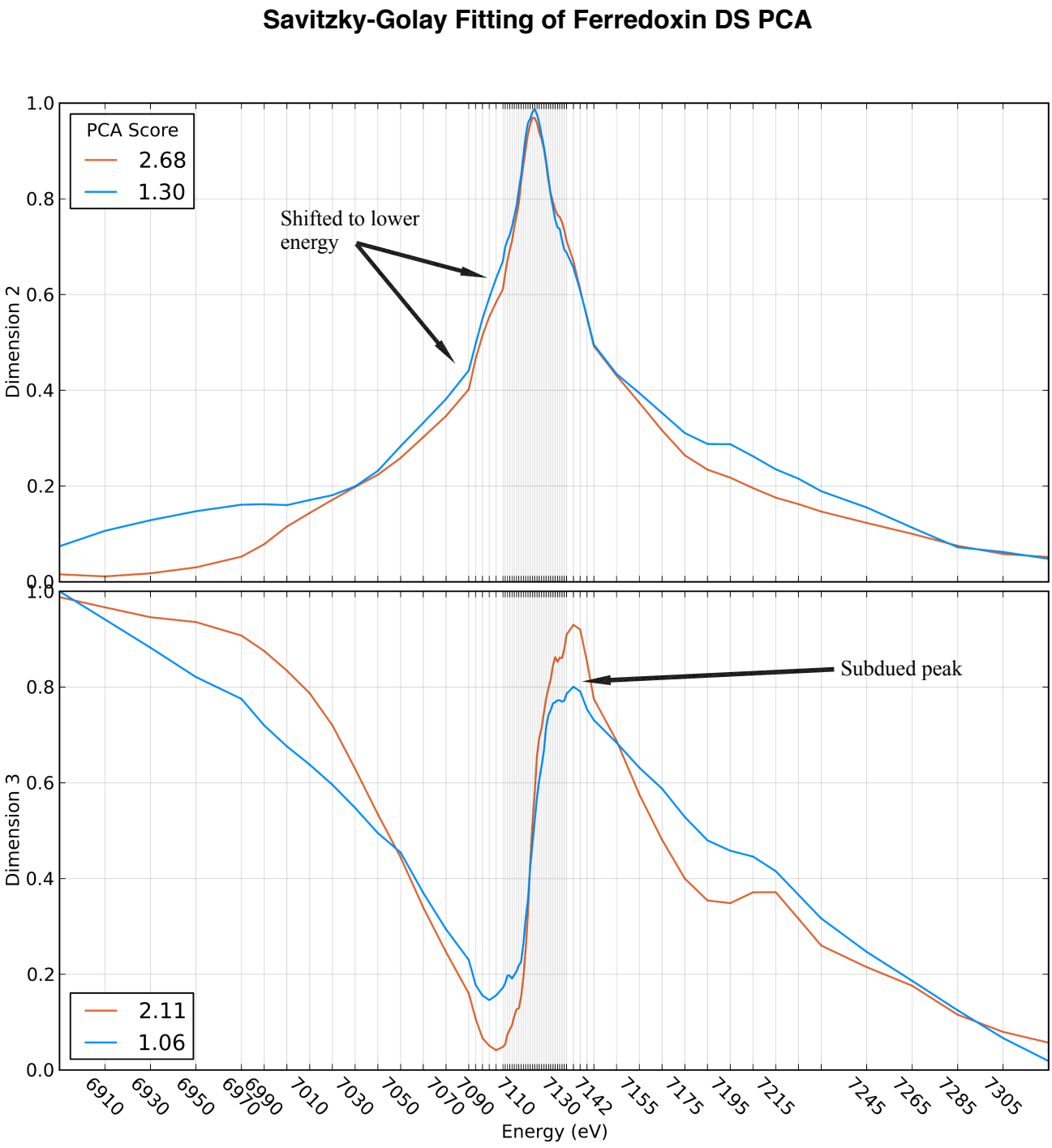
8.7 Conclusion

With noisy diffractions numbering in the low hundreds and over this limited spectrum there is good agreement between the more stable inner iron and the Kramers-Kronig pair of XAS Fe^{3+} spectra associated with it. The the XAS of Fe^{2+} originates from a mixed Fe^{2+}/Fe^{3+} solution. The outer iron diffraction spectra is also less clear. It has definitely suffered from not having as many diffractions to process and shares many traits with the inner iron. This was expected in that this iron would be less stably bound. The outer iron has neighbours in a broader range of configurations and is a candidate

for a mixed redox potential. Crystals are also exposed for relatively long periods of time (2-4 times longer than a 'normal' solution set) so there is a high expectation of radiation damage, the effects of which are unclear. All of these make this spectra more complicated to interpret.

Site-selective absorption experiments have now been conducted on large unit cell crystals at a third generation beamline, exclusively using existing equipment. The result supports prior evidence of the surface iron of ferredoxin being the reducing iron though it is not fully conclusive, the spectra generated are limited and noisy but are an excellent proof of concept for the methodology chosen. Improvements in crystal quality, collection strategy and analysis will elucidate more details over time.

Figure 8-13


Savitzky-Golay Fitting of Ferredoxin DS PCA
PCA Score
2.68
1.30
Shifted to lower energy
Dimension 2
Subdued peak
2.11
1.06
Dimension 3
Energy (eV)
6910
6930
6950
6970
6990
7010
7030
7050
7070
7090
7110
7130
7142
7155
7175
7195
7215
7245
7265
7285
7305

CHAPTER 9

DISCUSSION

Improvements in the understanding of the region between crystallography and anomalous dispersion can help discover the electro-chemical configurations of specific atoms within macromolecules. An accurate knowledge of the transition metals physical and electronic structures is essential for understanding the metal complex's role within the larger protein structures. Macromolecules are by definition large and will crystallize with large unit cells. The focus of this research is to differentiate between metals of the same element within these large crystals. The challenges of unit cells containing vast biological samples with thousands or diffractions has been tackled in this research. It is an escalation of the work done in the 1990's on small, well-ordered crystals whose research arose because the ability to inspect individual spectra from elementally identical atoms is very difficult by any other method. Producing XAS-like spectra for macromolecular crystallographers will help clarify the role of the metals in these much larger systems.

The methodologies designed and demonstrated in this research are the residual of a number of failed choices: bad crystal choices, inferior collection strategies, mishandling of data processing and analysis. The choices that returned the largest benefits occurred when the beamline and software strengths were utilized. Macromolecular beamlines work best for macromolecules; small molecules have a separate set of challenges all their own. Collecting multiple runs as opposed to collecting multiple crystals. Using processing software that was designed with auto processing in mind (XDS). Employing the maxim "shoot first ask questions later" gave many more datasets to analyze which in turn honed the analysis. Relentlessly testing different ways of discriminating diffractions (from data-sets) led to strictly dense

matrices, a modified Dixon-Q and broad outlier rejection. Combined with principal component analysis the techniques and the instrumentation has demonstrated that macromolecular crystal can be interrogated for details about individual metal atoms.

The research should be considered a qualified success, the results are encouraging but not 100% unambiguous. With every step of the process, the goal of sharply different spectra from site separated metals gets closer. The process needs improvements, DS would benefit from moving away from using PCA as a tool, and directly detecting the small signals. This should clarify results however it would require geometry, speed and sensitivity improvements in detectors. Smaller, more sensitive pixels in the CCD, proximity to the diffracting crystal and an increase in the solid angle occluded would all help improve the signal to noise ratio.

In data collection strategies the rotation wedge at different energies collects a significantly different set of hkls leading to a sparse matrix. This can be mitigated for in two ways: 1) Rotating the crystal for a different wedge width at each energy, and 2) To have the same reflections at different energies recorded at the same positions of an X-ray detector the detector should have been moved to appropriate distance for each energy, respectively. This will produce a more-similar dataset at each energy (less sparse matrix) and reduce the amount of time the crystal is exposed.

When more data has been taken, with marginally better crystals, an improved scaling and outlier rejection protocol is needed. This area of the research needs an investigation all its own. A rigorous, systematic, rejection theory would include appropriate diffractions and exclude egregious ones; positively effecting outcomes of the analysis. These experiments need to be performed on various elements and crystal forms to test for effectiveness.

Predicting the true intensity of a real diffracted spot may top the list of improvements that need attention. There is great disparity between modelled intensity and what is recorded. Inclusion of all the factors given in Chapter 3 as well as counting

statistics (repetition/exposure time), mosaicity (long range disorder) and anisotropic diffraction (crystal packing, unit cell vibration) might narrow the discrepancy to a more acceptable range.

9.1 Short-Term Upgrades

9.1.1 PCA Investigation

One simple upgrade stems from PCA randomly applying signs to each component. The helix in the cylinder projection has definite handedness and results from PCA should be forced to reflect this. Currently a considerable amount of time is wasted re-orientating results from PCA to fit expectations. The handedness needs to be set automatically. Also, the cause of the 'dimension 1 anomaly', Section 7.4, needs an investigation to determine the underlying physical origin, which should also betray the sign associated with its PCA component.

9.1.2 Simultaneous XAS

The crystals are absorbing and fluorescing while they are being exposed for diffraction. Data regarding absorption spectra should be collected simultaneously via fluorescence and transmission during the experiment. Currently this source of relevant information is being wasted. MX beamlines are equipped with a florescence detectors suitable for a crude XAS spectrum but most are not designed for this data to be recorded during diffraction. At some beamlines this might be impossible due to physical restrictions but if it can be implemented this information can only go to improve the analysis. Future beamline designs may include more sophisticated fluorescence detectors[1] without these physical restrictions. Additional information in the form transmission/absorption spectra could also be collected using ion chambers and/or a beamstop diode [74].

[1] Fluorescence detectors with optional Soller slits and filters similar to an XAS experimental set up.

9.1.3 Continuous Collection

Newer detectors have become available for crystallography, amongst the most interesting are Pilates [67] and Taurus-1 [68]. They are able to collect data in an open-shutter mode by reading data from the CCD continuously using CMOS technology. Improvements in this area eliminate the need for fine-phi slicing[1] which enhances spot profiling and speeds up collection times. Continued progress in CCD sensitivity reduces the exposure time required to achieve the same count rate on the detector.

9.1.4 Multiple Target Atoms

Currently *dev* has only been used to separate two target atoms but it should be able to be expanded to include 3 or more atoms. The number of diffractions collected would need to be increased and the complexity would increase but there is no reason why Equations 5.3. can not be generalized:

$$I_{\mathbf{h},E} = \left| \left[F_Z + \Delta_{Fe_1} + \Delta_{Fe_2} + \cdots + \Delta_{Fe_M} \right] \right|^2 \tag{9.1}$$

$$I^0_{\mathbf{h}} = \left| \left[F_Z \right]_{\mathbf{h}} \right|^2 \tag{9.2}$$

$$I^{Fe_x}{}_{\mathbf{h},E} \overset{def}{\equiv} \left| \left[F_Z + \Delta_{Fe_x} \right]_{\mathbf{h},E} \right|^2 \tag{9.3}$$

$$\sigma_{\mathbf{h}}{}^{Fe_x} = \sqrt{\frac{\sum_{E=1}^{N} \left(I_{\mathbf{h},E}{}^{Fe_x} - I_{\mathbf{h}}{}^0 \right)^2}{N-1}} \tag{9.4}$$

$$dev\left(Fe_x\right)_{\mathbf{h}} \overset{def}{\equiv} \left(\frac{\sigma_{\mathbf{h}}{}^{Fe_x}}{\sum_{i=1}^{M} \sigma_{\mathbf{h}}{}^{Fe_i}} \right) \tag{9.5}$$

[1] Sub 1° oscillations.

$$\sum_{i=1}^{M} dev\left(Fe_i\right)_{\mathbf{h}} = 1 \tag{9.6}$$

Simulated data collection of multi-metal crystals could be investigated using the same techniques applied in Chapter 6. Theoretical results from separating more than two target atoms could be calculated prior to actual experiments to judge the validity of the technique expanding in this way. The maximum values for $dev(Fe_x)_{\mathbf{h}}$ need to be calculated. It is not clear under which circumstances (if any) that a single hkl would have a large bias toward a single atom over 2, or more others; making that hkl viable for further analysis.

9.1.5 Data Mining the PDB

As this technique requires solved proteins[1] (successfully crystallized), it would be foolish to apply it to only new proteins. As mentioned in the introduction, the rate-limiting step for protein structure determination is the crystallization step. Therefore writing a script that scrapes the entire PDB archive looking for appropriate structures (past and current) for this experiment should be written. Once this technique is more fully developed, the PDB or data processing software (XDS, d*TREK etc) could alert an investigator whether DS is an appropriate experiment for their structure.

9.1.6 Temperature and Normal Polarization

The calculations for dev will be greatly improved if the simulated diffraction more closely simulates real diffraction; expansion of dev to include temperature, normal polarization and the Lorentz factors should be included. Once crystal orientation as it relates to the orbit of the synchrotron ring (and the detector face) is factored in, it should also be possible to include anisotropic temperature factors. Incremental steps for improving simulations should fractionally improve results. As mentioned in the

[1] Steps for a successful career in molecular biology: 1) Have crystals. 2) Don't not have crystals.

introduction to this chapter, there is great disparity between simulated and real diffraction however incremental improvements should incrementally improve results.

9.1.7 Cluster DS

'Cluster DS' has been discussed as a possible expansion of DS whereby, dev biases diffractions by a volume element of the unit cell instead of an atom. This would be beneficial to target atoms that are separated either by a good distance or possibly by clusters of target atoms. For example in nitrate reductase (narGHI) there are four distinctly separate iron clusters that transport an electron over a very large distance. Cluster DS may be able to look at each cluster's anomalous dispersion spectrum separately.

9.1.8 High Contrast Collection

Here the definition of 'high contrast' with regards to DS is given as diffractions that strongly bias one atom being in a region with other diffractions that favour the other atom of interest. Once a solution dataset is taken and the crystal is solved it is possible to calculate the wedge with the highest density of high contrast diffractions. The ability to orientate the crystal such that the thinnest wedge with greatest contrast of hkls that bias both atoms is calculable would reduce the exposure required. Currently in order to assess the same quality of data, a wedge with a visibly high frequency of diffractors is chosen by the experimenter.

9.1.9 Relationship to the Rees Method of Separation

As mentioned in the Introduction, another new method for site separation has been developed by Prof. Doug Rees at CalTech. The method uses many more diffractions but a lot fewer energy points [9]. Atomic form factors of individual atoms are calculated by solving the whole crystal structure at a limited number of energies. There

could be a good interstitial area between the two methods and this should be investigated.

9.1.10 Software Upgrade

At every stage in the development of this research, new software was written to handle and manipulate data. The new software is effective but it needs to be formalized and packaged for wider distribution as it may be impenetrable to an outside observer in its current state.

9.2 Middle Distance Upgrades

9.2.1 Limits

DS needs more mathematical and experimental rigour applied to it. Experiments on smaller unit cells, possibly even starting over again with small molecules and progressing is advisable. The ability to calculate the limitations of the process and the hardware due to physical constraints, such as detector distance and saturation needs to be explored. These types of experiments and calculations along with increasing the number of target atoms should bound the method and allow a better understanding of the flaws of the technique, and consequently its strengths.

9.2.2 Diffraction Anomalous Fine Structure

The methods presented in this thesis stem from research into the fine structure of diffraction from small molecule experiments in the 1990s. A lot of interesting work has been done in this field [75] and a few novel techniques such as George and Pickering's iterative Kramers-Kronig were employed [15]. Collecting fine structure should be possible with large macromolecules however the signal-to-noise and current detector's limited dynamic range will be an obstruction. Observing fine structure from site separated atoms within a large macromolecule still remains the goal. The width of the

spectrum needs to extended to accommodate this. Fine structure may be useful with phasing especially if site-separate spectra are collected at the same time as a solution dataset.

9.2.3 Bond Polarization

The direction of the impinging E-vector from the synchrotron is highly polarized and the orientation of that vector as it encounters an absorbing atom has an effect on the quantity of the absorption. If the X-ray's E-vector is parallel with an atom-bond there is a significant increase in absorption in the near edge related to transition to bound states. For absorption these effects are averaged out in randomly orientated sample such as a powder [13] or aqueous solution . This is not true for diffraction by crystals where the orientation of all the target atoms are fixed. The target iron for the Myoglobin might be a great candidate for this effect if the protein crystallized in a favourable symmetry, such that the plane of the porphyrin rings from each protein aligned. The absorption effects related to bonds of the porphyrin ring could be accentuated and those to the histidine and CO depressed (or vice versa) depending on orientation of the crystal. There is a branch of XAS that utilizes this phenomena, Polarized X-ray Absorption Spectroscopy (PXAS). The bond-polarization effect is expected in DS but is not yet compensated for. In the two crystals that are investigated here, symmetries mask the effect by having multiple orientations within a single unit cell or bonds are similar in a multitude of directions. Care must be given to future experiments with regards to this. It would be advantageous not only to calculate the expected size of the effect but in appropriate crystals one could see a way of utilizing the phenomena to collect more detail: similar to a PXAS experiment .

9.2.4 Phasing

MAD crystallography already utilizes the dispersion phenomena to help calculate phases, by investigation this region in detail more light could be shed and possibly help improve the phasing.

9.3 Long Term

Detectors are starting to do open-shutter collection, there is no longer a need to close the shutter, count the counts, reorient the crystal and repeat. Speeding up the collection times at beamlines significantly. DS runs currently have to stop and wait for an energy change before rotating through the same 10-12°, in light of this there are two possible future options:

1) Though mathematically challenging, one could rotate the monochromator simultaneously with rotation of the crystal. This would require using a detector that reads out at several thousand hertz. In theory, if the crystal has been solved, then this effect is calculable. This method would have the effect of fragmenting the spectrum across different diffractions that are only measured once, massively complicating not only processing diffractions and identifying which hkls are which but an individual spots intensity would change non-linearly across its own profile. A complicated and inelegant approach.

2) A number of early DAFS experiments [citeXX] used dispersive optics, in which the target crystal was not bathed in a single wavelength of light, but a full spectrum. Due to the nature of diffraction the position on the detector upon which a diffracted spot lands is related to the wavelength of the light (Bragg's Law). By using a full spectrum of wavelengths of a single diffraction a 'spread' of intensities can be collected simultaneously. There are two immediate physical limitations that would need to be addressed to implement dispersive optics: a) synchrotron beamlines for crystallography are not designed for this at the present moment and b) a dispersed spot on the detector face takes up a lot of space. In conjunction with large unit cell crystals that have closely spaced diffraction: a very large detector placed at a great distance would be needed to collect in this regime. If budget wasn't an issue then these technical difficulties could be overcome.

9.4 Outro

Experiments at the liminal space of two well researched areas of synchrotron science have been conducted. The methodology and equipment employed in this research demonstrates that the dispersion spectra from two different target atoms can be separated in a large unit cell. Old questions have been asked of new things. This area of research has lain dormant for years and in its return, the power of new instruments and computing have been used effectively. Initial goals have been tested and have returned encouraging results. These new techniques and this area of research (though it needs more attention and focused experiments) is a good place to start.

APPENDIX I

SCATTERING THEORY

This section is a faithful reproduction of Charles Kittle's *Introduction to Solid State Physics,* 5th edition [1971, John Wiley and Sons]. This material is reproduced with permission of John Wiley & Sons, Inc. The construction in Kittle's book is simple and clear, superior diagrams and the avoidance of Brillouin Zones are given here. This clarifies the books introduction to scattering theory and the electron concentration function.

Elastic scattering implies that the magnitude of the wavevector, $w=ck,$ scattered by the crystal satisfies:

$$w' = w; \quad k' = k. \tag{AI.1}$$

Where w' and k' are the diffracted wavevector and wavenumber respectively. The electric field of a plane wave in free space has the form:

$$E(\bar{x}) = E_0 e^{i(\bar{k}\bullet\bar{x}-wt)} \tag{AI.2}$$

Ignoring an angular offset and where E_0 is the maximum amplitude. The form of a single scattered wave in response to a scattering centred at ρ is:

$$E_{sc} = CE(\bar{\rho})\frac{e^{ikr}}{r} = \frac{CE_0 e^{-iwt}}{r} e^{i(\bar{k}\bullet\bar{\rho}+kr)} \tag{AI.3}$$

C is a constant of proportionality and $1/r$ preserves the flow of scattered energy. The point scatterer ρ is the sum of integer multiples of three basis vectors, if the direction of $\boldsymbol{R}$ is that of k' then:

$$E_{sc}(r) = \frac{CE_0 e^{i(kR-wt)}}{R} e^{-i\bar{\rho}_{mnp}\bullet\Delta\bar{k}} \tag{AI.4}$$

The total scattering from all lattice points is a sum over mnp. The interesting part being the sum of phase contributions, which is called the scattering amplitude, A:

$$\mathrm{A} \equiv \sum_{mnp} e^{-i\bar{\rho}_{mnp}\cdot\Delta\bar{k}} \tag{AI.5}$$

If $\boldsymbol{\rho}_{mnp}=m\boldsymbol{a}+n\boldsymbol{b}+n\boldsymbol{c}$ then maximum amplitude is when the three Laue equations are satisfied simultaneously:

$$\bar{a}\cdot\Delta\bar{k} = 2\pi h; \quad \bar{b}\cdot\Delta\bar{k} = 2\pi k; \quad \bar{c}\cdot\Delta\bar{k} = 2\pi l. \tag{AI.6}$$

Where hkl are integers. These conditions are met by the reciprocal lattice vectors:

$$\bar{G} \equiv \Delta\bar{k} = h\bar{A} + k\bar{B} + l\bar{C} \tag{AI.7}$$

such that:

$$. \bar{G}\cdot\bar{\rho}_{mnp} = (h\bar{A} + k\bar{B} + l\bar{C})\cdot(m\bar{a} + n\bar{b} + p\bar{c}). \tag{AI.8}$$

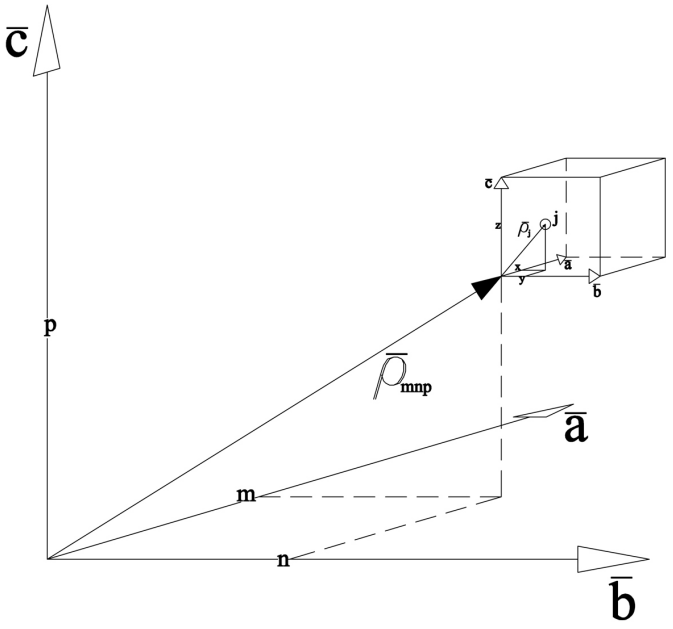
$$\bar{G}\cdot\bar{\rho}_{mnp} = 2\pi(hm + kn + lp) \tag{AI.9}$$

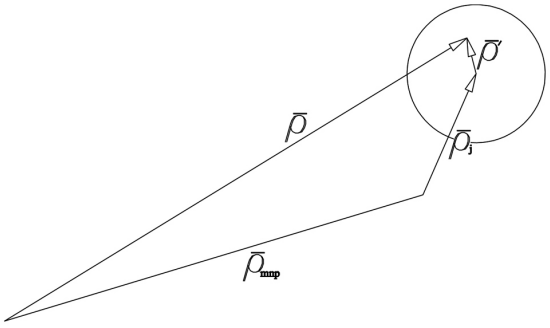
gives the allowed diffractions due to the crystal symmetry however the variation of intensity of each diffracted spot is governed by the contents of the unit cell and the relative positions of the atoms and their electronic distribution. If we consider the vector $\boldsymbol{\rho}$ as the sum of three vectors, $\boldsymbol{\rho}_{mnp}$ designating the unit cell, $\boldsymbol{\rho}_j$ the position of the nucleus of atom j and $\boldsymbol{\rho}'$ as the position of the electron relative the nucleus.

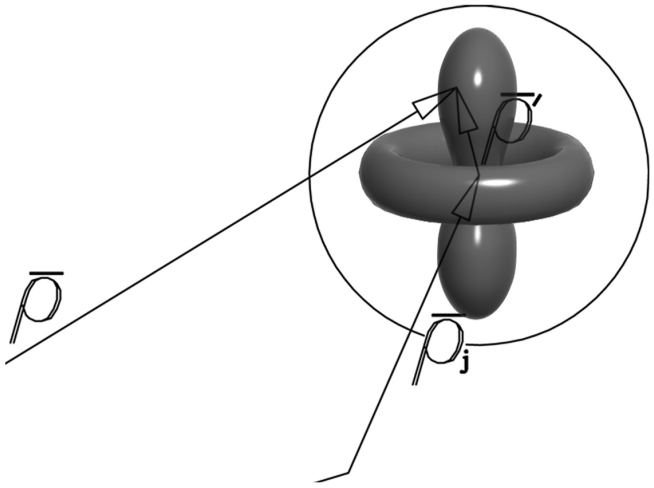
$$\bar{\rho} = \bar{\rho}_{nmp} + \bar{\rho}_j + \bar{\rho}' \tag{AI.10}$$

The electron concentration function for each atom, *j*, is then

$$c_j(\bar{\rho} - \bar{\rho}_j - \bar{\rho}_{mnp}) \tag{AI.11}$$


c
p
m
n
a
b
j
z
x
y





For s number of atoms in a unit cell the total electron concentration of the crystal $n(\boldsymbol{\rho})$ is a sum over all cells and all atoms within them.

$$n(\bar{\rho}) = \sum_{mnp} \sum_{j=1}^{s} c_j(\bar{\rho} - \bar{\rho}_j - \bar{\rho}_{mnp}) \tag{AI.12}$$

The scattering amplitude is then not over a discrete set but over two sums and a continuous volume element relating the electronic distribution.

$$\mathrm{A}_{\Delta\bar{k}} = \int dVn(\bar{\rho}) e^{-i\bar{\rho}\cdot\Delta\bar{k}} = \sum_{mnp} \sum_{j} \int dVc_j(\bar{\rho} - \bar{\rho}_j - \bar{\rho}_{mnp}) e^{-i\bar{\rho}\cdot\Delta\bar{k}} \tag{AI.13}$$

$$\mathrm{A}_{\Delta\bar{k}} = \sum_{mnp} \sum_{j} \int dVc_j(\bar{\rho}') e^{-i\bar{\rho}'\cdot\Delta\bar{k}} e^{-i(\bar{\rho}_j + \bar{\rho}_{mnp})\cdot\Delta\bar{k}} \tag{AI.14}$$

and defining the *atomic form factor* f_j:

$$f_j = \int dVc_j(\bar{\rho}') e^{-i\bar{\rho}'\cdot\Delta\bar{k}} \tag{AI.15}$$

$$\mathrm{A}_{\Delta\bar{k}} = \left(\sum_{mnp} e^{-i\bar{\rho}_{mnp}\cdot\Delta\bar{k}} \right) \left(\sum_{j} f_j e^{-i\bar{\rho}_j\cdot\Delta\bar{k}} \right) \tag{AI.16}$$

$A_{\Delta\boldsymbol{k}}$ is non-zero when $\Delta\boldsymbol{k}$ satisfies the Laue equations:

$$\mathrm{A}_{\Delta\bar{k}} = M^3 F_{\bar{G}} \tag{AI.17}$$

Introducing M^3 (See Section 3.2) and the structure factor $F_{\boldsymbol{G}}$:

$$F_{\bar{G}} = \sum_{j} f_j e^{-i\bar{\rho}_j\cdot\bar{G}} \tag{AI.18}$$

which also satisfies the Laue equations:

$$\bar{\rho}_j \cdot \bar{G} = (x_j\bar{a} + y_j\bar{b} + z_j\bar{c}) \cdot \left(h\bar{A} + k\bar{B} + l\bar{C}\right) = 2\pi(x_j h + y_j k + z_j l) \tag{AI.19}$$

which gives us:

$$F(hkl) = \sum_{j} f_j e^{-i2\pi(x_j h + y_j k + z_j l)} \tag{AI.20}$$

The intensity of diffracted X-rays by a crystal is proportional to the square of the scattering amplitude. The hkl reflections are selected by the crystal lattice and their respective strength is due to the structure factor. The structure factor is the sum of the

atomic form factor (the integral of the electronic distribution) and a phase term that relates the location of an atom to a particular diffraction *hkl*. The diffraction is proportional to the squared modulus of the structure factor:

$$I \propto \left|F_{\bar{G}}^{*} F_{\bar{G}}\right|^2 = \left|\sum_j f_j(\bar{G},E) e^{i\bar{\rho}_j \cdot \bar{G}}\right|^2 \qquad \text{(AI.21)}$$

The atomic form factor is dependent on both the $\boldsymbol{G}$ and the energy of the incident beam, E. The momentum transfer vector, $\boldsymbol{G}$, is perpendicular to the scattering plane *hkl*. The basis vector in coordinates *xyz* can be re-written as a phase with respect to an origin.

$$\bar{G} \bullet \bar{\rho}_j = 2\pi(hx_j + ky_j + lz_j) = 2\pi\delta_j \qquad \text{(AI.22)}$$

The sum is now:

$$\sum_j f_j(\bar{G},E) e^{i2\pi\delta_j} \qquad \text{(AI.23)}$$

The intensity of the diffraction is then a function of the scatter amplitude and a phase term related to its position. If the incident photon is in the vicinity of absorption energy of atom j, the scattering factor is a product of the Thompson scattering and a resonant scattering correction term.

$$f_j(\bar{G},E) = f_{0,j}(\bar{G}) + \Delta f_j(\bar{G},E) \qquad \text{(AI.24)}$$

APPENDIX II

SIMULATED FERREDOXIN XAS FROM PDB STRUCTURES

A1.1 *Analyzing PDB Structures*

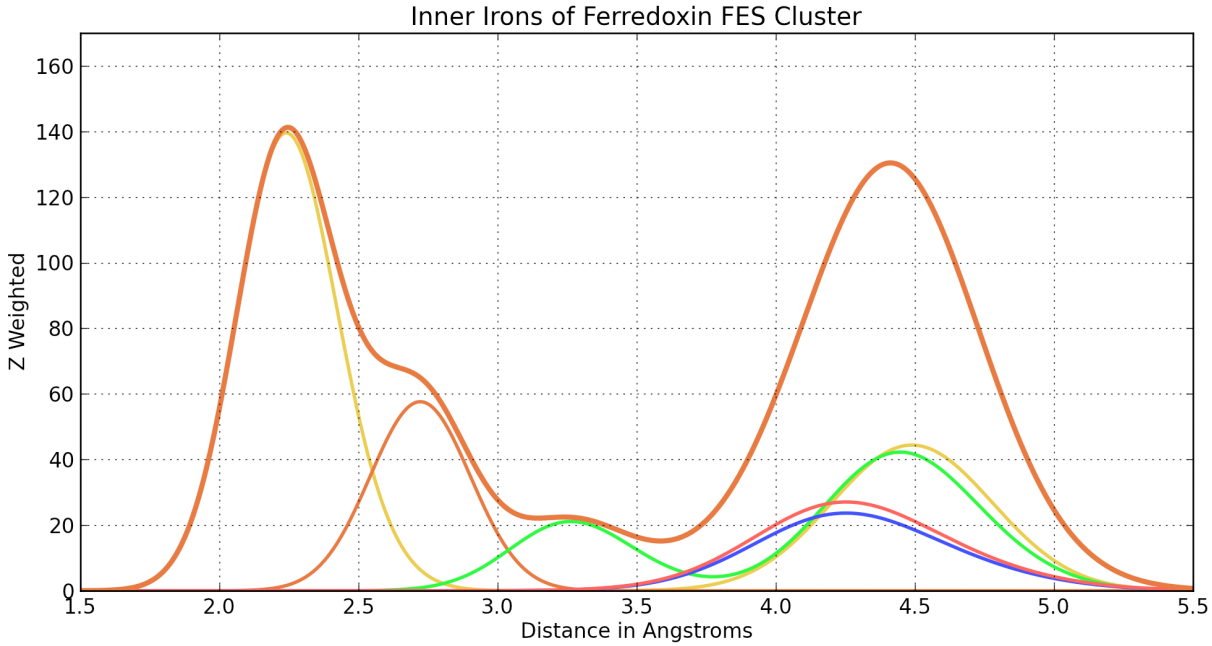
There were 43 Ferredoxin structures in the protein databank that contain a 2Fe-2S cluster with a resolution less than 2Å and are solved using X-ray crystallography. Each protein was systematically checked for errors and creative labelling. One structure was rejected for having an unusual format and one for having physically and chemically unlikely bonds. 15 more were rejected for not having each iron coordinated with 4 sulphurs. The irons in the 2Fe-2S cluster is either labeled FE1 or FE2, each protein structure is labeled independently of the others (so these could be different), also with a PDB file there may be more than one protein in each asymmetric subunit. Each PDB must be inspected to gauge which of the irons is closest to the surface, we applied a script for Pymol to make it easier:

surface_check.pml

```
cmd.hide("everything")
cmd.select("het", "het")
cmd.show("sticks", "het")
cmd.show("cartoon")
cmd.show("mesh")
cmd.set("mesh_color", "grey")
```

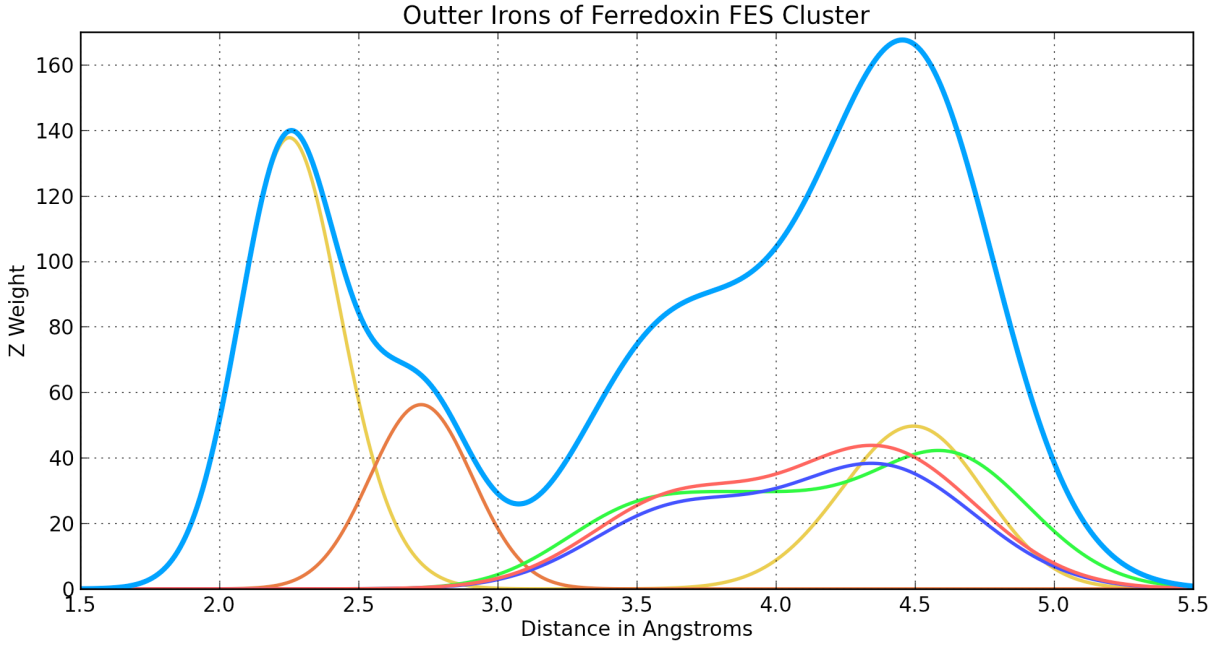
Obviously if its FE1 inner then its FE2 outer however there is one PDB file (2y5c.pdb) that has subunit A with FE1 as the outer iron and subunit B with FE1 as the inner iron! Besides these niggling issues we were able to categorize almost one hundred irons as some PDB files have multiple subunits. After the correct assignment of an iron's position to the surface it was necessary to make individual input files for the the program FEFF. In the process of writing the script pdb2FEFF.py (see Appendix IV) it became apparent that we could visualize all the pdb file radial distributions using the

Figure A2.1



A Gaussian is shown for each atom radially from the innermost iron in a Ferredoxin iron-sulfur cluster. The standard deviation is calculated from 26 PDB structures.

Figure A2.2



A Gaussian is shown for each atom radially from the outermost iron in a Ferredoxin iron-sulphur cluster. The standard deviation is calculated from 26 PDB structures.

FEFF input files. Each atom was labeled by distance from the target atom and a Gaussian was placed there whose sigma value were created from the standard deviation of positions from all the viable ferredoxins. Just by inspection you can see that the inner iron has a lot less disorder that the outer iron, which is to be expected as the surface iron takes on many more morphologies for different proteins. This looks a lot like the Fourier transforms of the EXAFS oscillations. They are closely related but this is just a crude version for visual inspection. We toyed with it a little so that it would be more representative of FEFF was actually calculating. We weighted each gaussian by the elements Z value and draped a summation line over it. It would be more accurate to get the average positions and apply the average temperature factor however the objective was not to slowly rewrite FEFF but give an idea of the spread of possible spectra for all the various 2Fe-2S Ferredoxins. Fig. A2.3 shows the spread in the radial distribution but Fig. A2.4 gives the detail of each FEFF calculation for all 92 Irons and the two chosen to represent the reduced and oxidized atoms in the rest of the thesis. An average of the lighter lines in A2.4 being an approximate Fourier transform of those in A2.3.

A2.2 *The Case 1CZP (and 1QT9)*

1CZP and 1QT9 are the reduced and partially oxidized version of the same protein. For a full description read the associated paper by Morales et al. [61]. Most of the Ferredoxins in the PDB are reduced. 1CZP, the partially oxidized version, is actually of mixed valence. The oxidation states have a occupancy ratios 60:40 and 45:55 depending on the sub unit. There is characteristic difference in the morphology that Morales et al. believe is associated with the oxidized iron and that is a CYS46 oxygen ‘flip’. It goes from ‘CO out’ to ‘CO in’ depending on whether the FE1 (iron nearest surface) is reduced or not, “CO out” is the reduced, “CO in” is the oxidized. As above this is the approximate radial distribution of electrons. I used the same standard deviation for each Gaussian, not the temperature factors. The XAS of 1CZP are good approximations for the lighter bands of the orange and blue of the meta-data from the all the other ferredoxins, Fig A2.4. The two irons picked from this protein, in

Figure A2.3

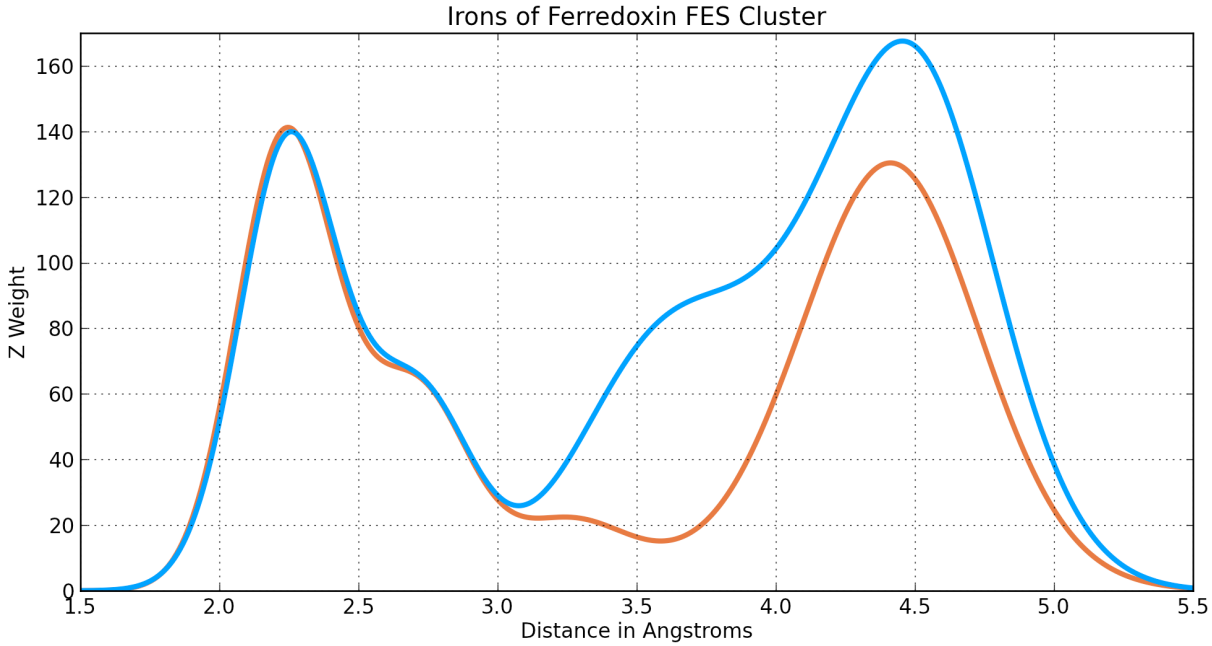
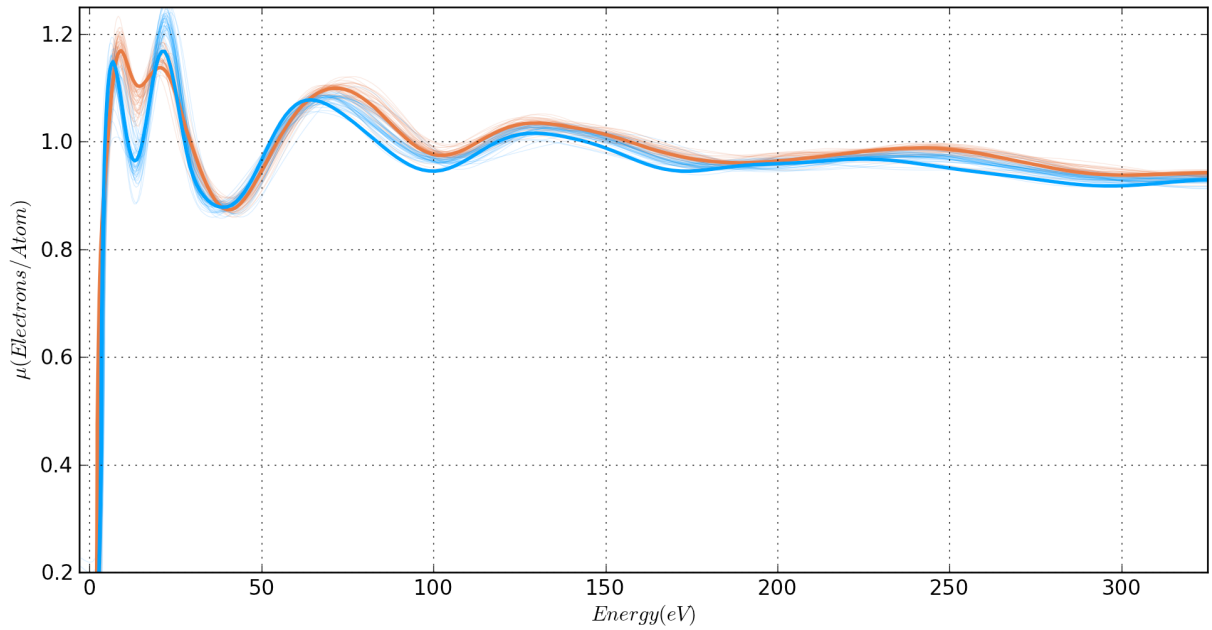


Figure A2.4



their oxidation states, are a good representation of all iron atoms throughout ferredoxins in the PDB. The oxidized iron is more of an outlier but this protein is only one captured with a reported Fe^{3+} in this subset of proteins.

APPENDIX III

FAST FOURIER TRANSFORM OF KRAMERS KRONIG IN PYTHON

fftkk.py

```
import os
import numpy as np
import matplotlib.pyplot as plt
from scipy.interpolate import interp1d
from scipy.interpolate import UnivariateSpline

def make_gauss(sig, mu):
    return lambda x: 1 / (sig * (2*np.pi)**.5) * np.e ** (-(x-mu)**2/(2 * sig**2))

def make_lrntz(gam, mu):
    return lambda x: gam / (np.pi*((x-mu)**2 + gam**2))

def make_voigt(sig, gam, mu, x):
    # Homemade.  A convolution of a Gauss and Lorentz function
    # interpolated back onto the original axis.
    gauss = make_gauss(sig, mu)(x)
    lrntz = make_lrntz(gam, mu)(x)
    voigt = np.convolve(gauss, lrntz)
    voigt /= abs(voigt.max())
    r = np.linspace(x[0], x[-1], len(voigt))
    voigt_intrp = interp1d(r, voigt, kind='linear')(x)
    return voigt_intrp

def get_file(fid):
    print fid
    energy, spectrum = [], []
    f = open(fid, 'r')
    for line in f.readlines():
        entry = line.split()
        energy.append(float(entry[0]))
        spectrum.append(float(entry[1]))
    return np.array(energy), np.array(spectrum)

def get_xmu(fid):
    print fid
    omega, e, k, mu, mu0, chi = [],[],[],[],[],[]
    f = open(fid, 'rU')
    for line in f.readlines()[1:]:
        if line.startswith('#'):
            continue
        entry = line.split()
        omega.append(float(entry[0]))
        e.append(float(entry[1]))
        k.append(float(entry[2]))
        mu.append(float(entry[3]))
        mu0.append(float(entry[4]))
        chi.append(float(entry[5]))
    return np.array(omega), np.array(mu)

def get_theo(element, grain, shift):
    print 'Retreaving element file,', element + '.dat'
    f = open('/Users/darrensherrell/Documents/das/EDGE/' + element + '.dat', 'r')
    FA = np.array([[float(line.split()[0]), float(line.split()[1]),
float(line.split()[2])] for line in f ])
    v  = FA[:,0] + shift
    f1 = FA[:,1]
    f2 = FA[:,2]
```

```
    theo_energy = np.arange(v[0], v[-1], grain)
    theo_fp = interp1d(v, f1, 'linear')(theo_energy)
    theo_fpp = interp1d(v, f2, 'linear')(theo_energy)
    return theo_energy, theo_fp, theo_fpp

def fftkk(dF_in, E, switch, grain):
    #Pad out the spectrum by 5000 points
    k = 5000
    E_dn = np.arange(E[0] - grain, E[0] - (k+1)*grain, -grain)
    E_dn = np.fliplr(E_dn.reshape(1, E_dn.shape[0])).reshape(E_dn.shape[0],)
    E_up = np.arange(E[-1] + grain, E[-1] + (k+1)*grain, grain)
    _E_ = np.hstack((E_dn, E, E_up))
    #Take edge of spectrum gently to zero using a quarter of a sine wave
    range = np.linspace(-np.pi/2, np.pi/2, k)
    dn = (np.sin(range) / 2) + 0.5
    up = 1 - dn
    dF_dn = dF_in[0] * dn
    dF_up = dF_in[-1] * up
    _dF_ = np.hstack((dF_dn, dF_in, dF_up))
    #This is based on fftkk.f by Graham George which
    #in turn is based on a paper.
    npts = _E_.shape[0]
    Hz = np.fft.fft(_dF_)
    mn = Hz[0]
    front = Hz[1:(len(Hz)/2)]
    tmp = front.copy()
    #I believe the next line is the magic.  Its the convolution with the signum or
    #how to bypass the difficult integration.
    tmp.real, tmp.imag = front.imag, front.real
    front = switch * tmp
    back = Hz[(len(Hz)/2):]
    tmp2 = back.copy()
    tmp2.real, tmp2.imag = back.imag, back.real
    back = -switch * tmp2
    new_Hz = np.hstack((mn, front, back))
    dF = np.fft.ifft(new_Hz).reshape((1, npts))
    dF_out = np.fliplr(dF).reshape((npts,))
    return dF_out[k:-k].real

def main(fid, element, shift, vert, scale, grain, switch):
    #Retrieve file energy and spectrum
    if fid.endswith('raw'):
        file_energy, file_spectrum = get_xmu(fid)
    if fid.endswith('edg'):
        file_energy, file_spectrum = get_file(fid)

    #The X-axis of energy that everything will be working from
    E = np.arange(file_energy[0], file_energy[-1], grain)

    #Scale and/or move spectrum vertically.  Interpolate to E
    file_spectrum = (scale * file_spectrum) + vert
    spectrum = interp1d(file_energy, file_spectrum, 'linear')(E)

    #Get theoretical values
    theo_energy, theo_fp, theo_fpp = get_theo(element, grain, shift)
    #Broaden theoretical values
    x = np.arange(-20.0, 20.0+grain, grain)
    sigma, gamma , mu= 2.0, 1.61, 0.0
    voigt = make_voigt(sigma, gamma, mu, x)
    theo_fp_cnv = np.convolve(voigt/voigt.sum(), theo_fp, 'same')
    theo_fpp_cnv = np.convolve(voigt/voigt.sum(), theo_fpp, 'same')
    #Chop theoretical values at incoming file values, because of pythons
    #list slicing its easier to add 1 to 'hi' at this stage
    lo = np.searchsorted(theo_energy, E[0])
    hi = np.searchsorted(theo_energy, E[-1]) + 1
    F1 = theo_fp_cnv[lo:hi]
    F2 = theo_fpp_cnv[lo:hi]

    #Start graph
    Figure1 = plt.Figure()
    Figure1.subplots_adjust(left   = 0.03,
                            bottom = 0.03,
                            right  = 0.97,
```

```
                        top    = 0.97)
    ax = Figure1.add_subplot(111, axisbg='beige')
    plt.title('FFT Kramers-Kronig of %s' %fid)
    plt.xlabel('Energy')
    plt.ylabel('Electrons')

    #switch to Hz, butterfly, switch back.
    if switch == 'DAFS':
        delta_fp = spectrum - F1
        delta_fpp = fftkk(delta_fp, E, -1, grain)
        fpp = F2 + delta_fpp
        fp = F1 + delta_fp    #same as 'spectrum'
        ax.plot(E, fpp, 'b', label='Calculated fpp')
    if switch == 'XAFS':
        delta_fpp = spectrum - F2
        delta_fp = fftkk(delta_fpp, E, 1, grain)
        fp = F1 + delta_fp
        fpp = F2 + delta_fpp  #same as 'spectrum'
        ax.plot(E, fp, 'b', label='Calculated fp')

    ax.plot(theo_energy, theo_fp, 'y', label='Cromer fp')
    ax.plot(theo_energy, theo_fpp, 'y', label='Cromer fpp')
    ax.plot(theo_energy, theo_fp_cnv, 'r', label='Convolution fp')
    ax.plot(theo_energy, theo_fpp_cnv, 'r', label='Convolution fpp')
    ax.plot(E, spectrum, 'k', label=fid)
    ax.plot(E, delta_fp, 'c', label='delta fp')
    ax.plot(E, delta_fpp, 'm', label='delta fpp')

    plt.xlim(theo_energy[0], theo_energy[-1])
    plt.xlim(6950, 7300)
    #plt.ylim(-10, 5)
    plt.legend(loc='best')

    fine_fp = np.hstack((theo_fp_cnv[:lo], fp, theo_fp_cnv[hi:]))
    fine_fpp = np.hstack((theo_fpp_cnv[:lo], fpp, theo_fpp_cnv[hi:]))
    coarse_range = np.arange(theo_energy[0], theo_energy[-1], 1)
    print coarse_range
    output_fp = interp1d(theo_energy, fine_fp, 'linear')(coarse_range)
    output_fpp = interp1d(theo_energy, fine_fpp, 'linear')(coarse_range)

    #output_fid = fid[:5] + fid[14:18] + '_fftkk.out'
    output_fid = fid[:-4] + '_fftkk.out'
    g = open(output_fid, 'w')
    for a, b, c in zip(coarse_range, output_fp, output_fpp):
        l = '\t'.join([str(a), str(b), str(c)]) + '\n'
        g.write(l)
    g.close()

if __name__ == '__main__':
    #####(fid,          element, shift, vert, scale, grain, switch)
    main('pf-rd-red.edg', 'Fe',    6.0,   0.5, 3.4,   0.05, 'XAFS')
    main('pf-rd-ox.edg',  'Fe',    6.0,   0.5, 3.4,   0.05, 'XAFS')
plt.show()
```

APPENDIX IV

PDB FILE TO FEFF.INP

pdb2FEFF.py

```
#!/usr/bin/python
import os
import re
import sys
import math
import glob
import datetime
import subprocess
import matplotlib.pyplot as plt

per_tab_dict = {'Ru': 44, 'Re': 75, 'Rf': 104, 'Rg': 111, 'Ra': 88, 'Rb': 37, \
'Rn': 86, 'Rh': 45, 'Be': 4, 'Ba': 56, 'Bh': 107, 'Bi': 83, 'Bk': 97, \
'Br': 35, 'Uuh': 116, 'H': 1, 'P': 15, 'Os': 76, 'Es': 99, 'Hg': 80, 'Ge': 32, \
'Gd': 64, 'Ga': 31, 'Uub': 112, 'Pr': 59, 'Pt': 78, 'Pu': 94, 'C': 6, 'Pb': 82, \
'Pa': 91, 'Pd': 46, 'Cd': 48, 'Po': 84, 'Pm': 61, 'Hs': 108, 'Uuq': 114, 'Uup': \
115, 'Uuo': 118, 'Ho': 67, 'Hf': 72, 'K': 19, 'He': 2, 'Md': 101, 'Mg': 12, \
'Mo': 42, 'Mn': 25, 'O': 8, 'Mt': 109, 'S': 16, 'W': 74, 'Zn': 30, 'Eu': 63, \
'Zr': 40, 'Er': 68, 'Ni': 28, 'No': 102, 'Na': 11, 'Nb': 41, 'Nd': 60, 'Ne': 10, \
'Np': 93, 'Fr': 87, 'Fe': 26, 'Fm': 100, 'B': 5, 'F': 9, 'Sr': 38, 'N': 7, \
'Kr': 36, 'Si': 14, 'Sn': 50, 'Sm': 62, 'V': 23, 'Sc': 21, 'Sb': 51, 'Sg': 106, \
'Se': 34, 'Co': 27, 'Cm': 96, 'Cl': 17, 'Ca': 20, 'Cf': 98, 'Ce': 58, 'Xe': 54, \
'Lu': 71, 'Cs': 55, 'Cr': 24, 'Cu': 29, 'La': 57, 'Li': 3, 'Tl': 81, 'Tm': 69, \
'Lr': 103, 'Th': 90, 'Ti': 22, 'Te': 52, 'Tb': 65, 'Tc': 43, 'Ta': 73, 'Yb': 70, \
'Db': 105, 'Dy': 66, 'Ds': 110, 'I': 53, 'U': 92, 'Y': 39, 'Ac': 89, 'Ag': 47, \
'Uut': 113, 'Ir': 77, 'Am': 95, 'Al': 13, 'As': 33, 'Ar': 18, 'Au': 79, 'At': 85,
\
'In': 49}

def off_with_their(head_fids):
    header_list = []
    for fid in head_fids:
        head = []
        f = open(fid, 'r')
        for line in f:
            if line.startswith('TITLE') or line.startswith('#'):
                continue
            if line.startswith('POTENTIALS'):
                break
            else:
                head.append(line)
        header_list.append(''.join(head))
        f.close()
    return header_list

def get_xmu(fid):
    omega, e, k, mu, mu0, chi = [],[],[],[],[],[]
    f = open(fid, 'rU')
    for line in f.readlines()[1:]:
        if line.startswith('#'):
            continue
        entry = line.split()
        omega.append(float(entry[0]))
        e.append(float(entry[1]))
        k.append(float(entry[2]))
        mu.append(float(entry[3]))
        mu0.append(float(entry[4]))
        chi.append(float(entry[5]))
```

```
    return omega, e, k, mu, mu0, chi

def ordered_set(seq):
   # order preserving
   checked = []
   for e in seq:
       if e not in checked:
           checked.append(e)
   return checked

def ASU(xtal_fid, target_elem):
    asu = []
    tag_list = []
    el_list = []
    f = open(xtal_fid, 'rU')
    for line in f:
       if line.startswith('ATOM') or line.startswith('HETATM'):
           sl = line.split()
           tag = '-'.join([sl[2], sl[1], sl[5], sl[4]])
           xyz_position, element = re.findall('([-]?\d+\.\d+)',line), \
                                   re.findall(r'.*(\b\w+[+-]?)', line)
           element[0] = element[0].title()
           xyz_position.pop(-1)
           if len(xyz_position) != 4:
               continue
           if element[0] == target_elem:
               tag_list.append(tag)
               asu.append([tag] + xyz_position + element)
           else:
               el_list.append(element[0])
               asu.append([tag] + xyz_position + element)
    f.close()
    el_list = list(set(el_list))
    return asu, tag_list, el_list

def write_FEFF(xtal_fid, target_elem, radius, asu, central_atoms_list, \
               head_fid, header):
    FEFF_fids_list = []
    for central_atom in central_atoms_list:
        output_fid = '_'.join([xtal_fid[:-4], target_elem, head_fid[:-5], \
                                              central_atom, 'FEFF.inp'])
        FEFF_atoms = []
        [[X, Y, Z]] = [[float(atom[1]), float(atom[2]), float(atom[3])] \
                       for atom in asu if atom[0] == central_atom]
        for atom in asu:
            tag_name = atom[0]
            x, y, z = float(atom[1]), float(atom[2]), float(atom[3])
            occu = float(atom[4])
            elem = atom[5]
            dist = math.sqrt((X-x)**2 + (Y-y)**2 + (Z-z)**2)
            new_ver = [tag_name, elem, x, y, z, occu, dist]
            if 0 <= dist <= radius:
                FEFF_atoms.append(new_ver)
        FEFF_atoms = sorted(FEFF_atoms, key=lambda a: a[6])

        elems = [x[1] for x in FEFF_atoms[1:]]
        elems = ordered_set(elems)

        g = open(output_fid, 'w')
        title0 = 'TITLE Date          : ' + str(datetime.date.today()) + '\n'
        title1 = 'TITLE Xtal PDB      : ' + xtal_fid + '\n'
        title2 = 'TITLE Target Element: ' + target_elem  + '\n'
        title3 = 'TITLE Central Atom  : ' + central_atom  + '\n'
        title4 = 'TITLE Radius        : ' + str(radius)  + '\n'
        title5 = 'TITLE Header File   : ' + head_fid + '\n'
        g.write(title0)
        g.write(title1)
        g.write(title2)
        g.write(title3)
        g.write(title4)
        g.write(title5)
        g.write(header)
        g.write('POTENTIALS\n')
```

```
130         ipot_line = '0 %i %s\n' %(per_tab_dict[target_elem], target_elem)
131         g.write(ipot_line)
132 
133         ipot_dict = {}
134         ipot_dict[target_elem] = 0
135         for i, elem in enumerate(elems):
136             ipot_dict[elem] = i+1
137             ipot_line = '%i %i %s\n' %(i+1, per_tab_dict[elem], elem)
138             g.write(ipot_line)
139 
140         g.write('ATOMS\n#   X       Y       Z    \tIpot Elem Occu  Dist  Tag\n')
141         for tag_name, elem, x, y, z, occu, dist in FEFF_atoms:
142             if tag_name == central_atom:
143                 line = '    %1.3f  %1.3f  %1.3f\t0    %s\t  %1.3f %1.3f %s\n' \
144                                     %(x, y, z, elem, occu, dist, tag_name)
145                 g.write(line)
146             else:
147                 line = '    %1.3f  %1.3f  %1.3f\t%i    %s\t  %1.3f %1.3f %s\n' \
148                         %(x, y, z, ipot_dict[elem], elem, occu, dist, tag_name)
149                 g.write(line)
150         g.write('END')
151         g.close()
152         FEFF_fids_list.append(output_fid)
153     return FEFF_fids_list
154 
155 def plot(xtal_fid):
156     Figure = plt.Figure()
157     Figure.subplots_adjust(left   = 0.05,
158                            bottom = 0.05,
159                            right  = 0.95,
160                            top    = 0.95,
161                            wspace = 0.00,
162                            hspace = 0.00)
163     ax = Figure.add_subplot(111, axisbg='beige')
164     prefix = xtal_fid[:4]
165     fid_list = glob.glob('%s*xmu*' %prefix)
166     for i, fid in enumerate(sorted(fid_list)):
167         omega, e, k, mu, mu0, chi = get_xmu(fid)
168         ax.plot(e, mu, lw=0.75, label=fid)
169         #ax.plot(e, mu0, color=colors[i], lw=0.75)
170         ax.plot(e, chi, lw=0.75)
171     plt.xlim(min(e)-15, max(e)+15)
172     #leg = plt.legend(loc='best', fancybox=True)
173 
174 def main(head_fids = None, \
175          xtal_fids = None, \
176          target_elems = None, \
177          radius = None, \
178          choices = None):
179 
180     #Head files
181     heads_list = [x for x in os.listdir('.') if x.endswith('head')]
182     head_dict = {}
183     for i, fid in enumerate(heads_list):
184         head_dict[i+1] = fid
185         print i+1, fid
186     if head_fids:
187         print head_fids
188     else:
189         head_choice = raw_input('Choose header, headers separated by a space or
"all": ').rstrip(' ')
190         if 'all' in head_choice:
191             head_fids = heads_list
192         else:
193             head_fids = [head_dict[int(x)] for x in head_choice.split(' ')]
194     header_list = off_with_their(head_fids)
195 
196     #Files
197     PDB_fids = [x for x in os.listdir('.') if x.endswith('PDB')]
198     PDB_fids_dict = {}
199     for i, PDB_fid in enumerate(PDB_fids):
200         PDB_fids_dict[i+1] = PDB_fid
201     for k, v in PDB_fids_dict.items():
```

```
202         print k, v
203     if xtal_fids:
204         print '\n\nFile   ', xtal_fids[:-4]
205     else:
206         xtal_choice = raw_input('Choose file, files separated by a space or "all":
').rstrip(' ')
207         if 'all' in xtal_choice:
208             xtal_fids = PDB_fids
209         else:
210             xtal_fids = [PDB_fids_dict[int(x)] for x in xtal_choice.split(' ')]
211         print xtal_fids
212 
213     #Element
214     print
215     if target_elems:
216         print 'Target(s) ', target_elems
217     else:
218         maybes = list(set([a.split()[-1] for a in \
219                            open(xtal_fids[0], 'r').readlines()
if a.startswith('HETATM')]))
220         for fid in xtal_fids:
221             listy = list(set([a.split()[-1] for a in \
222                               open(fid, 'r').readlines()
if a.startswith('HETATM')]))
223             print fid, listy
224             maybes = list(set(listy) & set(maybes))
225         maybes = [x.title() for x in maybes]
226         print 'Common elements', maybes
227         target_elem = raw_input('\nType target element or elements separated by
space: ').rstrip(' ')
228         target_elems = [x.title() for x in target_elem.split(' ')]
229     print 'Target(s) ', target_elems
230 
231 
232     #Go get atoms from file, unique element list, individual tags for each atom
233     asu_dict = {}
234     cent_atms_dict = {}
235     for xtal_fid in xtal_fids:
236         for j, target_elem in enumerate(target_elems):
237             print '\n---------------------->', xtal_fid, target_elem
238             asu, tag_list, el_list = ASU(xtal_fid, target_elem)
239 
240             tag_dict = {}
241             for i, tag in enumerate(tag_list):
242                 tag_dict[i+1] = tag
243 
244             if choices:
245                 if choices[j] == 'all':
246                     central_atoms_list = tag_dict.values()
247                 else:
248                     central_atoms_list = [tag_dict[x] for x in choices[j]]
249             else:
250                 for k, v in tag_dict.items(): print k, v
251                 choices = raw_input('Select the central atom or atoms separated by
space or type "all": ').rstrip(' ')
252                 if 'all' in choices:
253                     central_atoms_list = tag_dict.values()
254                     choices = None
255                 else:
256                     central_atoms_list = [tag_dict[int(x)] for x
in choices.split(' ')]
257                     choices = None
258             asu_dict[xtal_fid] = asu
259             cent_atms_dict[xtal_fid + target_elem] = central_atoms_list
260 
261     #Radius
262     if radius:
263        print 'Radius ', radius
264     else:
265         radius = float(raw_input('\nChoose radius from central atom (in
Angstroms): '))
266 
267     verbose_fid = 'FEFFlog.data'
```

```
268     screen_out = open(verbose_fid, 'w')
269
270     FEFF_choice = raw_input('\nWould you like to try and run FEFF8 on these atoms
(y/n) ? ')
271     if FEFF_choice == 'y':
272         for xtal_fid in xtal_fids:
273             asu = asu_dict[xtal_fid]
274             for target_elem in target_elems:
275                 central_atoms_list = cent_atms_dict[xtal_fid + target_elem]
276                 for head_fid, header in zip(head_fids, header_list):
277                     print '\n', head_fid
278                     FEFF_fids_list = write_FEFF(xtal_fid, target_elem, radius, \
279                                                 asu, central_atoms_list, head_fid,header)
280                     for FEFF_fid in FEFF_fids_list:
281                         print '-', FEFF_fid
282                         prefix =  FEFF_fid[:-8]
283                         subprocess.call(['cp', FEFF_fid, 'FEFF.inp'])
284                         retcode = subprocess.call(['FEFF8', 'FEFF.inp'], \
285                                 stdout=screen_out, stderr=subprocess.STDOUT)
286                         subprocess.call(['mv','xmu.dat', prefix + 'xmu.data'])
287                         del_list = glob.glob('*dat') + glob.glob('ldos*') + \
288                    glob.glob('FEFF.inp') + glob.glob('*.bin') + glob.glob('mod*')
289                         for fid in del_list:
290                             subprocess.call(['rm', fid])
291         #Plot
292         pq = raw_input('\nWould you like to plot? (y/n) ')
293         if pq == 'y':
294             plot(xtal_fid[:4])
295         else:
296             print 'Ok, thats cool too.'
297
298     else:
299         for xtal_fid in xtal_fids:
300             asu = asu_dict[xtal_fid]
301             for target_elem in target_elems:
302                 central_atoms_list = cent_atms_dict[xtal_fid + target_elem]
303                 for head_fid, header in zip(head_fids, header_list):
304                     print '\n', head_fid
305                     FEFF_fids_list = write_FEFF(xtal_fid, target_elem, radius, \
306                                                 asu, central_atoms_list, head_fid, header)
307                     for FEFF_fid in FEFF_fids_list:
308                         print '-', FEFF_fid
309
310     print '\n', 60*'-', '\n\n'
311
312 if  __name__ == '__main__':
313     #Normal mode
314     main()
```

BIBLIOGRAPHY

1. Friedrich, W and von Laue, M , Annalen der Physik, 345 (1913) 971-988
2. Hendrickson, W.A., Trans. Am. Crystallogr. Assoc., 21 (1985)
3. Judith Flippen-Anderson, RCSB PDB Statistics, The Protein Databank, 01 July 2013. Web. http://www.pdb.org/pdb/static.do?p=general_information/pdb_statistics/index.html
4. Stragier, H., Cross, J. O., Rehr, J. J., Sorensen, L. B., Bouldin, C. E., Woicik., J. C., Phys. Rev. Lett., 69 (1992) 3064-4067
5. Pickering, I. J., Sansone, M, Marsch, J., George, G. N., J. Am. Chem. Soc., 115 (1993) 6302-6311
6. Vacinova, J., Hodeau, J. L., Wolfers, P., Lauriat, J. P., ElKaim, E., J. Sync. Rad., 2 (1995) 236-244
7. Yeh, A.P., Ambroggio, X. I., Andrade, S. L. A., Einsle, O., Chatelet, C., Meyer, J., Rees, D.C., J. Bio. Chem., 277 (2002) 34499
8. Einsle, O., Andrade, S. L. A., Dobbek, H., Meyer, J., Rees, D.C., J. Am. Chem. Soc. 129 (2007) 2210-2211
9. Rupp, B., Biomolecular Crystallography (2009) Garland Science
10. Cotelesage, J. J. H., Pushie M. J., Grochulski, P., Pickering, I. J., George, G. N., J. Inorg. Biochem. 115 (2012) 127-137
11. Sayers, D.E., Stern, E.A., Lytle, F. W., Phys. Rev. Let., 27 (1971) 1204-1207
12. Pickering, I. J. and George, G. N., In. Chem., 34 (1995) 3142-3152
13. De Groot, F. M. F., Fuggle, J. C., Thole, B. T., Sawatzky, GA, Phys. Rev. B, 42 (1990) 5459
14. Pickering, I. J., George, G. N., Yu, E. Y., Brune, D. C., Tuschak, C., Overmann, J., Beatty, J. T., Prince, R. C., Biochem. 40 (2001) 8138-8145
15. Korbas, M., Blechinger, S. R., Krone, P. H., Pickering, I. J., George, G. N., Proc. Nat. Ac. Sci., 105 (2008)
16. Bragg, W. L., Proc. Cam. Phil. Soc., 17 (1913) 43-57

17. Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., Yonath, A., Cell 107 (2001) 679-688
18. Berman, H. M., Westbrook, J., Feng, Z, Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Philip E. B., Query, The Protein Databank, 05 Jan 2014. Web. http://www.pdb.org/pdb/results/results.do?qrid=9B937A17&tabtoshow=Current
19. Ewald, P. P., Acta. Cryst. Sec. A, 25 (1969) 103-108
20. Toll, J. S., Phys. Rev., 104 (1956) 1760
21. Lambert, J. H., Eberhardt Klett (1760)
22. Beer, J, Annalen der Physik und Chemie, 86 (1852) 78-88
23. Templeton, D. H., Handbook of Synchrotron Radiation, Vol 3 (1991) 205
24. Cross, J. O., Analysis of Diffraction Anomalous Fine Structure, Dissertation (1996)
25. Baym, G., Lectures on Quantum Mechanics, 3rd Printing (1974) 251
26. Griffiths, D. J., Introduction to Quantum Mechanics, (1995) WA Benjamin New York
27. Stragier, H. J., DAFS: A New Structural Technique, Dissertation (1993)
28. Sorensen, L. B., Cross, J. O., Newville, M., Ravel, B., Rehr, J. J., Straiger, H., Bouldin, C. E., Woicik, J. C., Res. Anom. X-Ray Scat. Theo. App. (1994) 389-420
29. Nave, R., Transition Probablilities and Fermi's Golden Rule, HyperPhysics, (2004), Web. 10 Mar 2014, http://hyperphysics.phy-astr.gsu.edu/hbase/quantum/fermi.html
30. Tuckerman, M. E., Fermi's Golden Rule, Stat. Mech. II, 28 Apr. 2004, Web. 10 Mar 2014, http://www.nyu.edu/classes/tuckerman/stat.mechII/lectures/lecture_16/node4.html
31. Cross, J. O., Analysis of Diffraction Anomalous Fine Structure, Dissertation (1996)
32. Cromer, D. T. and Liberman, D., J. Chem. Phys., 53 (1970) 1891
33. Brennan, S. and Cowan, P. L., Rev. Sci. Inst., 63 (1992) 850-853
34. Evans, G. and Pettifer, R. F., J. App. Cryst. 34 (2001) 82-86
35. Kittle, C., Introduction to Solid State Physics, 4th Ed. (1971) 56-60
36. James, R. W. The Crystalline State, Vol II, (1965) *Editor: Sir Lawrence Bragg*
37. Stout, G. H., Jensen, L. H., X-ray Structure Determination, 2nd Ed. (1989) 205
38. Renevier, H., Grenier, S., Arnaud, S., Berar, J. F., Caillot, B., Hondeau, J. L., Letoublon, A., Proietti, M. G., Ravel, J. Sync. Rad., 10 (2003) 435-444

39. Pickworth, G., Trueblood, K. N., Crystal Structure Analysis, 2nd Ed. (1985) Oxford University Press
40. Buerger, M. J., Proc. Nat. Ac. Sci., 26 (1940) 637
41. Drenth, J., Principles of Protein X-Ray Crystallography, 2nd Ed. (1999) Springer
42. Cullity, B. D., Stock, S. R., Elements of X-ray Diffraction, 2nd Ed. (1978) Adderson-Wesley Publishing Company
43. Azaroff, L. V., Acta Cryst. 8 (1955) 701-704
44. Azaroff, L. V., Kaplow, R., Kato, N., Weiss, R. J., Wilson, A. J. C., Young, R. A., X-ray Diffraction (Pure & Applied Physics) (1974) McGraw-Hill New York
45. Gwyndaf, E., The Method of Multiple Wavelength Anomalous Diffraction using Synchrotron Radiation at Optimal X-ray Energies: Application to Protein Crystallography, University of Warwick, Dissertation (1994)
46. Kahn, R., Fourme, R., Gadet, A., Janin, J., Dumas, C., Andre, D., J. App. Cryst. 15 (1982) 330-337
47. Templeton, D. H., Templeton, L. K., Acta Cryst. A36 (1980) 237-241
48. Templeton, D. H., Templeton, L. K., Acta Cryst. A38 (1982) 62-67
49. Templeton, D. H., Templeton, L. K., Acta Cryst. A41 (1985) 133-142
50. Pickering, I. J., George, G. N., Inorg. Chem. 34 (1995) 3142-3152
51. Westbrook, E., Naday, I., Meth. Enzym. 276 (1997) 244-368
52. Newville, M., J. Synch. Rad. 8 (2001) 96-100
53. Kabsch, W., Acta Cryst. Sec. D 66 (2010) 125-132
54. Thompson, A., Attwood, D., et al, X-ray Data Booklet 3rd Ed. (2009) 44
55. Creagh, D. C., McAuley, W. J., International Tables of Crytallography, Vol. C (1999) 242-258, Kluwer Academic Publishers
56. Cross, J. O., Newville, M., Rehr, J. J., Sorensen, L. B., Bouldin, C. E., Watson, G., Gouder, T., Lander, G. H., Bell, M. I., Phys. Rev. B 58 (1998) 215-219
57. Stuhrmann, S., Bartels, K. S., Braunwarth, W., Doose, R., Dauvergne, F., Gabriel, A., Knochel, A., Marmotti, M., Stuhrmann, H. B., Trame, C., J. Sync. Rad. 4 (1997) 298-310
58. Dean, R. B., Dixon, W. J., Ana. Chem. 23 (1951) 636-638

59. Bose, M. L., Mathematical Methods in the Physical Sciences, 2nd Ed. (1983) John Wiley & Sons
60. Templeton, L. K., Templeton, D. H., J. Appl. Cryst. 21 (1988) 558-561
61. Morales, R., Frey, M., Mouesca, J-M., J. Am. Chem. Soc. 124 (2002) 6714-7722
62. Malinowski, E. R., Howery, D. G., Factor Analysis in Chemistry (1980) John Wiley & Sons
63. Marsland, S., Machine Learning: An Algorithmic Perspective (2009) CRC Press
64. Dugad, L. B., La Mar, G. N., Banci, L., Bertini, I., Biochem 29 (1990) 2263-2271
65. Pflugrath, J. W., Acta Cryst. Sec. D 55 (1999) 1718-1725
66. George, G. N., Pickering, I. J., Stan. Synch. Rad. Lab. (1995)
67. Kraft, P., Bergamaschi, A., J., Broennimann, C. H., Dinapoli, R., Eikenberry, E. F., Henrich, B., Johnson, I., Mozzanica, A., Schleputz, C. M., Willmott, P. R., et al., Sync. Rad. 16 (2009) 368-375
68. Thompson, A. C., Westbrook, E. M., Nix J. C., J. Phys. Conf. (2013) Ser. 425
69. Springer, B. A., Sligar, S. G., Proc. Nat. Ace. Sc. 84 (1987) 8961-8965
70. McPhillips, T. M., McPhillips, S. E., Chiu, H-J, Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerly, R. P., J. Synch. Rad. 9 (2002) 401-406
71. Cohen, A. E., Ellis, P., Kresge, N., Soltis, S. M., Acta. Cryst. Sec. D 57 (2001) 233-238
72. Smith, C. A., Card, G. L., Cohen, A. E., Doukov, T. I., Eriksson, T., Gonzalez, A. M., McPhillips, S. E., Dunten, P. W., Mathews, I. I., Song, J., et al, J. App. Cryst. 43 (2010) 1261-1270
73. Yeh, A. P., Ambroggio, X. I., Andrade, S. L. A., Einsle, O., Chatelet, C., Meyer, J., Rees, D. C., Biol. Chem. 277 (2002) 34499-34507
74. Ellis, P. J., Cohen, A. E., Soltis, S. M., J. Synch. Rad. 10 (2003) 287-288
75. Newville, M., Cross, J. O., DAFS Bibliography, Diffraction Anomalous Fine Structure, 12 July 2008, Web, 10 Mar 2014, http://cars9.uchicago.edu/dafs/bib/
76. Cohen, A. E., Ellis, P., Kresge, N., Soltis, S. M., Acta. Cryst. Sec. D 57 (2001) 233-238
77. DeLano, W. L., The Pymol Molecular Graphics System (2002) Software