

© Copyright 2023

Samantha Tetef

Information Content and Analysis of X-ray Absorption Spectroscopy  
and X-ray Emission Spectroscopy Using Machine Learning

Samantha Tetef

A dissertation

submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy

University of Washington  
2023

Reading Committee:  
Gerald T. Seidler, Chair  
Joshua J. Kas  
Miguel Morales

Program Authorized to Offer Degree:  
Physics

University of Washington

## **Abstract**

# Information Content and Analysis of X-ray Absorption Spectroscopy and X-ray Emission Spectroscopy Using Machine Learning

Samantha Tetef

Chair of the Supervisory Committee:

Professor Gerald T. Seidler

Physics

Data science and machine learning (ML) methods are revolutionizing scientific analysis and data processing. As a case in point, ML applied to X-ray spectroscopies has recently exploded, showcasing its effectiveness in fields such as electrical energy storage and chemical catalysis. Here, I include comprehensive computational studies of ML techniques applied to X-ray spectra, including X-ray absorption near edge structure (XANES) and valence-to-core X-ray emission spectra (VtC-XES). First, I utilized unsupervised ML to extract import chemical fingerprints and information content in sulfororganics and phosphorganics. We compared different unsupervised ML techniques, namely principal component analysis (PCA), variational autoencoder (VAE), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP). Additionally, we developed open-source tools for future researchers to utilize,

including an API that interacts with PubChem to efficiently download and store metadata. Next, I used ML to improve the reliability of data analysis and decrease computational time in the context of imaging XANES experiments. To do so, we utilized dimensionality reduction and clustering to perform image segmentation and then identified phase composition using linear combination fitting. By decoupling the domain identification from the phase identification, we provided a more robust way to handle noise that was not reliant on obtaining appropriate linear combination fitting results. Finally, I used feature selection to speed up high-throughput experimental design. Specifically, we used Recursive Feature Elimination to select the most important energies in XANES spectra to measure in the context of a reference library. Then we demonstrated appropriate analysis on these reduced-energy spectra using a nano-XANES image. All these approaches and tools are broadly applicable to X-ray spectroscopy on other systems or using other spectroscopy techniques.

# TABLE OF CONTENTS

|  |      |
|--|------|
| Acknowledgements   | v    |
| List of Acronyms   | vi   |
| List of Figures  | vii  |
| List of Tables   | xxvi |
| Chapter 1 – Introduction and Overview of X-ray Spectroscopy            | 1    |
| 1.1    X-ray Absorption Spectroscopies                                 | 2    |
| 1.1.1    K-, L-, and M- Edges  | 5    |
| 1.1.2    Regions of XAS spectra: XANES                                 | 6    |
| 1.1.3    XANES Processing  | 7    |
| 1.1.4    Regions of XAS spectra: EXAFS                                 | 9    |
| 1.1.5    EXAFS Processing  | 9    |
| 1.1.6    Wavelets  | 12   |
| 1.2    X-ray Emission Spectroscopy                                     | 16   |
| 1.3    Are XANES and VtC-XES Complimentary?                            | 20   |
| 1.4    References  | 22   |
| 2    Chapter 2 – Survey of Theoretical Methods                         | 24   |
| 2.1    Density Functional Theory (DFT)                                 | 24   |
| 2.2    Time-Dependent Density Functional Theory (TDDFT)                | 25   |
| 2.3    Implementations of DFT and TDDFT                                | 26   |
| 2.4    NWChem: A Closer Look   | 27   |
| 2.4.1    XANES Calculations  | 27   |
| 2.4.2    VtC-XES Calculations  | 29   |
| 2.5    References  | 32   |
| 3    Chapter 3 – Interpreting XAFS and the Bane of the Inverse Problem | 34   |
| 3.1    Classical Inverse Problems                                      | 35   |
| 3.2    Goals of the XAFS Inverse Problems                              | 36   |
| 3.3    Alternatives to the Inverse Problem in XAFS                     | 36   |
| 3.3.1    Repeating the Forward Problem (XANES & EXAFS)                 | 36   |
| 3.3.2    Bayesian Approach (EXAFS)                                     | 37   |

|       |  |    |
|-------|--|----|
| 3.3.3 | Linear Combination Fitting (LCF) to a reference library (XANES)                          | 39 |
| 3.3.4 | Machine Learning   | 42 |
| 3.4   | References   | 44 |
| 4     | Chapter 4 – Introduction to Machine Learning   | 45 |
| 4.1   | Machine Learning Basics  | 45 |
| 4.1.1 | Bayes Theorem and Maximum Likelihood Estimation  | 45 |
| 4.1.2 | Bias-variance Tradeoff and Model Complexity  | 47 |
| 4.2   | Supervised Machine Learning  | 50 |
| 4.2.1 | Regression   | 50 |
| 4.2.2 | Classification   | 53 |
| 4.2.3 | Models for Both Regression and Classification  | 54 |
| 4.2.4 | Interpretability Versus Effectiveness  | 60 |
| 4.2.5 | Metrics  | 61 |
| 4.2.6 | Uncertainty Estimation   | 63 |
| 4.3   | Unsupervised Machine Learning  | 66 |
| 4.3.1 | Dimensionality Reduction   | 67 |
| 4.3.2 | Clustering   | 71 |
| 4.3.3 | Feature Selection  | 74 |
| 4.4   | References   | 76 |
| 5     | Chapter 5 – A Survey of New Developments Between X-ray Spectroscopy and Machine Learning | 78 |
| 5.1   | Solving the Inverse Problem  | 78 |
| 5.1.1 | Supervised Machine Learning Approaches   | 78 |
| 5.1.2 | Unsupervised machine learning approaches   | 85 |
| 5.2   | Solving the forward problem  | 86 |
| 5.3   | References   | 88 |
| 6     | Chapter 6 – Information Content of VtC-XES versus XANES spectra of Sulforganics          | 89 |
| 6.1   | Introduction   | 90 |
| 6.2   | Methods  | 96 |
| 6.2.1 | Electronic Structure Calculations  | 96 |
| 6.2.2 | Supervised ML Methods  | 98 |

|         |   |     |
|---------|---|-----|
| 6.2.3   | Unsupervised ML Methods   | 100 |
| 6.3     | Dimensionality Reduction Algorithms                                       | 101 |
| 6.4     | Results and Discussion  | 107 |
| 6.4.1   | Dataset and Dimensionality Reduction                                      | 107 |
| 6.4.1.1 | Principal Component Analysis  | 108 |
| 6.4.1.2 | Variational Autoencoder   | 113 |
| 6.4.1.3 | t-SNE, Clustering Without Mapping   | 120 |
| 6.4.2   | Classification  | 123 |
| 6.4.3   | Summary and Outlook   | 126 |
| 6.5     | Conclusions   | 129 |
| 6.6     | References  | 132 |
| 6.7     | Supporting Information  | 137 |
| 6.7.1   | Explanation of VAE loss   | 137 |
| 6.7.2   | Increasing latent space dimension   | 139 |
| 6.7.3   | Hyperparameter tuning   | 140 |
| 6.7.4   | FastICA, FA, and NMF  | 141 |
| 6.7.5   | References  | 155 |
| 7       | Chapter 7 – Clustering and Classification of Organophosphates             | 156 |
| 7.1     | Introduction  | 157 |
| 7.2     | Methods   | 162 |
| 7.3     | Results and Discussion  | 165 |
| 7.3.1   | Unbiased verification of heuristic classes                                | 166 |
| 7.3.2   | Emergent chemical fingerprints from clusters                              | 174 |
| 7.3.3   | Validation of chemical fingerprints from cluster analysis                 | 178 |
| 7.4     | Conclusions   | 180 |
| 7.5     | References  | 184 |
| 7.6     | Supplementary Information   | 188 |
| 7.6.1   | References  | 217 |
| 8       | Chapter 8 – Manifold Projection Image Segmentation for Nano-XANES Imaging | 218 |
| 8.1     | Introduction  | 219 |
| 8.2     | Methods   | 221 |

|       |   |     |
|-------|---|-----|
| 8.2.1 | Experimental Methods  | 221 |
| 8.2.2 | Computational Methods   | 222 |
| 8.3   | Results and Discussion  | 223 |
| 8.4   | Conclusions   | 233 |
| 8.5   | References  | 234 |
| 8.6   | Supplementary Information   | 236 |
| 8.6.1 | Linear combination fitting objective function                                 | 236 |
| 9     | Chapter 9 – Recursive Feature Elimination for Nano-XANES Imaging              | 245 |
| 9.1   | Introduction  | 246 |
| 9.2   | Methods   | 248 |
| 9.3   | Results and Discussion  | 250 |
| 9.3.1 | Recursive Feature Elimination (RFE) Training, Recommendations, and Validation | 250 |
| 9.3.2 | Reliability of Inferences Using Measurements Chosen by RFE                    | 255 |
| 9.4   | Conclusions   | 261 |
| 9.5   | References  | 262 |
| 9.6   | Supplementary Information   | 263 |

## Acknowledgements

Thank you to my advisor, Professor Jerry Seidler, who supported me in countless ways, and to the amazing people in the Seidler Lab for all your support, specifically Diwash Dhakal, Jared Abramson, Charles Cardot, Helen Chen, and Anthony Gironda. You all helped make work more enjoyable, even during the roughest patches, and provided invaluable insight and assistance.

Also, a special thank you to my collaborators and mentors, especially Dr. Niri Govind, who taught me everything I know about NWChem and DFT; Dr. Maria Chan, who guided me through machine learning problems and exposed me to the other amazing research going on in her group; and Danica Hendrickson, Ed.D., who taught me so much about education and communication with youth through my role as a MEM-C Education and Training Fellow.

I also want to recognize Tharindu Fernando and the amazing group of students in STEM Pals and their never-ending enthusiasm for outreach and science in general. I learned so much, and seeing your passions was truly inspiring. I am honored to have been able to participate so much in Tharindu's creation and I am excited to see all the cool events STEM Pals will continue to do, especially under guidance of Nicole Gregorio, Al Snow, and Ariana Frey. STEM Pals was a large part of my PhD experience and helped me stay excited about science.

Finally, thank you to all my friends and family for your never-ending love and support. I could not have done this without you all, especially my partner Dr. Alex Prossnitz. There are honestly too many people to list here, but to name a few – my parents Sue and Herb Tetef, my brother Tyler Tetef, and my friends Kaila Martin, Ari Hasbrouck, Jordan Fonseca, Adina Ripin, Jeremy Hartse, David Bell, Tharindu Fernando, Robert Pecoraro, Heather Harrington, Mia Kumamoto, David Sharp, and Emma Johnston. You know how amazing each of you are! I love you all.

# List of Acronyms

*XAS* – X-ray Absorption Spectroscopy

*XAFS* – X-ray Absorption Fine Structure

*XANES* – X-ray Absorption Near Edge Structure

*EXAFS* – Extended X-ray Absorption Fine Structure

*XPS* – X-ray Photoelectron Spectroscopy

*XRD* – X-ray Diffraction

*HERFD* – High Energy Resolution Fluorescence Detected

*RIXS* – Resonant Inelastic X-ray Scattering

*XES* – X-ray Emission Spectroscopy

*IUPAC* – International Union of Pure and Applied Chemistry

*CtC-XES* – Core to Core X-ray Emission Spectroscopy

*VtC-XES* – Valence to Core X-ray Emission Spectroscopy

*ML* – Machine Learning

*PCA* – Principal Component Analysis

*SVD* – Singular Value Decomposition

*VAE* – Variational Autoencoder

*t-SNE* – t-distributed Stochastic Neighbor Embedding

*UMAP* – Uniform Manifold Approximation and Projection

*RFE* – Recursive Feature Elimination

*SVM* – Support Vector Machine

*NMF* – Non-negative Matrix Factorization

# List of Figures

|  |    |
|--|----|
| <b>Figure 1.1</b> The cross section of various interaction processes of X-rays for carbon. Taken from the X-ray Data Booklet .....   | 2  |
| <b>Figure 1.2</b> Absorption coefficient for XAS. Taken from Rehr and Albers. [Rehr, 2000 #95] .....   | 3  |
| <b>Figure 1.3</b> The Auger-Meitner [Matsakis, 2019 #255;Meitner, 1922 #256] effect is a two-electron process, where an inner-shell electron falls to fill the core hole, thus emitting a photon, but the rest of the extra energy is dissipated by emitting a valence electron.....     | 5  |
| <b>Figure 1.4</b> Absorption edge of an element can be broken down into <i>K</i> , <i>L</i> , and <i>M</i> edges. Taken from Wikipedia [, #280]. .....   | 6  |
| <b>Figure 1.5</b> The XANES and EXAFS regions of XAS spectra. Taken from cei.washington.edu [, #281]. .....  | 7  |
| <b>Figure 1.6</b> An example of the normalization process for XAFS, where the pre-edge line is set to be along the $y = 0$ line, and the post edge line is set to be along the $y = 1$ line such that the edge step $\Delta\mu$ is one. Taken from Newville. [Newville, 2014 #246] ..... | 9  |
| <b>Figure 1.7</b> Smoothed experimental $\chi(E)$ EXAFS data for (a) crystalline Ge and (b) amorphous Ge. Only the oscillatory part $\chi$ of the absorption edge is shown. Figure taken from Sayers, Stern, and Lytle. [Sayers, 1971 #241].....   | 11 |
| <b>Figure 1.8</b> Fourier transform of the data in Figure 1.7. $\varphi(r)$ , a radial structure function, compares amorphous and crystalline Ge. Numbers over the peaks indicate the measured distances in Å. Figure taken from Sayers, Stern, and Lytle. [Sayers, 1971 #241] .....     | 12 |
| <b>Figure 1.9</b> Wavelet analysis can separate contributions based on the atomic number Z of the scatterer. Taken from Munoz, et al. [Muñoz, 2003 #247] .....   | 13 |

**Figure 1.10** Wavelet analysis is like Fourier transforming using different frequency (or conversely time) sampling rates to create a multiresolution transform. Figure from ML fundamentals [, #284].

..... 14

**Figure 1.11** Wavelet U-Acet, U-Form and U-Glyc within the range  $r = 3.2 - 4.1 \text{ \AA}$ . Taken from Funke, et al., 2005 [Funke, 2005 #283]. ..... 16

**Figure 1.12** Conventional naming of florescence lines. From Glatzel and Bergmann [Glatzel, #58]. ..... 17

**Figure 1.13** The Mn K-fluorescence lines for MnO. From Glatzel and Bergmann [Glatzel, #58].  
..... 18

**Figure 1.14** Molecular orbital perspective of  $K\beta$  and its satellite lines. Here the M represent the target metal and the L is the ligand. From Pollock, et al., 2013 [Pollock, 2013 #257]. ..... 19

**Figure 1.15** VtC-XES probes occupied electronic states while XANES probes the unoccupied electronic states. .... 21

**Figure 2.1** Experimental and simulated Ru L<sub>3</sub> edge XANES spectra (2p - 4d orbitals). Taken from Biasin, et al. [Biasin, 2021 #267] and Nascimento and Govind [Nascimento, 2022 #264]. ..... 29

**Figure 2.2** Experimental and simulated Ru VtC-XES spectra (4d - 2p orbitals). Taken from Biasin, et al. 2021 [Biasin, 2021 #267] and Nascimento and Govind, 2022 [Nascimento, 2022 #264]. . 30

**Figure 3.1** Can you hear the shape of a drum? These two drum shapes, although different, produce the same eigenvalues, as indicated by the  $\lambda$  values. Figure from Wikipedia [, #309]. ..... 35

**Figure 3.2** If there is a good enough model, then repeating the forward problem to obtain target observations, or spectra, from different input structures is one way to combat the inverse problem. .... 37

**Figure 3.3** Bayes analysis involves specifying a prior probability distribution and formalizing the probability of observations (evidence) and thus the probability of the model given the evidence (posterior). From Analytics Vidhya [, #310]. ..... 38

**Figure 3.4** The top panel shows the experimental data  $\chi(k)$  and respective errors. The solid (red) curve is the most probable curve resulting from the fit. The bottom panel shows the prior and post values of  $\chi_{\mu0,L3}$  as well as the *a posteriori* error band. Taken from Krappe and Rossner. [Krappe, 2009 #77] ..... 39

**Figure 3.5** Linear combination fitting XANES spectra to reference spectra is the most common analysis for XANES. However, choosing reference spectra must be done well or any inferences are unreliable..... 40

**Figure 3.6** Linear combination fitting to phosphorus XANES spectra. Taken form Werner and Prietzel. [Werner, 2015 #263]..... 41

**Figure 3.7** Machine learning can encode the structure to spectra relationship into the training set and this the model can learn how to invert that relationship. ..... 43

**Figure 4.1** Model complexity and total error – the bias versus variance trade-off. ..... 48

**Figure 4.2** The bias-variance tradeoff impacts the generalizability of the model..... 49

**Figure 4.3** Linear regression is finding a line (or hyperplane) that most follows the linear trends in the data, and as a bonus, it has an analytical solution..... 50

|   |    |
|---|----|
| <b>Figure 4.4</b> The regularization term affects the sparsity of the solution. From Penn State Statistics [ , #292].....   | 52 |
| <b>Figure 4.5</b> The SVM maximizes the distance from the decision boundary between all data points <i>and</i> it maximizes the margin around this decision boundary. The margin ensures that the solution is unique.....   | 54 |
| <b>Figure 4.6</b> A decision tree for a certain objective, in this case determining whether to take a walking break outside, is composed of decision nodes (composed of a question about the data), branches (which depend on the answer to the question, which are typically interpreted as a “yes” or “no” answers), and leaf nodes (the target variables). ..... | 55 |
| <b>Figure 4.7</b> A random forest is a collection of decision trees, where the overall decisions are a majority voting or averaged result from each tree. From TIBC [ , #295].....  | 56 |
| <b>Figure 4.8</b> The composition of a perceptron – linear combination of inputs via a weight vector which then get summed and passed to a nonlinear activation function. From DeepAI [ , #296].  | 57 |
| <b>Figure 4.9</b> Diagram of a fully connected MLP, or standard neural network, with three hidden layers. From IBM [ , #297]......  | 58 |
| <b>Figure 4.10</b> A generative adversarial network (GAN) can make new, realistic-looking samples by balancing a discriminator and generator during training. From Thalles' blog [Silva, #298].....   | 60 |
| <b>Figure 4.11</b> Summary of the interpretability versus accuracy (or strength) of each machine learning model. From Duval, 2019 [Duval, 2019 #299].....   | 61 |
| <b>Figure 4.12</b> A Gaussian process gives estimates of uncertainty depending on the location of the training data. From Leclercq, 2018 [Leclercq, 2018 #300]. .....   | 63 |

|   |    |
|---|----|
| <b>Figure 4.13</b> Monte Carlo dropout during test predictions is the easiest and most common way to estimate uncertainty because it does not require special model architecture. From AWS documentation [, #301]. .....        | 64 |
| <b>Figure 4.14</b> A Bayesian neural network (right) learns distributions of weights instead of just the weights themselves, as with standard neural networks (left). From Sanjay Thakur's Blog [Thakur, #302]. .....           | 65 |
| <b>Figure 4.15</b> A mixed density network is another easy implementation of a neural network that can formally account for uncertainty. From Vossen, 2018 [Vossen, 2018 #303]. .....   | 65 |
| <b>Figure 4.16</b> PCA tried to maximize explained variance, or equivalently minimize the distances needed for the data points to be projected onto that eigenvector (or basis vector). From Bits of DNA [Pachter, #304]. ..... | 68 |
| <b>Figure 4.17</b> The benefits of nonlinear dimensionality reductions algorithms. [Tetef, 2021 #140] .....   | 69 |
| <b>Figure 4.18</b> K-means clustering balances center of mass. From Towards Data Science [Introduction, #305]. .....  | 72 |
| <b>Figure 4.19</b> dbscan uses the hyperparameter epsilon $\epsilon$ to determine the radius of an expected cluster. From Khater, et al., 2020 [Khater, 2020 #306]. .....   | 73 |
| <b>Figure 4.20</b> Dendograms represent similarity of points using the distances between them as a metric. From Displayr [Bock, #307]. .....  | 73 |
| <b>Figure 4.21</b> Recursive feature elimination picks the best features by correlating them with the best score (e.g., accuracy) of predicting the target variables. [Tetef, 2023 #232] .....                                  | 75 |

**Figure 5.1** (a) shows the test nanoparticle structures. (b) and (c) show the true versus predicted coordination numbers from the neural network for both the first and fourth coordination shell, respectively, on the test dataset. Taken from Timoshenko, et al., 2017 [Timoshenko, 2017 #30].

..... 79

**Figure 5.2** Training dataset of each descriptor versus property, where the color reflects the CN values in (a), (b), the average Fe-O distance in (c), (d), and iron valence in (e), (f). Taken from Guda, et al., 2021 [Guda, 2021 #148]..... 81

**Figure 5.3** Featurization of XANES spectra included the pointwise energy-intensity values and fitting to third order polynomials for regions with varying energy resolutions. Taken from Torrisi, et al., 2020 [Torrisi, 2020 #122]. ..... 82

**Figure 5.4** Their workflow for classifying spectra into the three coordination environments: tetrahedral (*T*4), and square pyramidal (*S*5), and octahedral (*O*6). Of note, their structural database was the Materials Project. Taken from Carbone, et al., 2019 [Carbone, 2019 #99]..... 84

**Figure 5.5** (a) Structure to spectra correlation (b) Clustering spectra using hierarchical clustering (Wasserstein distance as a similarity metric) (c) Decision tree to determine the underlying properties distinguishing the three clusters. Taken from Mizoguchi and Kiyohara, 2020 [Mizoguchi, 2020 #23]..... 86

**Figure 5.6** Mean percentage error between the target (from simulations) and predicted spectra on the test set. Taken from Rankine and Penfold, 2020 [Rankine, 2020 #123]. ..... 87

**Figure 6.1** Schematic representation of the five types of sulphorganics investigated, along with sub-categories. ..... 95

**Figure 6.2** Schematic depiction of the data generation pipeline. ..... 96

|  |     |
|--|-----|
| <b>Figure 6.3</b> (a) Clusters where nonlinear dimension reduction routines, such as from a neural network, might yield better clustering than a linear dimension reduction like PCA. (b) Architecture of a simple autoencoder (AE) with one hidden layer, demonstrating the dimension reduction utility of the AE via its nonlinear latent space. (c) Schematic of how t-SNE uses the probability that data points are sampled from the same distribution to determine their similarity. .... | 103 |
| <b>Figure 6.4</b> VtC-XES (left) and XANES (right) spectra for all organosulphur compounds, displayed by compound type. Some spectra have been arbitrarily scaled or randomly removed for display purposes.....  | 108 |
| <b>Figure 6.5</b> Scree plot of PCA effectiveness for both VtC-XES and XANES. The vertical axis is the fraction of variance explained by each PC, e.g., the 10 <sup>th</sup> PC.....   | 109 |
| <b>Figure 6.6</b> Spectra reconstructed with increasing number of principal components (PCs) kept, for both VtC-XES and XANES of 2-thiazolidinone sulphone (Type 5) (top two panels) and 4-thiazoleaceticacid (Type 1) (bottom two panels).....  | 110 |
| <b>Figure 6.7</b> Principal Component Analysis (PCA) projection for two dimensions, color-coded by the two different property classification schemes: Scheme 1 is by oxidation and Scheme 2 is by sulphur bond type.....   | 112 |
| <b>Figure 6.8</b> Latent space representation in two dimensions via a Variational Autoencoder (VAE), color-coded by the two different property classification schemes: Scheme 1 is by oxidation and Scheme 2 is by sulphur bond type. ....   | 113 |
| <b>Figure 6.9</b> Reconstruction of XES (left) and XANES (right) spectra from a two-dimensional latent space via a VAE. From bottom to top, the compounds are from Type 1, 2, 3, 4, and 5. The black dashed line represents the original inputted spectra, and the solid-colored line is the decoded spectra after it has been passed through the VAE.....   | 114 |

|  |     |
|--|-----|
| <b>Figure 6.10</b> Chemically similar compounds are nearby in the latent space. (a) The latent space location of tetrabromothiophene and tetrachlorothiophene, with the corresponding XES spectra on the right. (b) The same structures but oxidized to form tetrabromothiophene oxide and tetrachlorothiophene oxide.....   | 116 |
| <b>Figure 6.11</b> A closer look at the outliers: the two “neutrally oxidized” compounds distinctly in the sulphone (+2 oxidation) cluster. ....   | 117 |
| <b>Figure 6.12</b> Compounds with aromatic sulphur versus aliphatic sulphur, in the latent space (VAE) for both VtC-XES (left) and XANES (right). ....   | 119 |
| <b>Figure 6.13</b> Residuals between the average of the aromatic and aliphatic spectra of Type 3 (thiols). ....  | 120 |
| <b>Figure 6.14</b> t-SNE for VtC-XES (left) and XANES (right). (a) is color-coded by Type, while (b) is color-coded by aromaticity within each Type.....   | 121 |
| <b>Figure 6.15</b> (Main) A closer look the the subclustering in the XANES t-SNE plot. (a) Separation of Type 1 aromatic compounds based on inclusion of chlorine or bromine in the aromatic system. (b) Separation of Type 5 aliphatic compounds based on bond strain via the inclusion of sulphur in a ring versus a chain. (c) Type 4 compounds with one R group aromatic and the other aliphatic share characteristics of both and thus form the bridge between the two clusters. .... | 123 |
| <b>Figure 6.16</b> Accuracy of KNN classification schemes on all dimensionally reduced spaces for both VtC-XES (top) and XANES (bottom). ....  | 124 |
| <b>Figure 6.17</b> As shown in (a), the evolution from goitrin (oxidation -2) to thiophene oxide (oxidation 0). (b) The linear combination of the spectra of thiophene oxide (top) and goitrin (bottom) that correspond to the points along the track in (a). (c) Tracks of 3000 different species evolutions.....   | 128 |

|   |     |
|---|-----|
| <b>Figure 6.S1</b> Loss plotted against number of epochs for the VAE model for both the XANES data (blue) and the VtC-XES data (green).....   | 142 |
| <b>Figure 6.S2</b> Reconstructed VtC-XES spectra with increasing latent space dimension. ....   | 143 |
| <b>Figure 6.S3</b> Classification via NN: confusion matrices for XES and XANES for both categorization schemes: 1) oxidation and 2) bond type.....  | 144 |
| <b>Figure 6.S4</b> Classification via NN: confusion matrices for VtC-XES for classification of aromatic versus aliphatic compounds within Types 1 to 5.....   | 145 |
| <b>Figure 6.S5</b> Classification via NN: confusion matrices for XANES for classification of aromatic versus aliphatic compounds within Types 1 to 5.....   | 146 |
| <b>Figure 6.S6</b> Unsupervised dimension reduction: VAE, t-SNE, FastICA, PCA, FA, and NMF for VtC-XES (left) and XANES (right), color-coded by sulfur bonding Type.....                              | 147 |
| <b>Figure 6.S7</b> KNN classification for Oxidation for VtC-XES. ....   | 148 |
| <b>Figure 6.S8</b> KNN classification for Oxidation for XANES.....  | 149 |
| <b>Figure 6.S9</b> KNN classification for sulfur bond Type on VAE, t-SNE, FastICA, PCA, FA, and NMF for VtC-XES.....  | 150 |
| <b>Figure 6.S10</b> KNN classification for sulfur bond Type on VAE, t-SNE, FastICA, PCA, FA, and NMF for XANES.....   | 151 |
| <b>Figure 6.S11</b> KNN classification for Aromaticity for VtC-XES. ....  | 152 |
| <b>Figure 6.S12</b> KNN classification for Aromaticity for XANES. ....  | 153 |
| <b>Figure 6.S13</b> Accuracy of KNN classification schemes on the PCA, VAE, and t-SNE reduced spaces for VtC-XES (top) and XANES (bottom) while increasing the latent or embedding dimension, D. .... | 154 |

|  |     |
|--|-----|
| <b>Figure 7.1</b> Flowchart of an analysis framework that uses unsupervised machine learning (such as cluster analysis) as a precursor to predictions on spectra via supervised machine learning, which can then inform experimental design and data creation.....   | 158 |
| <b>Figure 7.2</b> UMAP representation of VtC-XES (top) and XANES (bottom), color-coded by coordination, with some example spectra (as calculated by NWChem) shown to the right. ....   | 167 |
| <b>Figure 7.3</b> UMAP representation of VtC-XES (top) and XANES (bottom) for tricoordinate phosphorus and tetracoordinate phosphorus compounds, color-coded by number of oxygens bonded to the phosphorus within each coordination. The same example spectra as before are shown to the right, as calculated by NWChem. ..... | 169 |
| <b>Figure 7.4</b> UMAP representation of VtC-XES (left) and XANES (right) for compounds with sulfur ligands, color-coded by number of sulfurs. The pair of bottom insets on each panel are enlargements of the shown sub-regions to make it easier to see violations of cluster chemical classes, i.e., outlier compounds..... | 172 |
| <b>Figure 7.5</b> UMAP representation of the VtC-XES of compounds with consecutively more R groups (if bonded to an oxygen) replaced with an H atom (to create hydroxyl groups), color-coded by chemical class. ....   | 174 |
| <b>Figure 7.6</b> UMAP representation of XANES of phosphates, color-coded by sub-clusters. Cluster-averaged spectra and a summary structural motif for each cluster are also shown. ....   | 177 |
| <b>Figure 7.7</b> Gaussian Process Classifier prediction accuracies with corresponding average probability (“confidence”) for all chemically driven and cluster-driven classification schemes. ....  | 180 |

|  |     |
|--|-----|
| <b>Figure 7.S1</b> Experimental spectra versus theoretically calculated VtC-XES spectra using NWChem[Apra, #5]. The experimental procedure follows the same protocol as Holden et al. [Holden, 2020 #37] There is relatively good agreement in the existence and location of resonances, except a modest edge shift for GaP (bottom left) .....  | 188 |
| <b>Figure 7.S2</b> Experimental spectra versus theoretically calculated XANES spectra using NWChem[Apra, #5]. Experimental data is from Persson et al.[Persson, #144] There is relatively good qualitative agreement in the existence and energy of near-edge features.....  | 189 |
| <b>Figure 7.S3</b> PCA preprocessing step to keep only 95% of the variance of the dataset. ....  | 190 |
| <b>Figure 7.S4</b> Reconstructed VtC-XES spectra of randomly selected compounds (PubChem CIDS shown in top left for each) after being passed through the PCA pre-processing step to keep 95% of the variance of the dataset. The black lines are the theoretically calculated spectra, while the dashed colored lines are the reconstructed spectra after 95% variance, according to PCA, is retained..... | 191 |
| <b>Figure 7.S5</b> Reconstructed XANES spectra of randomly selected compounds (PubChem CIDS shown in top left for each) after being passed through the PCA pre-processing step to keep 95% of the variance of the dataset. The black lines are the theoretically calculated spectra, while the dashed colored lines are the reconstructed spectra after 95% variance, according to PCA, is retained.....   | 192 |
| <b>Figure 7.S6</b> Average spectra for each chemical class within the two coordination geometries.   | 196 |
| <b>Figure 7.S7</b> Cluster averages of compounds as they appear in the embeddings in Figures 2 and 3 in the main text.....   | 197 |

|   |     |
|---|-----|
| <b>Figure 7.S8</b> The XANES embedding corresponding to Figure 7.5, i.e., substitution of O-R with hydroxyl groups. The XANES does not cluster as well as the VtC-XES for this classification scheme.....   | 198 |
| <b>Figure 7.S9</b> Example compounds and their corresponding spectra and transitions in phosphate sub-cluster I.....  | 199 |
| <b>Figure 7.S10</b> Example compounds and their corresponding spectra and transitions in phosphate sub-cluster II. ....   | 200 |
| <b>Figure 7.S11</b> Example compounds and their corresponding spectra and transitions in phosphate sub-cluster III. ....  | 201 |
| <b>Figure 7.S12</b> Example compounds and their corresponding spectra and transitions in phosphate sub-cluster IV.....  | 202 |
| <b>Figure 7.S13</b> Compounds belonging phosphate sub-cluster I.....  | 204 |
| <b>Figure 7.S14</b> Compounds belonging phosphate sub-cluster II. ....  | 208 |
| <b>Figure 7.S15</b> Compounds belonging phosphate sub-cluster III.....  | 210 |
| <b>Figure 7.S16</b> Compounds belonging phosphate sub-cluster IV.....   | 211 |
| <b>Figure 7.S17</b> Phosphate sub-clusters <b>I</b> , <b>III</b> , and <b>IV</b> and their weak correlation to the energy of the absorption edge, defined as the spectral point with the greatest first derivative. (a) UMAP representation with points scaled to be different sizes based on the location of the edge, i.e., a higher energy edge yields a bigger data point. (b) Correlation matrix between the two UMAP axes and the edge energy. .... | 212 |
| <b>Figure 7.S18</b> Two-dimensional visualization of phosphate clustering in 10-dimensions. The dbscan clustering algorithm generally clusters the phosphates in the same groups as clustering applied directly to the 2D representation (as shown in Fig. 6). Instead, here the Cluster <b>III</b>   |     |

compounds are divided into two separate groups. Thus, the 2D representation of the phosphates is retaining the great majority of the information in the spectra, except likely a detail in Cluster **III** compounds that gets thrown away when reduced to so few dimensions. However, the overall classes are very similar and robust against UMAP dimension..... 213

**Figure 7.S19** Three-dimensional UMAP projections for various classification schemes for both the VtC-XES (left three panels) and XANES (right three panels). Most clustering in 3D seems to be the same as the 2D embeddings in the main text, indicating that a two-dimensional embedding captures most of the useful cluster information as three dimensions. .... 214

**Figure 7.S20** Two-dimensional UMAP projections for both VtC-XES (top) and XANES (bottom) of tricoordinate P and tetracoordinate P compounds when varying the hyperparameter number of expected neighbors in a cluster. The number of neighbors balances the global versus local structure preserved by UMAP, and it is similar to the perplexity hyperparameter in t-SNE. For small n\_neighbors values, local similarities are stressed, while large values stress global similarities (at the cost of losing fine details) [McInnes, #142]. ..... 215

**Figure 7.S21** Two-dimensional UMAP projections for both VtC-XES (top) and XANES (bottom) of tricoordinate P and tetracoordinate P compounds when varying the hyperparameter minimum distance between points. The minimum distance hyperparameter controls how tightly packed points in the reduced space can be. Generally low minimum distance values focus on more detailed topological structure, while large values stress broad topological structure. Smaller values of minimum distance are thus better for clear clusters for our analysis [McInnes, #142]..... 216

|   |     |
|---|-----|
| <b>Figure 8.1</b> Nano-XANES map, color-coded by the maximum spectral intensity of the Fe K-edge XANES spectra (to indicate the most likely places with sample due to the high photon counts). Each pixel is 150 nm. Note that background spectra are filtered out. ....  | 224 |
| <b>Figure 8.2</b> Our manifold projection image segmentation (MPIS) and linear combination fitting (LCF) pipeline for analyzing our nano-XANES image. ....  | 226 |
| <b>Figure 8.3</b> (top) k-means clustering on the first two principal components. (bottom) dbSCAN clustering on a two-dimensional UMAP embedding. The clusters and labeling in the two-dimensional UMAP representation not only match expectations, but they are easier to see and thus interpret than the k-means clusters on the top two PCA components. ....   | 228 |
| <b>Figure 8.4</b> Reference chemical classes. Often, LCF results are reported using the chemical class of the references. These classes are usually created using chemical knowledge of the system. Instead, we offer a completely data-driven way one can generate these classes, specifically by projecting references onto the UMAP space determined by the experimental spectra. ....   | 229 |
| <b>Figure 8.5</b> (a) Effects on the clusters when encoding XRF data and spatial location into the MPIS pipeline. (b) The resulting 2D phase maps, colored by cluster. (c) Score of linear combination fitting (LCF) predictions via the standard pixel-by-pixel analysis (“Standard”), pixel-by-pixel LASSO regression (“LASSO”), and LASSO regression via MPIS (“MPIS”). The upper leftmost panel shows no joint information encoding. .... | 231 |
| <b>Figure 8.6</b> Adding noise (as a percentage of spectral intensity) to the experimental spectra causes pixel-by-pixel analysis to have small unphysical fluctuations in the phase maps, resulting from uncertain LCF fits (top row). By applying MPIS (second row), the phase maps (third row) are more robust to noise, demonstrated by the consistent LCF results (bottom row). ....   | 232 |

|  |     |
|--|-----|
| <b>Figure 8.S1</b> UMAP spaces with and without PCA processing. UMAP applied to the first 6 principal components produces clusters that are very similar to the clusters made when UMAP is applied directly to the spectra, both on the raw experimental data and with augmented noise added to the spectra. ....  | 237 |
| <b>Figure 8.S2</b> PCA triangle plot of all two-dimensional projects of the six-dimensional hypercube of the top principal components of the spectral dataset. Six dimensions were chosen because it takes the top six principal components (PCs) to explain 97% of the variance. ....   | 238 |
| <b>Figure 8.S3</b> Scree plot of experimental spectra. It takes 6 PCs to explain 97% variance (dashed line). ....  | 239 |
| <b>Figure 8.S4</b> First four principal components of the measured spectra.....  | 240 |
| <b>Figure 8.S5</b> Four-dimensional UMAP hypercube (applied to the top PCA components), with two-dimensional projections (color-coded by density) shown in the bottom left corner. The upper right corner is composed of the same projections (transposed so that the upper and lower triangles match), except instead color-coded by the dbSCAN clustering labels. ....   | 241 |
| <b>Figure 8.S6</b> Correlation of reference spectra. (a) 1000 randomly sampled experimental spectra (which passed the background filter) normalized following the standard procedure in Athena [Ravel, 2005 #97]. (b) Reference spectra used in this study (11 total). (c) Variation of both the 1000 experimental spectra and the reference spectra. (e) Mean squared error between the original spectra and the fitted (via least squares) spectra of both the experimental data and true linear combinations of references (with random normal noise with a variance of 3% of the spectral intensity to model true experimental noise)..... | 242 |

**Figure 8.S7** Augmented MPIS versus noise. When noise is set to 10%, only three clusters appear (left column). However, adding XRF and spatial encoding recovers the lost cluster (right column).

..... 243

**Figure 8.S8** Error in predicted concentrations versus noise (on a dataset composed of true linear combinations of references) for both our MPIS cluster-averaged pipeline and the standard individual spectrum analysis. The variance in predictions decreases when spectra are averaged together in an informed way via MPIS. .... 244

**Figure 9.1** Recursive Feature Elimination (RFE) optimizes the feature subset to measure, in this case energies, by training a base machine learning model, such as linear regression, to predict target variables from spectra. .... 248

**Figure 9.2** Comparing RFE results with forced linear independence in the basis set and thus the training dataset. (a) The spectral reference library. (b) Training dataset of linear combinations of references. (c) The RFE results, trained on the spectral linear combinations, where the basis set can have linear dependence. (d) Reference spectra projected onto the first six principal components from PCA. PCA forces the basis set to be linear independent and thus the linear combinations are unique. (e) The same training data as before, which are also projected using PCA. (f) The RFE results trained on the PCA projections. .... 252

**Figure 9.3** Characterization and validation of RFE algorithm. (a) Collection of energies chosen by different feature selection algorithms: random selection (Rand), recursive feature elimination (RFE), random forest (RF), decision tree (DT), and linear regression (LR). The dark bars include the three default energies (white) to ensure normalization. (b) Corresponding errors in LCF predictions on both a generated test dataset and the experimental spectra for all models. (c)

Energies consecutively removed by the RFE as fewer energy points are kept, which shows consistency in training. (d) Error in reconstructing spectra using normalization parameters from reduced energy point (sub)spectra of different sizes compared to normalized full energy spectra. (e)  $R^2$  score of the linear regression (LR) base estimator in the RFE. (f) Error in LCF predictions versus (sub)spectra size on both simulated test data and the experimental spectra..... 255

**Figure 9.4** Fully measured experimental XANES spectra (left) compared to the reduced energy point (sub)spectra (right), with energies recommended by the RFE algorithm (vertical gray lines). The vertical white lines indicate energies we subsequently added for normalization purposes. 256

**Figure 9.5** Linear Combination Fitting (LCF) results via standard pixel-by-pixel analysis and Manifold Projection Image Segmentation (MPIS). (a) The “true” results, using the full energy spectra via pixel-by-pixel NNLS-LCF onto the four known reference phases. (b) Pixel-by-pixel NNLS-LCF applied to the full energy spectra. (c) Pixel-by-pixel NNLS-LCF applied to the reduced energy point (sub)spectra. (d) LASSO-LCF via MPIS applied to the full energy spectra. (e) LASSO-LCF via MPIS applied to the reduced energy point (sub)spectra. ..... 258

**Figure 9.6** MPIS and LASSO-LCF results using (a) the full spectra and multimodal encoding, (b) the (sub)spectra and multimodal encoding, and (c) the (sub)spectra by themselves *without* augmented information. Here, the total XRF intensity of sulfur, phosphorus, and chromium and the spatial location of pixels are multimodal information. ..... 260

**Figure 9.S1** (a) Random sampling of experimental data to act as a basis for linear combinations of spectra. Note that there is no guarantee that the basis spectra span or equally sample the experimental domain. (b) 1000 linear combinations generated from the corresponding basis. (c) The compiled results of an ensemble of 10 RFEs trained on the spectra. (d) The equivalent dataset

except projected onto the first six principal components. (e) The compiled results of an ensemble of 10 RFEs trained on the principal components. When N (the basis set size) is 50, there is so much linear dependence in the basis set that the RFE fails because it chooses points in the pre-edge, which has no variation. .... 263

**Figure 9.S2** Test of the RFE by using three gaussians used as basis spectra (left-most column) to make linear combinations (middle column). The RFE clearly picks features (i.e., “energies”) that correspond to the highest variation. We used linear regression as our base estimator. However, after the regions corresponding to the three distributions are filled, the RFE must rank areas in between peaks where there is no signal. The peaks in importance between the Gaussians represent random selections in these regimes..... 264

**Figure 9.S3** Comparing results of linear and nonlinear inputs to the RFE. The RFE results, where linear combinations of spectra are the input and the concentrations that created those spectra are output, is shown in purple. Instead, using projections onto the first few principal components as input (with the same output) is shown in pink. The green shows the RFE results using both linear input and outputs. The total results for an ensemble of 10 RFE algorithms for each are shown at the bottom, along with the mean squared error (MSE) of predictions using LASSO linear combination fitting (LASSO-LCF) on a generated dataset of linear combinations of reference spectra. .... 265

**Figure 9.S4** The number of principal components (PCs) needed to explain 99% variance of the reference set. Starting with the four known references, we randomly selected additional references from the set of 11 total references used in this study. After the 11 references were chosen, we randomly selected additional references from another larger set of 64 Fe K edge XANES to constitute the reference library. We reselected these random additions 50 times and show the

|  |     |
|--|-----|
| average and standard deviation of the calculated number of principal components for that reference library size..... | 266 |
|--|-----|

|   |     |
|---|-----|
| <b>Figure 9.S5</b> Correlation, or similarity matrices, of the reference set for both the entire spectra and the 14-energy (sub)spectra. The correlation coefficient ( $R^2$ ) qualitatively looks the same for both, although the quantitative range for the (sub)spectra is larger, indicating global correlations (and information) is retained..... | 267 |
|---|-----|

|   |     |
|---|-----|
| <b>Figure 9.S6</b> Scree plot of experimental data on full spectra (top) versus (sub)spectra (bottom).<br>..... | 268 |
|---|-----|

|  |     |
|--|-----|
| <b>Figure 9.S7</b> First four principal components of the spectral subset. These components, in theory, should match with the principal components from the full spectral dataset, if all information is retained..... | 269 |
|--|-----|

|   |     |
|---|-----|
| <b>Figure 9.S8</b> PCA triangle plot of experimental data on (sub)spectra. .... | 270 |
|---|-----|

|  |     |
|--|-----|
| <b>Figure 9.S9</b> UMAP and dbSCAN on energy subset..... | 271 |
|--|-----|

|  |     |
|--|-----|
| <b>Figure 9.S10</b> Strength of spatial encoding (S) versus UMAP's number of neighbors (N). The minimum distance in UMAP is 0 and dbSCAN epsilon is 1 for all. The top section shows the UMAP space color-coded by dbSCAN clusters, the middle shows the same clusters but on the 2D map, and the bottom shows the max contributions from the LCF fits. Pink = Pyrite, magenta = LFP, blue = Hematite, red = SS, and gray = all other references. .... | 272 |
|--|-----|

## List of Tables

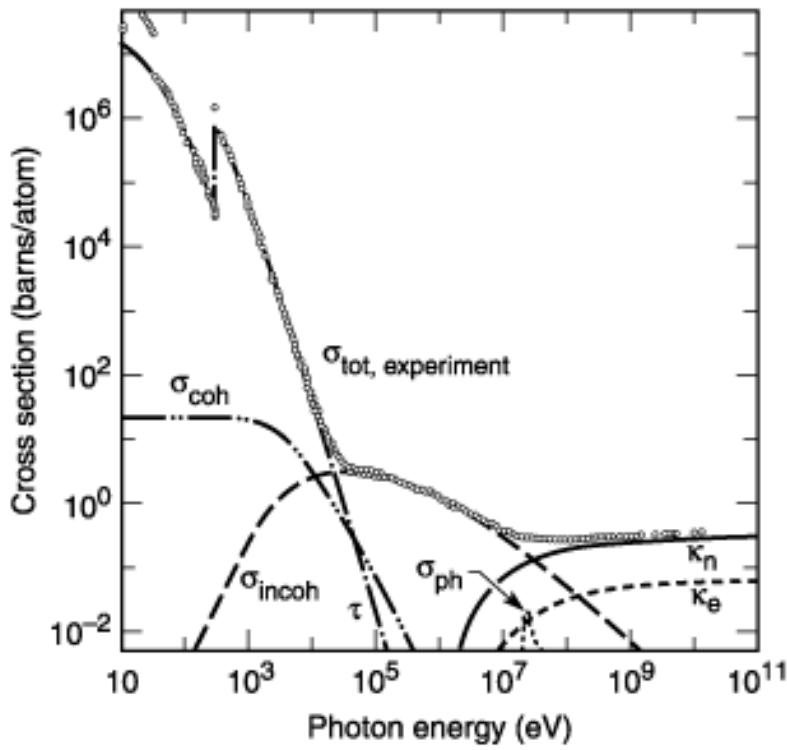
|   |     |
|---|-----|
| <b>Table 7.S1</b> Summary of all classification schemes developed with UMAP, with references to the figures in the main text where they are first displayed. These classifications are equivalent to the ones predicted in Figure 7 in the main text..... | 193 |
| <b>Table 7.S2</b> Structure, CID, and chemical class for each compound <b>a</b> to <b>h</b> labeled in Figures 2 and 3 in the main text.....  | 195 |

# Chapter 1 – Introduction and Overview of X-ray Spectroscopy

This chapter gives a brief overview of X-ray absorption and emission spectroscopies, including nomenclature and typical uses. For a more detailed discussion, see the many excellent review articles and key papers on XAFS, including but not limited to Rehr and Albers<sup>1</sup>, Sayers, Stern, and Lytle<sup>2</sup>, Bunker<sup>3</sup>, Ashley and Doniach<sup>4</sup>, Newville,<sup>5</sup> and Glatzel and Bergmann<sup>6</sup>.

Specifically, I will focus on the uses and chemical and electrical sensitivities of X-ray absorption near edge fine structure (XANES) and Valence-to-Core X-ray Emission Spectroscopy (VtC-XES). I will both qualitatively and quantitatively determine the strength of the information they encode in their respective spectra individually as well as comparing their sensitivities, framing the question of whether their information is complementary – a major contribution of this dissertation.

There are three different ways X-rays can interact with a sample – coherently, incoherently, or via the photoelectric effect, as shown in Fig. 1.1. Coherent scattering includes Rayleigh scattering while incoherent scattering includes Compton and Raman scattering. Here, I will focus on the photoelectric effect.



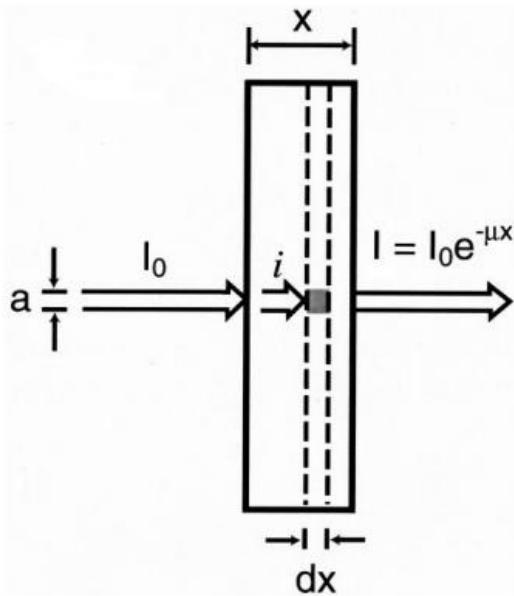
**Figure 1.1** The cross section of various interaction processes of X-rays for carbon. Taken from the X-ray Data Booklet.

## 1.1 X-ray Absorption Spectroscopies

The photoelectric cross section can be divided into either excitations or deexcitations. If in transmission mode, the measurement sums over all deexcitations as dictated by Fermi's Golden Rule. This measurement is called X-ray absorption spectroscopy (XAS) and is an important tool in many fields of science, such as materials science, physics, biology, chemistry, geosciences, and electronics. When in transmission mode, the shape of the background is important to identify the XAS spectra, given by the absorption coefficient  $\mu(E)$ . Transmission-mode XAS is governed by Beer's law

$$I(E) = I_0(E)e^{-\mu(E)x} \quad (1.1)$$

where  $x$  is the sample thickness,  $I(E)$  is the loss spectra (i.e., the photon counts after passing through the sample) and  $I_0(E)$  is the X-ray spectrum without sample. A cross-section is shown in Fig. 1.2.<sup>1</sup> We could solve for  $x*\mu(E)$  as  $-\ln (I(E) / I_0(E))$ . We will normalize out the sample thickness ( $x$ ) later.



**Figure 1.2** Absorption coefficient for XAS. Taken from Rehr and Albers.<sup>1</sup>

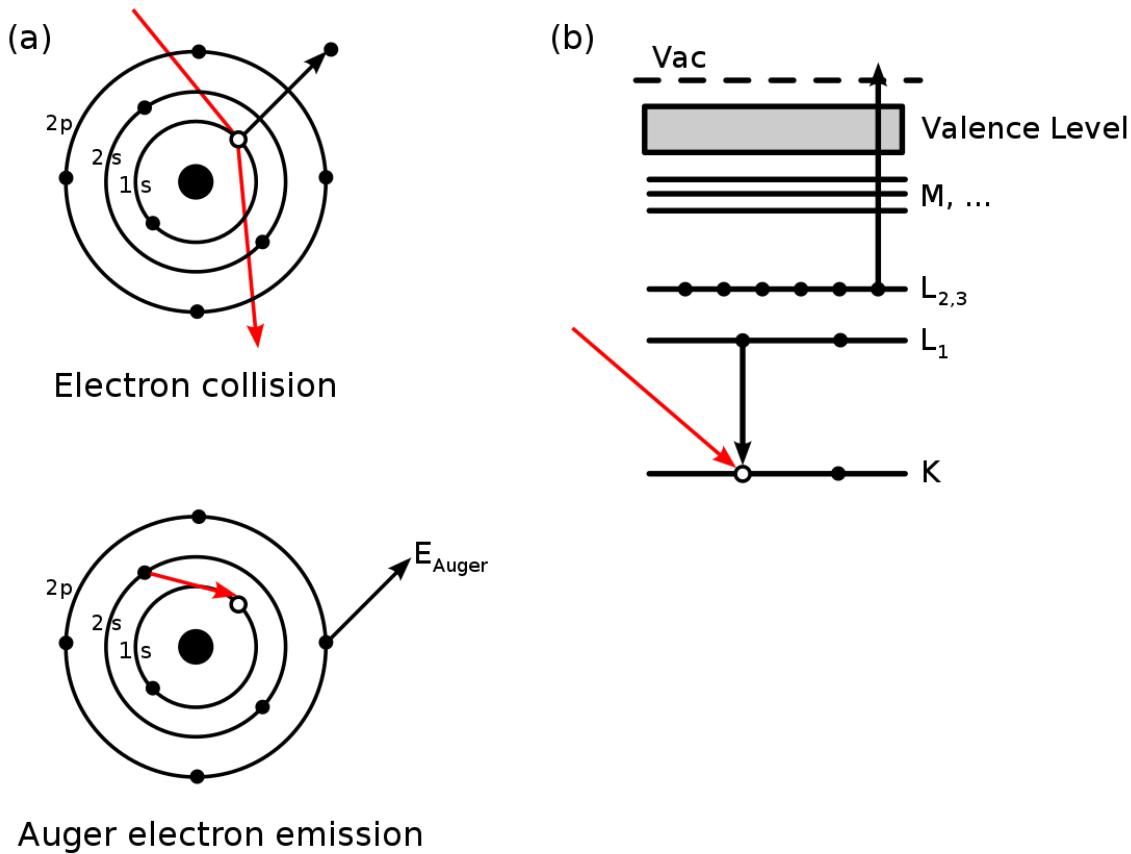
XAS, which produces spectra known as X-ray absorption fine structure (XAFS), is a bulk probe of both electronic and geometric structure around a chosen atomic species and is sensitive to properties such as oxidation state, valency, coordination, and bond length. When X-rays illuminate a sample, they can expel the electron out of the system (for X-ray Photoelectron Spectroscopy, or XPS). Or the X-rays push the system into an excited state by expelling the electron into a previously unoccupied but bound energy level for XAS. There is an intrinsic lifetime of this excited state unique to the atomic species and hole.

On the other hand, the photoelectric cross section can be detected from deexcitations. These measurements include fluorescence mode total fluorescence yield (TFY) XAFS measurements. In

florescence mode, there is no background shape (besides stray scatter, if there is any). In this case, the absorption coefficient  $\mu(E)$  can be solved for as  $I_f(E)/I_0(E)$ . Another fluorescence mode measurement includes a high-resolution version called High Energy Resolution Fluorescence Detected (HERFD), which has lifetime suppression. Finally, Resonant Inelastic X-ray Scattering (RIXS) is another measurement dictated by the Kramers-Heisenberg<sup>7</sup> equation

$$\frac{d^2\sigma}{d\Omega_k' d(\hbar\omega'_k)} = \frac{\omega'_k}{\omega_k} \sum_{|f\rangle} \left| \sum_{|n\rangle} \frac{\langle f | T^\dagger | n \rangle \langle n | T | i \rangle}{E_i - E_n + \hbar\omega_k + i\frac{\Gamma_n}{2}} \right|^2 \delta(E_i - E_f + \hbar\omega_k - \hbar\omega'_k). \quad (1.2)$$

Conversely, experiments that use X-rays simply to make a core hole and then measure deexcitations to fill that hole are not XAFS but rather fluorescence. These fluorescence measurements can be separated into two categories based on resolutions and thus use cases. Poor energy resolution is called X-ray Fluorescence and is usually over a much larger energy range and is thus used for qualitative identification used for elemental detection. However, if the energy resolution is on the scale of the core hole broadening such that there is distinct differentiation between the emission lines, the experiment is called X-ray Emission Spectroscopy (XES). However, rather than radiating, the core hole can instead fill via a nonradiative two-electron process, such as through the Auger-Meitner effect, as demonstrated in Fig. 1.3.

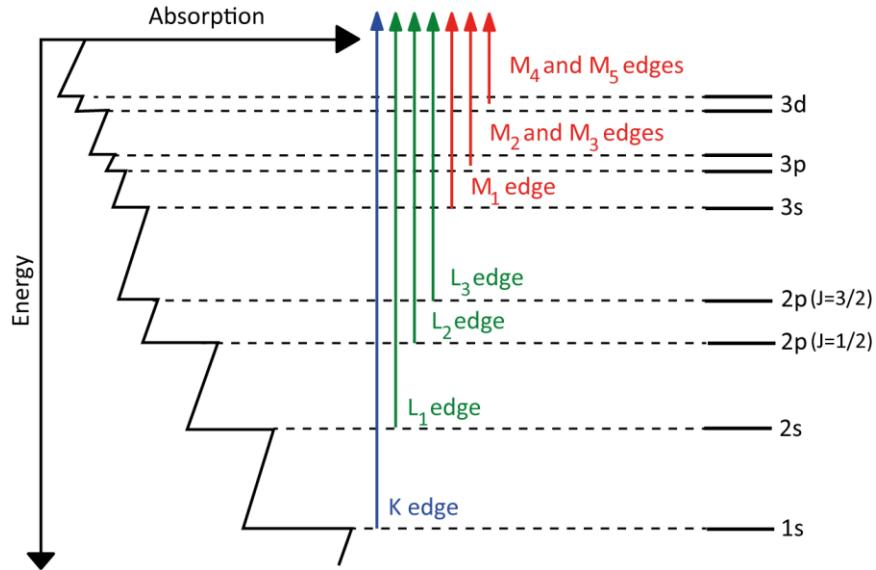


**Figure 1.3** The Auger-Meitner<sup>8,9</sup> effect is a two-electron process, where an inner-shell electron falls to fill the core hole, thus emitting a photon, but the rest of the extra energy is dissipated by emitting a valence electron.

### 1.1.1 K-, L-, and M- Edges

On a coarse energy scale,  $\mu(E)$  looks like a set of stairs, where each step, or edge, corresponds to an excitation energy of different inner-shell electrons. Each edge has a name and is commonly referred to using the IUPAC notation, as shown in Fig. 1.4. Exciting electrons from the  $1s$  ( $n = 1$ ) shell is called *K-edge* spectroscopy, while exciting electrons from the  $n = 2$  shell are the

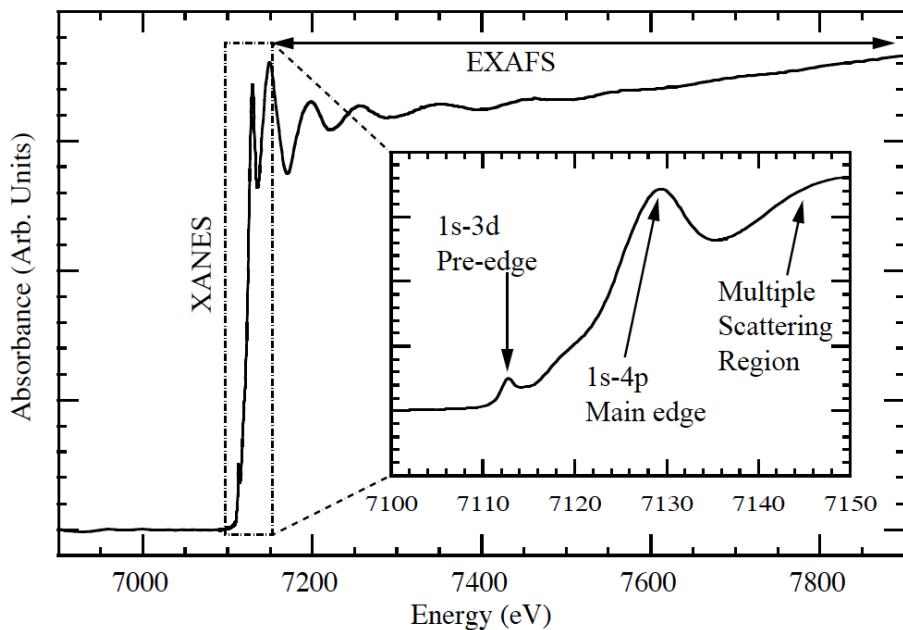
$L$  edges and the  $n = 3$  the  $M$  edges. Each edge is then broken down further based on the orbital angular momentum quantum number ( $\ell$ ) and the magnetic quantum number ( $j$ ). For example, exciting the  $2s$  shell is the  $L_1$  edge, exciting from the  $2p_{1/2}$  shell is the  $L_2$  edge, and from the  $2p_{3/2}$  shell is the  $L_3$  edge. An analogous pattern continues for the  $M$  edges.



**Figure 1.4** Absorption edge of an element can be broken down into  $K$ ,  $L$ , and  $M$  edges. Taken from Wikipedia <sup>10</sup>.

### 1.1.2 Regions of XAS spectra: XANES

Zooming in onto these edges, we see characteristic oscillations, as shown in Fig. 1.5. The XAS spectrum can be broken into two different regions – X-ray absorption near edge fine structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS). <sup>1,3</sup> XANES includes any pre-edge features, shoulders, and the region around the edge step, while the EXAFS includes the oscillations at higher energy.



**Figure 1.5** The XANES and EXAFS regions of XAS spectra. Taken from [cei.washington.edu](http://cei.washington.edu)<sup>11</sup>.

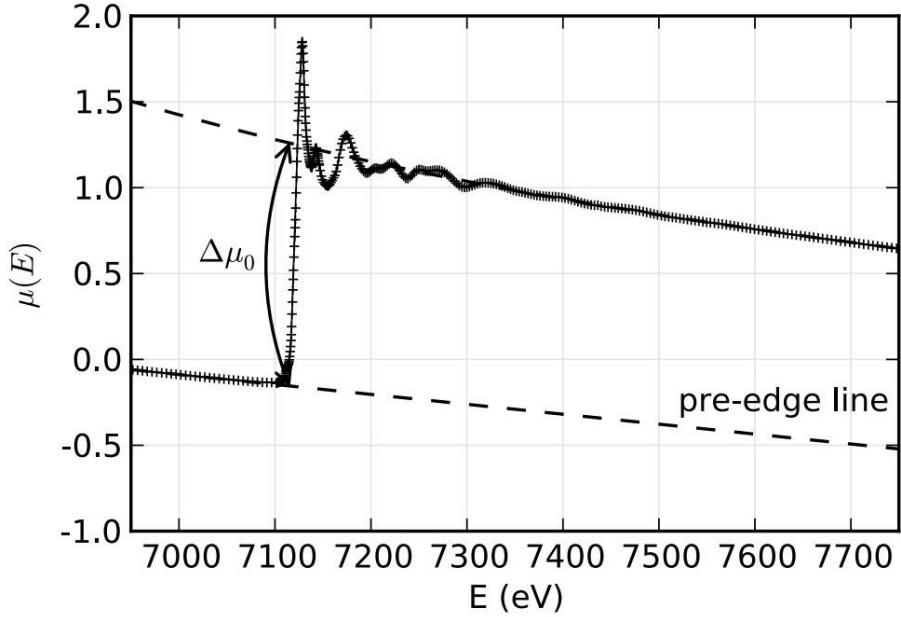
X-ray absorption near edge fine structure (XANES) is the region of XAS around the absorption edge, from any pre-edge features to about 50 eV past the edge, depending on the system. Because XANES probes electronic states near the Fermi level, it is sensitive to local electronic structure around the chosen atomic species, such as oxidation, spin, and valence as well as the local atomic structure, often including coordination symmetry. Calculating XANES spectra requires full multiple scattering theory and has broadening to transitions due to the lifetime of the core hole.

### 1.1.3 XANES Processing

Because spectral features in the XANES spectra are often correlated, meaning two chemical properties can cause the same spectral trends, the most common analysis for XANES

spectra is linear combination fitting onto reference spectra. Selecting reference spectra can be difficult because the set must encapsulate the experimental domain, so their choice relies on prior knowledge of the scientific system. Moreover, differences in the second and third coordination shells between the reference structures and experimental sample can have negative influences on the reliability of these fits.<sup>12</sup>

Furthermore, fitting requires proper “normalization,” as defined by the standard processing tools in Athena<sup>13</sup> and Larch<sup>14</sup>. Normalizing XANES spectra is defined as fitting a pre-edge line and a post-edge polynomial function (either constant, linear, or quadratic), where the edge is determined as the maximum of the derivative. The pre-edge fit is then subtracted from the entire XANES spectrum, and the post-edge is rescaled such that the fitted polynomial curve falls along the line  $y = 1$  after the edge. This process not only removes global scaling due to differences in sample thickness (i.e., number of atoms), but gives a consistent spectral shape.<sup>5</sup> An example of this process is shown in Fig. 1.6.



**Figure 1.6** An example of the normalization process for XAFS, where the pre-edge line is set to be along the  $y = 0$  line, and the post edge line is set to be along the  $y = 1$  line such that the edge step  $\Delta\mu$  is one. Taken from Newville.<sup>5</sup>

#### 1.1.4 Regions of XAS spectra: EXAFS

Extended X-ray Absorption Fine Structure (EXAFS) is the region of XAS spectra involving the higher energy oscillations past the absorption edge. EXAFS is sensitive to local geometric structure, such as coordination and bond length. Calculating EXAFS involves multiple scattering along enumerated, dominant paths and often assumes a muffin-tin potential as convenience. Much work has been done to utilize real space Green's functions to quickly calculate EXAFS via the FEFF program<sup>15</sup>, which is the *de facto* standard for EXAFS theory.

#### 1.1.5 EXAFS Processing

The goal in EXAFS is to obtain physical parameters like bond lengths, coordination, and disorder. An important aspect of EXAFS analysis thus derives from the parameters in the so-called

EXAFS equation, the earliest form of which appeared in Sayers, Stern, and Lytle <sup>2</sup>. The EXAFS equation is

$$\chi(k) = -kf(k) \sum_j N_j \exp[-k^2 \sigma_j^2 / 2] \exp[-\gamma r_j / r_j^2] \sin[2kr_j + 2\eta(k)] \quad (1.3)$$

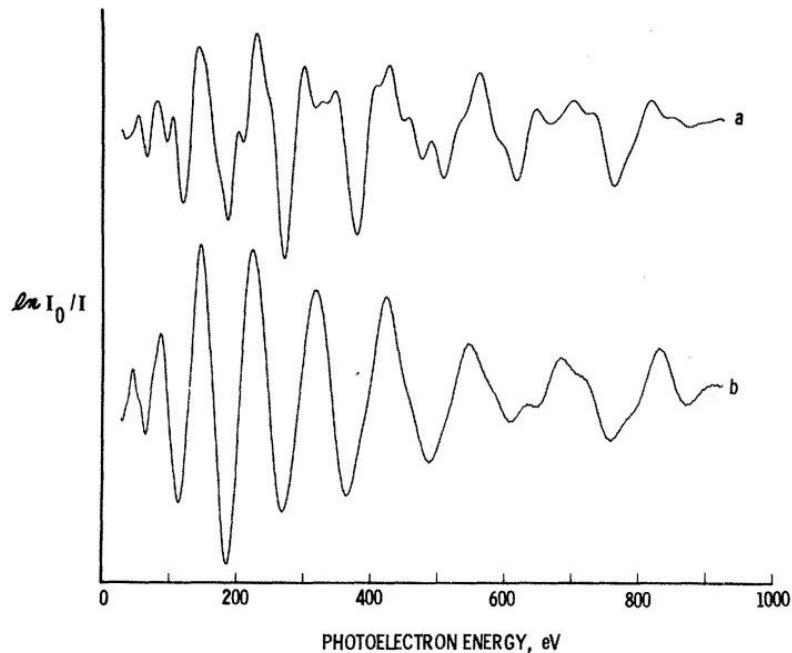
where  $k = 2\pi/\lambda$  is the photoelectron vector,  $N_j$  is the number of equivalent scatterers (or coordination number),  $f(k)$  is the scattering factor,  $\sigma_j^2$  is the Debye-Waller factor,  $r_j$  is the distance from the absorbing atom to the scatterer, and  $\eta(k)$  is the phase shift. <sup>2</sup> Because  $f(k)$  and  $\eta(k)$  can be calculated via the FEFF program <sup>15</sup> for a certain value of  $k$ , the only parameters left are  $N_j$ ,  $r_j$ , and  $\sigma_j^2$ .

This analysis is most often done using the community standard tools: Athena <sup>13</sup> and its newer cousin Larch <sup>14</sup>. Typically, you start with the normalized spectra following the same procedure in the XANES section of setting the edge to go from zero to one. Then, by removing the smooth post-edge background function, you isolate XAFS  $\chi(E)$ , where  $\chi(E) = (\mu - \mu_0) / \Delta\mu$ . Through a coordinate transform from Energy (E) to photoelectron momentum (k) and a Fourier transform,  $\chi(E)$  is transformed to  $\chi(k)$ , where k is

$$k = \sqrt{\frac{2m(E-E_0)}{\hbar^2}} \quad (1.4)$$

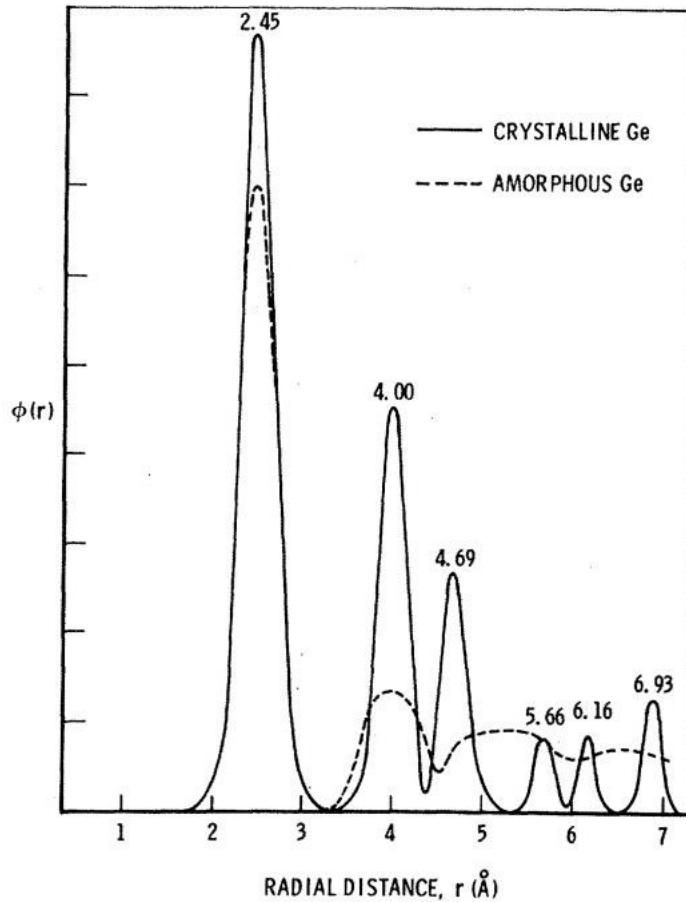
From  $\chi(k)$ , different fits and analysis can be performed to achieve the free parameters in the EXAFS equation. Often, because the high k features get washed out compared to the higher amplitude low k features, one will “k-weight”  $\chi(k)$ , i.e., multiplying  $\chi(k)$  by either  $k^2$  or  $k^3$  to pronounce the higher k features. Fourier transforming the k-weighted  $\chi(k)$  into R space then produces bond lengths. <sup>5</sup>

This basis for spectroscopy and analysis technique was first shown in Sayers, Stern, and Lytle <sup>2</sup> on crystalline and amorphous Ge, as shown in Fig. 1.7. Here, the crystalline Ge (Fig. 1.7a) clearly has the same first shell as amorphous Ge (Fig. 1.7b), demonstrating the locality of EXAFS.



**Figure 1.7** Smoothed experimental  $\chi(E)$  EXAFS data for (a) crystalline Ge and (b) amorphous Ge. Only the oscillatory part  $\chi$  of the absorption edge is shown. Figure taken from Sayers, Stern, and Lytle.<sup>2</sup>

The Fourier transform of the data in Figure 1.7 is shown in Figure 1.8. To make inferences on this data, usually fits to  $\chi(k)$  and  $g(R)$  are alternately obtained via Athena<sup>13</sup> or Larch<sup>14</sup> with the assumption that the parameter space is constrained enough that overfitting does not occur. However, there is recent progress in utilizing machine learning to perform these fits instead<sup>16</sup>.

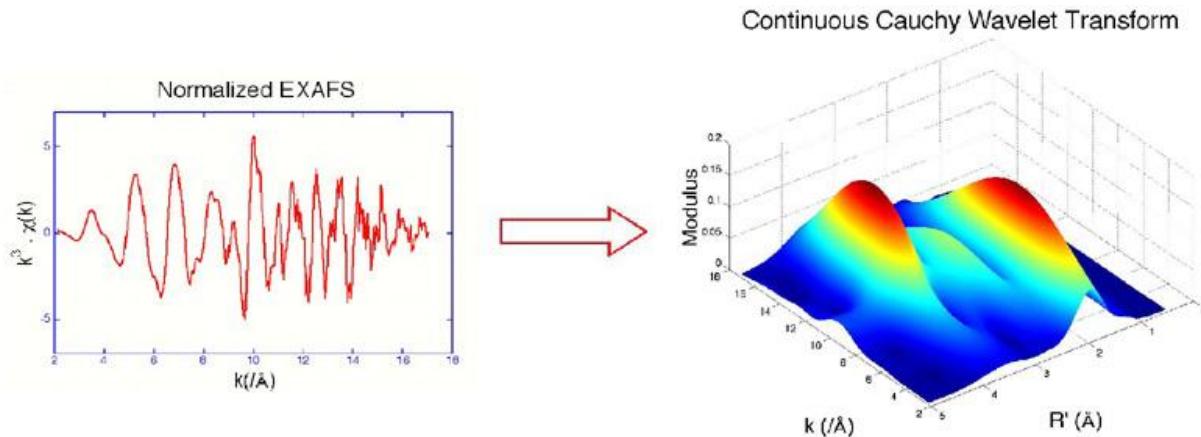


**Figure 1.8** Fourier transform of the data in Figure 1.7.  $\phi(r)$ , a radial structure function, compares amorphous and crystalline Ge. Numbers over the peaks indicate the measured distances in Å. Figure taken from Sayers, Stern, and Lytle.<sup>2</sup>

### 1.1.6 Wavelets

Another qualitative analysis of EXAFS that has seen recent traction is wavelet analysis. Wavelet transforms can filter EXAFS contributions by Z (atomic number) of the scattering species, e.g., atoms in the first coordination shell of the absorbing species. This benefit capitalizes on the fact that larger Z atoms have smaller (spatially) electron orbitals<sup>17</sup> and thus have a larger spread in momentum, which allows scattering in a larger k range than lighter Z atoms. A demonstration of the wavelet transform can be seen in Fig. 1.9, which shows the k-weighted  $\chi(k)$  data visualized

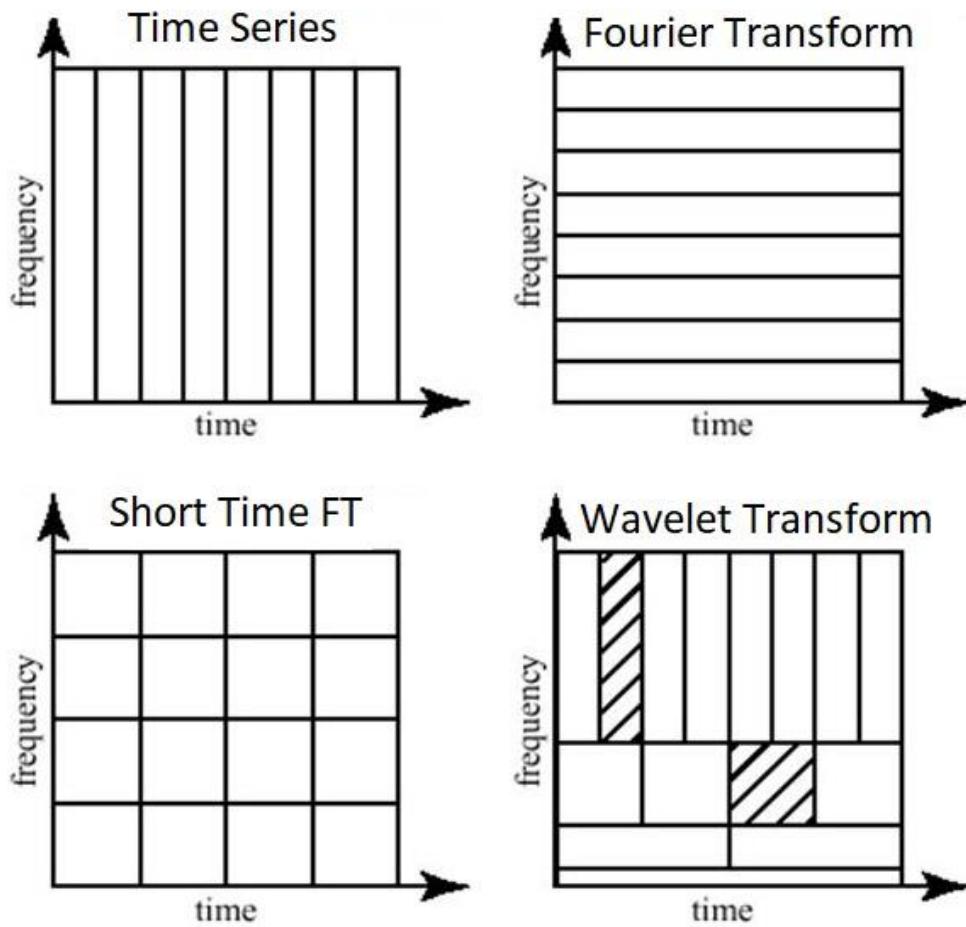
in three-dimensions, where one dimension is “distance” (from a k-weighted transform), the second dimension is momentum  $k$ , and the third dimension is the value of the modulus squared of the wavelet transform.



**Figure 1.9** Wavelet analysis can separate contributions based on the atomic number  $Z$  of the scatterer. Taken from Munoz, et al.<sup>17</sup>

Here is a brief mathematical motivation for using a wavelet transform. For a time-series dataset, where sampling occurs using the typical Shannon-Nyquist sampling rate, often one performs a Fourier transform to convert the data into frequency space. However, this analysis assumes signals are “stationary,” meaning they occur throughout the entire time of the signal, or there are no “events”.

One way to combat this assumption of stationary signals is to evenly divide time and frequency resolution. However, this approach can be an issue if, for example, signals very in time duration. Wavelet transforms use the idea that low frequency signals occur over longer periods of time and thus need better frequency resolution. Conversely, higher frequency signals occur over a short time scale and thus need better time resolution. The difference in resolution can be seen in Fig. 1.10.



**Figure 1.10** Wavelet analysis is like Fourier transforming using different frequency (or conversely time) sampling rates to create a multiresolution transform. Figure from ML fundamentals<sup>18</sup>.

This change in resolution is generated using a “mother” wavelet and adjusting the scale (and thus resolution). For example, this mother wavelet could look like

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (1.5)$$

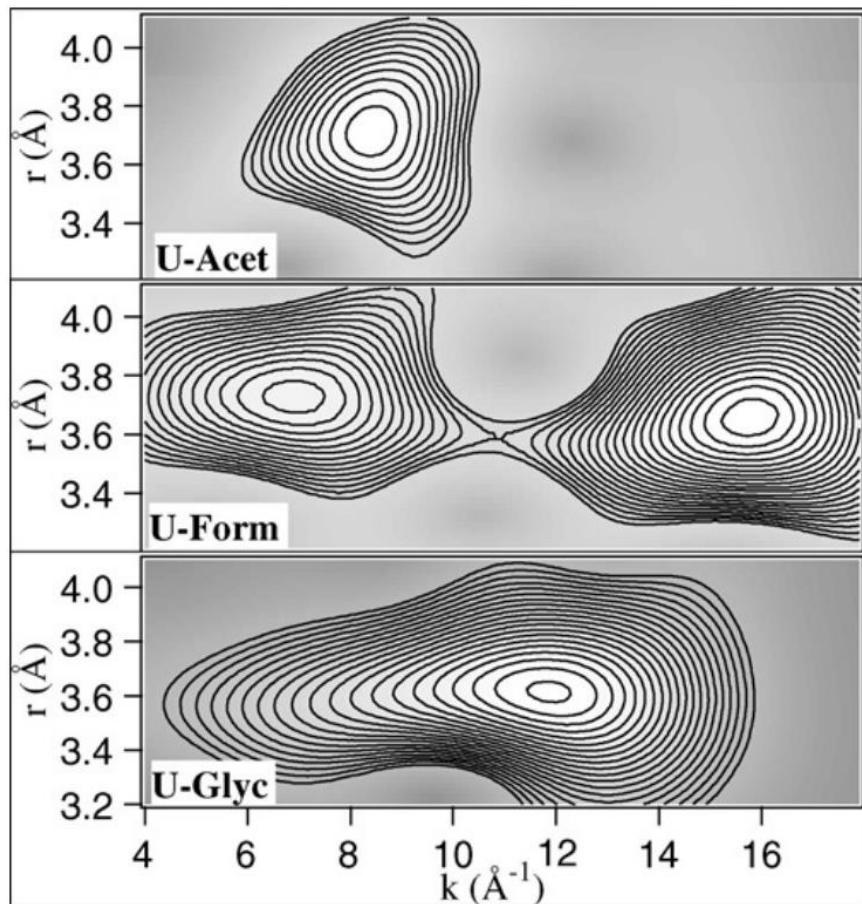
where  $a$  is the squishiness, or scale, of the wavelet and  $b$  is the time interval. Examples of specific wavelets  $\psi$  (which must be discrete) include the Haar, Mexican hat, Daubechies, and Coiflet

wavelets, depending on the type of data. The final wavelet transform is obtained by taking the inner product of the signal with all the wavelets (which form an orthogonal basis), given as

$$\mathcal{W}_\psi(f)(a, b) = \langle f(t), \psi_{ab}(t) \rangle \quad (1.6)$$

This inner product gives a complex value, where the modulus squared can then be interpreted, as is the case for XAFS wavelet analysis.

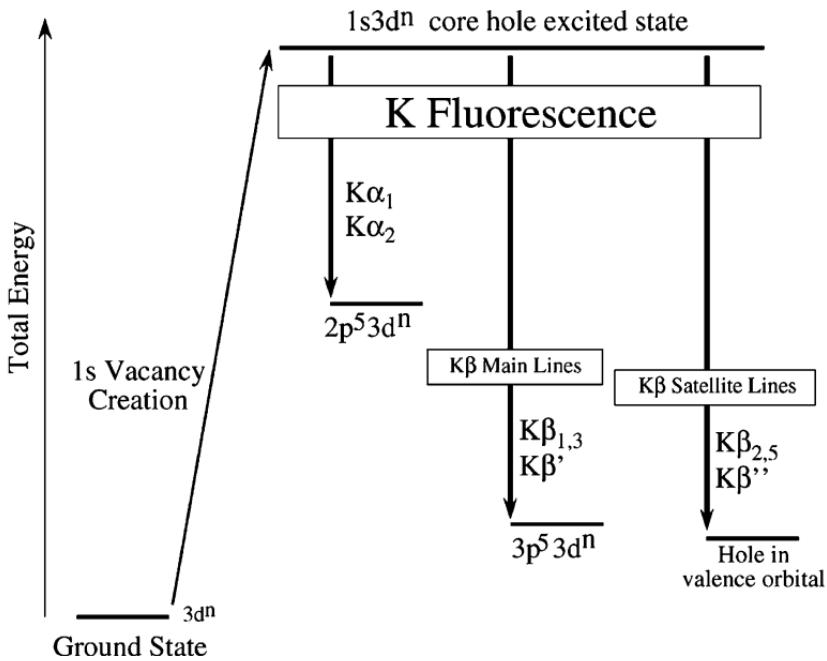
For example, Funke, et al. applied a wavelet transform to distinguish the complexation of Uranium (VI) with the carboxylic groups acetic-, formic-, and glycolic acid. While the typical Fourier transform could not distinguish the two different backscattering processes of uranium-uranium versus uranium-carbon-carbon in the uranium-formic acid complex, the wavelet transform could, as shown in Fig. 1.11<sup>19</sup>.



**Figure 1.11** Wavelet U-Acet, U-Form and U-Glyc within the range  $r = 3.2 - 4.1 \text{\AA}$ . Taken from Funke, et al., 2005<sup>19</sup>.

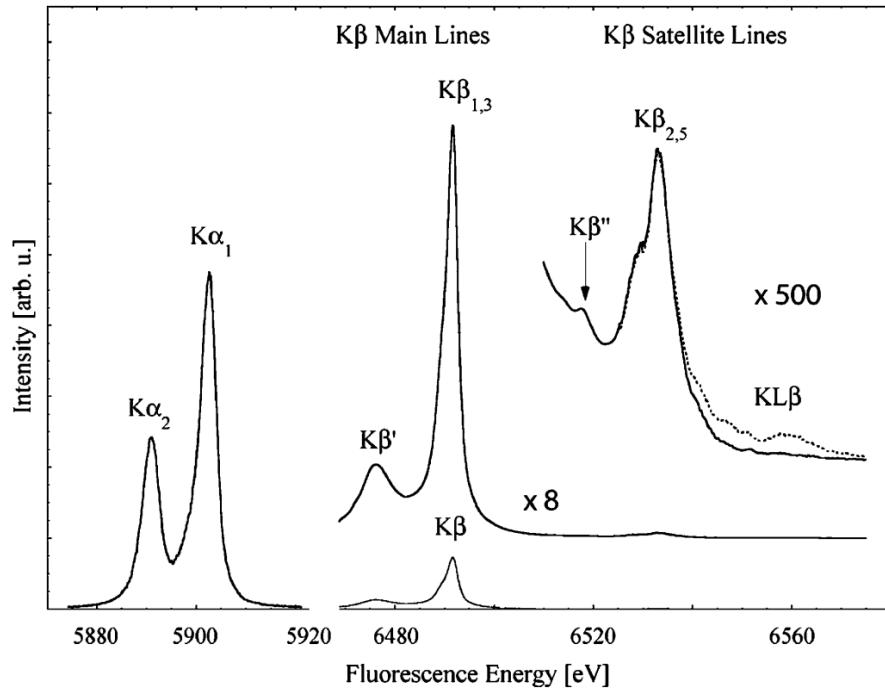
## 1.2 X-ray Emission Spectroscopy

X-ray fluorescence occurs when an excited atom emits a photon as an electron falls to fill the vacancy. Emission lines are similarly named as XAS by the hole location, or shell in which the electron was originally excited from. For example, having electrons fall into a  $1s$  core hole is  $K$  fluorescence. The initial shell of the falling electron then dictates whether the spectra are  $K\alpha$  ( $2p$  to  $1s$ ) or  $K\beta$  ( $3p$  to  $1s$ ). A breakdown of the K emission lines and their corresponding names can be seen in Fig. 1.12.



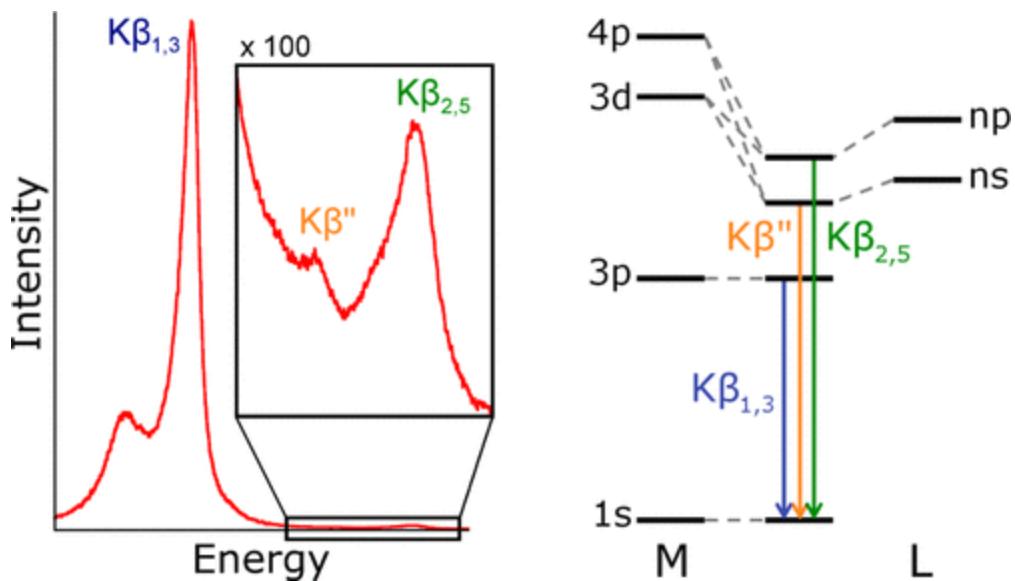
**Figure 1.12** Conventional naming of fluorescence lines. From Glatzel and Bergmann <sup>6</sup>.

The relative intensity of the different K emission lines is demonstrated in Fig. 1.12, where X-ray Emission Spectroscopy (XES) has good enough energy resolution to be able to distinguish between these lines. Here, the  $K\alpha$  lines are much more intense (by a factor of about 3) than the  $K\beta$  lines. Furthermore, the  $K\beta$  satellite lines (or electrons falling from the valence shell, i.e., the  $3d$  orbital for  $3d$  transition metals) are smaller by a factor of about 500. This difference in photon counts is exactly due to dipole selection rules and is important to consider when deciding on which emission line to probe.



**Figure 1.13** The Mn K-fluorescence lines for MnO. From Glatzel and Bergmann <sup>6</sup>.

Core-to-core X-ray Emission Spectroscopy (CtC-XES) includes the  $K\alpha$  and  $K\beta$  lines (at least for  $3d$  transition metals). An in-depth discussion of CtC-XES is found in Glatzel and Bergmann <sup>6</sup> and de Groot and Kotani <sup>20</sup>. The utility of XES was notably demonstrated in Bergmann, et al. <sup>21</sup>, which studied the oxidation state of Mn in Photosystem II via the  $K\beta$  emission lines.



**Figure 1.14** Molecular orbital perspective of  $K\beta$  and its satellite lines. Here the M represent the target metal and the L is the ligand. From Pollock, et al., 2013<sup>22</sup>.

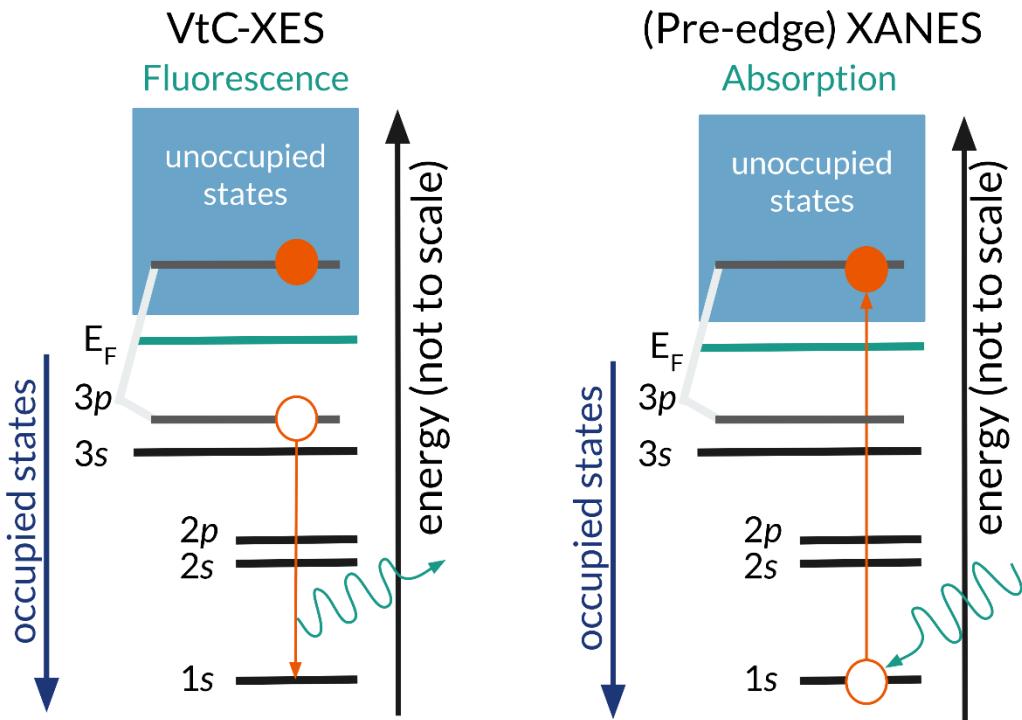
For 3d transition metals, the high energy  $K\beta$  lines are also called Valence-to-Core XES (VtC-XES) and are highly sensitive to ligand identity due to the hybridization of orbitals, as shown in Fig. 1.14. Study of VtC-XES is relatively new, enabled by the recent improvements in synchrotron and lab-based spectrometers. In addition, theoretical calculations were developed to be in good agreement with experiment<sup>23</sup>, where ground state density functional theory (DFT) was a sufficient approximation.

Notably, Pollock and coworkers have developed a theoretical approach to calculating VtC-XES<sup>24</sup> and used it to study the iron oxidation in the iron-molybdenum cofactor (FeMoco) in nitrogenase.<sup>25</sup> However, although VtC-XES is sensitive to ligand identity and other ligand properties, it is not a good quantitative method (rather, it qualitatively follows trends) to determining those properties.<sup>26</sup>

For elements with their valence shells in the  $3p$  orbital, such as phosphorus and sulfur, the Valence-to-Core XES (VtC-XES), which were the high energy  $K\beta$  lines for the  $3d$  transition metals, now becomes just the  $K\beta$  lines, meaning VtC-XES becomes dipole allowed and thus is stronger in intensity compared to the non-dipole allowed transitions for the  $3d$  shell. Several chapters this dissertation is focused on the VtC-XES of both phosphorus and sulfur. Prior and contemporary works show that VtC-XES for phosphorus and sulfur can help identify chemical classes.<sup>27-30</sup>

### 1.3 Are XANES and VtC-XES Complimentary?

XANES and VtC-XES are often seen as “complimentary”<sup>31-33</sup> as they are sensitive to similar properties even though they probe different states, specifically unoccupied and occupied states, respectively, as demonstrated in Fig. 1.15. Some chapters of this dissertation are focused on analyzing whether VtC-XES and XANES are indeed complementary in terms of the strength of the chemical information they encode, or if this information is just highly coincidental. Largely, we aimed to quantify the amount of scientifically independent information encoded in both the VtC-XES and very near-edge XANES and the computation ease at deriving desired properties from each spectroscopic technique, which had seen mostly qualitative discussion.



**Figure 1.15** VtC-XES probes occupied electronic states while XANES probes the unoccupied electronic states.

For example, Jahrman et al.<sup>34</sup> showed that the VtC-XES of vanadium of various vanadium oxides lacked distinct spectral features because of the high symmetry and simple bonding environment rather than the lack of sensitivity of VtC-XES. Additionally, it is expected that 3d transition metals without any 3d electrons (e.g., Cr<sup>6+</sup>) or with complete 3d shells (e.g., Zn<sup>2+</sup>) will have sensitivities to coordination in either XANES pre-edge features or VtC-XES, but not both. Another option is that one technique, such as XANES, will be too sensitive and thus encode too much information, creating correlated features and muddling experimental conclusions.<sup>12</sup> These examples bring up the question of the strength of chemically relevant information encoded in each spectroscopy technique.

## 1.4 References

1. J. J. Rehr and R. C. Albers, *Reviews of Modern Physics*, 2000, **72**, 621-654.
2. D. E. Sayers, E. A. Stern and F. W. Lytle, *Physical Review Letters*, 1971, **27**, 1204-1207.
3. G. Bunker, *Introduction to XAFS: A Practical Guide to X-ray Absorption Fine Structure Spectroscopy*, Cambridge University Press, Cambridge, 2010.
4. C. A. Ashley and S. Doniach, *Physical Review B*, 1975, **11**, 1279-1288.
5. M. Newville, *Reviews in Mineralogy and Geochemistry*, 2014, **78**, 33-74.
6. P. Glatzel and U. Bergmann, *Coordination Chemistry Reviews*, 2005, **249**, 65-95.
7. H. A. Kramers and W. Heisenberg, *Zeitschrift für Physik*, 1925, **31**, 681-708.
8. D. Matsakis, A. Coster, B. Laster and R. Sime, *Physics Today*, 2019, **72**, 10-11.
9. L. Meitner, *Zeitschrift für Physik*, 1922, **9**, 145-152.
10. X-ray absorption spectroscopy, (accessed April 4, 2023).
11. X-ray Absorption Near Edge Spectroscopy- XANES, (accessed April 1, 2023).
12. E. P. Jahrman, L. L. Yu, W. P. Krekelberg, D. A. Sheen, T. C. Allison and J. L. Molloy, *Journal of Analytical Atomic Spectrometry*, 2022, **37**, 1247-1258.
13. B. Ravel and M. Newville, *Journal of Synchrotron Radiation*, 2005, **12**, 537-541.
14. M. Newville, *Journal of Physics: Conference Series*, 2013, **430**, 012007.
15. J. J. Rehr, J. J. Kas, F. D. Vila, M. P. Prange and K. Jorissen, *Phys. Chem. Chem. Phys.*, 2010, **12**, 5503-5513.
16. J. Terry, M. L. Lau, J. Sun, C. Xu, B. Hendricks, J. Kise, M. Lnu, S. Bagade, S. Shah, P. Makhljani, A. Karantha, T. Boltz, M. Oellien, M. Adas, S. Argamon, M. Long and D. P. Guillen, *Appl. Surf. Sci.*, 2021, **547**, 149059.
17. M. Muñoz, P. Argoul and F. Farges, 2003, **88**, 694-700.
18. A guide for using the Wavelet Transform in Machine Learning, 2023).
19. H. Funke, M. Chukalina and A. Rossberg, *Physica Scripta*, 2005, **2005**, 232.
20. F. De Groot and A. Kotani, 2008, DOI: 10.1201/9781420008425.
21. U. Bergmann, M. M. Grush, C. R. Horne, P. DeMarois, J. E. Penner-Hahn, C. F. Yocom, D. W. Wright, Dubé, W. H. Armstrong, G. Christou, H. J. Eppley and S. P. Cramer, *The Journal of Physical Chemistry B*, 1998, **102**, 8350-8352.
22. C. J. Pollock, K. Grubel, P. L. Holland and S. DeBeer, *Journal of the American Chemical Society*, 2013, **135**, 11803-11808.
23. D. R. Mortensen, G. T. Seidler, J. J. Kas, N. Govind, C. P. Schwartz, S. Pemmaraju and D. G. Prendergast, *Physical Review B*, 2017, **96**, 125136.
24. C. J. Pollock, M. U. Delgado-Jaime, M. Atanasov, F. Neese and S. DeBeer, *Journal of the American Chemical Society*, 2014, **136**, 9453-9463.
25. K. M. Lancaster, M. Roemelt, P. Ettenhuber, Y. Hu, M. W. Ribbe, F. Neese, U. Bergmann and S. DeBeer, *Science*, 2011, **334**, 974-977.
26. C. J. Pollock and S. DeBeer, *Accounts of Chemical Research*, 2015, **48**, 2967-2975.
27. W. M. Holden, E. P. Jahrman, N. Govind and G. T. Seidler, *The Journal of Physical Chemistry A*, 2020, **124 (26)**, 5415-5434.
28. S. Yasuda and H. Kakiyama, *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.*, 1979, **35**, 485-493.
29. S. Yasuda, *Bulletin of the Chemical Society of Japan*, 1984, **57**, 3122-3124.

30. Z. Mathe, O. McCubbin Stepanic, S. Peredkov and S. DeBeer, *Chemical Science*, 2021, **12**, 7888-7901.
31. R. A. Mori, E. Paris, G. Giuli, S. G. Eeckhout, M. Kavčič, M. Žitnik, K. Bučar, L. G. M. Pettersson and P. Glatzel, *Inorganic Chemistry*, 2010, **49**, 6468-6473.
32. S. N. MacMillan, R. C. Walroth, D. M. Perry, T. J. Morsing and K. M. Lancaster, *Inorganic Chemistry*, 2015, **54**, 205-214.
33. M. Qureshi, S. H. Nowak, L. I. Vogt, J. J. H. Cotelesage, N. V. Dolgova, S. Sharifi, T. Kroll, D. Nordlund, R. Alonso-Mori, T.-C. Weng, I. J. Pickering, G. N. George and D. Sokaras, *Phys. Chem. Chem. Phys.*, 2021, **23**, 4500-4508.
34. E. P. Jahrman, W. M. Holden, N. Govind, J. J. Kas, J. Rana, L. F. J. Piper, C. Siu, M. S. Whittingham, T. T. Fister and G. T. Seidler, *Journal of Materials Chemistry A*, 2020, **8**, 16332-16344.

## 2 Chapter 2 – Survey of Theoretical Methods

This chapter is an introduction to the theoretical methods used in later chapters of this dissertation. Specifically, we present the context of NWChem<sup>1, 2</sup> and its use of both density functional theory (DFT) and linear response time-dependent density functional theory (LR-TDDFT) as well as the reliability of the resulting calculations.

### 2.1 Density Functional Theory (DFT)

Density functional theory (DFT) is a quantum mechanical theory that capitalizes on the idea of functionals to estimate the electron density of a many-body system. The most important idea of DFT is one of two Hohenberg-Kohn theorems – the ground state of a many-electron system is uniquely determined by the electron density (defined in three spatial coordinates). In essence,

$$\Psi = \Psi[n(\vec{r})] \quad (2.1)$$

where the ground state  $\Psi$  becomes a functional of the spatial electronic density  $n(\vec{r})$ <sup>3</sup>. This theorem is advantageous in that one can still use the variational principle. In other words, the electronic density  $n(\vec{r})$  that minimizes the functional also minimizes the ground state energy.

An extension of the original DFT approach is Kohn-Sham DFT (KS-DFT), which instead treats electrons as noninteracting and moving in an effective potential<sup>4</sup>. KS-DFT is now a standard approach. The solution to a KS many-electronic system is a wavefunction that satisfies antisymmetric requirements called a Slater determinant.

The Kohn-Sham equation (in which a Slater determinant is a solution) is the eigenvalue equation

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v_{eff}(r)\right)\phi_i(r) = \epsilon_i\phi_i \quad (2.2)$$

where  $\epsilon_i$  is the energy corresponding to the  $\phi_i$  KS orbital.

Historically, the exchange-correlation in KS-DFT was formulated within the local-density approximation (LDA). Over the years, this has been generalized to broader classes of exchange-correlation functionals like the generalized gradient approximation (GGA), orbital dependent global hybrid functionals which mix in a fraction of Hartree-Fock exchange, range-separated hybrids, and double hybrids which include the unoccupied orbitals.<sup>5</sup> It is worthwhile to note that extensive research is still ongoing to improve DFT exchange-correlation functionals, including machine learning approaches<sup>6</sup>.

## 2.2 Time-Dependent Density Functional Theory (TDDFT)

The problem with modeling excited states is the many-body problem, so a large factor in choosing a level of theory depends on whether it can do a sufficient job approximating an excited state. Although ground state DFT can work reasonably well for VtC-XES calculations because of the final state rule, TDDFT is a more approachable choice as it helps capture the multideterminant character of an excited state within the space of single excitations. TDDFT is formally exact, meaning that the time dependent electronic density can be used to exactly describe a system undergoing a time dependent perturbation. Specifically linear response time-dependent density functional theory (LR-TDDFT) uses perturbation theory to adiabatically approximate electronic response (in contrast to using real time propagation, or RT-TDDFT). Moreover, LR-TDDFT uses conventional ground-state Kohn-Sham (KS) DFT functionals and has been implemented in several leading computational chemistry programs.<sup>7</sup>

## 2.3 Implementations of DFT and TDDFT

The following section is an overview of available codes and their levels of theory; however, it is neither a comprehensive list nor a complete discussion of the applications of each code. For a more complete discussion, see Nascimento, et al.<sup>8</sup> and Rana, et al.<sup>9</sup>.

Starting with the “simplest” theory and going to the most sophisticated, we will start with atomic model codes, which include XCLAIM, quanty, and crispy. These codes use atomic cross-sections, multiplet theory with (mostly) fitted parameters, and model Hamiltonians. However, recent work has been done to remove these empirical parameters and replace them with *ab initio* calculated ones<sup>10</sup>. Core-to-core XES (CtC-XES) is commonly calculated using multiplets.

Next are the DFT level theories, which include WIEN2k, ABINIT, VASP, and CASTEP. These codes are accurate for ground state properties but are less reliable for excited states because they utilize the final state rule. A more sophisticated code is FEFF, which uses quasi-particle Green’s function theory. FEFF is appropriate for excited states and is very efficient, but it is missing some many-body physics. These properties, along with its historical context and supported GUI, make it one of the most popular choices when calculating condensed phases.

Then there are the Bethe-Salpeter Equation (BSE) codes, which include Exc!ting and OCEAN. These are the most accurate but the most computationally demanding and are thus less established. Finally, the most sophisticated quantum chemistry codes are MRCI, MRCC, CASPT2, and QMC, which utilize multi-configurational self-consistent field (MCSCF). Although these codes are highly accurate for systems where DFT is no longer adequate, their complexity is almost intractable<sup>11</sup>.

We used NWChem for our work to perform the DFT and TDDFT calculations with Gaussian basis sets as they are used for molecular systems and thus the following section presents a focused discussion on NWChem.

## 2.4 NWChem: A Closer Look

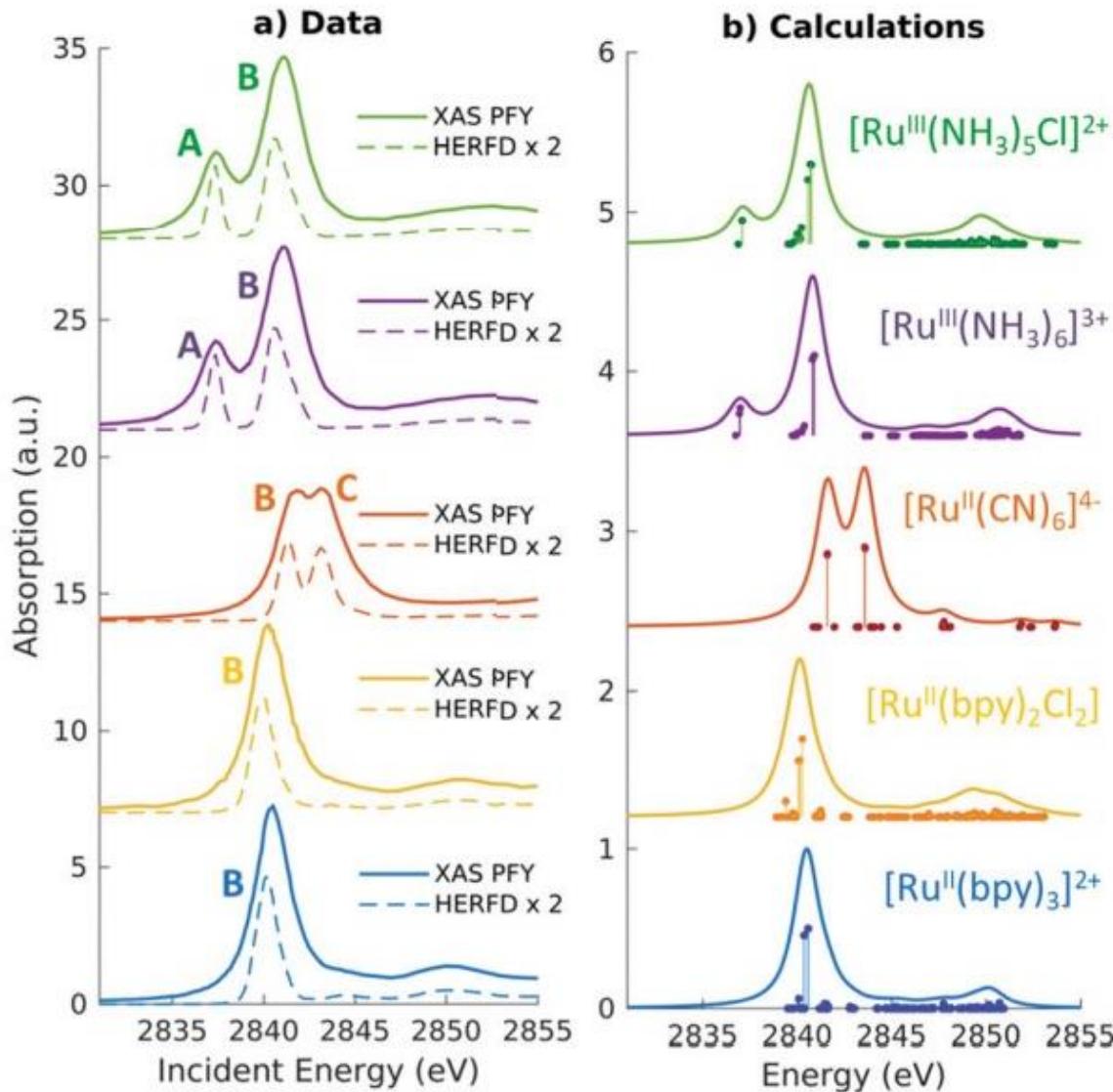
NWChem<sup>1, 2</sup> is a computational chemistry software package that was developed by a consortium of scientists and contains modules that address various electronic structure calculations which were designed to run on parallel processing cluster systems and supercomputers. It is also available as open source under the Educational Community License. NWChem has many functionalities, including Hartree-Fock self-consistent field (SCF) and post-SCF correlated many-body approaches. NWChem includes single- and multi-reference (MR) theories, ground and excited-state theories, and linear-response (LR) coupled-cluster (CC) theories. Moreover, NWChem offers ground and excited-state molecular dynamics (MD) and both LR-TDDFT and RT-TDDFT.<sup>2</sup> Specifically, we utilize the Gaussian basis set DFT and LR-TDDFT modules in NWChem for XANES and VtC-XES calculations, respectively.

### 2.4.1 XANES Calculations

NWChem uses Gaussian basis sets to represent atomic orbitals. The basis sets are atom-centered and are comprised of character from every orbital (*s*, *p*, *d*, *f*, and *g*). The coefficients in front of these basis orbitals then determine how strong the character of each orbital an electron has, i.e., a *1s* orbital will still have *g* orbital components, albeit it should be extremely small. These Gaussian basis sets are ideal approximations for molecular systems, i.e., not condensed matter.

Many studies have found good agreement between theoretically calculated spectra via NWChem and experiment<sup>8, 12-14</sup>. K. Lopata, et al.<sup>12</sup> has a good overview with case studies comparing RT-TDDFT and LR-TDDFT calculations of core-level excitations, specifically of the oxygen *K*-edge of water and carbon monoxide, the carbon *K*-edge of carbon monoxide, the ruthenium *L*<sub>3</sub>-edge for the hexaammineruthenium(III) ion, and the carbon and fluorine *K*-edges for a series of fluorobenzenes.

Recently, Biasin, et al.<sup>14</sup> and Nascimento and Govind<sup>8</sup> both feature Ru *L*<sub>3</sub>-edge XANES spectra, as shown in Fig. 2.1. We see good agreement, with the transitions (shown as “sticks” under the spectra) depicting spectral features quite well. Note that a broadening is applied to these sticks, representing electronic transitions; the broadening corresponds to both the core-hole lifetime and experimental resolution.



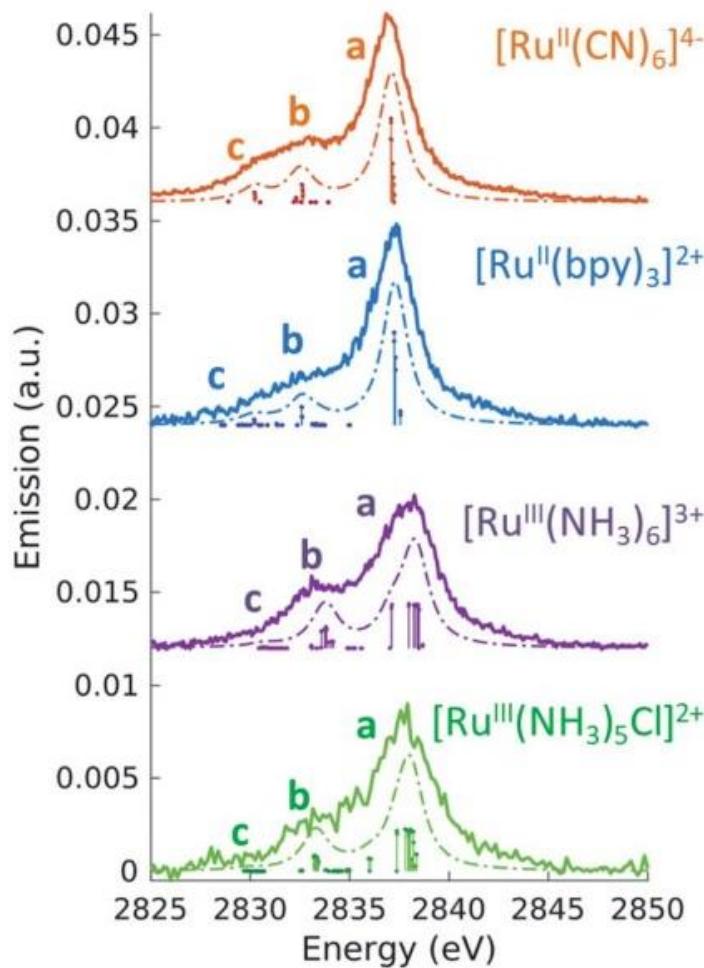
**Figure 2.1** Experimental and simulated Ru L<sub>3</sub> edge XANES spectra (2p - 4d orbitals). Taken from Biasin, et al.<sup>14</sup> and Nascimento and Govind<sup>8</sup>.

#### 2.4.2 VtC-XES Calculations

As stated earlier, VtC-XES calculations in NWChem utilize LR-TDDFT. Here, the core-hole and valence-ionized states are approximated using only one Slater determinant of ground-state Kohn-Sham (KS) orbitals each. The energies of the transitions are calculated as the difference

in the valence and core orbital energies, while the corresponding transition probabilities are proportional to the dipole transition moment between the respective valence and core KS orbitals.

There are a variety of studies which validate NWChem calculations with experimental data for VtC-XES<sup>8, 15, 16</sup>. A case study of VtC-XES calculations from NWChem is shown in Fig. 2.2. Here, the VtC-XES of the 4d transmission metal Ru is calculated, which is the 4d to 2p transition (rather than the 3d to 2p transition, as is the case for 3d transmission metals).



**Figure 2.2** Experimental and simulated Ru VtC-XES spectra (4d - 2p orbitals). Taken from Biasin, et al. 2021<sup>14</sup> and Nascimento and Govind, 2022<sup>8</sup>.

Of special note, Holden, et al.<sup>17</sup> compared experimentally measured sulfur VtC-XES spectra with those calculated by NWChem on a wide range of both organic and inorganic sulfur compounds and found relatively good agreement. My first project (Chapter 6) capitalized on this agreement to expand the library of sulforganics with both XANES and VtC-XES theoretically calculated spectra using NWChem.

## 2.5 References

1. M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. De Jong, *Computer Physics Communications*, 2010, **181**, 1477-1489.
2. E. Apra, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Attafynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauet, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Fruchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Gotz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jonsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. M. del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tippuraju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, A. Vazquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Wolinski, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison, *J. Chem. Phys.*, 2020, **152**, 26.
3. P. Hohenberg and W. Kohn, *Physical Review*, 1964, **136**, B864-B871.
4. W. Kohn and L. J. Sham, *Physical Review*, 1965, **140**, A1133-A1138.
5. J. P. Perdew and K. Schmidt, *AIP Conference Proceedings*, 2001, **577**, 1-20.
6. R. Pederson, B. Kalita and K. Burke, *Nature Reviews Physics*, 2022, **4**, 357-358.
7. M. Huix-Rotllant, N. Ferré and M. Barbatti, in *Quantum Chemistry and Dynamics of Excited States*, 2020, DOI: <https://doi.org/10.1002/978111941774.ch2>, pp. 13-46.
8. D. R. Nascimento and N. Govind, *Phys. Chem. Chem. Phys.*, 2022, **24**, 14680-14691.
9. R. Rana, F. D. Vila, A. R. Kulkarni and S. R. Bare, *ACS Catalysis*, 2022, **12**, 13813-13830.
10. J. K. Charles Cardot, Jared Abramson, John Rehr, Gerald T. Seidler, *arXiv*, 2023.
11. K. McCardle, *Nature Computational Science*, 2021, **1**, 777-777.
12. K. Lopata, B. E. Van Kuiken, M. Khalil and N. Govind, *JOURNAL OF CHEMICAL THEORY AND COMPUTATION*, 2012, **8**, 3284-3292.
13. D. Boglaienko, A. Andersen, S. M. Heald, T. Varga, D. R. Mortensen, S. Tetef, G. T. Seidler, N. Govind and T. G. Levitskaia, *Journal of Alloys and Compounds*, 2022, **897**, 162629.
14. E. Biasin, D. R. Nascimento, B. I. Poulter, B. Abraham, K. Kunus, A. T. Garcia-Esparza, S. H. Nowak, T. Kroll, R. W. Schoenlein, R. Alonso-Mori, M. Khalil, N. Govind and D. Sokaras, *Chemical Science*, 2021, **12**, 3713-3725.
15. E. P. Jahrman, W. M. Holden, N. Govind, J. J. Kas, J. Rana, L. F. J. Piper, C. Siu, M. S. Whittingham, T. T. Fister and G. T. Seidler, *Journal of Materials Chemistry A*, 2020, **8**, 16332-16344.

16. Y. Zhang, S. Mukamel, M. Khalil and N. Govind, *Journal of Chemical Theory and Computation*, 2015, **11**, 5804-5809.
17. W. M. Holden, E. P. Jahrman, N. Govind and G. T. Seidler, *The Journal of Physical Chemistry A*, 2020, **124 (26)**, 5415-5434.

### 3 Chapter 3 – Interpreting XAFS and the Bane of the Inverse Problem

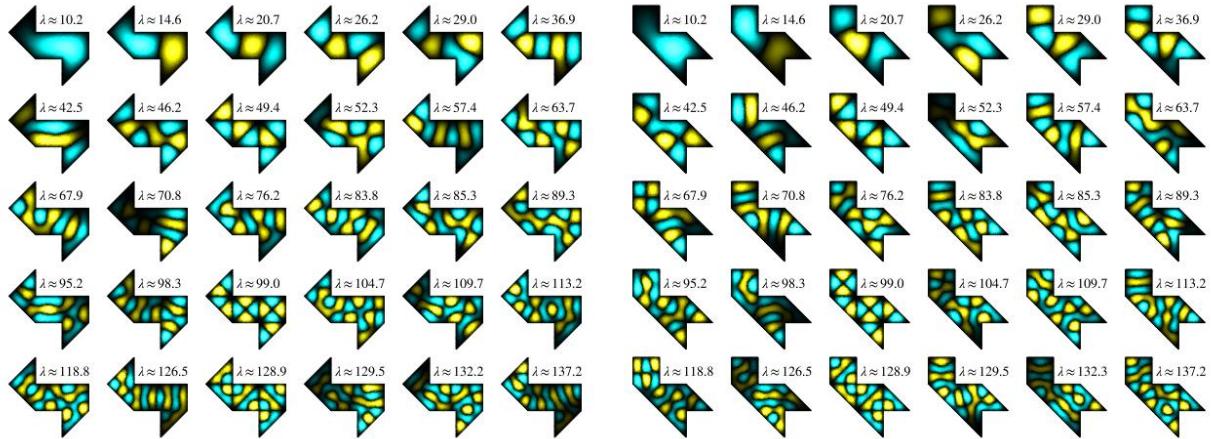
In science, the “forward” problem means starting with properties or structures and obtaining experimental measurements or observations. This framework is equivalent to calculating expected responses given certain inputs and is relatively easy to do, as long as a good theoretical model is available. For example, the forward problem could be calculating the UV-Vis spectrum, the magnetic susceptibility, or the stress-strain curve of a certain material.

On the other hand, let’s say we want to come up with a material that has a certain property, like a specific band gap. This problem would be classified as an “inverse” problem because it is inverting the structure to measurement effects. In other words, the general task of the inverse problem is using observations to infer causal parameters. Inverse problems are often termed “ill-posed”, which means they either (1) don’t have a solution, (2) the solution is not unique, or (3) the solution’s behavior does not change continuously with the initial conditions. In X-ray spectroscopy, the “inverse” problem – going from spectra to structure – often runs into problems with both (2) and (3).

In this chapter, I start with an example of a classical inverse problem, namely using the acoustic modes to effectively “hear” the shape of a drum. Then I will discuss the ill-posed nature of the XAFS inverse problem, for both EXAFS and XANES, including some alternative methods. Particularly, I will focus on different ways people have incorporated prior knowledge to interpret XAFS spectra, both at an individual level and as ensemble, to provide context to my approach throughout this dissertation.

### 3.1 Classical Inverse Problems

Consider whether you can hear the shape of the drum, whether it is round, square, annular, etc. The shape of a drumhead dictates the frequencies at which the drum can vibrate via the Helmholtz equation. However, when this question was originally posed, it was unknown whether the same frequencies could be produced from different shapes. Finally, in the early 1990s, Gordon, Webb, and Wolpert determined that the spectrum frequencies did not uniquely determine the shape of the drum. For example, Fig. 3.1 shows two drum shapes that produce the same eigenvalues.



**Figure 3.1** Can you hear the shape of a drum? These two drum shapes, although different, produce the same eigenvalues, as indicated by the  $\lambda$  values. Figure from Wikipedia <sup>1</sup>.

In addition to source reconstruction from acoustics, calculating the density of the Earth from measurements of its gravitational field and reconstructing a three-dimensional object from its two-dimensional shadows are other examples of classical inverse problems where there are multiple solutions, just like hearing the shape of the drumhead.

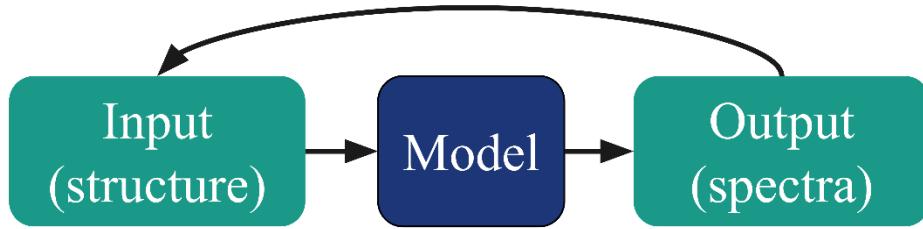
## 3.2 Goals of the XAFS Inverse Problems

The goal of much XAFS inference is to obtain chemical or atomic structure from XAFS spectra. The XAFS inverse problem is similar to hearing the shape of a drumhead because there are multiple solutions; different structures may cause the same spectral features, or distinct spectral features are correlated with the same structural properties <sup>2</sup>. The difficulties of analyzing XAFS spectra arise from both the quantum mechanical nature of the electronic states that XAFS probes as well as the loss of information from both experimental resolution and collective disorder or correlations, such as from phonons. Thus, it is impossible to analyze spectra independently without any background knowledge of the system. Instead, I discuss some ways to incorporate prior knowledge into XAFS analysis to tackle the inverse problem in XAFS, turning the inverse problem into an “informed” inverse problem instead.

## 3.3 Alternatives to the Inverse Problem in XAFS

### 3.3.1 Repeating the Forward Problem (XANES & EXAFS)

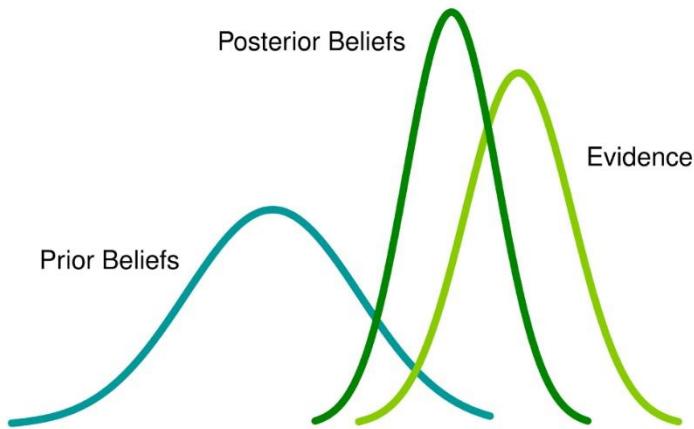
One generic alternative to the inverse problem – which occurs when the forward problem is solved, as is the case for XAFS – is to repeatedly perform the forward problem, each time changing the input structure. One can then adjust the input structures depending on how well the output matches the target, as shown in Fig. 3.2. However, this approach only works within a good model, i.e., one has obtained enough prior knowledge of the system that the structural parameter space is small enough to be tractable. In the context of XAFS, repeating the forward problem often means theoretically calculating spectra using DFT and comparing them to experimental spectra. Because DFT calculations involve many parameters and can be time consuming, this process is only tractable for small explorations.



**Figure 3.2** If there is a good enough model, then repeating the forward problem to obtain target observations, or spectra, from different input structures is one way to combat the inverse problem.

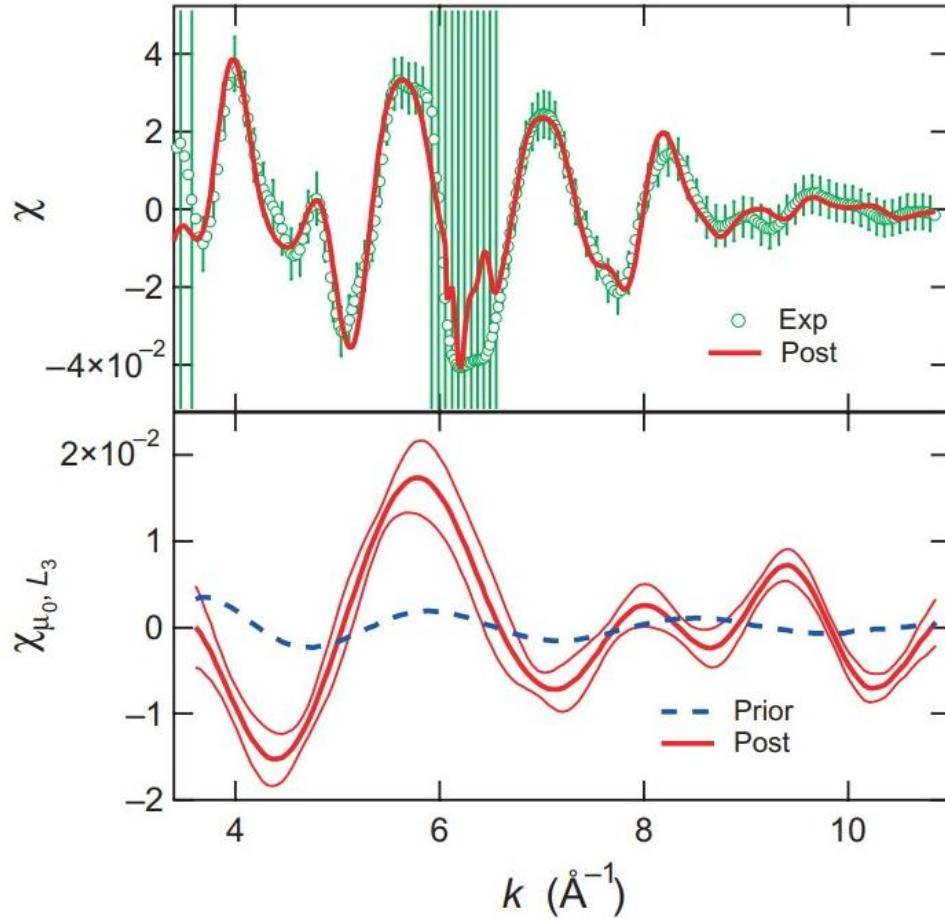
### 3.3.2 Bayesian Approach (EXAFS)

Another alternative is using formal Bayesian statistics, as shown in Fig. 3.3, specifically the Bayes–Turchin approach to fit X-ray absorption fine-structure (EXAFS) and thus calculate EXAFS parameters, such as bond length and disorder. Because EXAFS involves taking the Fourier transform, the loss of phase information is a particularly prominent issue in EXAFS analysis. Moreover, both thermal and structural disorder can cause broadening in EXAFS data; this broadening is encompassed in the Debye-Waller factor in the EXAFS equation. Both effects can make inferring EXAFS parameters difficult and thus using Bayesian statistics is beneficial.



**Figure 3.3** Bayes analysis involves specifying a prior probability distribution and formalizing the probability of observations (evidence) and thus the probability of the model given the evidence (posterior). From Analytics Vidhya <sup>3</sup>.

Although this process has been shown to successfully work on multiple systems <sup>4-7</sup>, as shown in Fig. 3.4 where the experimental data in the top panel matches the posterior well, it has not gained much traction in the community. This unpopularity is due the complexity of the formalism and the large computational cost of the high-dimensional numerical integration. While highly desirable from the basis of modern statistics, this formalism is inaccessible to many researchers utilizing X-ray spectroscopy experiments who use the measurement as a characterization technique and do not have formal statistical training.

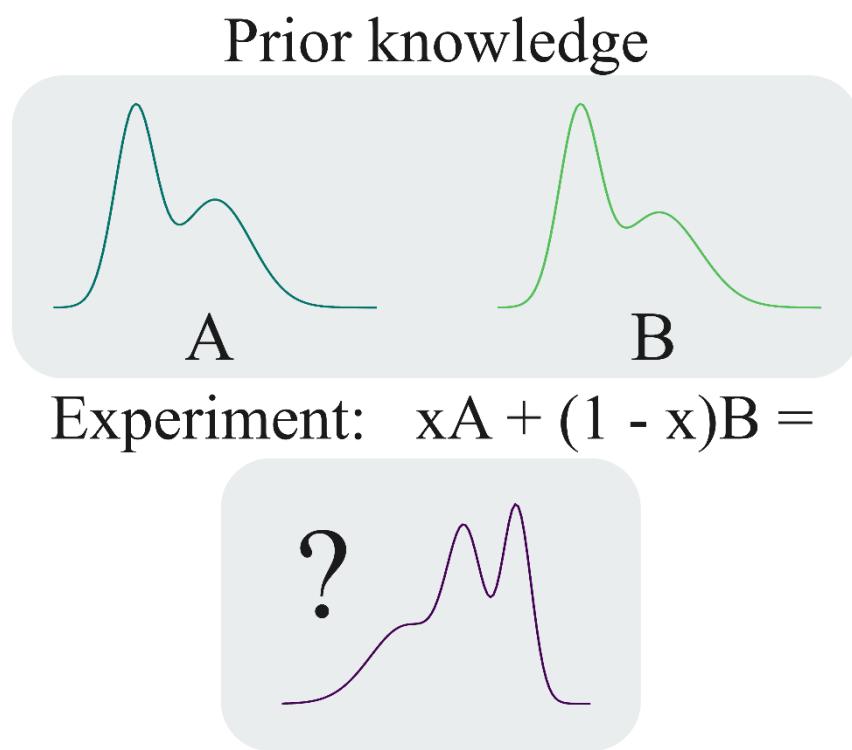


**Figure 3.4** The top panel shows the experimental data  $\chi(k)$  and respective errors. The solid (red) curve is the most probable curve resulting from the fit. The bottom panel shows the prior and post values of  $\chi_{\mu_0, L_3}$  as well as the *a posteriori* error band. Taken from Krappe and Rossner.<sup>5</sup>

### 3.3.3 Linear Combination Fitting (LCF) to a reference library (XANES)

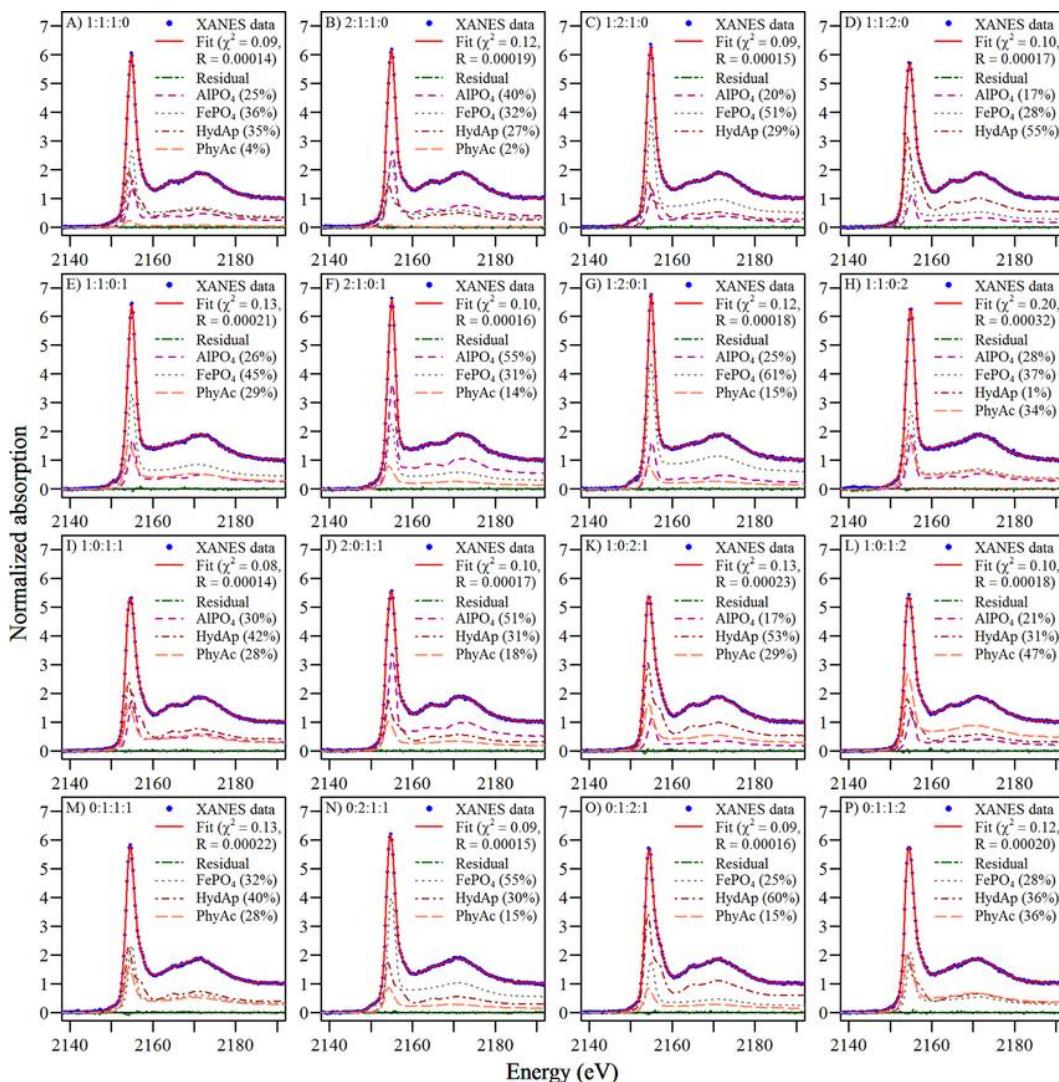
Tackling the inverse problem for XANES becomes more difficult than EXAFS due to the requirement for full multiple scattering and thus the inability to perform a Fourier transform to obtain physically meaningful information. Rather, XANES analysis is largely done on the spectra itself, not on a transform.

Additionally, sources of information loss are different for XANES rather than EXAFS. For example, information is lost partly due to the Heisenberg uncertainty principle, which states that you cannot know exactly both momentum and position, or energy and time in the case of X-ray spectra. Because excited states have inherent lifetimes, electronic transitions are broadened in energy. Moreover, limits on experiment apparatuses, such as the resolution of your monochromator, have inherent resolution. Thus, any transition too close in energy will be smoothed out and indistinguishable from each other. Other types of spectral broadening can occur from more classical phenomena, such as plasmons and thermal vibrations.



**Figure 3.5** Linear combination fitting XANES spectra to reference spectra is the most common analysis for XANES. However, choosing reference spectra must be done well or any inferences are unreliable.

The most common method for analyzing XANES spectra is linear combination fitting (LCF) onto reference spectra, as shown in Fig. 3.5. This procedure capitalizes on the fact that XAS is an average bulk probe, so any components of different structures will contribute directly proportion to their concentration, or percentage of makeup, assuming experimental spectra are composed of varying amounts of reference spectra added together. Then, normalization ensures these coefficients add to zero.



**Figure 3.6** Linear combination fitting to phosphorus XANES spectra. Taken from Werner and Prietzel.<sup>8</sup>

For example, an experimental sample with 2/3 the iron atoms in a 2+ oxidation state and the other 1/3 of the iron atoms in a 3+ oxidation state will result in a spectrum composed of two parts of an iron 2+ oxidation state reference and one part of an iron 3+ oxidation state reference. While this approach is more a data analysis technique rather than directly attacking the inverse problem (because it still runs into issues with overlapping and correlated spectral features), it is another way to incorporate prior knowledge of the system.

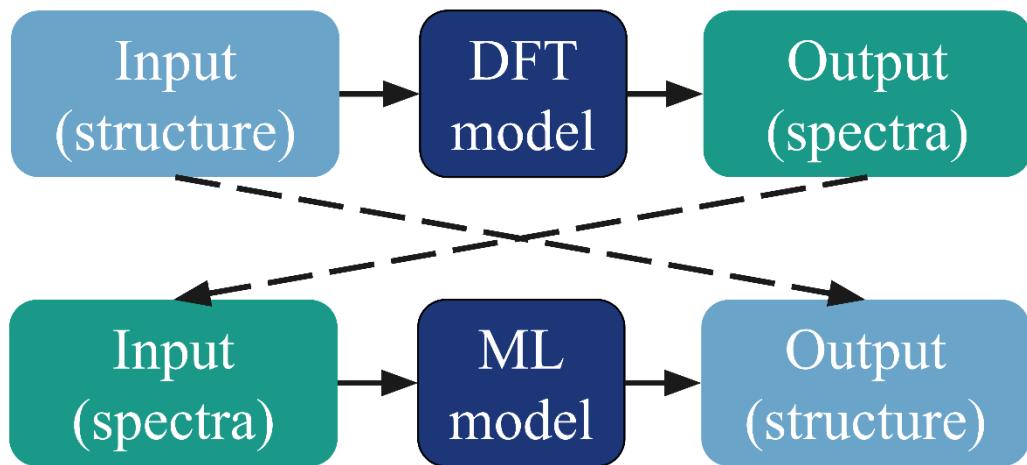
Obviously, things can get complicated quickly as the number of desired properties and unknown parameters increases. Moreover, performing many fits over and over is computationally expensive, especially for large experimental datasets like the one shown in Fig. 3.6. Furthermore, choosing an appropriate library, or reference set, is critical in that it must find a balance between spanning a large enough domain to cover the experimental space but also not have redundant or correlated spectra. This issue becomes especially problematic with the highest uncertainty, or least prior knowledge, of the system.

Finally, this method is prone to propagating errors, especially when fitting to theoretical spectra. It can also propagate any systematic errors in the experiment or normalization. It is especially unreliable if all your reference compounds have different second or third coordination shells, which is often the case for solution studies where reference compounds are usually crystalline.

### 3.3.4 Machine Learning

The final form of tackling the inverse problem in XAFS, which has seen a rise in popularity over the last few years, is utilizing machine learning. Machine learning encodes prior information

in a training dataset, much like collecting reference spectra into a library creates a domain of possible structure-to-spectra relationships, as shown in Fig. 3.7.



**Figure 3.7** Machine learning can encode the structure to spectra relationship into the training set and this the model can learn how to invert that relationship.

Instead of focusing on fitting to spectra, machine learning models can be trained to learn structures instead. Moreover, accurate and relatively easy theoretical calculations have made the creation of large and accurate (enough) datasets for machine learning possible. Chapter 4 will give an overview of machine learning methods before delving into how those methods are used in the context of XAS and XES in Chapter 5.

### 3.4 References

1. Hearing the shape of a drum).
2. E. P. Jahrman, L. L. Yu, W. P. Krekelberg, D. A. Sheen, T. C. Allison and J. L. Molloy, *Journal of Analytical Atomic Spectrometry*, 2022, **37**, 1247-1258.
3. Power of Bayesian Statistics & Probability).
4. H. J. Krappe and H. H. Rossner, *Physical Review B*, 2002, **66**, 184303.
5. H. J. Krappe and H. H. Rossner, *Physica Scripta*, 2009, **79**, 048302.
6. J. J. Rehr, J. Kozdon, J. Kas, H. J. Krappe and H. H. Rossner, *Journal of Synchrotron Radiation*, 2005, **12**, 70-74.
7. H. H. Rossner, D. Schmitz, P. Imperia, H. J. Krappe and J. J. Rehr, *Physical Review B*, 2006, **74**, 134107.
8. F. Werner and J. Prietzel, *Environmental Science & Technology*, 2015, **49**, 10521-10528.

# 4 Chapter 4 – Introduction to Machine Learning

Machine learning (ML) has seen recent explosions in popularity and applications, catalyzed by advances in algorithms, access to computational power, and the ubiquity of large data sets. In the most general sense, ML can be described as data-driven pattern recognition by performing linear decomposition onto nonlinear transformations; these transformations are often called *basis functions*. ML can be divided into three main categories: (1) supervised machine learning, (2) unsupervised machine learning, and (3) reinforcement learning.

Here, I will focus on supervised and unsupervised learning, which are often discussed together, as reinforcement learning is an exciting and complicated field unto itself and has not been implemented in the context of XAFS, yet. This chapter gives a brief overview of the most popular supervised and unsupervised methods as well as context to the more specialized methods used in later chapters.

## 4.1 Machine Learning Basics

In this section, I will cover the basics of machine learning. However, it is helpful to start with a statistical approach to model generation, as machine learning models and algorithms often pull from both frequentist and Bayesian approaches.

### 4.1.1 Bayes Theorem and Maximum Likelihood Estimation

It is helpful to first consider the case where we observe a random variable, or a sampling from some inherent probability distribution. The goal of our model is to then use observations to predict new outcomes based on initial conditions. To accurately do so, we want our model,

characterized by the *parameters*  $\vec{\theta} = \theta_1, \theta_2, \dots$ , to have the best  $\vec{\theta}$  given our *data*, which is comprised of target variables  $\vec{y}$  and input parameters  $\vec{x}$ .

It is helpful to approach this problem from a Bayesian point of view using Bayes theorem

$$P(\theta|X) = P(X|\theta)P(\theta)/P(X). \quad (4.1)$$

Rephrased in words, this equation states the posterior (probability of the parameters given the data) equals the likelihood (probability of the data given the parameters) times the prior (probability of the parameters), divided by the evidence (probability of the data). Let's say we have no prior knowledge and can normalize the distribution later. Now we can ignore the prior and evidence, respectively. Thus, the likelihood just represents the probability of observing the data given the model parameters. Maximizing this likelihood, called *Maximum Likelihood Estimation (MLE)*, is a frequentist approach but is easily motivated by Bayesian inference (if using naive or uniform priors). However, if you have prior knowledge about the form of the solution, then you would want to use the purely Bayesian approach to solving this problem, called *Maximum a Posteriori (MAP)*.

Thus, given a set of observations  $\vec{x}_i$  for  $i = 1 \dots N$ , the likelihood we want to maximize is given by  $\max_{\theta} \prod_{i=1}^N P(\vec{x}_i | \vec{\theta})$ . However, the maximum of any monotonic function is the same as the maximum of the log of that function. Thus, we can take the log of the entire equation and use log rules to transform this *cost, or objective, function* into  $\max_{\theta} \sum_{i=1}^N \ln P(\vec{x}_i | \vec{\theta})$ . Instead of maximizing this function, often algorithms minimize the negative of the log likelihood. Here, we can also start to make assumptions about the form of the probability distribution. If we assume  $\vec{x}_i$  is normally distributed around the random variable  $X$ , then we get an ordinary least squares solution (OLS). This solution is also equivalent to  $\chi^2$  minimization. However, again, this is only true for Gaussian noise.

### 4.1.2 Bias-variance Tradeoff and Model Complexity

If data is sampled from some inherent distribution and is thus a random variable, the expected error in your fit, i.e.,  $E[(y - \hat{f}(x))^2]$ , can be decomposed into three terms: the irreducible error, the bias-squared terms, and the variance term. The irreducible error comes from the noise in the data – here, we assume your target data  $y$  can be represented by  $\vec{y} = f(\vec{x}; \vec{\theta}) + \epsilon$ , where  $f(\vec{x}; \vec{\theta})$  is a deterministic function that depends on the input parameters  $\vec{x}$  and model parameters (or weights)  $\vec{\theta}$ , added to random fluctuations via  $\epsilon$ , which we can model as  $\sim N(0, \sigma_\epsilon^2)$ , or sampling from a Gaussian distribution.<sup>1</sup>

Given this definition for  $\epsilon$ , then

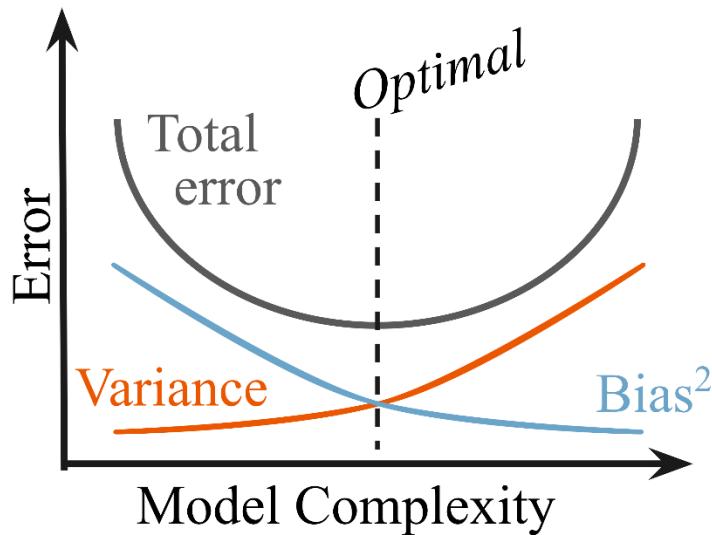
$$E[(y - \hat{f}(x; \theta))^2] = E[(f(x; \theta) - \hat{f}(x; \theta))^2] + \sigma_\epsilon^2, \quad (4.2)$$

where  $\sigma_\epsilon^2$  is the irreducible error. Expanding the other term and performing some algebra, we get

$$\begin{aligned} E[(f(x; \theta) - \hat{f}(x; \theta))^2] &= \left(E[\hat{f}(x; \theta)] - f(x; \theta)\right)^2 + E[(\hat{f}(x; \theta) - E[\hat{f}(x; \theta)])^2] \\ &= \text{bias}[\hat{f}(x; \theta)]^2 + \text{var}(\hat{f}(x; \theta)) \end{aligned} \quad (4.3)$$

The bias-squared term is the error, or deviation, of the data points from the model's predictions, while the variance term represents the deviation of predictions from the real answer with each new draw (or realization) of the random variable. Essentially, the bias-squared term represents error in a single sample, while the variance represents the error from multiple samples.

<sup>2</sup> As can be expected, the bias-squared and variance terms come into play when choosing the complexity of the model.

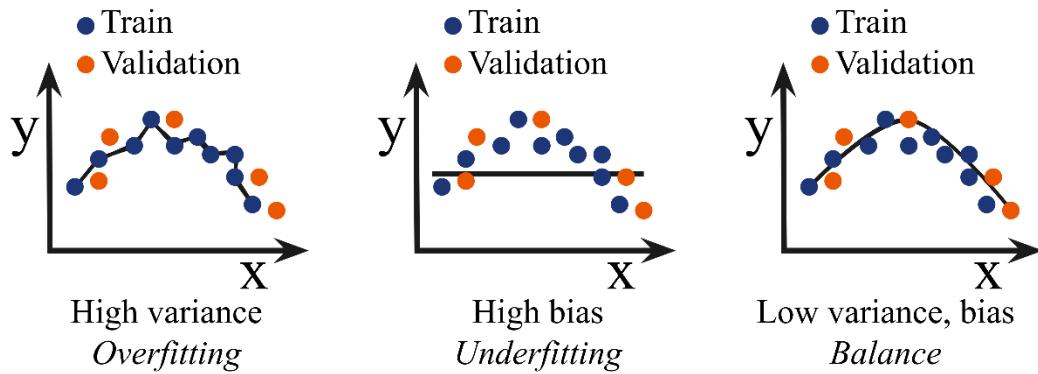


**Figure 4.1** Model complexity and total error – the bias versus variance trade-off.

The importance of model complexity, i.e., the number of hyperparameters, is illustrated in Fig. 4.1. Here, we have plotted the total error (the sum of the irreducible, bias-squared, and variance terms) versus model complexity. A simple model might predict every point as the mean of all points; thus, the errors from draw to draw will be similar (low variance) but within a draw might be high (high variance). On the other hand, a complex model might perfectly predict every data point it has seen (low bias) but will fail in regions where the model has not seen data before (high variance error). Thus, the Goldilocks complexity occurs at the minimum of the total error, which represents the best balance between the bias-squared error and the variance error term.

The reason why bias-variance tradeoff is important is because it indicates how generalizable your predictions are when predicting a new draw of your random variable. Fig. 4.2 shows how the complex models with a high variance and low bias will overfit, meaning that even though accuracy on the training data is great, the accuracy on new or validation data is bad. Underfitting occurs when simple models can't predict well (training and validation accuracies are

both the same, but low). Ideally, both the training and validation accuracies are high and the same, which occurs when a good balance between both bias and variance error terms is met.



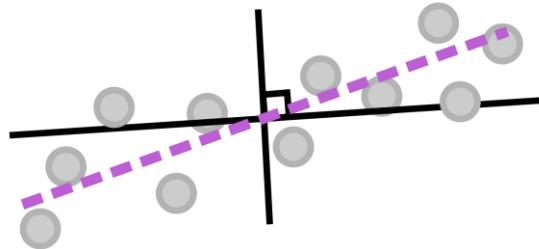
**Figure 4.2** The bias-variance tradeoff impacts the generalizability of the model.

To deal with this overfitting problem, a validation set is often set aside during model training. Throughout training, one can check results using this validation dataset, which the model has not seen before (and is therefore not training data). This validation dataset is *distinct* from test set. *The test set should only be used as a final metric and should not impact any decisions about model architecture*; otherwise, predictions become biased, and the test set is no longer an accurate metric for the generalizability of the trained model. A validation set is used to determine hyperparameters, model complexity, or any other training decision before a final evaluation with the test set. Selecting both the validation and test set should be random unless preserving equal sampling from classes (to avoid biasing results) is needed.

## 4.2 Supervised Machine Learning

### 4.2.1 Regression

A common goal is to obtain a quantitative prediction of a target variable, or output, given a set of inputs, or features. Predicting a quantitative target is called *regression* and is one of the two types of supervised machine learning. The simplest form of regression is linear regression. Again, consider a target variable  $\vec{y} = f(\vec{x}, \vec{\theta}) + \epsilon$ , where  $f(\vec{x}, \vec{\theta})$  is a deterministic function that depends on the input parameters  $\vec{x}$  and model parameters (or weights)  $\vec{\theta}$ . We incorporate random fluctuations via  $\epsilon$ , which we can model as  $\sim N(0, \sigma^2)$ , or Gaussian noise. The analytic solution to finding the optimal weights, or  $\min_{\vec{\theta}} \|\vec{y} - \phi \vec{\theta}\|$ , is  $\theta_{ML} = (\phi^T \phi)^{-1} \phi^T \vec{y}$ , where  $\phi$  is a matrix composed of all observed  $\vec{x}$ . This solution, a line or hyperplane that can be visualized in Fig. 4.3, is thus determined by all the input and output variables and is linear because it is purely matrix multiplication and inversion.

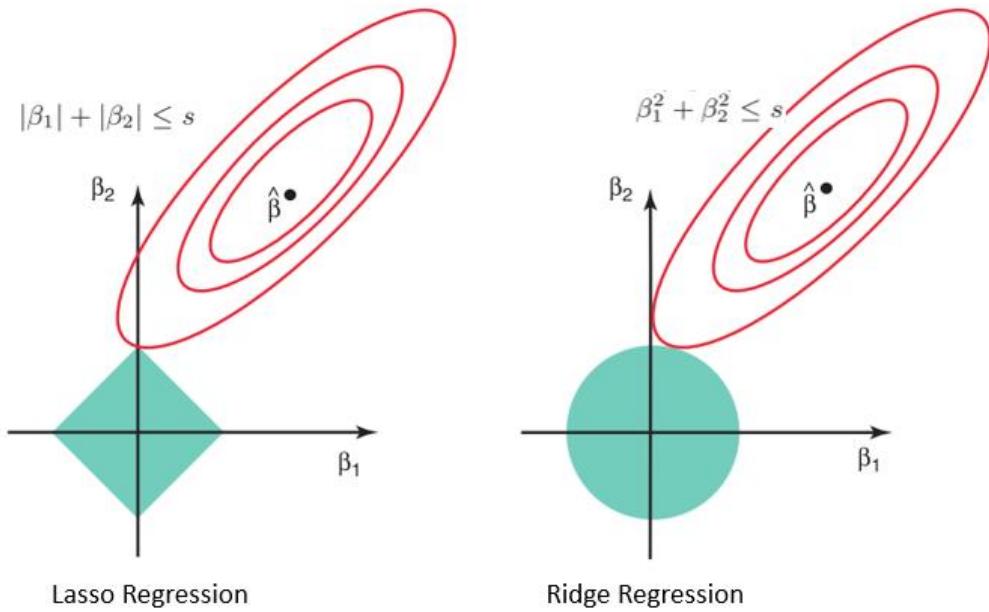


**Figure 4.3** Linear regression is finding a line (or hyperplane) that most follows the linear trends in the data, and as a bonus, it has an analytical solution.

Multivariate regression is the same as linear regression, except when the output variable ( $\vec{y}$ ) is multidimensional (and thus becomes a matrix  $Y$ ). Thus, the epsilon term becomes a matrix

(instead of a vector) and the weight vector  $w$  becomes a matrix  $W$  as well. However, the solution looks the same:  $W_{ML} = (\phi^T \phi)^{-1} \phi^T Y$ .

Linear (and multivariate) regression can be further expanded to incorporate other constraints on the learned weights. Including a regularization term in fits, i.e., penalizing high weights by including a new term in the objective function (which is essentially a Lagrangian the algorithm is trying to minimize), has been shown to have better generalizability. First, a general norm looks like  $\|x\|_p \equiv (\sum |x_i|^p)^{1/p}$ , where  $p$  identifies the type of norm. Thus, to add a regularization term to an objective function, the solution would look something like  $\hat{w} = \min_w \| \vec{y} - \phi w \| + \lambda \|w\|_p$ , where the second term (the regularization term) is modified in scale by the Lagrange multiplier  $\lambda$ . Setting  $p$  to be two is called the  $L_2$  norm, which represents the standard Euclidean distance, and is called *ridge regression*. Setting  $p$  to be one is called the  $L_1$  norm, which is essentially taking the absolute value, and is called LASSO (Least Absolute Shrinkage and Selection Operator) regression.



**Figure 4.4** The regularization term affects the sparsity of the solution. From Penn State Statistics 3.

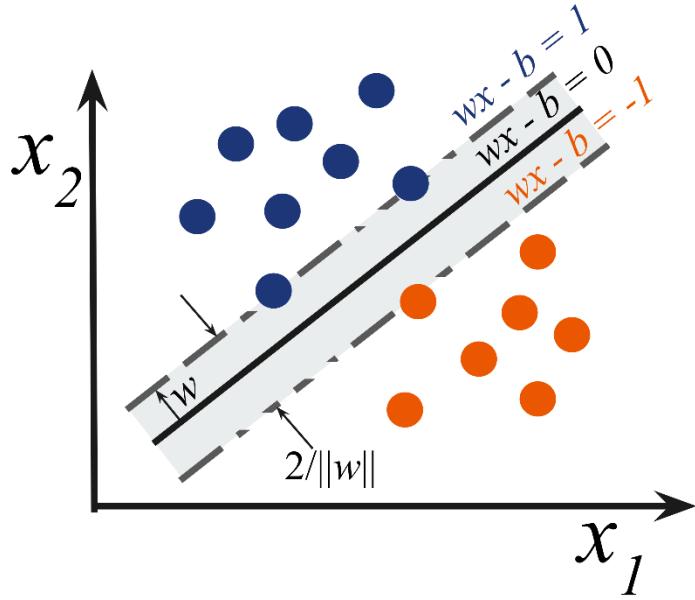
The type of regularization is important because it impacts the sparsity of the solution. For example, let's there are two features, or input variables, and thus the weight vector is also two-dimensional. The equipotential surfaces – in other words, solutions where the overall loss (given by the norm) is the same – look like the green shapes in Fig. 4.4 – LASSO gives diamond shaped surfaces while Ridge Regression gives circular surfaces. Solutions to the first term (i.e., the data-driven term) are represented by the concentric ellipses in red. All these solutions (i.e., any combination of the two weight components that fall along one of those red ellipses) may equally explain the data, but the individual elements of the weight vector will be different.

An overall solution that minimizes the regularization term can be found by slowly growing the green equipotential surface from the origin until it intersects with one of the red curves. By

doing so, the  $L_1$  norm will more likely converge on a solution with sparse weights (or solutions along one axis of the weight vector) because most of the volume is distributed along the axis, whereas the Ridge Regression equipotential surfaces jut out in all directions.

#### 4.2.2 Classification

Instead of learning a (multidimensional) line that follows the trends in the data as in regression, classification finds a *decision boundary*, which is often a line, or in higher dimensions, a (hyper)plane, that best separates data points. Thus, classification predicts categorical targets. One model that performs classification (of linearly separable data) well is a support vector machine (SVM). SVMs generate *unique* decision boundaries by including some leeway, or margin, in the fit in which they try to maximize. This unique solution is called the maximum-margin hyperplane. A demonstration of a SVM is seen in Fig. 4.5. Note that, in order to maximize the margin (the width of which is given by  $2 / \|w\|$ ), SVM must minimize  $w$ . Because the solution is completely determined by points on or within the margin (indicated by the points on the dashed line in Fig. 4.5), these points are called *support vectors*. Moreover, the SVM activation function can be modified to be soft, meaning it can allow for a few outliers to either get close or cross the decision boundary, a beneficial property for nonlinearly separable data.<sup>4</sup>



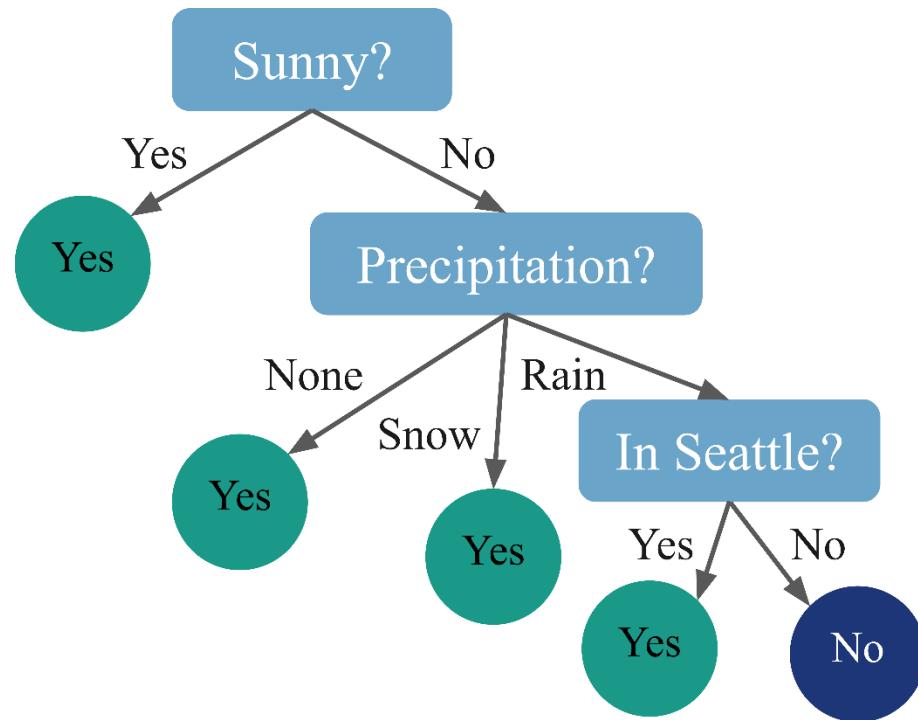
**Figure 4.5** The SVM maximizes the distance from the decision boundary between all data points and it maximizes the margin around this decision boundary. The margin ensures that the solution is unique.

#### 4.2.3 Models for Both Regression and Classification

Most machine learning models can be adapted for both types of supervised machine learning. The next section will give an overview of the most popular models. The easiest supervised classification model to interpret is a decision tree, which is one of their biggest strengths. However, decision trees easily overfit, which is their biggest downfall, because they look for the strongest correlations between input and output features. Correlation does not necessarily mean causation though, so having a good representative set of features is critical when using a decision tree.<sup>4</sup> A diagram of a decision tree can be seen in Fig. 4.6. Decision trees make their branches by recursively looking for the strongest correlation (via entropy) between input features and the target variables. The strongest correlations appear near the top (or root) node,

while further categorizations occur on lower branches. Thus, decision trees are most naturally used for classification problems.

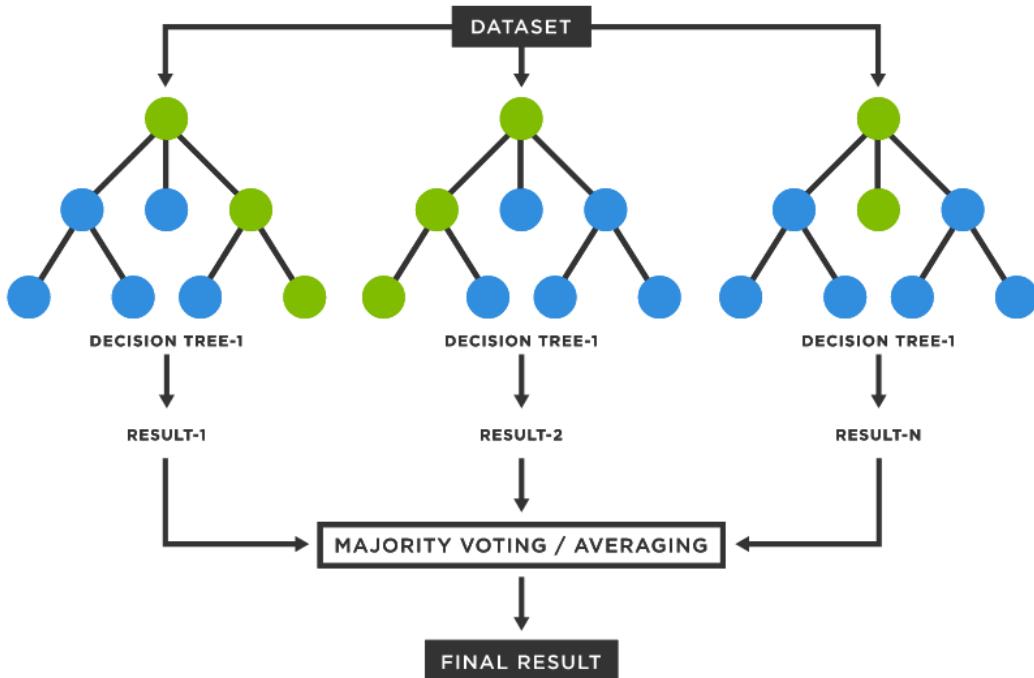
## Should I take a walking break outside?



**Figure 4.6** A decision tree for a certain objective, in this case determining whether to take a walking break outside, is composed of decision nodes (composed of a question about the data), branches (which depend on the answer to the question, which are typically interpreted as a “yes” or “no” answers), and leaf nodes (the target variables).

A random forest is a collection of decision trees and gains its strength from “ensemble” learning. Essentially, random forests try to avoid the overfitting issues with decision trees by combining many models (decision trees) together and taking a majority vote of each tree. The

individual trees are then “weakly” trained, meaning they are trained on different subsets of the training dataset, or on predicting different classes.<sup>5</sup>

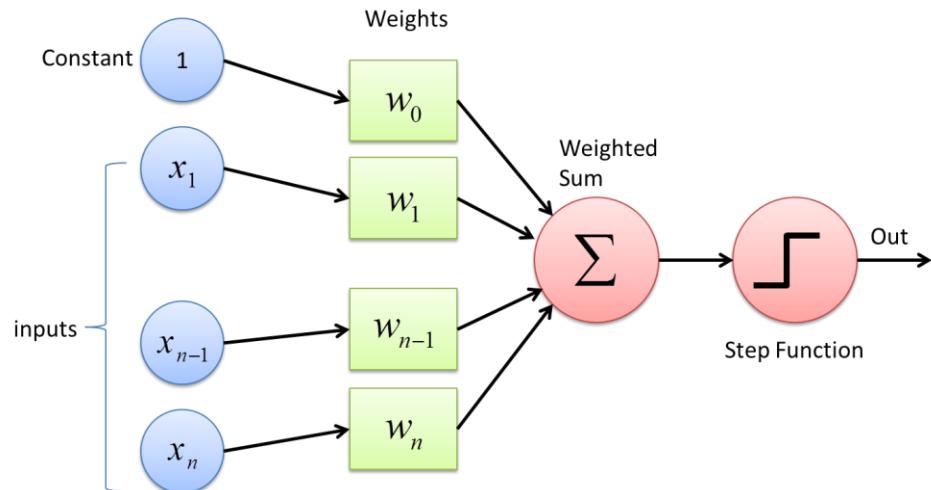


**Figure 4.7** A random forest is a collection of decision trees, where the overall decisions are a majority voting or averaged result from each tree. From TIBC<sup>6</sup>.

Finally, neural networks are the most versatile machine learning model, and the hardest to interpret, which is why they are notoriously labeled as “black boxes.” (They are not.) However, their strengths are derived from the nonlinear transformations interspersed with linear transformations throughout the network. Here, nonlinear means that the output does not correlate 1:1 with the input and therefore the transformation cannot be inverted. Errors in the final predictions are then propagated throughout the model (starting from the output to the input) via *backpropagation*, which relies on the derivatives of the weights. However, modern python tools use automatic differentiation, which numerically solves for the derivative (instead of finding an

analytic solution), thus turning this issue into a quick and painless behind-the-scenes aspect of model training.

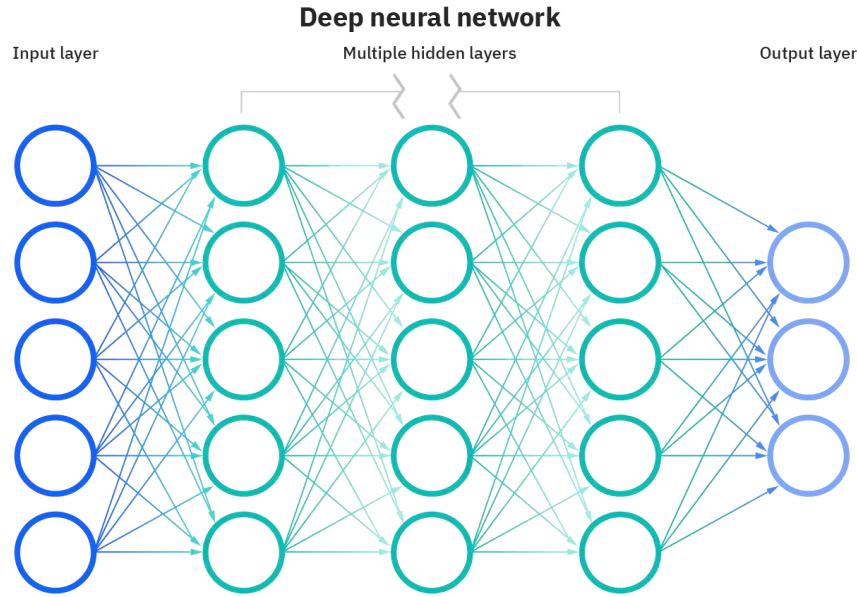
The basic building block of neural networks is the perceptron. Perceptrons were originally invented in 1943 by McCulloch and Pitts <sup>7</sup> but were horrible at generalizing predictions and were thus ignored for a long time. Instead, people focused on kernel-based models. That is, until 1986, when David Rumelhart, Geoffrey Hinton, and Ronald Williams <sup>8</sup> had the idea of combining them in “layers” and machine learning algorithms and their applications exploded.



**Figure 4.8** The composition of a perceptron – linear combination of inputs via a weight vector which then get summed and passed to a nonlinear activation function. From DeepAI <sup>9</sup>.

Perceptrons are composed of four different steps: 1. Input, 2. Weights, 3. Summation, and 4. Activation function. The first three steps are the linear transformation, where inputs are weighted and summed together (the weight vector is what is “learned”). Then, that summation is passed to a nonlinear activation function. Activation functions that are commonly used are sigmoid, tanh, ReLu (rectified linear unit), step, and hinge. All these activation functions are zero for negative

values and then (typically) move towards one around the origin (when the input is around zero). Activation functions like sigmoid, tanh, and the step function max out at one, making them good for output layers (where they could be interpreted as a probability, for example).

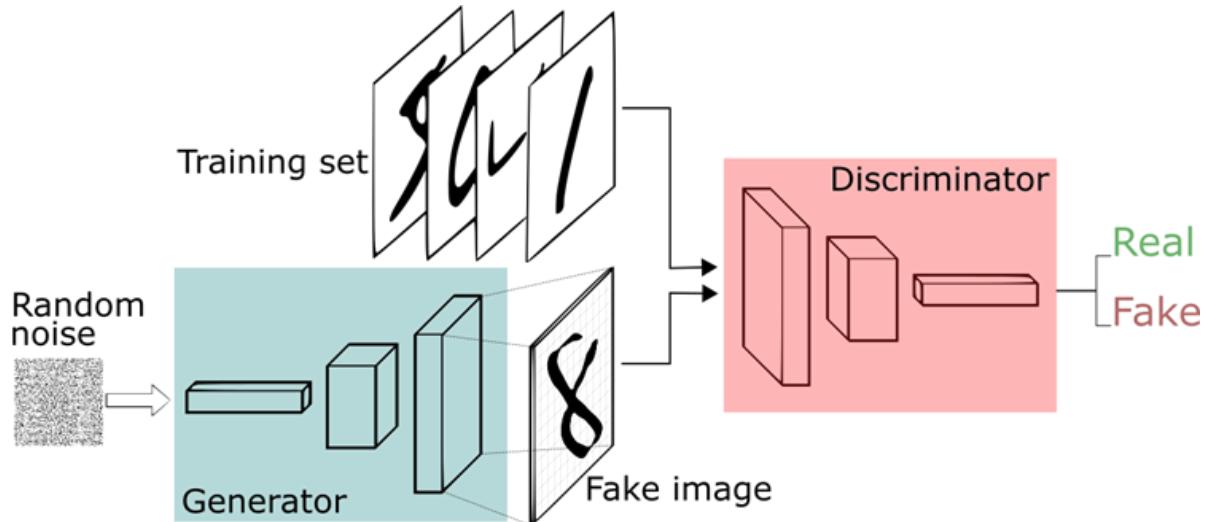


**Figure 4.9** Diagram of a fully connected MLP, or standard neural network, with three hidden layers. From IBM <sup>10</sup>.

Perceptrons that are strung together in layers make for a powerful model and are called multilayer perceptrons (MLPs). MLPs are just another name for your standard neural networks.<sup>4</sup> A diagram of this model can be seen in Fig. 4.9. In an MLP, each input node corresponds to an input feature. Then, the inputs are passed to any number of hidden layers, or layers that are neither input nor output layers, where each input goes to every node in the next layer, and every node in the next layer receives every input (at least for a fully connected model). Finally, the weights are passed to an output layer, where the dimension of the target variables (often) dictates how many output nodes there are. A deep neural network is simply an MLP with many hidden layers. Dropout

(random chance that weights go to zero to control how connected the network is) and regularizers can also be added to the layer weights.

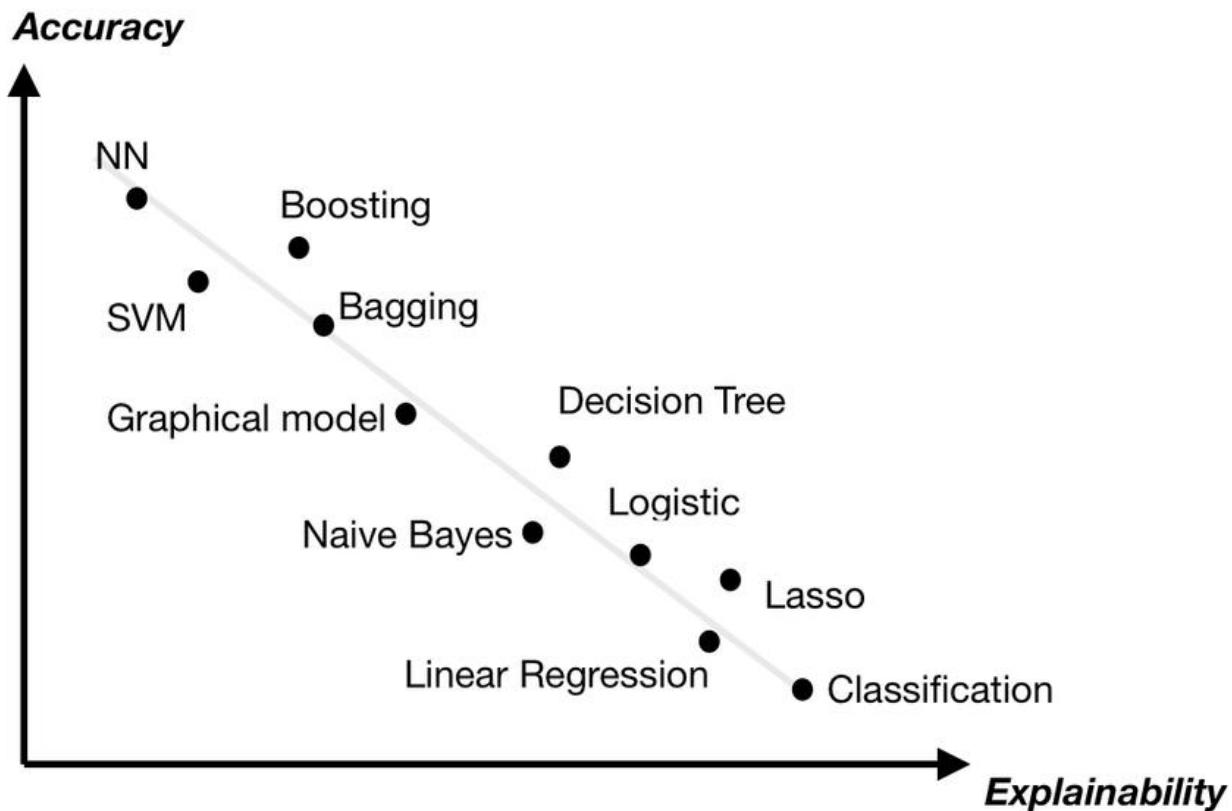
Neural networks start to get fancy very quickly. For example, there are recurrent neural networks, transformer neural networks, and convolutional neural networks (CNNs). While CNNs are often used for image classification and processing, the science community in particular benefits from generative models. One interesting generative network of note is generative adversarial networks (GANs).<sup>11</sup> GANs generate new data by having one model that creates data, seeded with random noise, called a generator. A parallel model, called the discriminator, compares real samples to the generated, or fake, samples and the entire model is penalized if the discriminator can distinguish between real and fake inputs. Obviously, this process requires an intricate balance between training the generator and discriminator, and many modifications have been made to encourage this, such as using the Wasserstein distance metric in a Wasserstein GAN (WGAN). However, GANs are known to be very finicky so other generative models like variational autoencoders, normalizing flows, and Generative Flow Networks (GFlowNets) have risen in popularity over GANs for their generative characteristics. Particularly, normalizing flows allow for an explicit representation of the likelihood function.



**Figure 4.10** A generative adversarial network (GAN) can make new, realistic-looking samples by balancing a discriminator and generator during training. From Thalles' blog <sup>12</sup>.

#### 4.2.4 Interpretability Versus Effectiveness

In general, deciding what model to use depends on the problem but also on the desired interpretability of the results. While linear regression is the most limited in its scope, only working well for linear data, it is the most interpretable as the importance of features can be directly obtained using the learned weights. Decision trees and SVMs are next easiest to interpret. The third tier in interpretability includes ensemble models, such as a random forest, as the majority voting hides the interpretation behind another layer. Finally, neural networks are the hardest to interpret as the vast number of weights and layers make it difficult to correlate features to specific outcomes. However, the versatility and accuracy of neural networks make them very attractive.



**Figure 4.11** Summary of the interpretability versus accuracy (or strength) of each machine learning model. From Duval, 2019<sup>13</sup>.

#### 4.2.5 Metrics

Regression metrics quantitatively compare predictions to true values. The most common are mean squared error (MSE), root mean squared error (RMSE), and mean average error (MAE). A good way to visualize regression predictions is by making a correlation curve by plotting predicted values versus true (target) values, where a perfect model would fall along the  $y = x$  line.

On the other hand, classification metrics can be more nuanced. Reporting an appropriate classification metric is critical as there are often drastically different interpretations of results depending on the metric used; for example, it is more important to identify the probability of false

negatives for cancer detection that it is to purely report accuracy. Here are some common metrics to consider.

Accuracy is the sum of true positives and true negatives divided by the total number of predictions. Accuracy is a good metric where distinguishing between false positives (Type 1 errors) and false negatives (Type 2 errors) does not matter. A good way to visualize all predictions is by forming a confusion matrix. Other metrics include precision (true positive divided by true positive plus false positive) and recall (true positive divided by true positive plus false negative). An important metric that is often reported is the  $F_1$  score as it combines precision and recall. It is defined as twice precision times recall divided by the sum of precision and recall, or

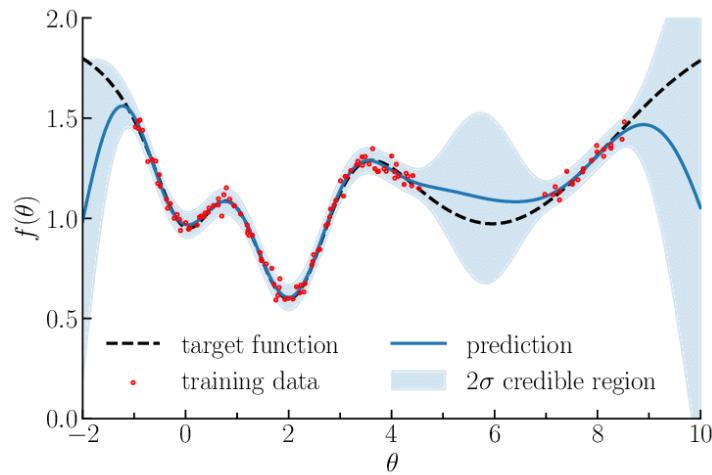
$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} . \quad (4.4)$$

This metric punishes extremes values more and places importance on false positives and false negatives. However, there is a tradeoff between precision and recall, so choosing which one to minimize or report must be handled with care.

As stated earlier, cross validation is important for determining any model hyperparameters and gauging the generalizability of a model. The most basic type of cross validation is k-fold cross validation. Here, the training data is split into k different chunks, where the model will be trained on  $k - 1$  of the sections, and the last section will be used as a validation set. Throughout training repetitions, called epochs in the case of neural networks, the validation set will systematically move between every  $k$  section. Then, an average and standard deviation can be calculated using the k predictions on the different validation sets.

#### 4.2.6 Uncertainty Estimation

There are a handful of ways to formally estimate uncertainty of predictions from machine learning models (which is essential for scientific interpretation) that are better than just rerunning the model and seeing how much its predictions change. First, let's understand the basics of uncertainty from a statistical perspective. Uncertainty can be divided into two different types – aleatoric and epistemic. Aleatoric uncertainty is from the internal randomness of phenomena while epistemic uncertainty is from the lack of knowledge of the system, i.e., hidden variables in your system that may be affecting the outcome of the event.<sup>14</sup>

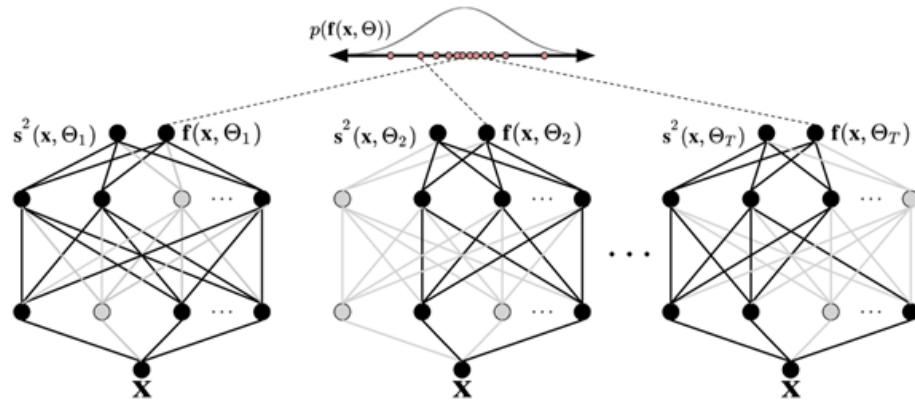


**Figure 4.12** A Gaussian process gives estimates of uncertainty depending on the location of the training data. From Leclercq, 2018<sup>15</sup>.

The following are a few models used in machine learning that formally incorporate uncertainty into their predictions. The first model to discuss is a Gaussian process (GP). GPs are non-parametric kernel-based methods that incorporate Bayes rule into their predictions.<sup>16</sup> Thus, they can give an estimate of uncertainty, or conversely confidence, to every output. A GP performs

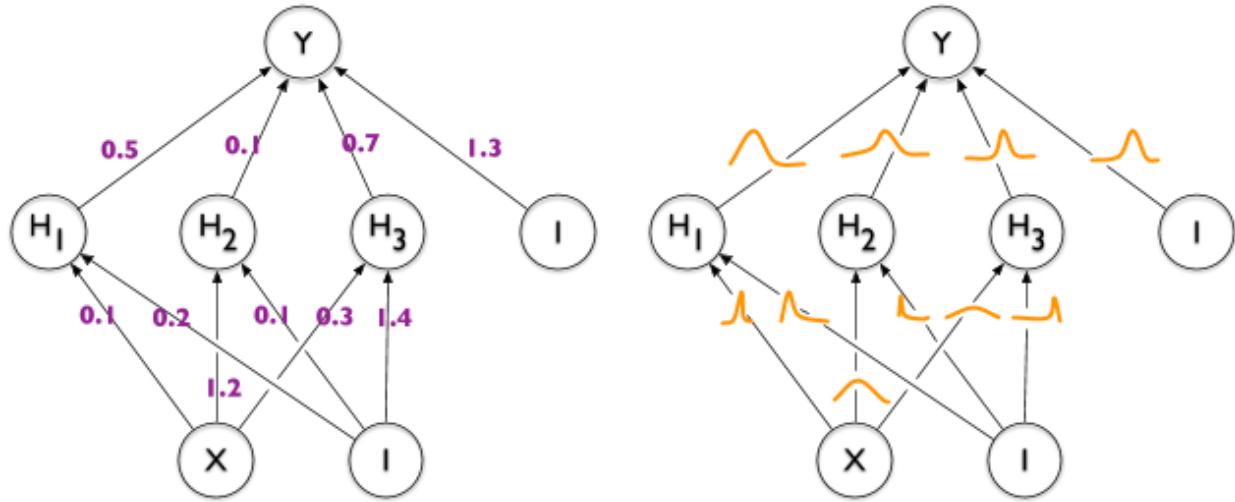
well on smoothly varying data, such as a multidimensional gaussian, but poorly on discontinuous or sharply featured data, such as a step function.

The second and most common way to estimate uncertainty of a prediction from a neural network is by using dropout during predictions. Dropout means randomly setting weights in the network to zero, thus randomly canceling out some correlations.<sup>17</sup> This process has been shown to estimate Bayesian uncertainty, and it is the easiest to implement as it does not require any architecture modifications.

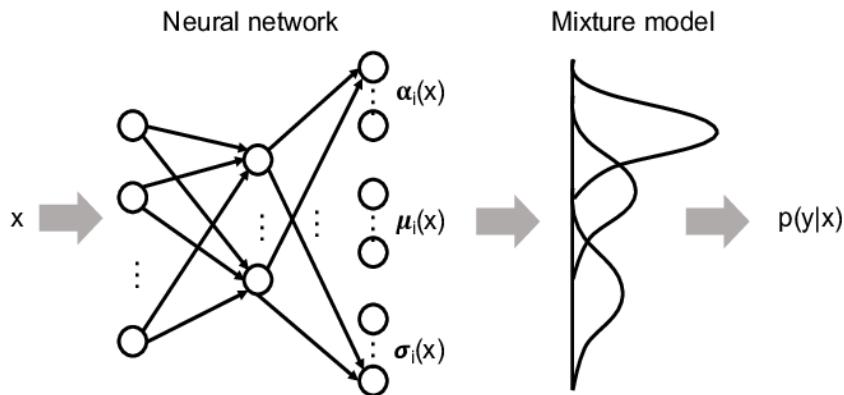


**Figure 4.13** Monte Carlo dropout during test predictions is the easiest and most common way to estimate uncertainty because it does not require special model architecture. From AWS documentation<sup>18</sup>.

The third way to incorporate uncertainties is to use a Bayesian neural network, which essentially learns distributions instead of weights. This difference means twice as many parameters to learn, assuming Gaussian distributions, which can be described by a mean and standard deviation.<sup>19</sup>



**Figure 4.14** A Bayesian neural network (right) learns distributions of weights instead of just the weights themselves, as with standard neural networks (left). From Sanjay Thakur's Blog<sup>20</sup>.



**Figure 4.15** A mixed density network is another easy implementation of a neural network that can formally account for uncertainty. From Vossen, 2018<sup>21</sup>.

Finally, the last model that incorporates uncertainty into predictions is a mixed density network. Here, the final output layer is expanded to include the parameters of the desired

distributions (such as mean and standard deviation, not just the mean) and the loss function is adjusted accordingly to reflect this interpretation.<sup>22</sup>

In summary, supervised machine learning models learn patterns in the context of a specific goal or target. There are a variety of models that range in their versatility and interpretability. Pre-processing data – either through normalization, feature selection, or cleaning – is integral to maintaining good practices by helping the model focus on what is important. Data is usually divided into two sets – training and test datasets. The training set is used for learning while the test set is used for final evaluation of the model *only*. Moreover, due to the large number of hyperparameters required to define machine learning models, the training dataset is often then divided further into training and validation sets to allow for tuning of those hyperparameters. While the test set is for evaluation, choosing appropriate metrics or estimating uncertainties in predictions can also assist in formalizing the effectiveness and reliability of models.

### 4.3 Unsupervised Machine Learning

Most of my work in this dissertation will focus on unsupervised machine learning. Unsupervised machine learning is searching for trends in data, but with no target variable. Instead, it is purely a data-driven pattern recognition technique. Often, unsupervised machine learning is used to disentangle correlated features (features selection) or look for the most important information in your data. This analysis can speed up computations or make predictions more interpretable.

### 4.3.1 Dimensionality Reduction

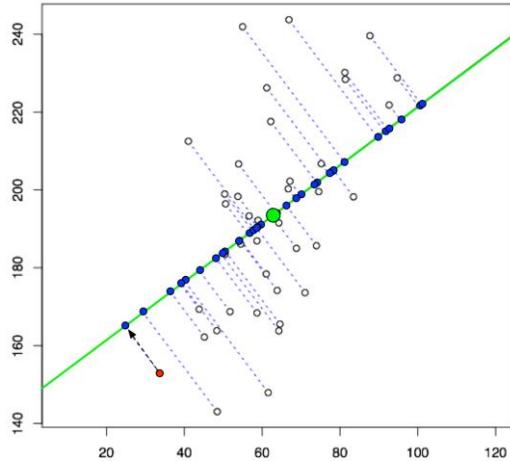
Dimensionality reduction is beneficial, in part, to combat the curse of dimensionality. The curse of dimensionality refers to the phenomenon that points that are far away in some small dimensional space look even farther away in higher dimensions, following an exponential explosion. The trend means that points that might have started close in low dimensions now look just as far away from every other point when in high dimensions, absolving any neighborhood occurring in the low dimensional space. This phenomenon is from the  $r^3$  law, or that the Jacobian in high dimensions depends on the radius  $r$  by increasing orders of magnitude. The curse is an issue when trying to identify similarities in high-dimensional data. Thus, dimensionality reduction tries to solve this issue by finding a lower-dimensional representation of the original dataset.<sup>4</sup>

First, I will discuss some linear dimensionality reduction transformations. The most common linear routine is Principal Component Analysis (PCA). PCA is an algorithm that finds the eigenvalue and eigenvector decomposition of a dataset such that the eigenvectors are specially ordered to be in order of importance, where importance is described as “explaining the most variance in the dataset” or, equivalently, minimizing the distance data points need to be projected onto that eigenvector.<sup>23</sup>

Singular value decomposition (SVD) is the more complete form of PCA and thus takes more computational power (because of a matrix inversion).<sup>23</sup> SVD represents the original matrix  $A$  as  $A = UDV^T$ , where  $U$  is composed of left singular vectors,  $D$  is a matrix with the singular values along its diagonal, and  $V$  is composed of right singular vectors. SVD is unique up to rearranging the order of the singular values.

Another popular linear transformation, especially for real values where is nonnegative matrix factorization (NMF). NMF is like PCA, except it forces the eigenvalues to be positive,

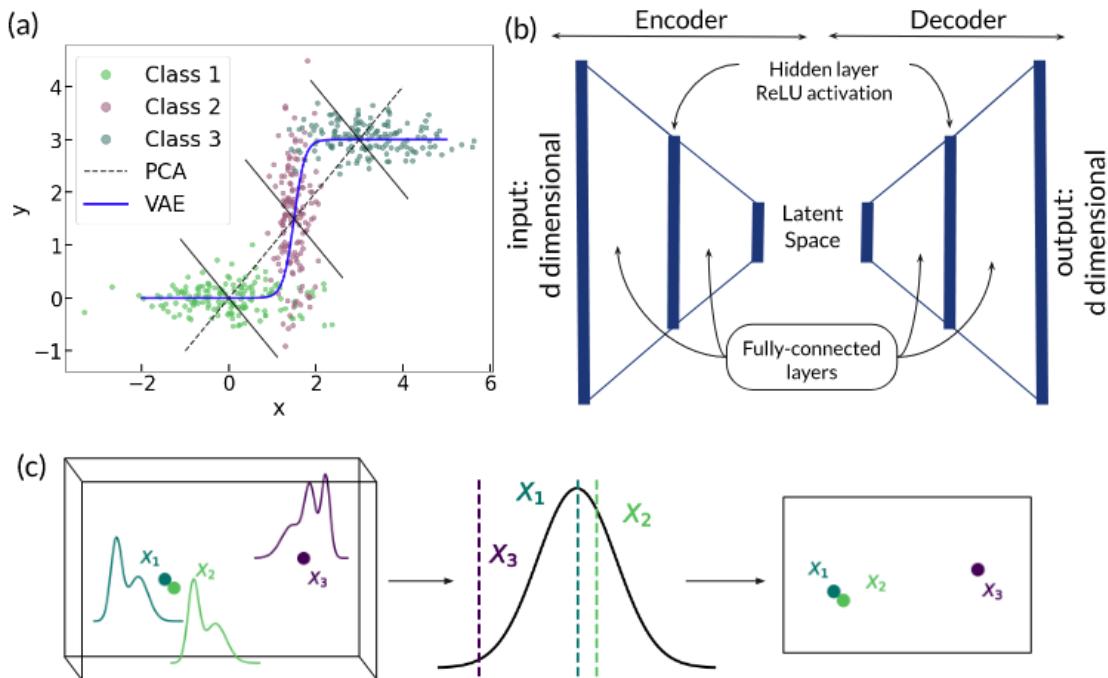
which is beneficial for physical variables that likewise cannot be negative.<sup>24</sup> NMF calculates two non-negative matrices whose product reproduces the original dataset, or  $A = WH$ , where the column vectors in  $W$  are linearly combined using the coefficients in the columns of  $H$ .



**Figure 4.16** PCA tried to maximize explained variance, or equivalently minimize the distances needed for the data points to be projected onto that eigenvector (or basis vector). From Bits of DNA<sup>25</sup>.

On the other hand, nonlinear transformations have strengths over linear methods because they can capture nonlinear trends in data. For example, if the data truly lies on a manifold, such as the Swiss Roll dataset, then nonlinear transformations may better capture the structure of the data. Effectively, nonlinear transformations are not constrained to basis vectors that are (hyper)planes but can have a weaving manifold to explain the data, as shown in Fig. 4.17a. One type of nonlinear transformation is using a neural network called an autoencoder, the architecture of which is shown in Fig. 4.17b. An autoencoder is a special type of neural network that is composed of two sequential

networks – an encoder and a decoder. The encoder takes data, in the original  $d$  dimensions, and reduces it to a lower dimensional representation called a *latent space*. The decoder then takes the latent space and expands the dimensionality back to the original  $d$  dimensions. By trying to reconstruct the original input in the output layer, the autoencoder iteratively learns how to maximize the information being squeezed through the bottleneck layer, or latent space.



**Figure 4.17** The benefits of nonlinear dimensionality reduction algorithms.<sup>26</sup>

An autoencoder can then be modified to be a variational autoencoder (VAE) by learning a distribution in the latent space instead of a deterministic embedding. This modification is done by learning two parameters for each latent space dimension (instead of one) and interpreting one as a mean and the other as a standard deviation. Then, the model will randomly vary the input going to the decoder by sampling from that learned mean and standard deviation. This property allows the

latent space to be generative and thus create new data that it hasn't seen before by interpolating through the latent space.<sup>27</sup> A final modification to autoencoders is making a conditional variational autoencoder, which includes target variables into the architecture (the condition), effectively performing supervised machine learning. The variational aspect of VAEs make them generative models and more generalizable than standard autoencoders.

Another nonlinear dimensionality reduction routine is t-distributed stochastic neighbor embedding (t-SNE)<sup>28</sup>. t-SNE is a tool that is often used for visualizing data in two or three dimensions (it does not work in higher dimensions) but can be used as a dimensionality reduction technique as well. t-SNE creates a graph-based similarity representation of the original data by calculating the joint conditional probability distribution between every pair of points to effectively get a similarity matrix. It then projects the data points to a lower dimension (either two or three) and tries to match the distances between each data point in the reduced space such that the subsequently calculated similarity matrix in the lower dimensions matches the one originally generated in the higher dimension. Ideally, this process means that there is no data compression or loss of information since the similarity between points is retained. This process is demonstrated in Fig. 4.17c.

However, t-SNE generates a “non-parametric” embedding, which means that it needs the entire dataset every time it makes a lower dimensional representation. There are no weight vectors it can save to do a transformation later. “Non-parametric” does not mean it has no hyperparameters. In fact, the one hyperparameter it does have is called perplexity, which represents the expected minimum cluster size. Moreover, t-SNE is excellent for looking at local similarities, but at the cost of loss of global structure. That means you must interpret clustering the t-SNE reduced space carefully. Data points in the same cluster can be interpreted as similar, but distances between

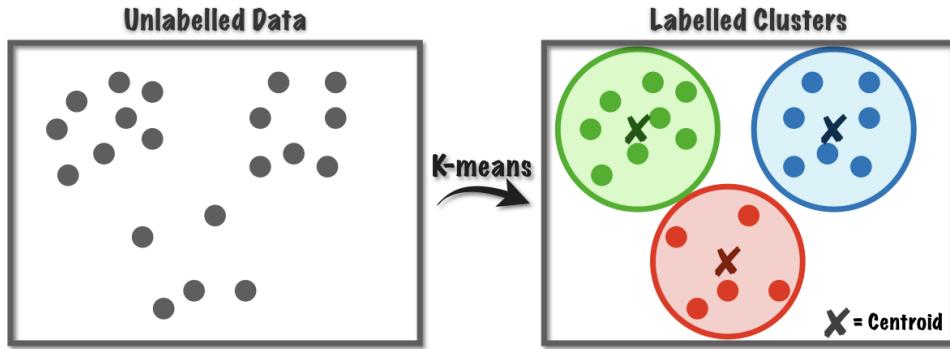
clusters cannot tell you exactly how similar or different the two clusters are, only that they are different.

Finally, Uniform Manifold Approximation and Projection (UMAP) <sup>29</sup> is very similar to t-SNE in that it constructs a graphic similarity representation of the original high-dimensional data and tries to match that similarity metric in the reduced dimensional embedding. However, because of a different choice in the cost function, it retains global similarity, unlike t-SNE. UMAP has two hyperparameters – the minimum distance between clusters and the average number of neighbors a data point in a cluster is expected to have. These hyperparameters let you tune how global versus local you want your similarity metric and thus how tightly your clusters appear in your reduced space.

#### 4.3.2 Clustering

Although a reduced representation of data is beneficial in and of itself, it can be helpful to apply a clustering algorithm to that reduced space. Clustering algorithms work better on fewer dimensions exactly because of the curse of dimensionality. Moreover, clustering can help identify unbiased classes in your dataset and be a precursor to supervised machine learning. Next, I will discuss some clustering algorithms, although it is certainly not comprehensive.

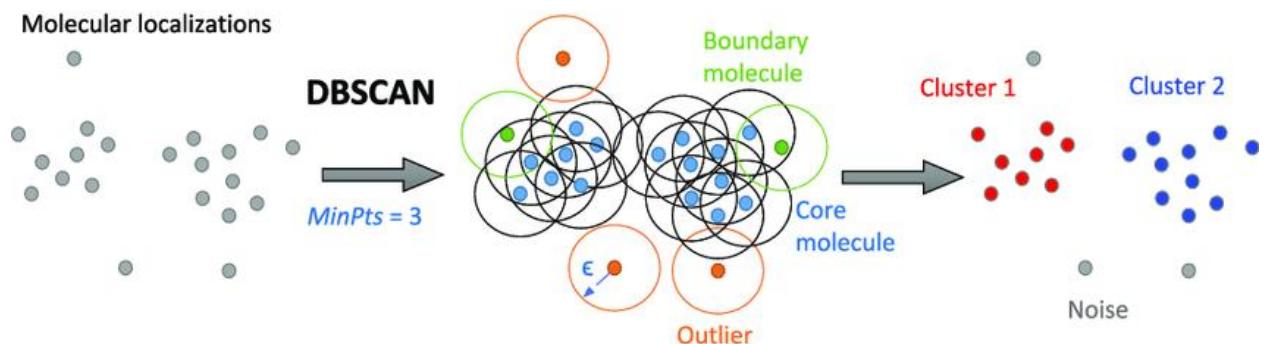
First, k-means clustering is a centroid-based clustering algorithm, i.e., it tries to balance the center of mass among all the clusters. K-means randomly places  $k$  centroids, where  $k$  is the number of clusters, and then it moves the centroid locations until they are the furthest apart from each other while being the center of mass for the nearby data points. Thus, k-means clustering works well if you know the expected number of clusters and the data has a Gaussian-like distribution (no Swiss rolls).



**Figure 4.18** K-means clustering balances center of mass. From Towards Data Science<sup>30</sup>.

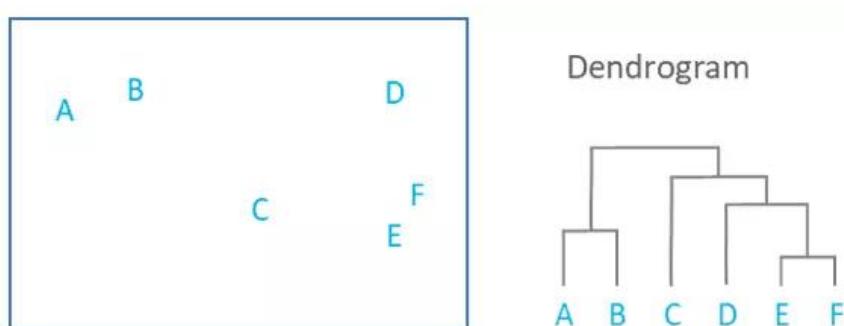
K-nearest neighbors (KNN)<sup>31</sup> clustering is similar to K-means in that it is centroid-based in that it calculates the probability of a new point belonging to a class based on the nearby data points, where the hyperparameter k represents the expected number of members in each cluster. However, KNN is a *supervised* clustering approach, unlike most other unsupervised clustering algorithms. The algorithm allows clusters to be more globular or amorphous in shape and is often used in semi-supervised applications.

Next, dbscan<sup>32</sup> is a density-based clustering algorithm, which means it looks at the density of nearby data to determine whether the points belong in the same cluster or not. It has one main hyperparameter, epsilon, which is effectively the expected radius of a cluster. The algorithm will group data points together with the expectation that almost all points in one cluster will fall within a radius of epsilon from another other point in the same cluster.



**Figure 4.19** dbSCAN uses the hyperparameter  $\epsilon$  to determine the radius of an expected cluster. From Khater, et al., 2020<sup>33</sup>.

Finally, agglomerative hierarchical clustering<sup>34</sup> is a recursive process that groups data points together one at a time until all data points belong to the same cluster. Using a divisive algorithm does the opposite – splitting the dissimilar data points from each other until every point belongs to its own cluster. Both result in a dendrogram, a tree-like graphical representation of the data, where “cuts” can be taken horizontally across them to determine clusters. Like decision trees, dendrograms create interpretability in their architecture but can have limited versatility.



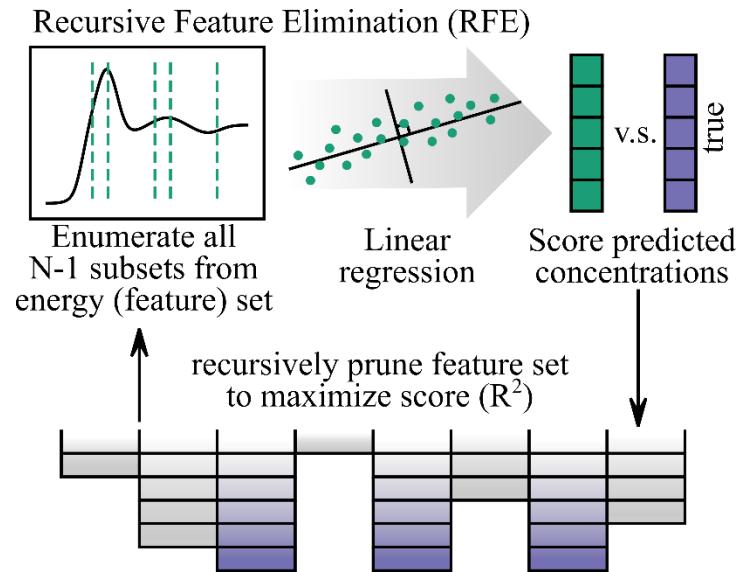
**Figure 4.20** Dendrograms represent similarity of points using the distances between them as a metric. From Displayr<sup>35</sup>.

### 4.3.3 Feature Selection

Feature selection is critical when input variables are correlated as it makes models more interpretable. The choice of an appropriate metric is essential for feature selection methods and often depends on the type of input and output variables, where each input and output can either be numerical or categorical. Because there is a target output, as is often the case, the metric chosen must be supervised. For example, Analysis of Variance (ANOVA) works well for numerical input variables and categorical outputs, Pearson's correlation works well for numerical inputs and outputs, and  $\chi^2$  works well for categorical inputs and outputs and is commonly used in scientific settings.  $\chi^2$  can be separated into three main types: tests for goodness of fit, tests for independence, and tests for homogeneity. The “goodness of fit”  $\chi^2$  test is its most common usage, which determines the probability that a variable comes from a specific distribution.

Machine learning models can additionally be used to perform feature selection, rather than a particular metric. For example, the weights learning through linear regression can be used to identify the importance of each feature in the context of a learning problem. Decision trees and random forests are likewise other model options for feature selection. However, a wrapper-based method uses one of these methods at its core, but it can have additional benefits. Recursive feature elimination (RFE)<sup>36</sup> is a wrapper-based feature selection method that, as the name suggests, recursively prunes the input, or feature space, such that the most important features remain. The algorithm decides the importance of features by training a base machine learning model, such as a linear regressor as shown in Fig. 4.22, and then calculating the prediction accuracy for that set of features. To remove a feature, it will repeat this process by enumerating through every combination of  $N - 1$  features, where  $N$  is the current size of the feature set. In each step, it removes the feature

that corresponded to the lowest accuracy (or importance). The algorithm continues until a desired number of features remains.



**Figure 4.21** Recursive feature elimination picks the best features by correlating them with the best score (e.g., accuracy) of predicting the target variables.<sup>37</sup>

## 4.4 References

1. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
2. G. Papachristoudis, The Bias-Variance Tradeoff).
3. <https://online.stat.psu.edu/stat508/book/export/html/749>).
4. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education, New Jersey, 3rd edn., 2010.
5. L. Breiman, *Machine Learning*, 2001, **45**, 5-32.
6. What is a Random Forest?).
7. W. S. McCulloch and W. Pitts, *The bulletin of mathematical biophysics*, 1943, **5**, 115-133.
8. D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533-536.
9. Perceptron).
10. What are neural networks?).
11. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Y. Bengio, *arXiv*, 2014.
12. T. S. Silva, A Short Introduction to Generative Adversarial Networks).
13. A. Duval, *Explainable Artificial Intelligence (XAI)*, 2019.
14. E. Hüllermeier and W. Waegeman, *Machine Learning*, 2021, **110**, 457-506.
15. F. Leclercq, *Physical Review D*, 2018, **98**, 063511.
16. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.
17. Yarin Gal and Z. Ghahramani, 2016.
18. Monte Carlo dropout).
19. E. Goan and C. Fookes, in *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, eds. K. L. Mengersen, P. Pudlo and C. P. Robert, Springer International Publishing, Cham, 2020, DOI: 10.1007/978-3-030-42553-1\_3, pp. 45-87.
20. S. Thakur, The very Basics of Bayesian Neural Networks).
21. J. Vossen, B. Feron and A. Monti, 2018.
22. D. J. C. MacKay, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1995, **354**, 73-80.
23. W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press, 2007.
24. D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788-791.
25. L. Pachter, What is principal component analysis?).
26. S. Tetef, N. Govind and G. T. Seidler, *Phys. Chem. Chem. Phys.*, 2021, **23**, 23586-23601.
27. G. E. Hinton, *Science*, 2006, **313**, 504-507.
28. L. van der Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579-2605.
29. L. McInnes, J. Healy and J. Melville, *arXiv*, 2020.
30. K.-m. A. C. Introduction, K-means: A Complete Introduction).
31. K. Fukunaga and P. M. Narendra, *IEEE Transactions on Computers*, 1975, **C-24**, 750-753.
32. M. Hahsler, M. Piekenbrock and D. Doran, *Journal of Statistical Software*, 2019, **91**, 1 - 30.
33. I. M. Khater, I. R. Nabi and G. Hamarneh, *Patterns*, 2020, **1**, 100038.

34. F. Murtagh, *The Computer Journal*, 1983, **26**, 354-359.
35. T. Bock, What is a dendrogram?).
36. H. Jeon and S. Oh, *APPLIED SCIENCES-BASEL*, 2020, **10**.
37. S. Tetef, 2023, in prep.

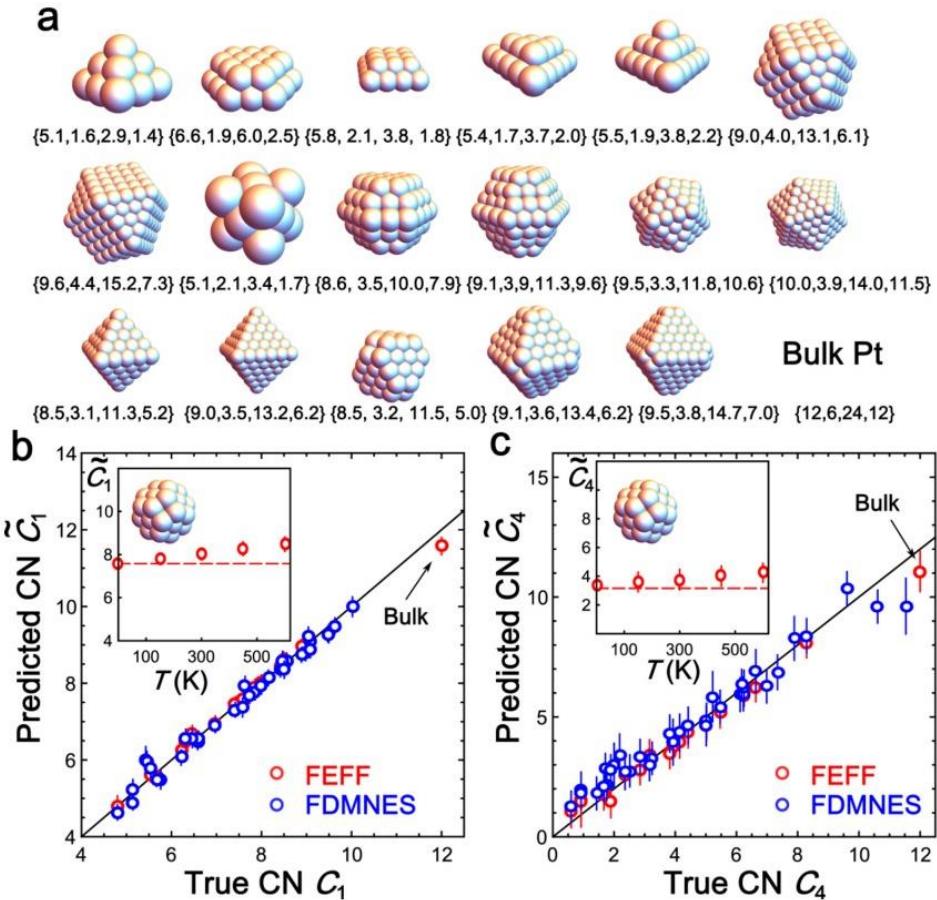
# 5 Chapter 5 – A Survey of New Developments Between X-ray Spectroscopy and Machine Learning

In this chapter, I give an overview of the state of the field utilizing machine learning in the context of X-ray absorption and emission spectroscopies. I focus on the inverse problem, where the majority of work is on XANES regression problems. A nice perspective on an overview of ML in X-ray spectroscopy is given in *Speciation of Nanocatalysts Using X-ray Absorption Spectroscopy Assisted by Machine Learning* by Routh, et al., 2023<sup>1</sup>.

## 5.1 Solving the Inverse Problem

### 5.1.1 Supervised Machine Learning Approaches

In 2017, seminal work which used machine learning to analyze XANES spectra was published in Timoshenko, et al.<sup>2</sup>, where they predicted coordination from XANES rather than EXAFS. Timoshenko and coworkers continued to publish papers along the same vein, showcasing how machine learning models could predict coordination from XANES spectra when humans required EXAFS to predict the same property. I will first give a brief overview of these papers, which largely focus on metallic nanoparticles.



**Figure 5.1** (a) shows the test nanoparticle structures. (b) and (c) show the true versus predicted coordination numbers from the neural network for both the first and fourth coordination shell, respectively, on the test dataset. Taken from Timoshenko, et al., 2017<sup>2</sup>.

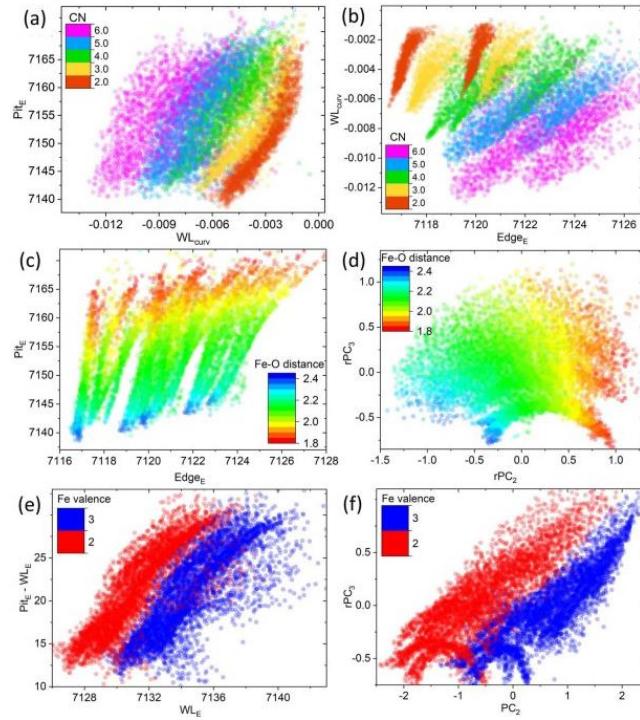
First, in the seminal paper, *Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles*, Timoshenko, et al., 2017<sup>2</sup>, they use Pt XANES of nanoparticles to predict the coordination of Pt nanoparticles from spectra using a neural network. Next, in *Probing Atomic Distributions in Mono- and Bimetallic Nanoparticles by Supervised Machine Learning*, Timoshenko, et al., 2019<sup>3</sup>, they used EXAFS of Pt and PdAu nanoparticles to predict partial radial distribution functions (RDF) using a neural network from the

wavelet-transformed EXAFS. Finally, “*Inverting” X-ray Absorption Spectra of Catalysts by Machine Learning in Search for Activity Descriptors*, Timoshenko and Frenkel, 2019<sup>4</sup> is a perspective on an overview of the ML methods one can use to solve the inverse problem for XANES, including decision trees, neural networks, PCA, and MCR-ALS (multivariate curve resolution-alternating least squares), applied to various datasets.

Another seminal paper was work with the Materials Project database. In particular, they utilized ensemble learning, which should (theoretically) be more generalizable than a neural network and has thus seen a rise in popularity, along with gradient boosting algorithms. Their aim was to match an unknown spectrum against all the spectra in the corresponding spectral database for the closest and thus likeliest candidates, using the corresponding properties in the Materials Project Database to infer characteristics rather than predicting them directly from spectra. In sum, in *Automated generation and ensemble-learned matching of X-ray absorption spectra*, Zheng, et al., 2018<sup>5</sup>, they generated XASdb, a database of over 800,000 k-edge XANES spectra of structures from the Materials Project. They then developed the Ensemble-Learned Spectra IdEntification (ELSIE) algorithm, which uses an ensemble of “weak” learners to compare similar spectra and thus identify oxidation state or coordination number.

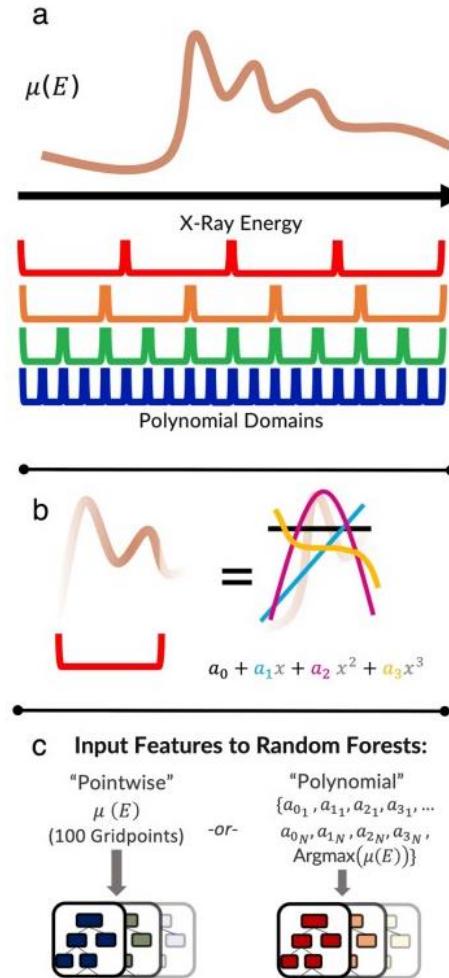
Instead of predicting properties directly from spectra, the following four papers explore other ways to generate features of XANES spectra to then use to perform predictions. In all four, both features (of the spectra) and properties were selected, and then correlations between those features and the properties of interest were identified. First, in *Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms*, Guda, et al., 2021<sup>6</sup>, they used XANES spectra of FeSiO<sub>2</sub> to correlate XANES features (edge position, intensities, positions, and

curvatures of minima and maxima) to predict properties (coordination numbers, bond distances and angles, and oxidation state) using Elastic Net (combining ridge and LASSO regression).



**Figure 5.2** Training dataset of each descriptor versus property, where the color reflects the CN values in (a), (b), the average Fe-O distance in (c), (d), and iron valence in (e), (f). Taken from Guda, et al., 2021<sup>6</sup>.

Second, in *Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships*, Torrisi, et al., 2020<sup>7</sup>, they used XANES of several different 3d transition metals to correlate XANES features to properties (coordination number, Bader charge, and mean nearest neighbor distance), where featurization of XANES spectra included both just pointwise energy values (normal way to featurize spectra) and third order polynomial fits to different sized energy regions using a random forest (RF).



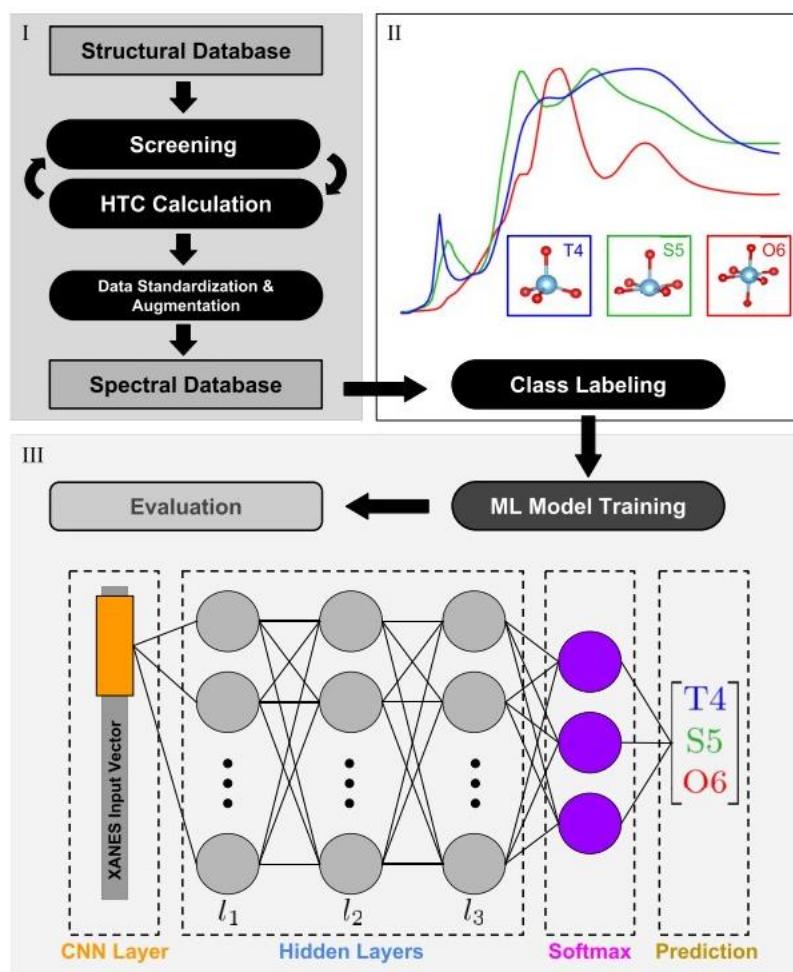
**Figure 5.3** Featurization of XANES spectra included the pointwise energy-intensity values and fitting to third order polynomials for regions with varying energy resolutions. Taken from Torrisi, et al., 2020<sup>7</sup>.

Third, in *Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure*, Zheng, et al., 2020<sup>8</sup>, they used K-edge XANES to predict coordination number from XANES spectra using a random forest, which was trained on a database of 190,000 spectra. Of note, they analyzed feature importance using the drop-variable technique. Finally, in *How Much Structural Information Could Be Extracted from XANES Spectra*

*for Palladium Hydride and Carbide Nanoparticles*, Usoltsev, et al., 2020<sup>9</sup>, they used Pd K-edge XANES of nanoparticles to correlate various structural descriptors and their combinations (such as Pd-Pd interatomic distances, hydrogen concentration, and adsorbed hydrocarbons) to the “pure” spectral components obtained from MCR (multivariate curve resolution), where they used PCA to determine number of MCR components.

While feature importance can help elucidate important spectral components, using machine learning to improve fits to spectra in the context of reference standards, such as linear combination fitting, has also been explored. The following four papers utilize other supervised machine learning applications, but in the context of fitting XANES spectra rather than direct prediction of properties. First, in *PyFitit: The software for quantitative analysis of XANES spectra using machine-learning algorithms*, Martini, et al., 2019<sup>9</sup>, they made a package for XANES fitting and demonstrated the tools by fitting Ce L<sub>3</sub> edge and Fe(terpy)<sub>2</sub> XANES as examples. The package included an array of ML algorithms (Gradient Boosting of Random Trees, Radial Basis Functions and Neural Networks) to fit spectra. They also used Latin hypercube sampling (LHS) to generate molecular deformations, which is generally most effective for sampling a high-dimensional parameter space. Second, in *Assessing arsenic species in foods using regularized linear regression of the arsenic K-edge X-ray absorption near edge structure*, Jahrman, et al., 2022<sup>10</sup>, they used As K-edge XANES and used LASSO regression to perform linear combination fitting onto a reference library of spectra. Finally, in *Solving the structure of “single-atom” catalysts using machine learning – assisted XANES analysis*, Xiang, et al., 2022<sup>11</sup>, they used Co XANES of “single-atom” catalysts to apply PCA to determine the number of species for linear combination fitting. They also used a neural network to predict the distance between Co and C from the carbonyl, d<sub>C</sub>, and the distance between Co and bottom O, d<sub>O</sub>.

The following paper is unique in that it performs classification, rather regression, as its form of supervised machine learning. In *Classification of local chemical environments from x-ray absorption spectra using supervised machine learning*, Carbone, et al., 2019<sup>12</sup>, they simulated K-edge XANES of eight 3d transition metals (Ti, V, Cr, Mn, Fe, Co, Ni, and Cu) and classified the local coordination environment using a neural network.



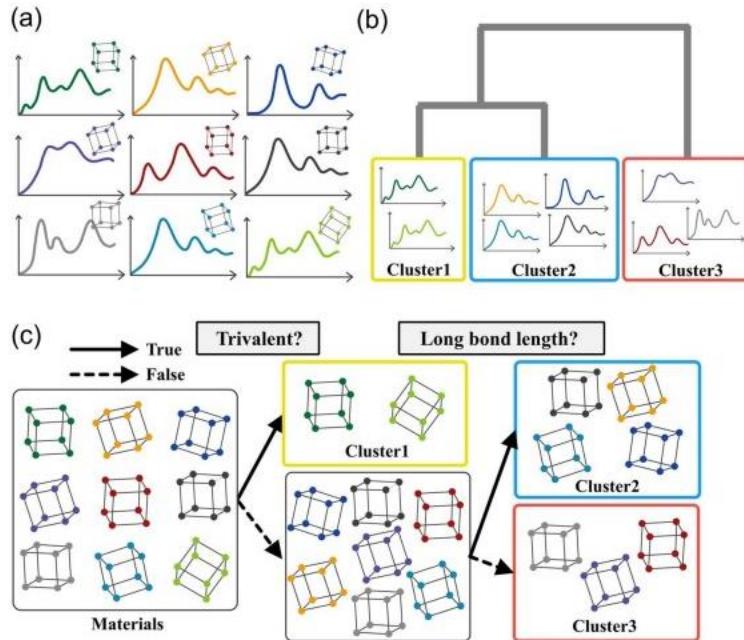
**Figure 5.4** Their workflow for classifying spectra into the three coordination environments: tetrahedral (T4), and square pyramidal (S5), and octahedral (O6). Of note, their structural database was the Materials Project. Taken from Carbone, et al., 2019<sup>12</sup>.

### 5.1.2 Unsupervised machine learning approaches

The following two papers utilize unsupervised machine learning. This approach is far less popular than using supervised machine learning. Not only does unsupervised machine require much more interpretation from the user, but it cannot “predict” properties by itself. Thus, it must be paired with other models to perform any predictions.

First, in *Latent Representation Learning for Structural Characterization of Catalysts, Routh, et al., 2021*<sup>13</sup>, they used Pd K-edge XANES and correlated the dimensions of the latent space of an autoencoder to physical properties like coordination number (N), interatomic distance (R), and hydrogen fraction (H). They found applying PCA to the latent space produced stronger correlations. Then they trained a neural network to predict N, R, and H from the transformed latent space (the PCA-on-latent space representation).

Second, in *Machine learning approaches for ELNES/XANES, Mizoguchi and Kiyohara, 2020*<sup>14</sup>, they used a dataset composed of 39 electron energy loss near edge structure (ELNES), a.k.a. O K-edge XANES spectra, 14 of which were mono-metal oxides and the other 25 were polymorphous SiO<sub>2</sub>. They used a decision tree to explain and predict hierarchical clustering of spectra, and then used a neural network to predict classes generated with that decision tree.

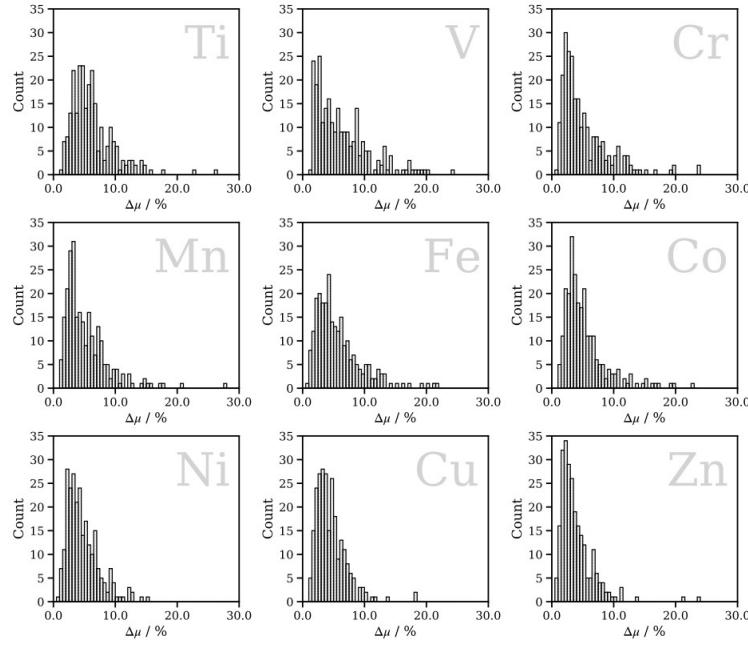


**Figure 5.5** (a) Structure to spectra correlation (b) Clustering spectra using hierarchical clustering (Wasserstein distance as a similarity metric) (c) Decision tree to determine the underlying properties distinguishing the three clusters. Taken from Mizoguchi and Kiyohara, 2020<sup>14</sup>.

## 5.2 Solving the forward problem

The forward problem, rather than the inverse problem, is instead predicting XANES spectra from structural parameters. While the inverse problem dominates the applications of machine learning in the X-ray spectroscopy community, there is some notable work to solve the forward problem and thus replace time-consuming DFT calculations. First, in *Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy*, Carbone, et al., 2020, they used a graph based neural network to predict XANES spectra from molecular structures in the QM9 database. Using a different encoding of structure, in *Accurate, Affordable, and Generalisable Machine Learning Simulations of Transition Metal X-ray Absorption Spectra using the XANESNET Deep Neural Network*, Rankine and Penfold, 2020<sup>15</sup>, they used a deep neural network to predict K-edge XANES

for nine first-row transition metals (Ti-Zn) from weighted atom-centered symmetry functions (wACSF), a featurization of the local coordination geometry.



**Figure 5.6** Mean percentage error between the target (from simulations) and predicted spectra on the test set. Taken from Rankine and Penfold, 2020<sup>15</sup>.

Finally, predicting Valence-to-Core X-ray emission spectroscopy (VtC-XES) spectra rather than XANES, *A deep neural network for valence-to-core X-ray emission spectroscopy, Penfold and Rankine, 2022*, extends the neural network Penfold and Rankine previously trained for XANES spectra, but instead use it to predict row transition metal K-edge VtC-XES spectra from weighted atom-centered symmetry functions (wASF).

### 5.3 References

1. P. K. Routh, N. Marcella and A. I. Frenkel, *The Journal of Physical Chemistry C*, 2023, **127**, 5653-5662.
2. J. Timoshenko, D. Y. Lu, Y. W. Lin and A. I. Frenkel, *Journal of Physical Chemistry Letters*, 2017, **8**, 5091-5098.
3. J. Timoshenko, C. J. Wrasman, M. Luneau, T. Shirman, M. Cargnello, S. R. Bare, J. Aizenberg, C. M. Friend and A. I. Frenkel, *Nano Letters*, 2019, **19**, 520-529.
4. J. Timoshenko and A. I. Frenkel, *Acs Catalysis*, 2019, **9**, 10192-10211.
5. C. Zheng, K. Mathew, C. Chen, Y. M. Chen, H. M. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson and S. P. Ong, *Npj Computational Materials*, 2018, **4**, 12.
6. A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot, J. A. van Bokhoven, C. Copéret and A. V. Soldatov, *npj Computational Materials*, 2021, **7**, 203.
7. S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram and L. Hung, *npj Computational Materials*, 2020, **6**, 109.
8. C. Zheng, C. Chen, Y. Chen and S. P. Ong, *Patterns*, 2020, **1**, 100013.
9. A. Martini, S. A. Guda, A. A. Guda, G. Smolentsev, A. Algasov, O. Usoltsev, M. A. Soldatov, A. Bugaev, Y. Rusalev, C. Lamberti and A. V. Soldatov, *Computer Physics Communications*, 2020, **250**, 107064.
10. E. P. Jahrman, L. L. Yu, W. P. Krekelberg, D. A. Sheen, T. C. Allison and J. L. Molloy, *Journal of Analytical Atomic Spectrometry*, 2022, **37**, 1247-1258.
11. S. Xiang, P. Huang, J. Li, Y. Liu, N. Marcella, P. K. Routh, G. Li and A. I. Frenkel, *Phys. Chem. Chem. Phys.*, 2022, **24**, 5116-5124.
12. M. R. Carbone, S. Yoo, M. Topsakal and D. Lu, *Physical Review Materials*, 2019, **3**, 033604.
13. P. K. Routh, Y. Liu, N. Marcella, B. Kozinsky and A. I. Frenkel, *The Journal of Physical Chemistry Letters*, 2021, **12**, 2086-2094.
14. T. Mizoguchi and S. Kiyohara, *Microscopy*, 2020, **69**, 92-109.
15. C. D. Rankine, M. M. M. Madkhali and T. J. Penfold, *The Journal of Physical Chemistry A*, 2020, **124**, 4263-4270.

## 6 Chapter 6 – Information Content of VtC-XES versus XANES spectra of Sulfororganics

Originally published as: S. Tetef, N. Govind and G. T. Seidler. *Physical Chemistry Chemical Physics* 2021 Vol. 23 Issue 41 Page 23586. S. Tetef wrote and conducted the majority of this work.

*We report a comprehensive computational study of unsupervised machine learning for extraction of chemically relevant information in X-ray absorption near edge structure (XANES) and in valence-to-core X-ray emission spectra (VtC-XES) for classification of a broad ensemble of sulphorganic molecules. By progressively decreasing the constraining assumptions of the unsupervised machine learning algorithm, moving from principal component analysis (PCA) to a variational autoencoder (VAE) to t-distributed stochastic neighbour embedding (t-SNE), we find improved sensitivity to steadily more refined chemical information. Surprisingly, when embedding the ensemble of spectra in merely two dimensions, t-SNE distinguishes not just oxidation state and general sulphur bonding environment but also the aromaticity of the bonding radical group with 87% accuracy as well as identifying even finer details in electronic structure within aromatic or aliphatic sub-classes. We find that the chemical information in XANES and VtC-XES is very similar in character and content, although they unexpectedly have different sensitivity within a given molecular class. We also discuss likely benefits from further effort with unsupervised machine learning and from the interplay between supervised and unsupervised machine learning for X-ray spectroscopies. Our overall results, i.e., the ability to reliably classify without user bias and to discover unexpected chemical signatures for XANES and VtC-XES, likely generalize to other systems as well as to other one-dimensional chemical spectroscopies.*

## 6.1 Introduction

The emergence of modern data science techniques, along with improved theoretical tools addressing physical observables and open access online databases, has led to new and insightful interpretation of experimental results. Thus, machine learning (ML) has proliferated throughout chemistry, materials science, and chemical engineering <sup>1, 2</sup>. Large databases, such as the Materials Project <sup>3</sup>, Inorganic Crystal Structure Database <sup>4, 5</sup>, and QM9 <sup>6</sup>, along with open access packages for ML, have all contributed to this rise in popularity and reliability of machine learning analysis of data <sup>7</sup>. Recent work includes the use of ML to develop a way to represent molecular structures <sup>8, 9</sup>, to study charge transport at the nanoscale level <sup>10</sup>, or to automate chemical predictions from atomistic simulations <sup>11</sup>.

X-ray absorption spectroscopy (XAS), an important chemical speciation technique, has seen impressive recent developments using ML <sup>12-32</sup>. Briefly, XAS encompasses both X-ray absorption near edge structure (XANES) and extended X-ray absorption fine structure (EXAFS) and involves interrogating the unoccupied electronic states by a core photoelectron. On the other hand, X-ray emission spectroscopy (XES) interrogates the occupied electronic density of states by relaxing from an excited state to a ground state <sup>33-35</sup>. Furthermore, recent developments of reliable lab-based spectrometers in multiple energy ranges have facilitated an increase in accessibility of both XAS and XES measurements <sup>36-40</sup>.

Both XAS and XES are manifestly element-specific, as either the excitation or the deexcitation energy, respectively, selects the species of interest. These methods appear in a plethora of subfields in chemistry, physics, materials science, and earth and planetary sciences, with representative contemporary research in renewable energy <sup>41</sup>, electrical

energy storage<sup>42, 43</sup>, protein structure and function<sup>44</sup>, terrestrial and lunar basalts<sup>45</sup>, chemical catalysis<sup>46</sup> in biomolecules<sup>47</sup>, and photochemical dynamics<sup>48</sup>. In such applications, the experimenter seeks to understand local electronic and atomic structure, elucidating properties of the selected species such as oxidation state, bond lengths, ligand identity, and coordination symmetry and numbers.

Several decades of effort has resulted in theoretical approaches that reliably solve the forward problem, i.e., the prediction of XAS and XES spectra from known structures<sup>33, 49, 50</sup>. However, the inverse problem of obtaining structural, electronic, or chemical information from spectra is ill-posed and demands the use of prior information. Although formal statistics have been occasionally applied to address the imposition of the experimenter's constraining physical knowledge on the system<sup>51-54</sup>, prior knowledge is more commonly implicit via the user interaction with the standard tools for interpretation of EXAFS<sup>55, 56</sup> or XES spectra<sup>57</sup>. However, the analysis of XAS – and of XES, as seen here – is seeing rapid development, which is both exciting for the XAS community and potentially informative for other spectroscopies. We propose that these efforts can address broader questions of the encoding of chemical information via physical measurement.

In a seminal work, Timoshenko, et al.<sup>27</sup> used supervised ML to train a neural network on an ensemble of differently coordinated nanoparticles to extract geometric information from merely the X-ray absorption near-edge structure (XANES), the first ~50 eV of XAS. This work exemplified how prior information could be encoded via the selection of structures for the training data set as well as showcasing a supervised machine learning model that performed better than human researchers, who would instead require the entire EXAFS spectrum to obtain similar information. Working contemporaneously,

Zheng et al.<sup>31</sup> took a different direction. Instead of seeking inferences about fine structural parameters, they developed an algorithm to match unknown materials with known materials in a large database, showcasing its effectiveness by predicting oxidation and coordination from the material's XAS spectra.

Subsequent ML work aimed at a better interpretation of XAS has sought to identify important energy regions or features of spectra that contribute most prominently to specific properties<sup>12, 20, 29</sup>. Moreover, supervised ML has seen use in classifying coordination and local chemical environments<sup>14, 16</sup> and the oxidation state<sup>19</sup> of 3d transition metals, and used to extract geometric properties<sup>30</sup>, especially during high-throughput experiments<sup>17</sup> in real-time<sup>26</sup>. As another example with a pragmatic application, ML has recently been implemented for fitting XANES spectra<sup>18</sup>. Further work utilizing artificial intelligence for fitting EXAFS data is also actively being developed<sup>24, 25</sup>. Finally, and by means of closure by returning to the forward problem, Rankine et al. utilized machine learning to quickly predict Fe XANES spectra given local geometric parameters<sup>22</sup>. Other efforts to utilize machine learning to predict XANES spectra, either from structural parameters or from the partial density of states, include Carbone et al.<sup>13</sup> and Kiyohara, et al.<sup>15</sup>, respectively.

In the present manuscript, we take a new direction in the use of ML methods in X-ray spectroscopies. Not only is this the first analysis of valence-to-core XES (VtC-XES) using ML methods, but we apply *unsupervised* ML to identify chemically relevant classes based on both XANES and VtC-XES. Furthermore, instead of using unsupervised ML to force a correlation of certain geometric regressive properties of a system of interest to *specific* dimensions of a reduced dimensional representation of XANES spectra, as seen in the recent work of Routh, et al.<sup>23</sup>, which we believe is the first application of unsupervised

ML in XAS, we *fully examine clustering in this reduced dimensional space for unbiased discovery of chemical classes* and thus the extent of encoded information in spectra.

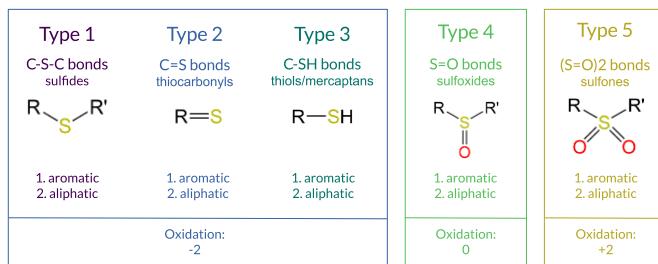
As a secondary consequence of our choice to investigate both XANES and VtC-XES, we are also able to test the common qualitative assertion that the methods are “complementary” because of their respective sensitivity to unoccupied and occupied electronic states<sup>58</sup>, here quantitatively addressing whether the *chemically relevant* information in XANES and VtC-XES is indeed complementary or is instead highly coincident<sup>59-61</sup>.

Based on our results, we propose that chemical classification problems are best addressed with unsupervised ML methods at least as a precursor analysis method<sup>11</sup>, an approach that may enrich or suggest refinement of prior structure-specific inferential work in XAS<sup>14, 16, 17, 26, 27, 31</sup> and similar work in a wide and rapidly growing range of other spectroscopies in chemical sciences<sup>62-64</sup>. This distinction is nontrivial. Subject only to the imposition of prior information through the choice of the training domain of materials or molecules, unsupervised learning serves to identify the extent of the underlying and scientifically useful chemical properties<sup>23</sup> for a given spectroscopy without user bias. These methods allow any spectral similarities, and thus classes, to emerge from the algorithm and then researchers can *a posteriori* interpret its chemical relevance. This ensures that unanticipated encodings of chemical information are not overlooked. An unsupervised ML approach is, we feel, especially suitable for X-ray spectroscopies exactly because of the challenges presented by the ill-posed nature of the inverse problem. Hence, both our motivations and our methods are distinct from prior work using data science and ML methods in X-ray spectroscopies.

We now define our system of interest and the methods that will be used for classification. Our training domain encompasses a very wide range of sulphorganic molecules chosen because of: (1) their rich diversity of bonding environments; (2) the considerable evidence for sensitivity of both XANES and VtC-XES of the S K-edge to chemical bonding in this family<sup>60, 65, 66</sup>; and (3) the prior demonstration of good agreement between experiment and time-dependent density functional theory (TD-DFT)<sup>66</sup> calculation of XANES<sup>67</sup> and VtC-XES<sup>66, 68-70</sup>.

For chemical context, the five “Types” of molecules used in our study are shown in Fig. 6.1. They are: (1) sulphides, (2) thiocarbonyls, (3) thiols, (4) sulfoxides, and (5) sulphones. Type 1, or sulphides, are compounds with C-S-C bonds. This includes S in cyclic sulphides, such as thiophenes and thiazoles, along with sulphides where the S is bonded to two separate functional groups. Type 2, or thiocarbonyls, have S double bonded to a single C. Type 2 includes variations such as isothiocyanates and thioureas. Type 3 are thiols, also known as mercaptans, and have an SH functional group bonded to a C atom in some radical. Types 1, 2, and 3 all have a sulphur oxidation of -2. Type 4, or sulfoxides, have S double-bonded to O and single bonded to two C atoms. Type 4 has a sulphur oxidation of 0. Finally, Type 5 are sulphones, which have S double-bonded to two oxygens and single bonded to two C atoms. Type 5 also includes sulphonamides. Type 5 has an oxidation of +2. Every Type is additionally divided into subcategories based on whether the S is a member of a conjugated system, e.g., in an aromatic ring, or not, i.e., is aliphatic. There are similarities and differences in these classifications compared to Yasuda and Kakiyama<sup>65</sup> and Holden, et al.<sup>66</sup>. Specifically, we have somewhat expanded the core

“Types” compared to that prior work but have retained the use of oxidation state and aromaticity as additional refining parameters.



**Figure 6.1** Schematic representation of the five types of sulphorganics investigated, along with sub-categories.

Here we investigate three different classification schemes that follow the general rubric of dimensionality reduction, followed by cluster identification. We report a critical comparison of (1) Principal Component Analysis (PCA), which is a fully linear method with an underlying Euclidean metric, (2) a Variational Autoencoder (VAE), which is a deeply nonlinear method that still has a local metric, and (3) t-distributed Stochastic Neighbour Embedding (t-SNE), a nonlinear, non-parametric embedding that is inherently non-metric. In all cases, the accrued benefit is the ability to see clustering in the reduced dimensional spaces from which we then assign chemical descriptors and, in turn, infer the general character of chemical information that is encoded within XANES and VtC-XES.

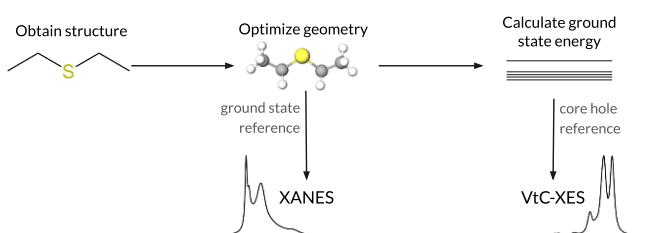
We find surprisingly strong absolute and comparative performance for t-SNE, which draws attention to a shared core weakness of PCA and VAE in the present context. In those methods, the similarity of spectra is only quantified after dimensionality reduction, i.e., only after information has necessarily been lost. This is in contrast with t-SNE, where the

original spectra drive the creation of a probabilistic description of similarity (with no necessary loss of spectral information) and then a subsequent embedding in a lower dimension is determined. t-SNE thus has significant heuristic benefits for classification, albeit at the cost of losing any meaningful metric properties in the resulting embedding. On the other hand, the retention of formal mappings and metrics for PCA and VAE allows for applications that require tracking the trajectory of evolving chemical systems, such as in high-throughput synchrotron experiments.

## 6.2 Methods

### 6.2.1 Electronic Structure Calculations

Our data generation pipeline is shown schematically in Fig. 6.2. A list of sulphorganic compounds was created from a wide variety of sources, starting with the compounds in Yasuda and Kakiyama<sup>65</sup> and Holden et al.<sup>66</sup>, so as to make best contact with those prior experimental studies of classification of VtC-XES. First, in all cases, structures (in the form of .mol files) were downloaded from the PubChem database<sup>71</sup> via the MolView API<sup>72</sup>. All ground state structures, XANES<sup>73</sup>, and VtC-XES<sup>74</sup> computations were performed with the open-source NWChem computational chemistry program<sup>75,76</sup>. In total, 769 molecules are included in this work.



**Figure 6.2** Schematic depiction of the data generation pipeline.

The existence of single, internally-consistent energy scales is due to the self-consistent field (SCF) DFT solution that is solved for each system, which serves as the reference for the TDDFT-based X-ray spectroscopy calculations. In the case of XANES, we compute the ground-state SCF solution as the reference, while for the XES we compute the core-hole SCF solution, as indicated in Fig. 6.2.

The geometry optimizations utilized the 6-31G\* basis sets<sup>73, 74, 77, 78</sup> and the B3LYP exchange correlation functional<sup>79</sup>. The XANES and VtC-XES spectra were then computed using the Sapporo QZP-2012 and Sapporo TZP-2012 basis set<sup>80</sup>, respectively, for S, while the remaining atoms were represented using 6-31G\* basis set, and PBE0 exchange correlation functional<sup>81</sup>. In cases where compounds contained heavier atoms than S, such as bromine and chlorine, an effective core potential was substituted for the atom, specifically the Stuttgart RLC ECP<sup>82</sup>.

Because our linear-response TDDFT-based XANES spectra are computed from stationary Kohn-Sham DFT states, a broadening must be applied to account for the finite lifetime of the electronic states. Thus, an energy-dependent linear broadening scheme was applied to the XANES transitions, similar the scheme in Mijovilovich et al.<sup>83</sup>. Pre-edge transitions until the whiteline were Lorentz broadened at a full-width half-maximum (FWHM) of 0.6 eV, to be consistent with the core-hole lifetime. Then a linear increase in the FWHM broadening was applied, starting from the whiteline at 0.6 eV and increasing to 4.0 eV FWHM at 15 eV past the whiteline, to account for inelastic scattering effects at higher energies. This broadening scheme reproduced spectral features well<sup>83</sup>. In this case, the energy-dependent broadening values of the transitions were chosen arbitrarily such that

they most accurately depicted experimental features<sup>60, 84</sup>. Finally, the spectra were individually normalized by dividing their total K $\alpha$  intensities and an energy shift of -53.3 eV was applied to all XANES transitions to align the theoretically calculated transitions with experiment.

For the VtC-XES, the calculated transitions were all shifted by -18.6 eV to align to experiment<sup>65, 66</sup>. Additionally, a Lorentz broadening of FWHM of 0.6 eV in addition to a Gaussian broadening of FWHM of 0.3 eV was added to each transition, which represents the core-hole lifetime and the best possible experimental resolution (limited by the bent crystal analyser), respectively. We found no significant changes in the clustering upon qualitative examination of the reduced-dimensional spaces using less broadening. This is likely due to the loss in information upon compression to just two dimensions, where sharpening features, or the emergence of small new peaks, will not compete with the most prominent characteristics of the spectra. Thus, we chose to use experimentally motivated broadening. The resulting spectra were also normalized by their total K $\alpha$  intensity to achieve a common intensity scale per S atom.

### 6.2.2 Supervised ML Methods

To pre-process our spectra, the intensity was represented pointwise with 1000 linearly spaced energy values along a consistent energy range across the entire ensemble. The training and test set consist of 717 and 52 molecules, respectively, and were both scaled such that they were peak normalized to the highest intensity value of the training set; this ensured spectra had intensity values between 0 and 1 in addition to preserving overall transition amplitudes.

All neural network models in this study were implemented in Python using the Keras<sup>85</sup> package with a Tensorflow backend<sup>86</sup>. As a benchmark for defining “good” accuracy when compared to the dimensionally reduced spaces, we performed classification via supervised machine learning by passing the original high-dimensional spectra into a fully connected neural network classifier. The fully connected neural network for the three classification schemes for the VtC-XES had one hidden layer with dimension 512, ReLU activation, L2 kernel regularization, and 5% dropout. It was optimized via Keras’s default ADAM using binary cross entropy loss, with a softmax output activation function. The network architecture for the XANES had all the same hyperparameters as the VtC-XES, except it had a hidden dimension of 1024 instead of 512. The resulting confusion matrices for VtC-XES and XANES for all classification schemes are given in Fig. 6.S3 (Scheme 1: Oxidation), Fig. 6.S3 (Scheme 2: Type), and Figs. 6.S4 and 6.S5 (Scheme 3: Aromaticity within each Type, henceforth simply “Aromaticity”). The benchmark accuracies for classifying the VtC-XES spectra were 100%, 96%, and 71% for Oxidation, Type, and Aromaticity, respectively, for the 52 compounds of the test set. And the benchmark test accuracies of classifying the XANES spectra were 100%, 85%, and 69% for Oxidation, Type, and Aromaticity, respectively.

We applied supervised machine learning on the reduced dimensional spaces by implementing K-Nearest Neighbours (KNN) classification with scikit-learn using 20 nearest neighbours for classification Schemes 1: Oxidation and 2: Type, and with 10 nearest neighbours for Scheme 3: Aromaticity (within each Type). KNN is a supervised classification algorithm that categorizes data points based on the other data points in the vicinity, specified by this number of neighbours ( $k$ ) hyperparameter. While it is perhaps

unfortunate that we are comparing accuracies obtained from different models – a neural network versus KNN – we chose KNN to evaluate the reduced spaces because it mimics the nearest neighbour behaviour of t-SNE and requires fewer hyperparameters to be tuned. Furthermore, the predicted classification boundaries on the reduced spaces between KNN and a neural network trained were similar and thus both methods are comparable.

### 6.2.3 Unsupervised ML Methods

Our VAE model took the spectra as input, where each spectrum was represented by 1000 points of intensity as indicated above. This model was also implemented in Python with Keras and Tensorflow. The network was trained using a batch size of 50 and had two hidden layers of dimension 512 and 128 respectively, with ReLU activation. Additionally, L2 kernel regularization was added to each layer, and a dropout of 10% was applied after every layer, both of which were implemented to help prevent overfitting and encourage generalizability. The encoder and decoder were then symmetric, although the output layer of the decoder had a sigmoid activation function. An almost identical model architecture and hyperparameters were used to train the VAE for both the VtC-XES and XANES spectra; however, the XANES model had a dropout of 15% and the second hidden layer had dimension 246 instead of 128. Both models were optimized via the default settings of the optimizer ADAM in Keras. The VAE and fully connected classifier neural networks were verified on a validation set via the model loss and reconstruction efficacy to check for overfitting. See Fig. 6.S1. The trained VAE models, analysis code, and datasets are available on GitHub<sup>87</sup>.

We applied Principal Component Analysis (PCA), along with the t-distributed stochastic neighbour embedding (t-SNE), independently to the XANES and VtC-XES spectra using the scikit-learn<sup>88</sup> package in Python. The optimal hyperparameter for t-SNE, perplexity (which roughly represents cluster size), was found by searching through perplexity values between 5 and 50, with perplexity equal to 18 yielding the qualitatively most distinguishable yet believable clusters on the training set. All two-dimensional reduced spaces were linearly scaled to be between 0 and 1 for each axis.

### 6.3 Dimensionality Reduction Algorithms

Given the novelty of unsupervised ML in the context of x-ray spectroscopies, it is useful to give a detailed overview and comparison of the methods used here. To begin, dimensionality reduction not only helps determine which features in data are most “evident” or variational, but by doing so in a data-driven matter, it also removes biases imposed by the researcher. Of central importance here, lower dimensional representations often yield better classification by addressing the curse of dimensionality, i.e., everything in a high dimensional space looks far away, so it may be difficult to quantify similarity of points in a high dimensional space<sup>89</sup>. However, selecting the best dimensionality reductional algorithm is, as investigated here, closely dependent on both the constraints inherent to the method and the underlying variance of the training data. The question is whether progressive weakening of constraints on the algorithm, such as by removing the requirements of linearity or a quasi-metric mapping, in fact better preserves information content and thus allows for more robust classification. While this is an appealing hypothesis, it is by no means a certain outcome: one might find that the constraints are

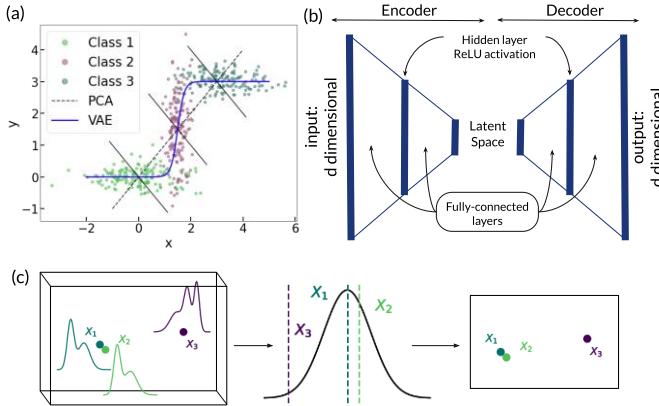
needed to suppress overamplification of spectral features that do not have physical importance.

To this end, we will compare linear and nonlinear forms of dimensionality reduction where both algorithms perform formal mappings between the original high-dimensional space (where the calculated ensemble of spectra live) and learn a mapping to a lower-dimensional representation. Then, we will compare these mapping-based algorithms to a probabilistic, non-parametric embedding algorithm that, instead of learning a formal mapping function from a higher- and lower-dimensional space, creates a lower-dimensional representation by preserving a similarity metric of the original spectra. The results of this work elucidate the chemically relevant information content in XANES and VtC-XES, allow a comparison of their relative information content, and suggest possible methods for real-time monitoring of high-throughput experiment.

We begin with the two mapping algorithms, as opposed to the embedding. The dominant linear method for dimensionality reduction is Principal Component Analysis (PCA).<sup>90</sup> Nonlinear dimensionality reduction can be achieved via unsupervised machine learning, specifically here, via the VAE neural network model<sup>91</sup>. Given that there is very scarce prior work using VAE's in spectroscopies, e.g., optical-wavelength spectroscopy in an astrophysical study<sup>92</sup>, we will especially discuss the key differences between PCA and VAE. For work detailing the use of just an autoencoder (AE) for XANES analysis, see Routh et al.<sup>23</sup>. With this in mind, we will additionally discuss the difference between an AE and VAE, and the additional properties inherent to a VAE.

To begin, in Fig. 6.3a, we envision a scenario of synthetic data in three different clusters in a parameter space of some unknown dimension, here shown in two dimensions

for ease of presentation. If the data distribution is well-represented by a simple N-dimensional (hyper)ellipsoid, PCA would successively choose orthogonal axes in a new coordinate system that consecutively encompassed the most variability contained within the high dimensional data set. Equivalently, PCA chooses an orthonormal basis to represent a lower dimensional (hyper)plane such that the distance the data travels to be projected onto this PCA (hyper)plane is minimized. Thus, data can be represented using only the first few basis vectors, or dimensions, that explain the most variation within the data.



**Figure 6.3** (a) Clusters where nonlinear dimension reduction routines, such as from a neural network, might yield better clustering than a linear dimension reduction like PCA. (b) Architecture of a simple autoencoder (AE) with one hidden layer, demonstrating the dimension reduction utility of the AE via its nonlinear latent space. (c) Schematic of how t-SNE uses the probability that data points are sampled from the same distribution to determine their similarity.

However, whether in two dimensions, as in Fig. 6.3a, or in some higher dimensional realization, dimensionality reduction for complex data that spans multiple qualitative classes is frequently poorly suited to decomposition via purely orthogonal axes and

Euclidean-preserving metrics in the host high-dimensional space. This is where less restrictive coordinate transformations often have superior dimensionality reduction *en route* to classification. VAEs have not previously been used in X-ray spectroscopies, although they have been shown to be superior to PCA in several other contexts<sup>92-95</sup>.

In Fig. 6.3b, a schematic of a simple autoencoder demonstrates how a coupling of two neural networks – an encoder and a decoder – performs nonlinear dimensionality reduction. The encoder takes in  $d$ -dimensional input, reduces it down to a nonunique lower dimensional representation called a *latent space*, and then the decoder expands the dimension back to the original  $d$  dimensions. The nonlinear activation functions in each neuron give the mathematical freedom for deforming the metric. The autoencoder learns, through iterative training, how to encode data to a lower dimension by trying to match the input and output – ensuring that maximal information is retained as the data is passed through this information bottleneck layer, or latent space. Because no predetermined classes or labels are given to the network, clustering in the latent space is inherently unsupervised – hence we neither impose prior knowledge that, for example, oxidation state will create useful spectral distinctions, nor limit ourselves to discovering only a few prescribed categories of chemical information.

Autoencoders, however, suffer from overfitting that reduces their ability to generalize or generate new data and thus have limited utility for classifying unseen data. To resolve this concern, an autoencoder can be modified into a variational autoencoder (VAE)<sup>91</sup>. VAEs have almost the same model architecture as autoencoders, except instead of learning an exact latent space encoding, they learn a latent space probability distribution, which is described in more detail in the SI. Points in the latent space are instead sampled

from a learned normal distribution. This sampling creates perturbations in the latent space, which helps prevent overfitting and allows the latent space to be complete, continuous, and regularized, leading to *the generation of new data*. Most importantly, the probabilistic sampling ensures that similar spectra are in fact mapped to similar locations in the latent space, and the decoder will be able to decode points in the latent space it has not previously seen, both of which are imperative for classification.

Returning to Fig. 6.3a, the benefits of the VAE’s nonlinear dimensionality reduction are illustrated by the thick blue line, representing a possible first coordinate axes of a VAE latent space. The nonlinearity of the VAE allow it to weave and thus, imagining the data in Fig. 6.3a in a higher dimensional space, create a manifold that would better capture variance of the data domain with fewer reduced dimensions. Hence, while the nonlinearity of the VAE prohibits its use for linear superposition analysis of composition – a common application of PCA in XAS – we posit that VAEs, or other nonlinear dimensionality reduction methods, might provide special advantages for classification problems, i.e., for grouping data with respect to the underlying chemically-relevant information in XANES and VtC-XES spectra.

We will demonstrate the utility of unsupervised methods, either linear (PCA) or nonlinear (VAE), to not only analyse the information retained by a reduced-dimensional representation, but most importantly, to generate a *mapping* to the reduced-dimensional space. That is, both PCA and VAE create a functional mapping from the high-dimensional space of spectra to the derived two-dimensional spaces that can be saved and used later, without modification, to subsequently map new data onto the derived spaces. Thus, they are tools to store data. Moreover, this ability allows us to quantify the quality of mapping

by calculating the accuracy of classification on a subsequent test set. However, if the final scientific goal is understanding the connection between spectral features and information content in an ensemble, then the imposition of a well-behaved mapping may be unnecessary and may in fact over-constrain and hence degrade performance toward chemical classification. This brings us to use of embedding algorithms.

The t-distributed Stochastic Neighbour Embedding (t-SNE)<sup>96</sup> is performed by calculating a pairwise similarity matrix over the entire dataset by creating a joint conditional probability distribution. For example, imagine the three points, called  $X_1$ ,  $X_2$ , and  $X_3$  in Fig. 6.3c, exist in the original high-dimensional space that fully characterizes the spectra, i.e., each such point corresponds to a full spectrum. Here,  $X_1$  and  $X_2$  are clearly more alike than  $X_3$ . When t-SNE compares similarities between high-dimensional points, it assumes all data points are sampled from an inherent Gaussian distribution such that data that are more similar have a higher probability of being sampled from the same distribution, while dissimilar data have a lower probability of being sampled from the same distribution.

Therefore, similar data points should be closer together in a reduced representation, i.e., closer to the assumed mean of the inherent joint distribution, and dissimilar data points are farther away. To obtain the lower dimensional embedding, t-SNE then randomly projects the data to a lower-dimensional space and computes an analogous pairwise conditional probability distribution function (now assuming points are sampled from a t-distribution to encourage spread). Through an iterative minimization process, t-SNE tries to match the pairwise conditional probabilities from the lower dimensional space to the one calculated in the high dimensional space.

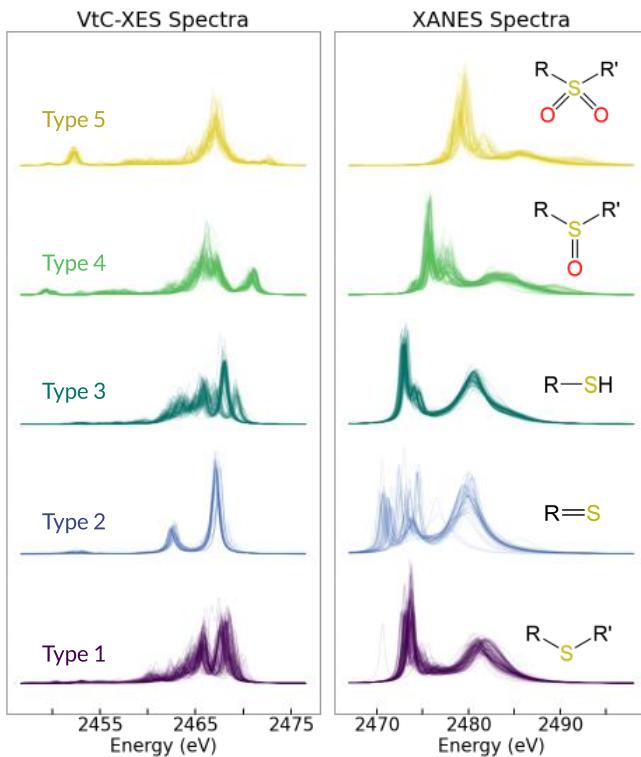
Thus, similarity relationships between data points in the original high-dimensional space should be maintained by t-SNE in this reduced space. This contrasts PCA and VAE, which project the spectra onto a low-dimensional space via a simple basis using a Euclidean metric (PCA) or else an adaptive metric (VAE), and for which the issue of the similarity of data is only addressed after this inherently lossy compression process.

## 6.4 Results and Discussion

### 6.4.1 Dataset and Dimensionality Reduction

It is useful to consider a qualitative presentation of variance of the XANES and VtC-XES spectra – both within and across compound Types. Hence, in Fig. 6.4, we show the VtC-XES and XANES spectra for a representative sampling of the molecules in this study. Beyond energy shifts, there are some interesting variations within Types for each of VtC-XES and XANES. For example, the Type 2 XANES has far more variation than the VtC-XES. Conversely, the Type 3 VtC-XES has far more variation than the XANES. Such details encourage the use of unsupervised learning *en route* to a chemical explanation.

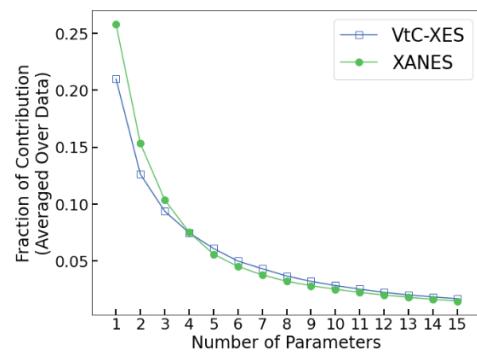
We now report on unsupervised dimensionality reduction for this data set. In this, we primarily focus on PCA, VAE, and t-SNE, but also include several competing linear algorithms for completeness. These results are then used for classification in Section 4.2.



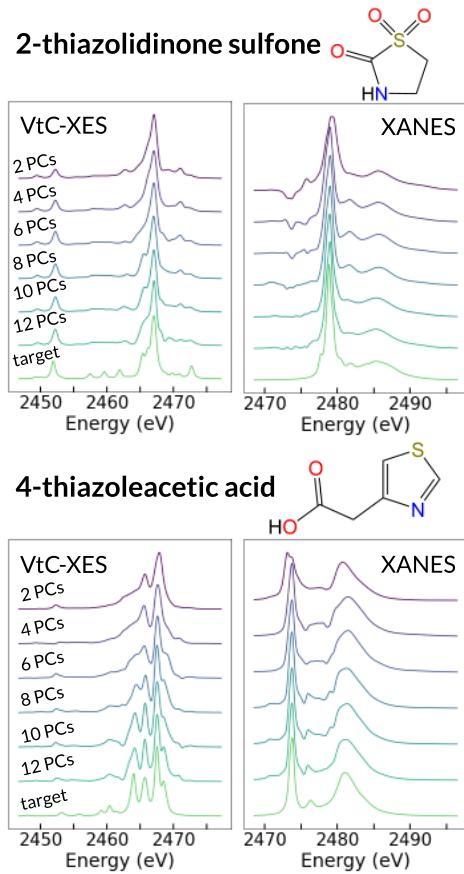
**Figure 6.4** VtC-XES (left) and XANES (right) spectra for all organosulphur compounds, displayed by compound type. Some spectra have been arbitrarily scaled or randomly removed for display purposes.

#### 6.4.1.1 Principal Component Analysis

The most important measure for the utility of PCA is the proportion of variance explained by a PCA basis, in order of most important principal component to least, which is shown in Fig. 6.5 (averaged over the entire dataset). The basis elements have been sorted so that the eigenvectors corresponding to the largest eigenvalues are considered first; in other words, the first principal component (PC) is the most important as it explains the most variance of the data. For both the XANES and VtC-XES data, a point of diminishing returns is found at  $\sim 6 - 8$  principal components.



**Figure 6.5** Scree plot of PCA effectiveness for both VtC-XES and XANES. The vertical axis is the fraction of variance explained by each PC, e.g., the 10<sup>th</sup> PC.

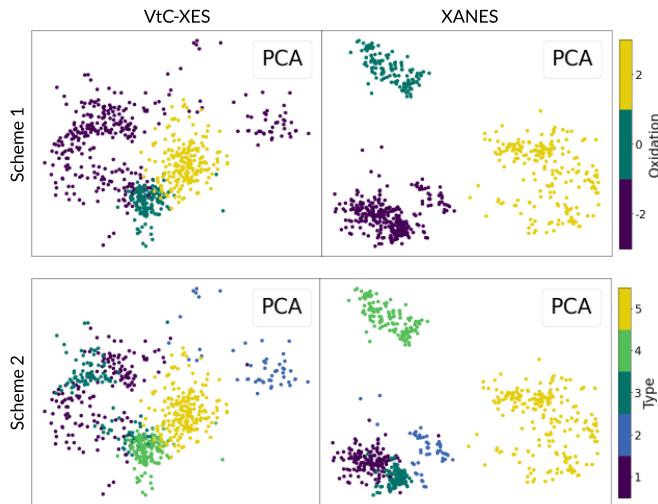


**Figure 6.6** Spectra reconstructed with increasing number of principal components (PCs) kept, for both VtC-XES and XANES of 2-thiazolidinone sulphone (Type 5) (top two panels) and 4-thiazoleaceticacid (Type 1) (bottom two panels).

To illustrate this fact, we show in Fig. 6.6 the gradual convergence with increasing number of PCA basis elements for two representative molecules, one from Type 5 and the other from Type 1. By increasing the number of PCs kept, more information is retained. For example, for 4-thiazoleacetic acid (bottom), starting at 2 PCs at the top and increasing downward to the original spectra at the bottom, the VtC-XES spectra clearly evolves from two peaks to three. For the XANES, the small peak in the valley at 2476 eV starts to appear around 8 PCs. However, the increase from 10 PCs to 12 PCs does not provide any

distinguishable change in the spectra. For 2-thiazolidinone sulphone (top), the XANES pre-edge features (or lack thereof) are not accurately represented until about 8 PCs, whereas just 2 PCs captures most of the spectral features for the VtC-XES. Again, the principal components were determined using the entire training data set for both XANES and VtC-XES.

The first two PCs can also be visualized by projecting the data onto a two-dimensional space using the corresponding eigenvectors, as shown in Fig. 6.7. Here, we color-coded the data via two chemically relevant classification schemes: “Scheme 1” (oxidation state) and “Scheme 2” (molecular moiety “Type”). Note how the oxidation state of the compounds clearly dominates the PCA of XANES (due to energy shifts, as expected), and thus the PCA of VtC-XES has better distinction between Types as it is not being over-dominated by oxidation. That said, there is considerable mixing of chemically different compounds in the XES projection – for example, the blue Type 2 thiocarbonyls mixing with the yellow Type 5 sulphones, and the purple Type 1 sulphides mixing with the dark green Type 3 thiols.



**Figure 6.7** Principal Component Analysis (PCA) projection for two dimensions, color-coded by the two different property classification schemes: Scheme 1 is by oxidation and Scheme 2 is by sulphur bond type.

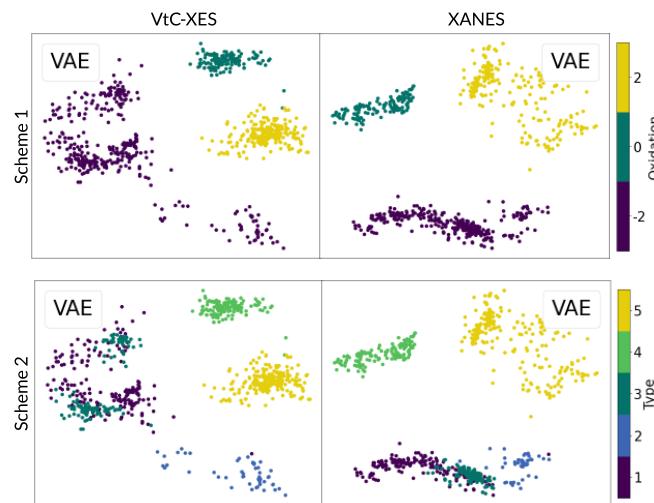
To summarize, PCA is a linear dimension reduction method that, when applied to both the XANES and VtC-XES of our ensemble on compounds, can accurately reconstruct spectra when a suitable number of PCs are retained. However, even just two PCs capture oxidation state, seen most obviously for XANES, and significant hints of sulphur bonding environment via the VtC-XES under the Type classification scheme.

However, the question now arises as to whether the orthogonalization and use of a Euclidean metric by PCA is optimal for the problem of chemical classification, especially if strongly limiting the number of principal components. This opens two questions. First, it is fair to ask if another linear algorithm could prove superior to PCA. This is investigated with Fast Independent Component Analysis (FastICA)<sup>97</sup>, Factor Analysis (FA)<sup>98, 99</sup>, and Non-negative Matrix Factorization (NMF)<sup>100</sup>, as shown in Fig. 6.S6. These three methods are other common linear dimensionality reduction routines and have been compared to PCA

in other systems<sup>101</sup>. See the SI for further information on those methods. By initial visual inspection, some seem to perform comparable PCA but are not categorically superior. Second, one must inquire, with linear dimensional reduction algorithms exhausted, if there is improved performance by using a nonlinear unsupervised method – either creating a nonlinear mapping (VAE) or merely a embedding (t-SNE).

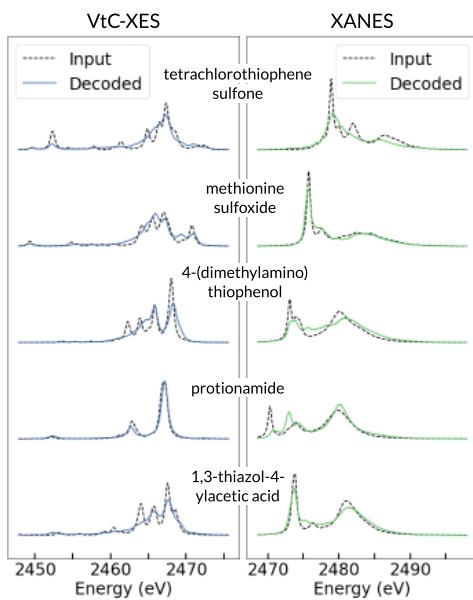
#### 6.4.1.2 Variational Autoencoder

We again present in Fig. 6.8 a reduction to a two-dimensional space, but now via the latent space of a trained VAE. Before comparing these results with the PCA-derived two-dimensional space in Fig. 6.7, it is useful to establish some basic properties of the VAE training and resulting latent space.



**Figure 6.8** Latent space representation in two dimensions via a Variational Autoencoder (VAE), color-coded by the two different property classification schemes: Scheme 1 is by oxidation and Scheme 2 is by sulphur bond type.

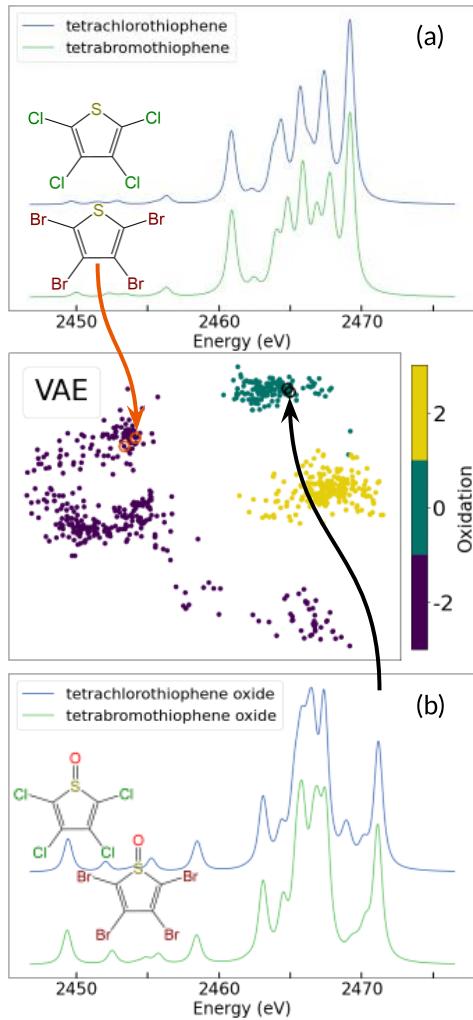
First, in Fig. 6.9 we demonstrate the agreement between input and decoded spectra – this is roughly analogous to the consideration of the number of retained PCs for PCA as shown in Fig. 6.6. The five spectra-pairs shown are for randomly selected compounds of each Type. Qualitative agreement is seen with a limited number of dominant spectral features, as would be expected given the inherent blurriness of decoded data from a VAE in two dimensions. Errors are largely restricted to features that are spectrally small or (especially) to spectra with numerous peaks. In some cases, this includes information-rich features, such as the first peak in the XANES of protonamide or the loss of the triple-peak structure in the immediate region near the Fermi level in the VtC-XES for 1,3-thiazol-4-ylacetic acid.



**Figure 6.9** Reconstruction of XES (left) and XANES (right) spectra from a two-dimensional latent space via a VAE. From bottom to top, the compounds are from Type 1, 2, 3, 4, and 5. The black

dashed line represents the original inputted spectra, and the solid-colored line is the decoded spectra after it has been passed through the VAE.

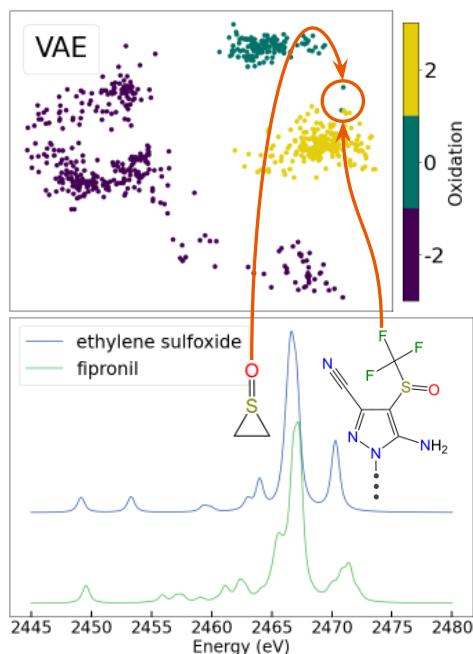
Second, while the VAE is nonlinear, the resulting mapping is still continuous and regular, such that similar spectra are mapped to nearby points in the latent space and, conversely, nearby points in the latent space decode to similar spectra. In Fig. 6.10a, the spectra for tetrabromothiophene and tetrachlorothiophene are very similar, and they are in fact mapped to a similar location in the latent space. Looking at the corresponding oxides in Fig. 6.10b, there is again a close location mapping of chemically related compounds of similar VtC-XES spectra. This indicates that the VAE is correctly mapping similar data to nearby locations, and therefore the latent space is in fact regularized, continuous, and complete. These three properties allow for data generation, where *the VAE can decode points in the latent space it has not previously seen*. We return to this subtle consequence of the good, if non-Euclidean, behaviour of the VAE latent space in section 4.3.



**Figure 6.10** Chemically similar compounds are nearby in the latent space. (a) The latent space location of tetrabromothiophene and tetrachlorothiophene, with the corresponding XES spectra on the right. (b) The same structures but oxidized to form tetrabromothiophene oxide and tetrachlorothiophene oxide.

As a final point of interest for the fidelity of the VAE latent space, it is interesting to investigate outliers in the VAE latent space, i.e., those molecules that substantially escape from the cluster associated with their oxidation state or Type. In Fig. 6.11 we identify both fipronil (only the relevant part of the structure is shown) and ethylene

sulphoxide as two Type 4 sulphoxides with nominally zero oxidation state that are unexpectedly in the sulphone +2 oxidation state cluster. The corresponding VtC-XES spectra and molecular structures are shown at the bottom of the figure. For fipronil, one of the carbons bonded to the S is special in that it is bonded to three fluorine, whose electronegativity also makes the carbon electronegative and thus the sulphur has an effective +1 oxidation, which might explain the grouping with the positive oxidation cluster. For ethylene sulphoxide, the abnormal triangle shape and unusual bond angles and lengths might contribute to its grouping with the +2 oxidation cluster.



**Figure 6.11** A closer look at the outliers: the two “neutrally oxidized” compounds distinctly in the sulphone (+2 oxidation) cluster.

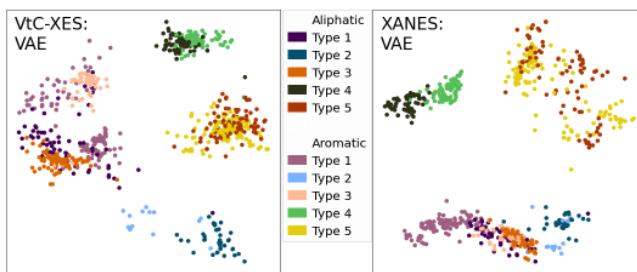
Moving now to the relative merits of the two-dimensional PCA representation (Fig. 6.7) and the VAE latent space (Fig. 6.8), the superior performance of the nonlinear method

is an important result of the present study, and there are three details that require further discussion. First, note how the latent space of the VtC-XES has very clear clustering of chemically related compounds in both classification schemes. In fact, the VtC-XES has better clustering than the XANES in Scheme 2 as Types 1, 2, and 3 are more distinguishable via VtC-XES. Also note that more similar compounds, such as Type 1 sulphides and Type 3 thiols, which have the same oxidation and very similar sulphur bonding environments, are closer together in the latent space for both XANES and VtC-XES when compared to the more chemically different Type 4 sulfoxides and Type 5 sulphones.

Second, the fact that there is better clustering of different oxidation states than for different sulphur bonding types is expected. The appearance of peaks due to the introduced oxygen bonds, in addition to the blueshift of the high energy tail, makes oxidation state correlate to the most pronounced differences in VtC-XES spectra. On the other hand, the XANES latent space is dominated by the oxidation state because of the multi-eV blue shift of the whiteline as oxidation state increases. However, the XANES has less-distinct clustering between Types 1, 2, and 3, all which have the same oxidation state, because the XANES spectra, in general, have less variation, both within individual Types and across them (recall Fig. 6.4). Hence, the fact that the VAE, at least when limited to a two-dimensional latent space, cannot as clearly distinguish sulphides (Type 1) from thiols (Type 3) in XANES, indicated by the large overlap in the purple and green dots, is expected; the sulphur local environment in both those Types is similar enough that there is large overlap.

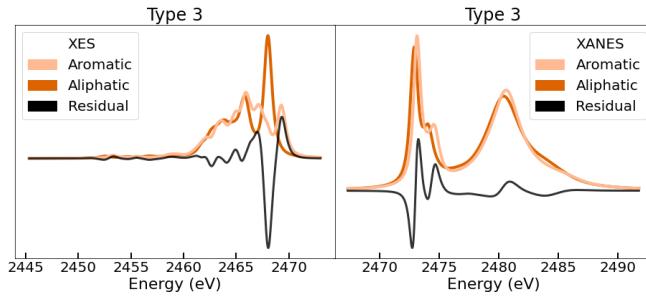
Third, the VAE latent space of the VtC-XES has two very distinct Type 3 clusters (not clearly seen in the PCA two-dimensional representation), whereas the XANES has grouped all Type 3 compounds together. These clusters in the VtC-XES spectra are directly

correlated to whether the sulphur in the thiol functional group belongs to a conjugated system (aromatic) or a non-conjugated one (aliphatic), as shown in Fig. 6.12. Here, we have color-coded spectra within types to indicate aromaticity, following Yasuda and Kakiyama<sup>65</sup>, who first noticed the sensitivity of sulphur VtC-XES to aromaticity. This separation is chemically reasonable as researchers have long known XAS to be sensitive to aromaticity for the carbon edges<sup>102</sup>, and have also observed sensitivity to aromaticity in a ligand, e.g. the sulphur K edge of sulphides<sup>60, 65</sup>.



**Figure 6.12** Compounds with aromatic sulphur versus aliphatic sulphur, in the latent space (VAE) for both VtC-XES (left) and XANES (right).

As shown in Fig. 6.13, the greatest difference in the VtC-XES spectra for Type 3 occurs at the highest energy peak, a consistent finding with the observations mentioned in Yasuda and Kakiyama<sup>65</sup>, which notes the aromaticity of the compound increases the energy but lowers the intensity of that peak, likely due to the presence of the  $\pi$  bonding system. Conversely, the XANES spectra, on average, have only a small ( $< 1$  eV) energy shift between the aromatic and aliphatic compounds for Type 3 without any substantial change in the overall spectral features.



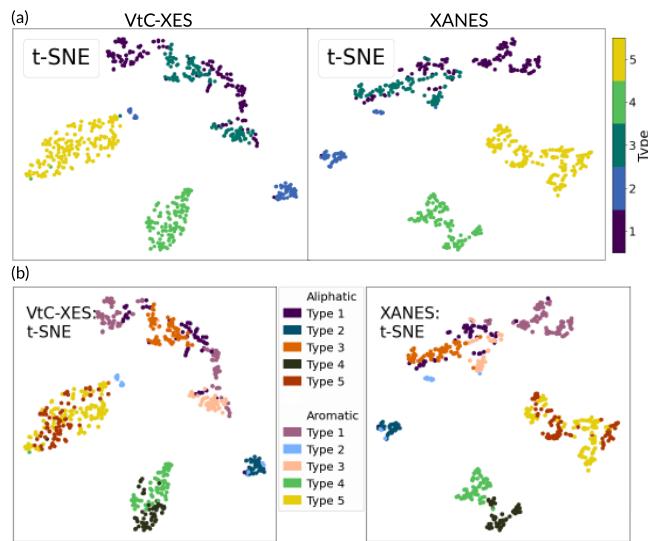
**Figure 6.13** Residuals between the average of the aromatic and aliphatic spectra of Type 3 (thiols).

This brings us naturally to the final section of raw results, where we use an algorithm that diverges even further from any metric constraint and instead emphasizes measuring similarity of the spectra prior to reducing the dimensionality of the problem.

#### 6.4.1.3 t-SNE, Clustering Without Mapping

In Fig. 6.14a, we show the two-dimensional embedding generated by the t-distributed Stochastic Neighbour Embedding (t-SNE), color-coded by Type, for the same training data sets as was used for PCA and the VAE, e.g., that resulted in the mappings in Fig. 6.7 and Fig. 6.8. Recall that although the closeness of points t-SNE embedding does correlate to similarity, the distances separating clusters in t-SNE does not necessarily represent the relative similarity of the clusters themselves – t-SNE is, again, inherently non-metric. The clustering is clearly tighter and, more importantly, there is less overlap between clusters corresponding to the different Types. In Fig. 6.14b we show the additional sub-classifications by conjugation of the radical group bonded to the sulphur, i.e., aromaticity. Notice that, as with the VAE, the VtC-XES clearly distinguishes the aromaticity of the

Type 3 thiols. Moreover, there is a clearer separation between aromatic and aliphatic compounds for all Types. Another observation in the t-SNE VtC-XES that was not present in PCA or VAE results is that the blue Type 2 group by the yellow Type 5 cluster consists of isothiocyanates, which are distinct from the other Type 2 thioketones.



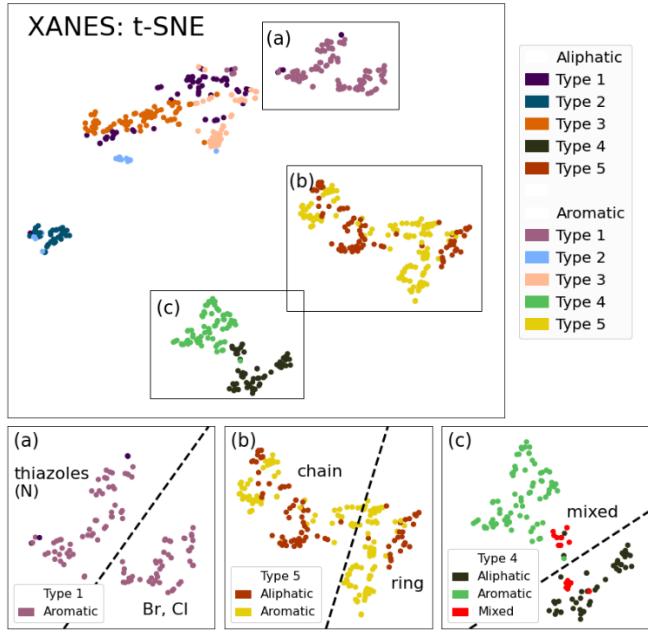
**Figure 6.14** t-SNE for VtC-XES (left) and XANES (right). (a) is color-coded by Type, while (b) is color-coded by aromaticity within each Type.

Some sensitivity to aromaticity could have been expected (although whether it would be seen in just a two-dimensional representation was definitely uncertain), given the prior work by Yasuda and Kakiyama<sup>65</sup> on VtC-XES and by Qureshi et al.<sup>60</sup> on XANES. Here, because t-SNE is unbiased, we can explore clustering in more detail to look for unexpected chemical classifications, an issue that we explore in Fig. 6.15 for XANES. First, we examine the further splitting of the Type 1 aromatic compounds as shown in Fig. 6.15a. On average, the spectra of the bottom cluster have about a 50% increase in the intensity of

the whiteline. These compounds all have either a chlorine or bromine bonded to the aromatic ring with the sulphur. On the other hand, the top cluster is typically thiazoles, or compounds where there is a nitrogen within the aromatic system containing the sulphur. Since chlorine and bromine are more electronegative than sulphur, it is chemically reasonable that they will dominate the compositions of the transitions close to the Fermi level and thus increase the whiteline intensity whereas the nitrogen in the ring will have the reverse affect.

Next, looking at the red aliphatic Type 5 compounds in Fig. 6.15b, it appears that they are grouped on either the left or right side of the overall Type 5 cluster. The cluster on the right, on average, has a slightly lower intensity and energy of the whiteline, with ~0.5 eV redshift. About 75% of the compounds in this cluster have the sulphur as part of a non-conjugated ring, compared to the sulphur being a member of chain-like compounds, as on the left side of the Type 5 cluster.

Finally, examining the split of the green Type 4 compound in Fig. 6.15c, we see clear partitioning based on aromaticity. However, upon identifying compounds in which one R group bonded to the sulphur is aromatic and the other R group is aliphatic, labelled as “mixed,” we see these in fact create the bridge between the two clusters as they share chemical characteristics with both groups. Thus, t-SNE has clearly identified real chemical (and thus spectral) trends in the XANES data.

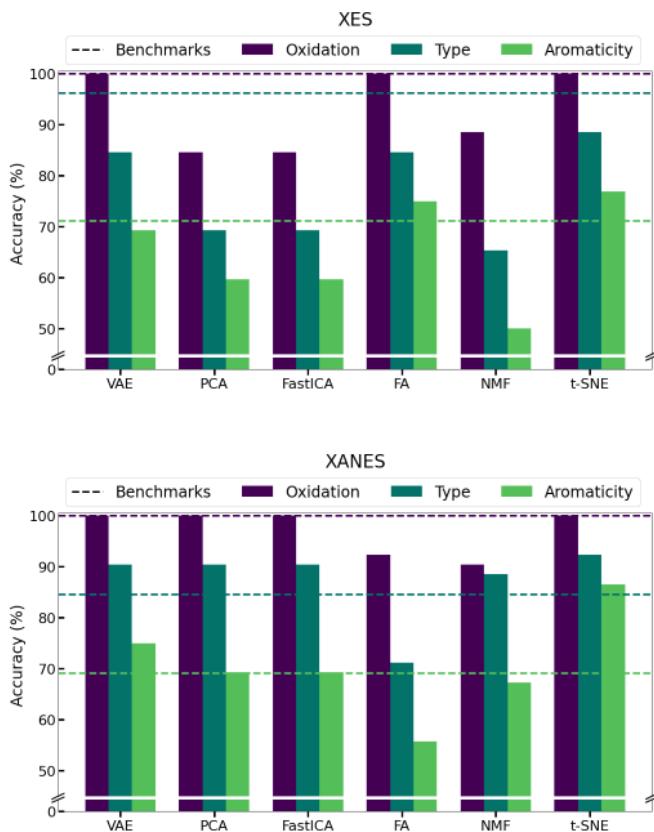


**Figure 6.15** (Main) A closer look the the subclusternig in the XANES t-SNE plot. (a) Separation of Type 1 aromatic compounds based on inclusion of chlorine or bromine in the aromatic system. (b) Separation of Type 5 aliphatic compounds based on bond strain via the inclusion of sulphur in a ring versus a chain. (c) Type 4 compounds with one R group aromatic and the other aliphatic share characteristics of both and thus form the bridge between the two custers.

#### 6.4.2 Classification

Hence, our initial qualitative inspection of the relative efficacy of PCA, VAE, and t-SNE for classification strongly supports the use of the least restrictive algorithm consistent with one's overall goals. We now seek quantitative assessment of the accuracy of classification via these algorithms. Based on K-Nearest Neighbours (KNN) partitioning on the reduced spaces for both VtC-XES and XANES, we derived the classification accuracies for the three primary methods of this study as well as the auxiliary linear methods FastICA, FA, and NMF, as shown in Fig. 6.16. For t-SNE, because of its nature

as a non-parametric embedding rather than a mapping, the test data was folded into the initial embedding, so the entire dimension reduction and test accuracy were applied in one step, although the KNN was only trained on the training dataset. For all other methods, training included both fitting the dimension reduction mapping to the training dataset, and then applying KNN on the two-dimensional space using that training data projection. To assess accuracy, the test data was then passed through the mapping to lower dimensional and subsequently through the fitted KNN partitioning.



**Figure 6.16** Accuracy of KNN classification schemes on all dimensionally reduced spaces for both VtC-XES (top) and XANES (bottom).

Regarding classification Scheme 1 (Oxidation), most methods performed extremely well (above 95% accurate) and were comparable to the benchmark accuracy obtained from the fully connected neural network classifier, as shown in purple in Fig. 6.16. Applying KNN to achieve classification accuracy using Scheme 2 (Type) on all reduced spaces for both XES and XANES is also shown in Fig. 6.16. For the VtC-XES spectra, VAE, FA, and t-SNE performed the best (with FA having surprisingly high accuracies) and closest to the benchmark, while for the XANES spectra, all methods (besides FA) performed comparably. Finally, we applied KNN to the spaces for classification Scheme 3 (Aromaticity). All methods performed comparatively to each other as they performed on the Type classification, and accuracies were comparable for both the VtC-XES and the XANES, despite the clear Type 3 separation in the VtC-XES. However, t-SNE applied on the XANES spectra clearly dominated, achieving a notable accuracy of 87% for aromaticity. Moreover, of the three classification schemes for both the VtC-XES and XANES, the VAE and t-SNE outperformed or matched the benchmark accuracy 75% of the time. This is extraordinary, as these reduced spaces were constrained to merely two dimensions.

Some other things to note overall: (1) t-SNE and the VAE were much more consistent and robust than the linear algorithms, whose accuracies greatly depended on both the chosen dataset and classification scheme and thus seem more volatile than the nonlinear methods (all KNN spaces can be viewed in Figs. S7 to S12); (2) the performance of VAE is comparable to t-SNE for oxidation state and Type (although not for aromaticity or finer speciation), but has an additional benefit in that it is a mapping and can thus be used to

efficiently store future spectra, discussed in more detail below; and (3) the VtC-XES and XANES had extremely similar overall categorical sensitivity to electronic structure.

#### 6.4.3 Summary and Outlook

We have focused here on three chemical classification schemes, determined from clusters in a reduced representation of the dataset. Although identifying similarities of XANES spectra via clustering was introduced in Kiyohara, et al.<sup>20</sup>, which used a decision tree to interpret the results of hierarchical clustering of small ensemble of XANES spectra, they could not directly obtain characteristic information corresponding to each cluster. On the other hand, our routines created clusters that were directly interpretable into chemical classes. It would be interesting in the future to evaluate more fully the VAE and t-SNE reduced spaces for other potential properties of interest, such as bond length, that can be used for prediction via regression. Furthermore, expansion of the dataset to include ligands other than carbon or oxygen would be another beneficial investigation, which has been shown to be challenging in other systems<sup>59</sup>. Additionally, the extension of our methods to other classes of organic and inorganic systems would not only help to understand the spectral encoding of chemically relevant information in those other systems but will also further illuminating the differences, or lack thereof, in the information content of VtC-XES and XANES.

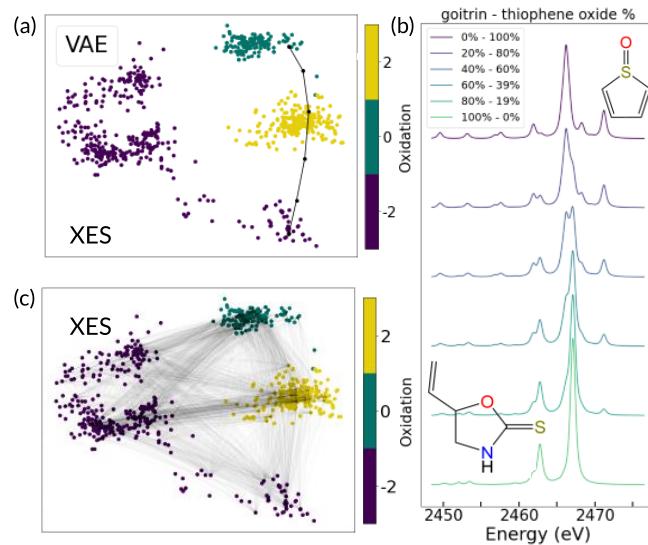
On a different point, the observation that some of the dimension reduction routines performed comparably to the benchmark accuracy indicates that they are ripe, either in their current condition or with some more tuning, for compressing high dimensional spectra with minimal informational loss, and thus provide classification accuracies close to an upper

bound, limited only by the aleatoric variation of the dataset itself. Moreover, classification accuracies can be further improved by keeping more dimensions when projecting onto these reduced spaces, along with more training data, if available, such as augmenting the dataset to include noise or impurities to better mimic experimental data. Further tuning of these methods, especially modelling spectral artifacts and realistic experimental conditions in the training dataset to increase robustness, would allow for potential use in encoding high dimensional spectral data in high throughput experiments.

As a case in point, recall that in section 4.1.2, and especially in Fig. 6.10, we discussed the regularized, continuous, and complete nature of the VAE latent space. These characteristics allow for both the encoding of additional spectra into the latent space and, conversely, allow the VAE to decode points in the latent space that do not correspond to previous observations. We propose that this capability might be useful for the growing number of high-throughput XAS experiments that require real-time data encoding, although the same may of course also hold for other one-dimensional spectroscopies. For example, *in operando* XAS catalysis studies are a high-throughput effort that observes progressive changes in spectral features and then seeks to understand the corresponding local chemical changes. A latent space mapping of such chemical evolution might be at least qualitatively useful to the experimenter.

In Fig. 6.17a we show the evolution from goitrin (oxidation state -2) to thiophene oxide (oxidation state 0). In Fig. 6.17b, we have the decoded spectra from the points in Fig. 6.17a along a trajectory corresponding to linear combination of mole fraction of the two molecules. A more complete depiction of latent space trajectories is shown in Fig. 6.17c, where we have over 3000 different combinations of randomly selected species evolutions.

Because the tracks cross over the regions between the clusters, generating or tracking in this region will be reliable, whereas the spaces outside these clusters will not yield any meaningful interpretation to the latent space encoding.



**Figure 6.17** As shown in (a), the evolution from goitrin (oxidation -2) to thiophene oxide (oxidation 0). (b) The linear combination of the spectra of thiophene oxide (top) and goitrin (bottom) that correspond to the points along the track in (a). (c) Tracks of 3000 different species evolutions.

A technical point worthy of mention here is that several prior ML studies in X-ray spectroscopy have augmented their training dataset by including linear combinations of basis spectra, e.g., Timoshenko, et al.<sup>27</sup> However, PCA and VAE inherently encode these linear combinations into the reduced mapping. This attribute is obvious based on how PCA constructs its components and was verified in the VAE, where training on an augmented dataset resulted in statistically the same latent space representation of the pure component

spectra. On the other hand, properly including linear combinations into a t-SNE training set would result in a multivariate t-distribution and completely detract from the purpose of applying t-SNE – obtaining clusters and identifying similarities. Moreover, our dataset included enough variation of our system of interest that we did not need to augment our training set to improve results.

## 6.5 Conclusions

Using a large family of sulphorganic molecules as a test case, we have performed a comprehensive survey of dimensionality reduction via unsupervised machine learning (ML) methods applied to X-ray absorption and X-ray emission spectroscopy as a means toward chemical classification. In this paper, we come to three main conclusions.

First, despite all algorithms being restricted to two dimensions, the unsupervised ML methods showed good accuracy for most of the relevant chemical information, with t-SNE somewhat outperforming the supervised benchmark and the other methods comparable to it. Particularly, t-SNE appears to have surpassed the other methods exactly because it retains the similarity measures initially calculated in the original high-dimensional space of the training data set, avoiding the lossy compression inherent to methods that map first and compare second.

One might ask if PCA or VAE could find improved performance by increasing their reduced dimensionality, where these two methods have the benefit over t-SNE of providing actual mapping functions, and thus they can more naturally be used for real-time interpretation of experimental results. Fig. 6.S13 shows the accuracies for PCA, VAE, and t-SNE for a latent or embedding dimension of three and four. This figure exemplifies the superiority of t-SNE at low dimensions, such as two or three, exactly because it solves the

“crowding problem”<sup>96</sup> that results from the curse of dimensionality. However, at four or more dimensions, t-SNE is not only more comparable to the VAE – the crowding problem becomes less of an issue then – but the computational cost greatly increases. Specifically, an exact solution (instead of the Barnes-Hut approximation) optimization algorithm must be used for dimensions greater than or equal to four. However, the slight increase in accuracy for all methods while increasing the reduced dimension (at least to four) suggests further tuning could yield even greater classification accuracies for all models. These results suggest multiple directions forward, particularly for their use not only across other chemical systems, but also other one-dimensional spectroscopies.

In Fig. 6.16, we have shown superior classification performance for t-SNE, and as stated earlier, this is likely because t-SNE performs a comparison between the full, original spectra prior to dimension reduction via embedding, whereas PCA and VAE are inherently lossy mappings.

Second, t-SNE not only had superior performance for classifying aromaticity, but also unexpectedly found new chemically relevant clusters not seen in any other method, such as distinguishing finer sub-classes within the aromaticity of sulphides (Type 1), sulphoxides (Type 4), and sulphones (Type 5). We see considerable future benefit to combining highly adaptive unsupervised ML algorithms, such as t-SNE, in tandem with supervised ML or with structural parameterization questions that have to date been only addressed in XAS using supervised ML.

Finally, the above results allow us to formally quantify and compare the chemical information content between XANES and VtC-XES, an issue which has only seen qualitative discussion. We find that XANES and VtC-XES methods each have strengths

for chemical classification, but that many are the same, at least for the question of chemical classification of sulphorganics.

### **Author Contributions**

Tetef led the effort and investigation in each of electronic structure calculations, machine learning calculations, and subsequent statistical analysis. Tetef led the writing effort, with strong contributions from the other two authors.

### **Conflicts of interest**

There are no conflicts to declare.

### **Acknowledgements**

We acknowledge funding from NRT-DESE: Data Intensive Research Enabling Clean Technologies (DIRECT) under grant no. NSF #1633216 and acknowledge funding from NSF CHE-1904437. NG acknowledges support from the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences under Award No KC-030105172685. This research benefited from computational resources (Cascade) provided by the Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at PNNL. PNNL is operated by Battelle Memorial Institute for the United States Department of Energy under DOE Contract No. DE-AC05-76RL1830. We would especially like to thank Dr. Fernando Vila for invaluable input and useful discussions.

## 6.6 References

1. D. A. C. Beck, J. M. Carothers, V. R. Subramanian and J. Pfaendtner, *AICHE Journal*, 2016, **62**, 1402-1416.
2. C. Ashraf, N. Joshi, D. A. C. Beck and J. Pfaendtner, *Annual Review of Chemical and Biomolecular Engineering*, 2021, **12**, 15-37.
3. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.
4. G. Bergerhoff, R. Hundt, R. Sievers and I. D. Brown, *Journal of Chemical Information and Computer Sciences*, 1983, **23**, 66-69.
5. A. Belsky, M. Hellenbrandt, V. L. Karen and P. Luksch, *Acta Crystallographica Section B*, 2002, **58**, 364-369.
6. L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *Journal of Chemical Information and Modeling*, 2012, **52**, 2864-2875.
7. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547-555.
8. S. Jaeger, S. Fulle and S. Turk, *Journal of Chemical Information and Modeling*, 2018, **58**, 27-35.
9. C. R. Collins, G. J. Gordon, O. A. von Lilienfeld and D. J. Yaron, *The Journal of Chemical Physics*, 2018, **148**, 241718.
10. F. Huang, R. Li, G. Wang, J. Zheng, Y. Tang, J. Liu, Y. Yang, Y. Yao, J. Shi and W. Hong, *Phys. Chem. Chem. Phys.*, 2020, **22**, 1674-1681.
11. M. Ceriotti, *The Journal of Chemical Physics*, 2019, **150**, 150901.
12. A. Aarva, V. L. Deringer, S. Sainio, T. Laurila and M. A. Caro, *Chemistry of Materials*, 2019, **31**, 9243-9255.
13. M. R. Carbone, M. Topsakal, D. Lu and S. Yoo, *Physical Review Letters*, 2020, **124**, 156401(156406).
14. M. R. Carbone, S. Yoo, M. Topsakal and D. Lu, *Physical Review Materials*, 2019, **3**, 033604.
15. S. Kiyohara, M. Tsubaki and T. Mizoguchi, *Npj Computational Materials*, 2020, **6**, 68.
16. L. Li, M. Lu and M. K. Y. Chan, *arXiv*, 2019.
17. Y. Liu, N. Marcella, J. Timoshenko, A. Halder, B. Yang, L. Kolipaka, M. J. Pellin, S. Seifert, S. Vajda, P. Liu and A. I. Frenkel, *The Journal of Chemical Physics*, 2019, **151**, 164201.
18. A. Martini, S. A. Guda, A. A. Guda, G. Smolentsev, A. Algasov, O. Usoltsev, M. A. Soldatov, A. Bugaev, Y. Rusalev, C. Lamberti and A. V. Soldatov, *Computer Physics Communications*, 2020, **250**, 107064.
19. I. Miyazato, L. Takahashi and K. Takahashi, *Molecular Systems Design & Engineering*, 2019, **4**, 1014-1018.
20. S. Kiyohara, T. Miyata, K. Tsuda and T. Mizoguchi, *Scientific Reports*, 2018, **8**, 13548.
21. T. Mizoguchi and S. Kiyohara, *Microscopy*, 2020, **69**, 92-109.
22. C. D. Rankine, M. M. M. Madkhali and T. J. Penfold, *The Journal of Physical Chemistry A*, 2020, **124**, 4263-4270.
23. P. K. Routh, Y. Liu, N. Marcella, B. Kozinsky and A. I. Frenkel, *The Journal of Physical Chemistry Letters*, 2021, **12**, 2086-2094.

24. J. Terry, M. L. Lau, J. Sun, C. Xu, B. Hendricks, J. Kise, M. Lnu, S. Bagade, S. Shah, P. Makhijani, A. Karantha, T. Boltz, M. Oellien, M. Adas, S. Argamon, M. Long and D. P. Guillen, *Appl. Surf. Sci.*, 2021, **547**, 149059.
25. J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans and A. I. Frenkel, *Physical Review Letters*, 2018, **120**, 225502.
26. J. Timoshenko and A. I. Frenkel, *Acs Catalysis*, 2019, **9**, 10192-10211.
27. J. Timoshenko, D. Y. Lu, Y. W. Lin and A. I. Frenkel, *Journal of Physical Chemistry Letters*, 2017, **8**, 5091-5098.
28. J. Timoshenko, C. J. Wräzman, M. Luneau, T. Shirman, M. Cagnello, S. R. Bare, J. Aizenberg, C. M. Friend and A. I. Frenkel, *Nano Letters*, 2019, **19**, 520-529.
29. S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram and L. Hung, *npj Computational Materials*, 2020, **6**, 109.
30. C. Zheng, C. Chen, Y. Chen and S. P. Ong, *Patterns*, 2020, **1**, 100013.
31. C. Zheng, K. Mathew, C. Chen, Y. M. Chen, H. M. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson and S. P. Ong, *Npj Computational Materials*, 2018, **4**, 12.
32. C. D. Rankine and T. J. Penfold, *The Journal of Physical Chemistry A*, 2021, **125**, 4276-4293.
33. G. Bunker, *Introduction to XAFS: A Practical Guide to X-ray Absorption Fine Structure Spectroscopy*, Cambridge University Press, Cambridge, 2010.
34. P. Glatzel and U. Bergmann, *Coordination Chemistry Reviews*, 2005, **249**, 65-95.
35. F. de Groot, *Chemical Reviews*, 2001, **101**, 1779-1808.
36. E. P. Jahrman, W. M. Holden, A. S. Ditter, D. R. Mortensen, G. T. Seidler, T. T. Fister, S. A. Kozimor, L. F. J. Piper, J. Rana, N. C. Hyatt and M. C. Stennett, *Review of Scientific Instruments*, 2019, **90**, 024106.
37. G. T. Seidler, D. R. Mortensen, A. J. Remesnik, J. I. Pacold, N. A. Ball, N. Barry, M. Styczinski and O. R. Hoidn, *Review of Scientific Instruments*, 2014, **85**, 113906.
38. W. M. Holden, O. R. Hoidn, A. S. Ditter, G. T. Seidler, J. Kas, J. L. Stein, B. M. Cossairt, S. A. Kozimor, J. Guo, Y. Ye, M. A. Marcus and S. Fakra, *Review of Scientific Instruments*, 2017, **88**, 073904.
39. W. Malzer, C. Schlesiger and B. Kanngießer, *Spectrochimica Acta Part B: Atomic Spectroscopy*, 2021, **177**, 106101.
40. P. Zimmermann, S. Peredkov, P. M. Abdala, S. DeBeer, M. Tromp, C. Müller and J. A. van Bokhoven, *Coordination Chemistry Reviews*, 2020, **423**, 213466.
41. N. Kornienko, J. Resasco, N. Becknell, C.-M. Jiang, Y.-S. Liu, K. Nie, X. Sun, J. Guo, S. R. Leone and P. Yang, *Journal of the American Chemical Society*, 2015, **137**, 7448-7455.
42. M. Cuisinier, P.-E. Cabelguen, S. Evers, G. He, M. Kolbeck, A. Garsuch, T. Bolin, M. Balasubramanian and L. F. Nazar, *The Journal of Physical Chemistry Letters*, 2013, **4**, 3227-3232.
43. D. Asakura, E. Hosono, H. Niwa, H. Kiuchi, J. Miyawaki, Y. Nanba, M. Okubo, H. Matsuda, H. Zhou, M. Oshima and Y. Harada, *Electrochemistry Communications*, 2015, **50**, 93-96.
44. A. Arcovito, M. Benfatto, M. Cianci, S. S. Hasnain, K. Nienhaus, G. U. Nienhaus, C. Savino, R. W. Strange, B. Vallone and S. Della Longa, *Proceedings of the National Academy of Sciences*, 2007, **104**, 6211.

45. M. Bounce, J. Boyce, F. M. McCubbin, J. Humphreys, J. Reppard, E. Stolper and J. Eiler, *Am. Miner.*, 2019, **104**, 307-312.
46. Y. Zhou, D. E. Doronkin, Z. Zhao, P. N. Plessow, J. Jelic, B. Detlefs, T. Pruessmann, F. Studt and J.-D. Grunwaldt, *ACS Catalysis*, 2018, **8**, 11398-11406.
47. C. Kupitz, S. Basu, I. Grotjohann, R. Fromme, N. A. Zatsepин, K. N. Rendek, M. S. Hunter, R. L. Shoeman, T. A. White, D. Wang, D. James, J.-H. Yang, D. E. Cobb, B. Reeder, R. G. Sierra, H. Liu, A. Barty, A. L. Aquila, D. Deponte, R. A. Kirian, S. Bari, J. J. Bergkamp, K. R. Beyerlein, M. J. Bogan, C. Caleman, T.-C. Chao, C. E. Conrad, K. M. Davis, H. Fleckenstein, L. Galli, S. P. Hau-Riege, S. Kassemeyer, H. Laksmono, M. Liang, L. Lomb, S. Marchesini, A. V. Martin, M. Messerschmidt, D. Milathianaki, K. Nass, A. Ros, S. Roy-Chowdhury, K. Schmidt, M. Seibert, J. Steinbrener, F. Stellato, L. Yan, C. Yoon, T. A. Moore, A. L. Moore, Y. Pushkar, G. J. Williams, S. Boutet, R. B. Doak, U. Weierstall, M. Frank, H. N. Chapman, J. C. H. Spence and P. Fromme, *Nature*, 2014, **513**, 261-265.
48. M. Maiuri, M. Garavelli and G. Cerullo, *Journal of the American Chemical Society*, 2020, **142**, 3-15.
49. J. J. Rehr and R. C. Albers, *Reviews of Modern Physics*, 2000, **72**, 621-654.
50. F. De Groot and A. Kotani, 2008, DOI: 10.1201/9781420008425.
51. J. J. Rehr, J. Kozdon, J. Kas, H. J. Krappe and H. H. Rossner, *Journal of Synchrotron Radiation*, 2005, **12**, 70-74.
52. H. J. Krappe and H. H. Rossner, *Physical Review B*, 2002, **66**, 184303.
53. H. J. Krappe and H. H. Rossner, *Physica Scripta*, 2009, **79**, 048302.
54. H. H. Rossner, D. Schmitz, P. Imperia, H. J. Krappe and J. J. Rehr, *Physical Review B*, 2006, **74**, 134107.
55. B. Ravel and M. Newville, *Journal of Synchrotron Radiation*, 2005, **12**, 537-541.
56. M. Newville, *Journal of Synchrotron Radiation*, 2001, **8**, 322-324.
57. E. Stavitski and F. M. F. De Groot, *Micron*, 2010, **41**, 687-694.
58. R. A. Mori, E. Paris, G. Giuli, S. G. Eeckhout, M. Kavčič, M. Žitnik, K. Bučar, L. G. M. Pettersson and P. Glatzel, *Inorganic Chemistry*, 2010, **49**, 6468-6473.
59. S. N. MacMillan, R. C. Walroth, D. M. Perry, T. J. Morsing and K. M. Lancaster, *Inorganic Chemistry*, 2015, **54**, 205-214.
60. M. Qureshi, S. H. Nowak, L. I. Vogt, J. J. H. Cotelesage, N. V. Dolgova, S. Sharifi, T. Kroll, D. Nordlund, R. Alonso-Mori, T.-C. Weng, I. J. Pickering, G. N. George and D. Sokaras, *Phys. Chem. Chem. Phys.*, 2021, **23**, 4500-4508.
61. C. J. Pollock and S. DeBeer, *Accounts of Chemical Research*, 2015, **48**, 2967-2975.
62. J. L. Lansford and D. G. Vlachos, *Nature Communications*, 2020, **11**, 1513.
63. X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov and Z. Chen, *Angewandte Chemie International Edition*, 2020, **59**, 10297-10300.
64. F. Lussier, V. Thibault, B. Charron, G. Q. Wallace and J.-F. Masson, *TrAC Trends in Analytical Chemistry*, 2020, **124**, 115796.
65. S. Yasuda and H. Kakiyama, *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.*, 1979, **35**, 485-493.
66. W. M. Holden, E. P. Jahrman, N. Govind and G. T. Seidler, *The Journal of Physical Chemistry A*, 2020, **124** (26), 5415-5434.
67. K. Lopata, B. E. Van Kuiken, M. Khalil and N. Govind, *Journal of Chemical Theory and Computation*, 2012, **8**, 3284-3292.

68. Y. Zhang, S. Mukamel, M. Khalil and N. Govind, *Journal of Chemical Theory and Computation*, 2015, **11**, 5804-5809.
69. E. P. Jahrman, W. M. Holden, N. Govind, J. J. Kas, J. Rana, L. F. J. Piper, C. Siu, M. S. Whittingham, T. T. Fister and G. T. Seidler, *Journal of Materials Chemistry A*, 2020, **8**, 16332-16344.
70. D. R. Mortensen, G. T. Seidler, J. J. Kas, N. Govind, C. P. Schwartz, S. Pemmaraju and D. G. Prendergast, *Physical Review B*, 2017, **96**, 125136.
71. S. Lee, M. Kwak, K. L. Tsui and S. B. Kim, *Eng. Appl. Artif. Intell.*, 2019, **83**, 13-27.
72. H. Bergwerf, MolView, <http://molview.org/>, (accessed February 16, 2021).
73. M. M. Francz, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. Defrees and J. A. Pople, *J. Chem. Phys.*, 1982, **77**, 3654-3665.
74. M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro and W. J. Hehre, *Journal of the American Chemical Society*, 1982, **104**, 2797-2803.
75. M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. De Jong, *Computer Physics Communications*, 2010, **181**, 1477-1489.
76. E. Apra, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Attafynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauet, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Fruchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Gotz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jonsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. M. del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, A. Vazquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Wolinski, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison, *J. Chem. Phys.*, 2020, **152**, 26.
77. P. C. Hariharan and J. A. Pople, *Theoretica chimica acta*, 1973, **28**, 213-222.
78. W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257.
79. A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648-5652.
80. T. Noro, M. Sekiya and T. Koga, *Theoretical Chemistry Accounts*, 2012, **131**, 1124.
81. C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158-6170.
82. A. Bergner, M. Dolg, W. Küchle, H. Stoll and H. Preuß, *Molecular Physics*, 1993, **80**, 1431-1441.
83. A. Mijovilovich, L. G. M. Pettersson, S. Mangold, M. Janousch, J. Susini, M. Salome, F. M. F. de Groot and B. M. Weckhuysen, *Journal of Physical Chemistry A*, 2009, **113**, 2750-2756.

84. S. B. Emilie Chalmin, Marine Cotte, Jean-Pierre Cuif, Koen Janssen, Laurence Lemelle, Magnus Sandström, and M. S.-B. Andréas Scheinost, Frances Westall, and Max Wilke, *Journal*.
85. F. a. o. Chollet, *Journal*, 2015.
86. A. A. Martín Abadi, Paul Barham, Eugene Brevdo,, C. C. Zhifeng Chen, Greg S. Corrado, Andy Davis,, M. D. Jeffrey Dean, Sanjay Ghemawat, Ian Goodfellow,, G. I. Andrew Harp, Michael Isard, Rafal Jozefowicz, Yangqing Jia,, M. K. Lukasz Kaiser, Josh Levenberg, Dan Mané, Mike Schuster,, S. M. Rajat Monga, Derek Murray, Chris Olah, Jonathon Shlens,, I. S. Benoit Steiner, Kunal Talwar, Paul Tucker,, V. V. Vincent Vanhoucke, Fernanda Viégas,, P. W. Oriol Vinyals, Martin Wattenberg, Martin Wicke, and a. X. Z. Yuan Yu, *Journal*, 2015.
87. stetef, *Journal*, 2021, June 11, DOI: <http://doi.org/10.5281/zenodo.4931519>.
88. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825-2830.
89. P. Indyk and R. Motwani.
90. S. Wold, K. Esbensen and P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37-52.
91. A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo and S. Severini, *npj Quantum Information*, 2018, **4**, 28.
92. S. K. N. Portillo, J. K. Parejko, J. R. Vergara and A. J. Connolly, *Astron. J.*, 2020, **160**, 17.
93. G. E. Hinton, *Science*, 2006, **313**, 504-507.
94. M. S. Mahmud, J. Z. Huang and X. H. Fu, *Int. J. Comput. Intell. Appl.*, 2020, **19**, 19.
95. M. Farrell, S. Recanatesi, R. C. Reid, S. Mihalas and E. Shea-Brown, *Neural Networks*, 2021, DOI: <https://doi.org/10.1016/j.neunet.2021.03.010>, 330-343.
96. L. van der Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579-2605.
97. A. Hyvärinen and E. Oja, *Neural Networks*, 2000, **13**, 411-430.
98. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
99. D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.
100. D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788-791.
101. S. Sun, J. Zhu, Y. Ma and X. Zhou, *Genome Biology*, 2019, **20**, 269.
102. J. Stöhr, *NEXAFS Spectroscopy*, Springer, 1992.

## 6.7 Supporting Information

### 6.7.1 Explanation of VAE loss

In Fig. 6.S1 we show the special loss, or objective function, used in VAEs as a function of training epoch for both the training and validation data sets for the XANES and VtC-XES datasets. This special loss is defined as the mean of the reconstruction loss (binary cross entropy) and the *Kullback-Leibler* (KL) divergence. KL divergence ensures the VAE is fully utilizing the latent space by penalizing lost information. In general, it is given by

$$\begin{aligned} D_{KL} [P(z|x)||Q(z||x)] &= E[ \log P(z|x) - \log Q(z|x) ] \\ &\equiv \sum P(z|x) \log \frac{P(z|x)}{Q(z|x)} \end{aligned} \quad (6.1)$$

where P is the probability distribution and Q is the approximation of P. Thus, KL divergence identifies how much information it lost using the approximation Q. In a VAE objective function, z is the latent space representation of our data x, Q is the encoder, and P is the decoder. Thus, the KL divergence is

$$-\frac{1}{2} \sum 1 + \log \sigma_z^2 - \mu_z^2 - \sigma_z^2 \quad (6.2)$$

where  $\log \text{var}(z)$  and  $\text{mean}(z)$  are the two parallel latent space layers of the VAE. Moreover, KL divergence encourages the latent space to be centered around zero with normal variance and is therefore regularized. For an in-depth derivation of VAE objective function, see Rocchetto et al.<sup>1</sup>

A plot such as Fig. 6.S1 is a useful heuristic for understanding training convergence and for evaluating the degree of overfitting or underfitting. To be specific, starting with XANES, the losses plateau at about 20 epochs and the validation loss does not increase. This indicates that the resulting neural network is generalizable and is not overfitting and thus is likely to have high

utility, i.e., it has not overfit such that it cannot address spectra outside the training data set but also enough detail has been encoded that most useful information has likely been incorporated. The VtC-XES shows a similar plateau in the VAE losses, which indicates this model is not overfitting as well.

### 6.7.2 Increasing latent space dimension

As with PCA, where one must wisely choose the number of PCs to get a good representation of the training data set, we are also free to modify the dimension of the latent space for the VAE. This is investigated in Fig. 6.S2 where representative XES spectra (one from each type) are compared to the corresponding decoded spectra as a function of the dimension of the latent space, starting with two dimensions on the left and proceeding to 50 dimensions at the right. Increasing the latent space dimension up to 50 dimensions does not drastically change the accuracy of the decoded spectra, as the most distinct features are obtained just from a two-dimensional latent space. Hence, for the VTC-XES for this broad collection of sulphorganics, a two-dimensional representation VAE is enough to capture the most distinct spectral features, giving a dramatically effective encoding and dimensionality reduction.

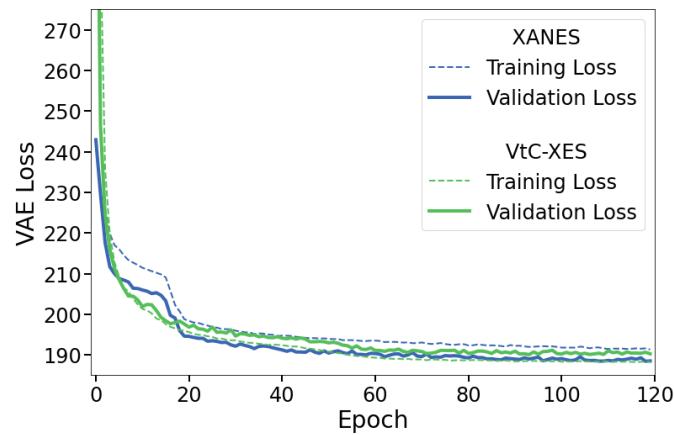
### 6.7.3 Hyperparameter tuning

Hyperparameters for all machine learning methods were selected using multiple validation sets separated from within the entire training set. The VAE was limited to one or two hidden layers for each the encoder and decoder with layer dimension sizes constrained to powers of two between 32 and 1024. The ANN classifier was also limited to one or two hidden layers with dimensions constrained to powers of two between 32 and 1024. Dropout was also constrained to be between 5% and 20% and implemented to encourage generalizability. The t-SNE perplexity value was selected from values between 5 and 50, where the smallest perplexity value was chosen such that (a) there did not appear to be spurious or artificial clusters and (b) yielded consistent embeddings upon recalculation, indicating a global minimum was reached. The k nearest neighbor hyperparameter for KNN was then selected from a neighborhood of values around the t-SNE perplexity value, since both approximately represent cluster or group size. This was determined to be between 10 and 30. All other hyperparameters not specified in the manuscript were set to default values.

#### 6.7.4 FastICA, FA, and NMF

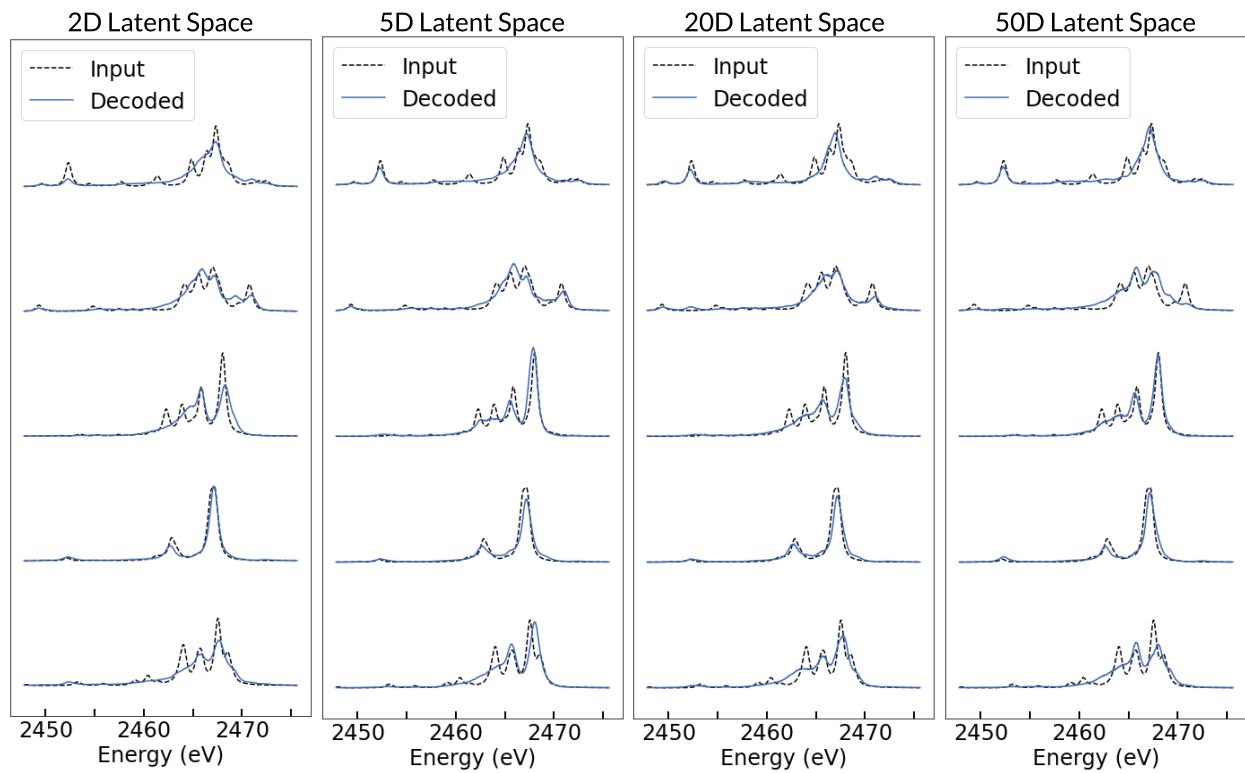
The three supplemental linear dimensionality reduction methods included in this study are Fast Independent Component Analysis (FastICA)<sup>2</sup>, Factor Analysis (FA)<sup>3, 4</sup>, and Non-negative Matrix Factorization (NMF)<sup>5</sup>. FastICA is an implementation of independent component analysis, which is a generalization of PCA. Often, independent component analysis is used to separate out independent signals, or components, contributing to data. However, because its aim is to calculate independent components, there is not a clear statistical method to reduce dimension, such as maximizing the explained variance as with PCA. Factor analysis (FA) is similar to PCA as well, except it calculates the eigenvalue decomposition on the reduced correlation matrix instead of the full correlation matrix. This analysis on the reduced correlation matrix helps identify latent, or hidden, features in the data, i.e., variables that cause correlated features in the original dataset. However, it has been shown that if the number of included datapoints is large enough (about 40), PCA and FA have similar results<sup>6</sup>. Finally, non-negative matrix factorization (NMF) is another linear dimensionality reduction algorithm that assumes the data is (as the name suggests) non-negative, which is true for both XANES and XES spectra. NMF calculates two non-negative matrices whose product reproduces the original dataset. This encourages factors to be positive and thus more physically interpretable.

Loss plotted against number of epochs



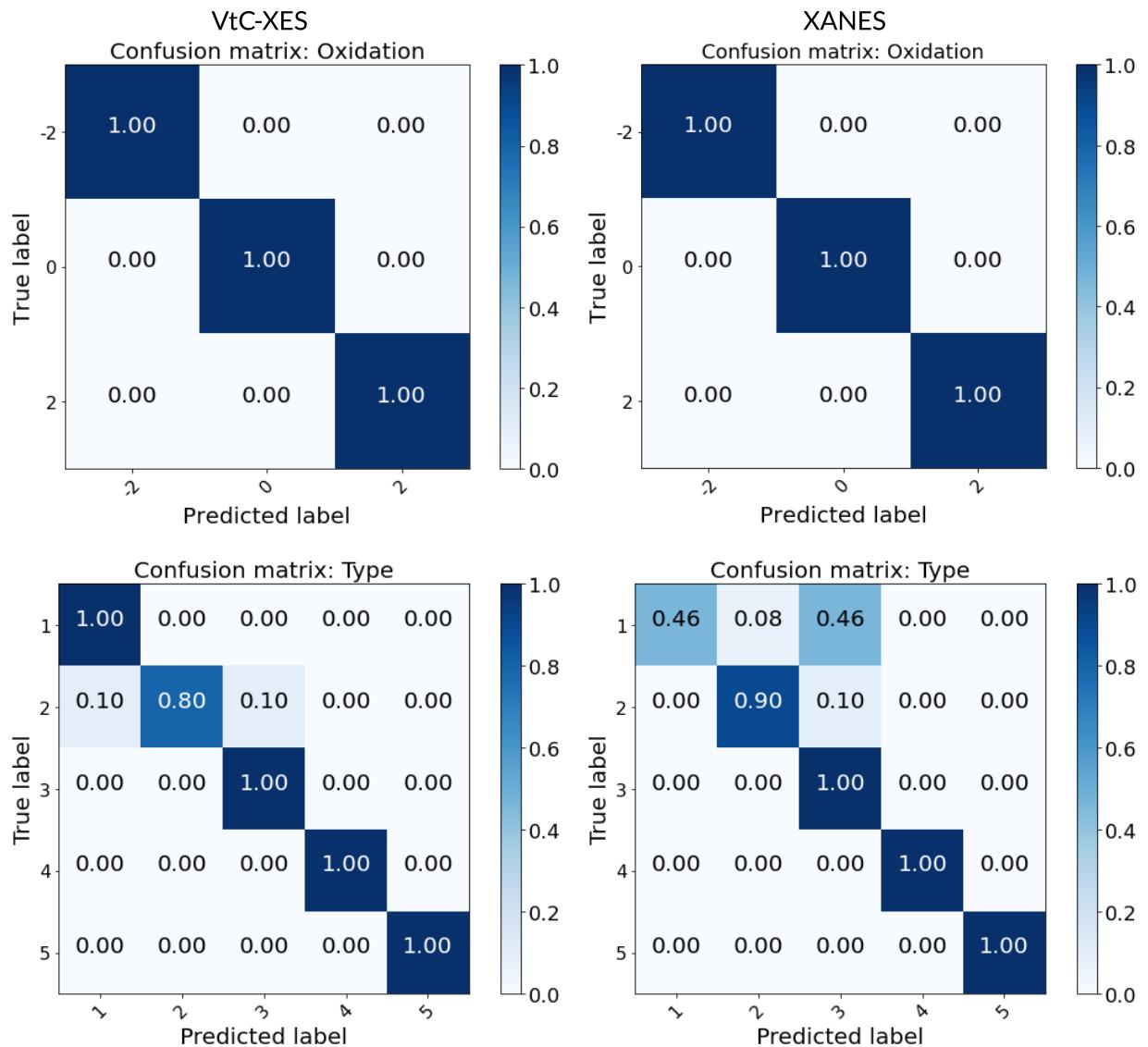
**Figure 6.S1** Loss plotted against number of epochs for the VAE model for both the XANES data (blue) and the VtC-XES data (green).

### Reconstructed VtC-XES spectra



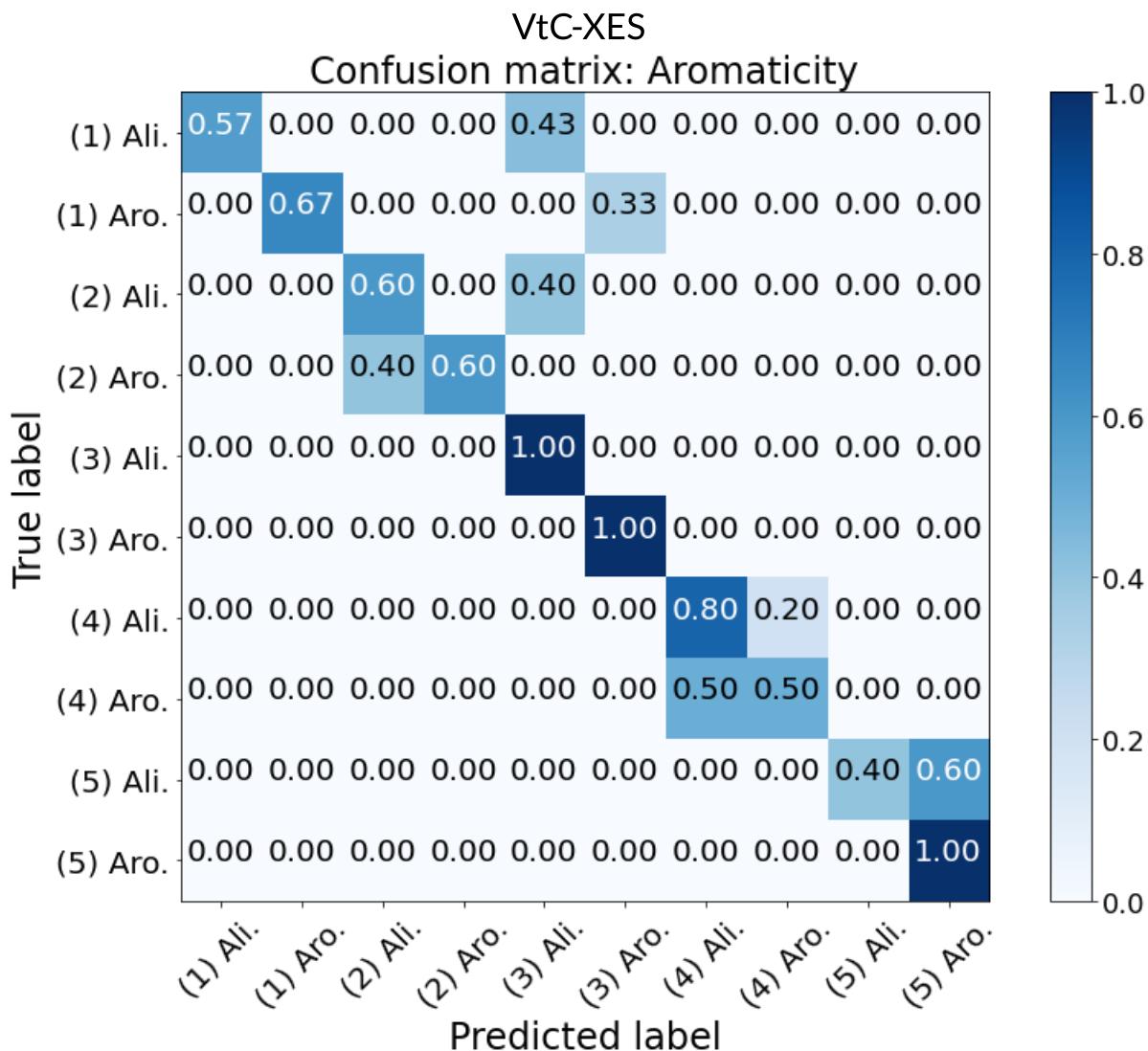
**Figure 6.S2** Reconstructed VtC-XES spectra with increasing latent space dimension.

### Schemes 1 and 2 confusion matrices



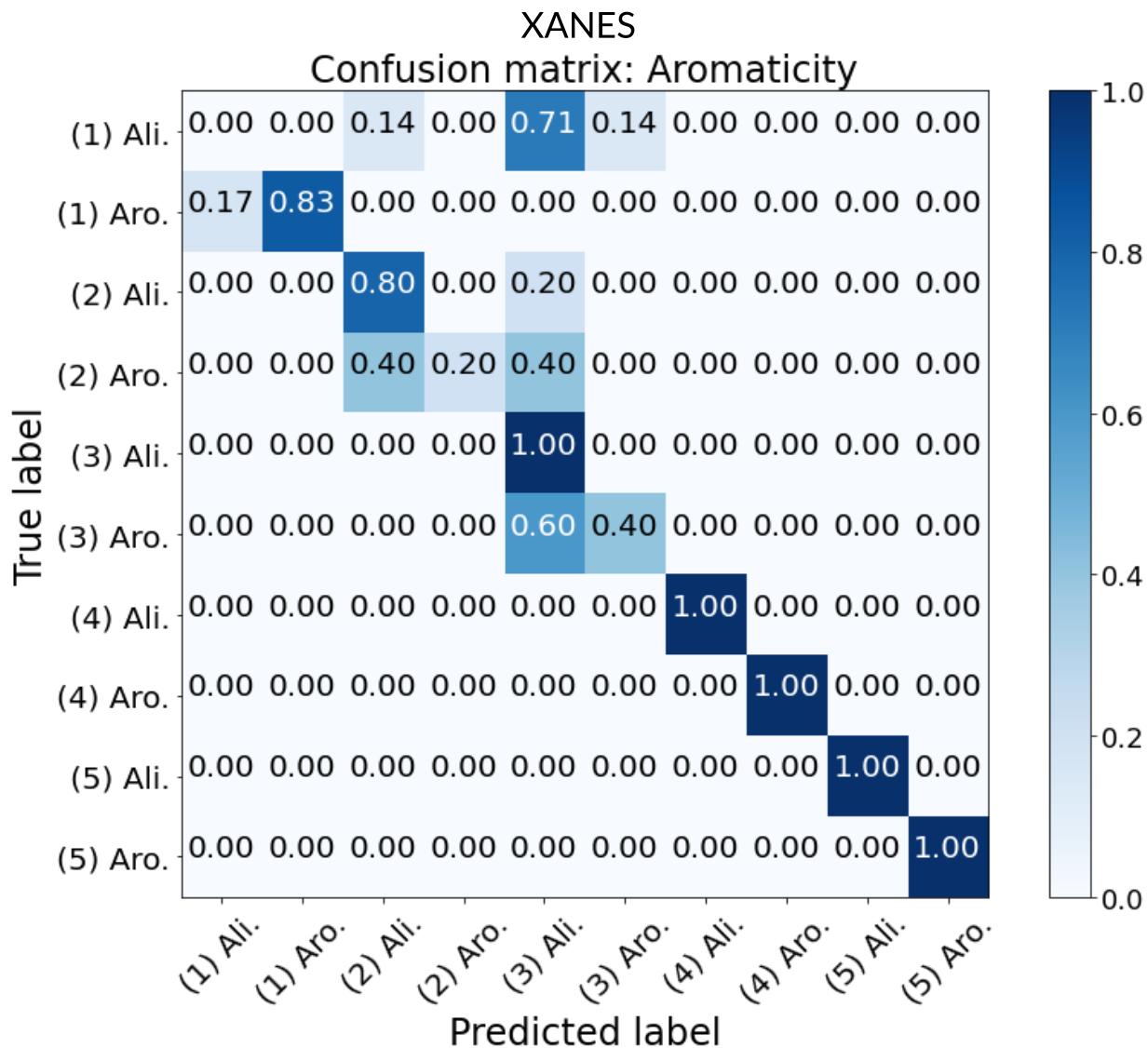
**Figure 6.S3** Classification via NN: confusion matrices for XES and XANES for both categorization schemes: 1) oxidation and 2) bond type.

Scheme 3 confusion matrix: VtC-XES



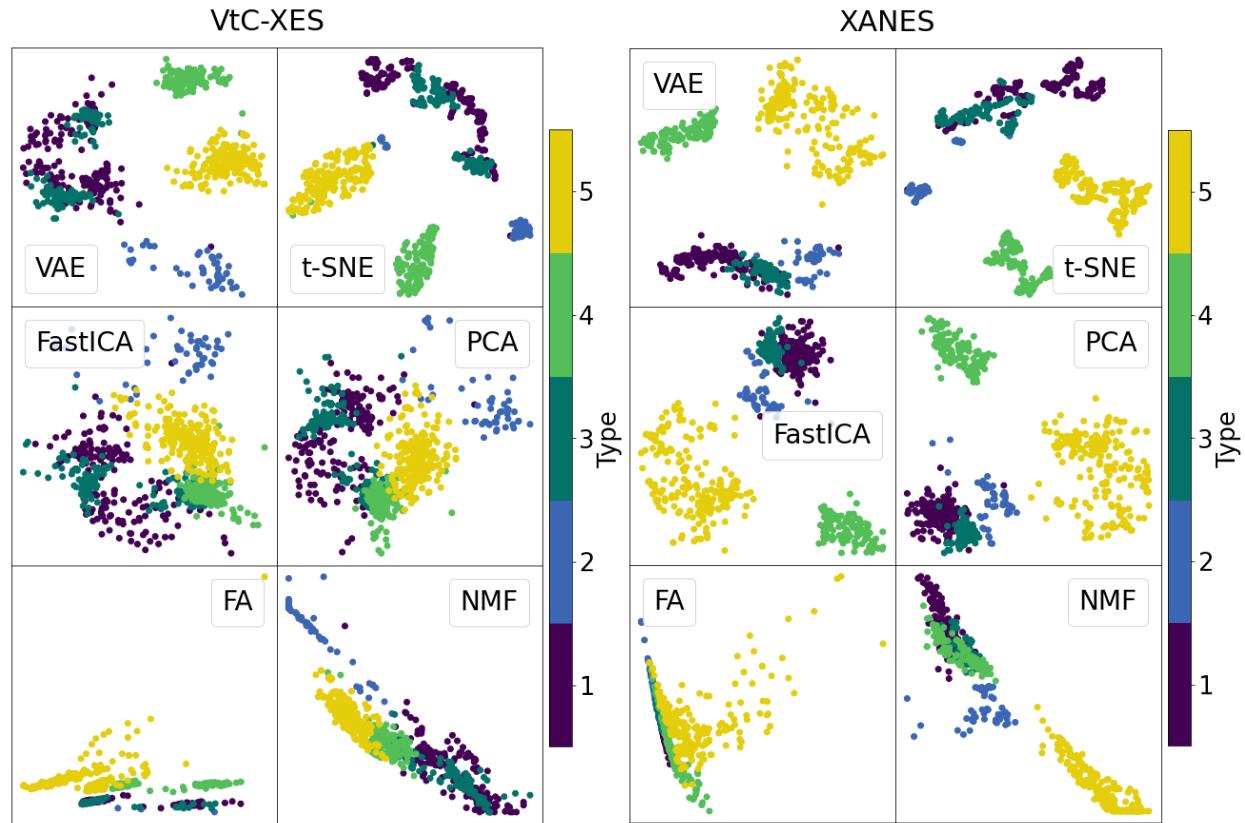
**Figure 6.S4** Classification via NN: confusion matrices for VtC-XES for classification of aromatic versus aliphatic compounds within Types 1 to 5.

Scheme 3 confusion matrix: XANES



**Figure 6.S5** Classification via NN: confusion matrices for XANES for classification of aromatic versus aliphatic compounds within Types 1 to 5.

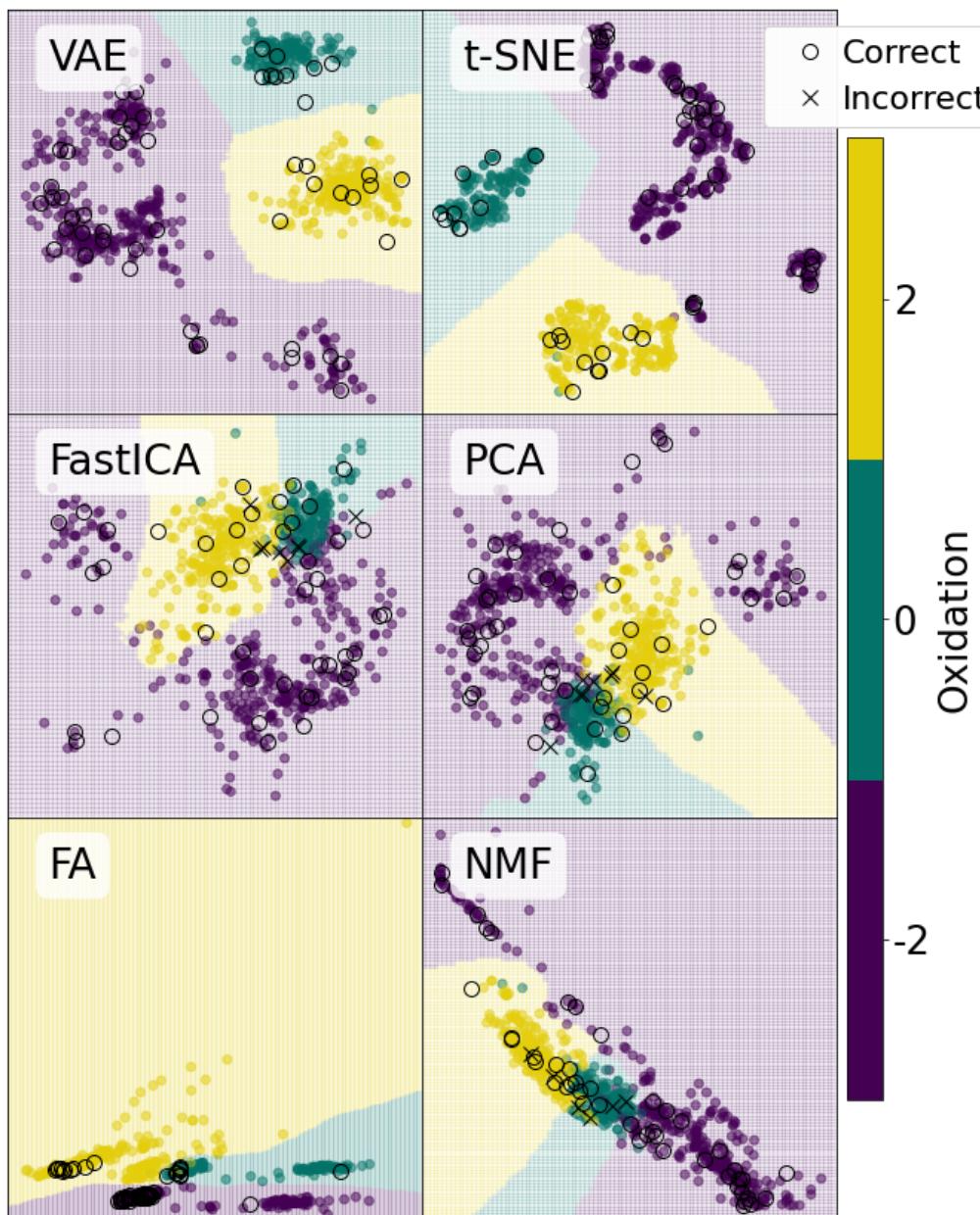
### Dimensionally reduced spaces



**Figure 6.S6** Unsupervised dimension reduction: VAE, t-SNE, FastICA, PCA, FA, and NMF for VtC-XES (left) and XANES (right), color-coded by sulfur bonding Type.

Oxidation KNN: VtC-XES

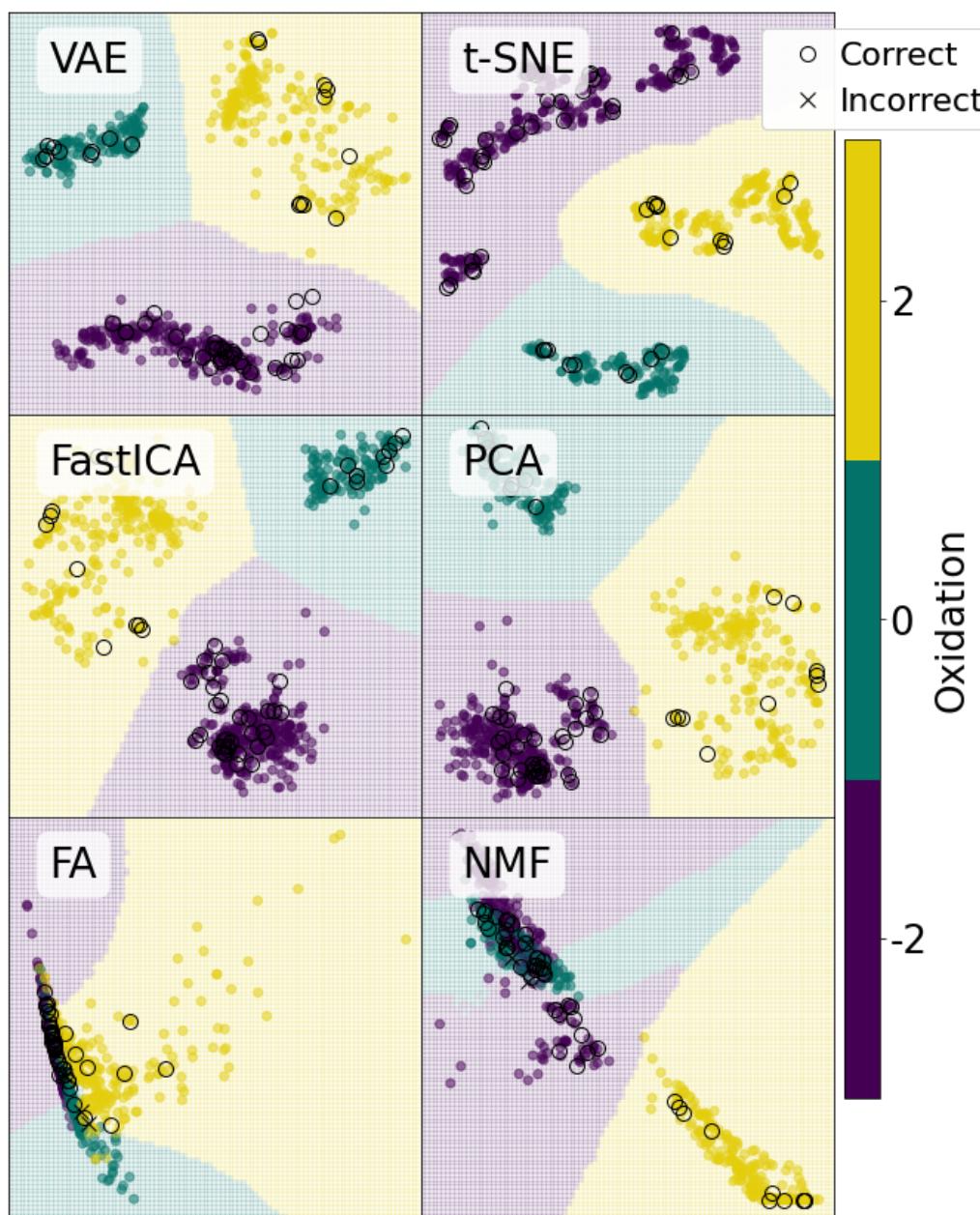
VtC-XES



**Figure 6.S7** KNN classification for Oxidation for VtC-XES.

Oxidation KNN: XANES

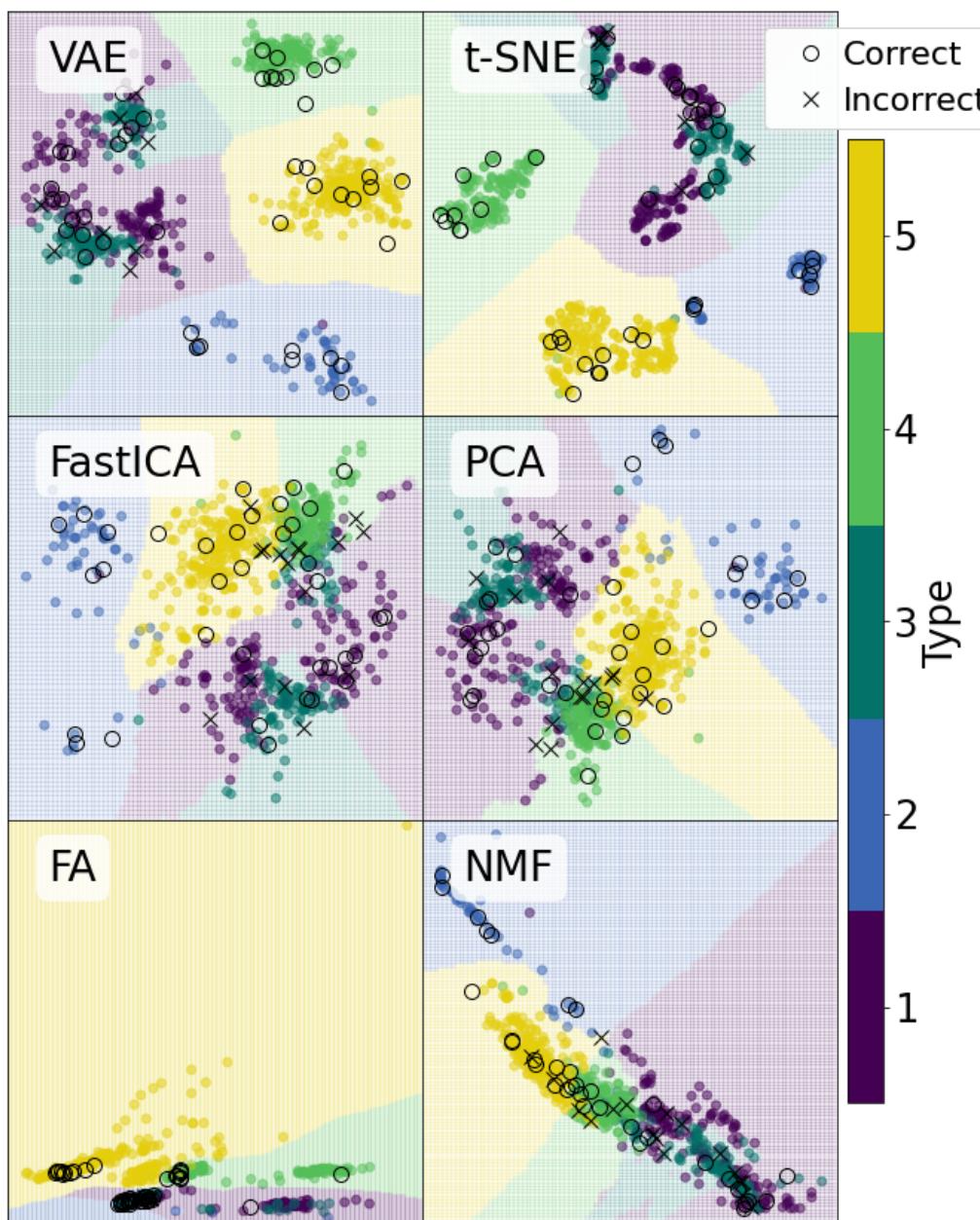
XANES



**Figure 6.S8** KNN classification for Oxidation for XANES.

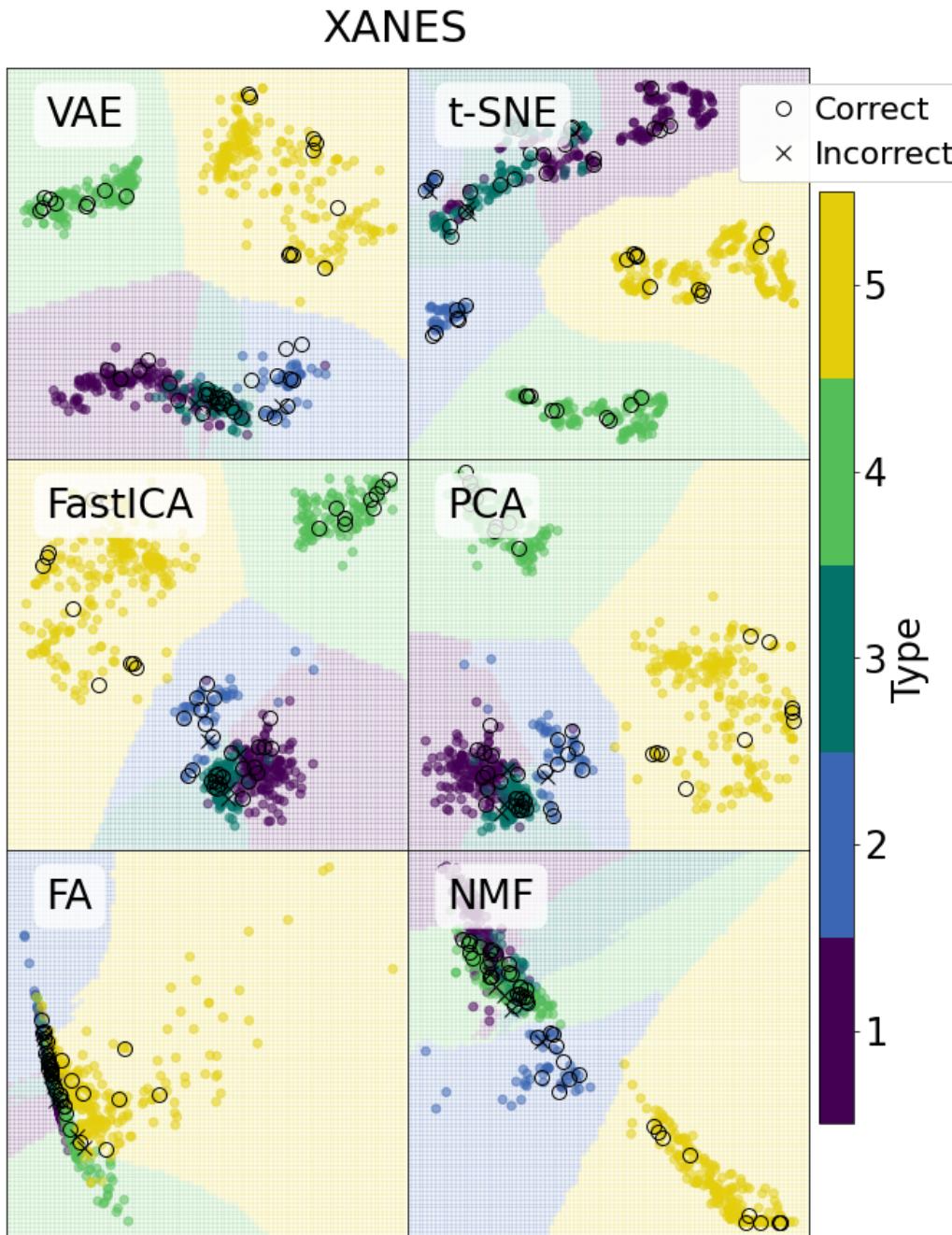
Sulfur Type KNN: VtC-XES

VtC-XES



**Figure 6.S9** KNN classification for sulfur bond Type on VAE, t-SNE, FastICA, PCA, FA, and NMF for VtC-XES.

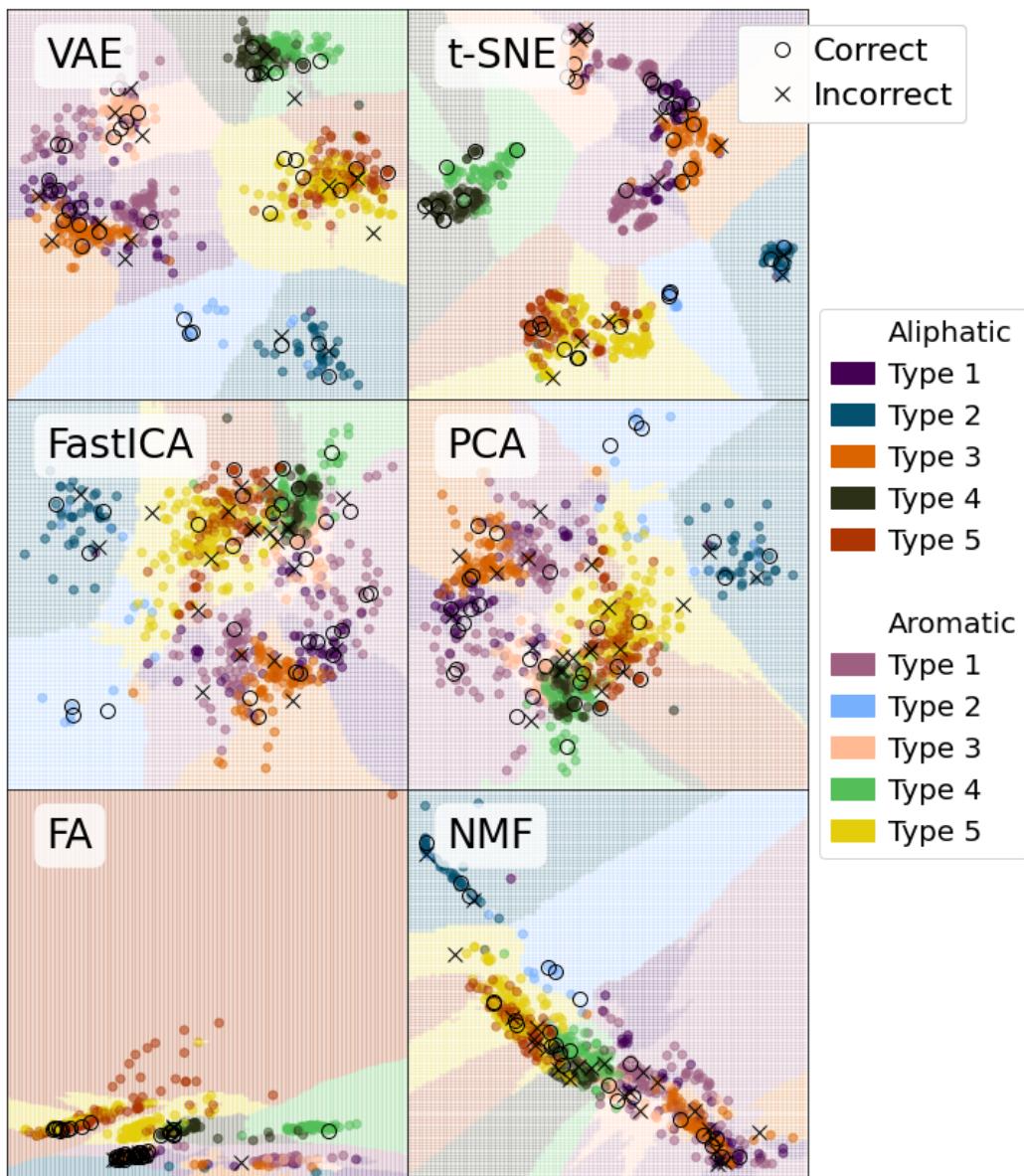
Sulfur Type KNN: XANES



**Figure 6.S10** KNN classification for sulfur bond Type on VAE, t-SNE, FastICA, PCA, FA, and NMF for XANES.

Aromaticity KNN: VtC-XES

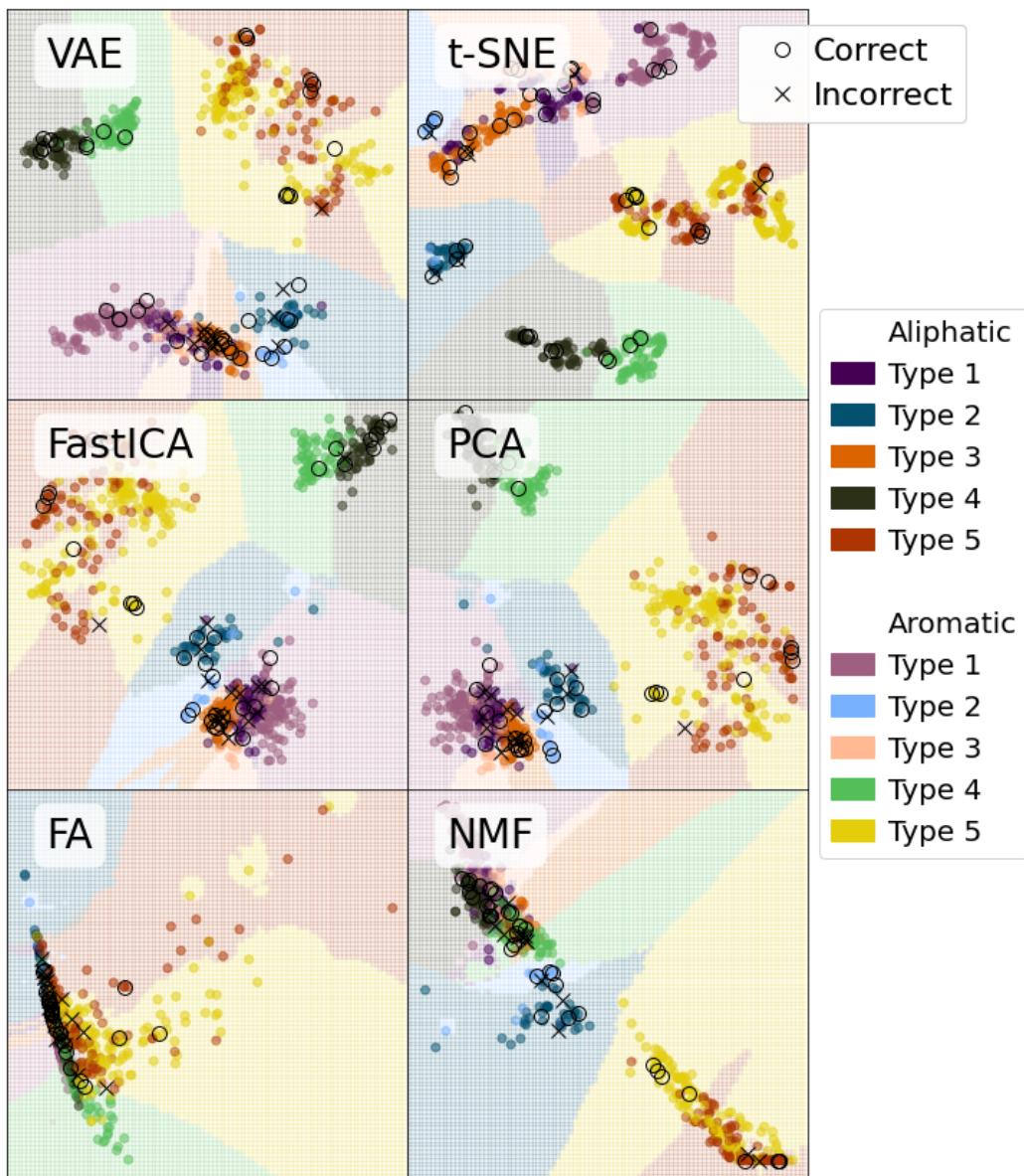
VtC-XES



**Figure 6.S11** KNN classification for Aromaticity for VtC-XES.

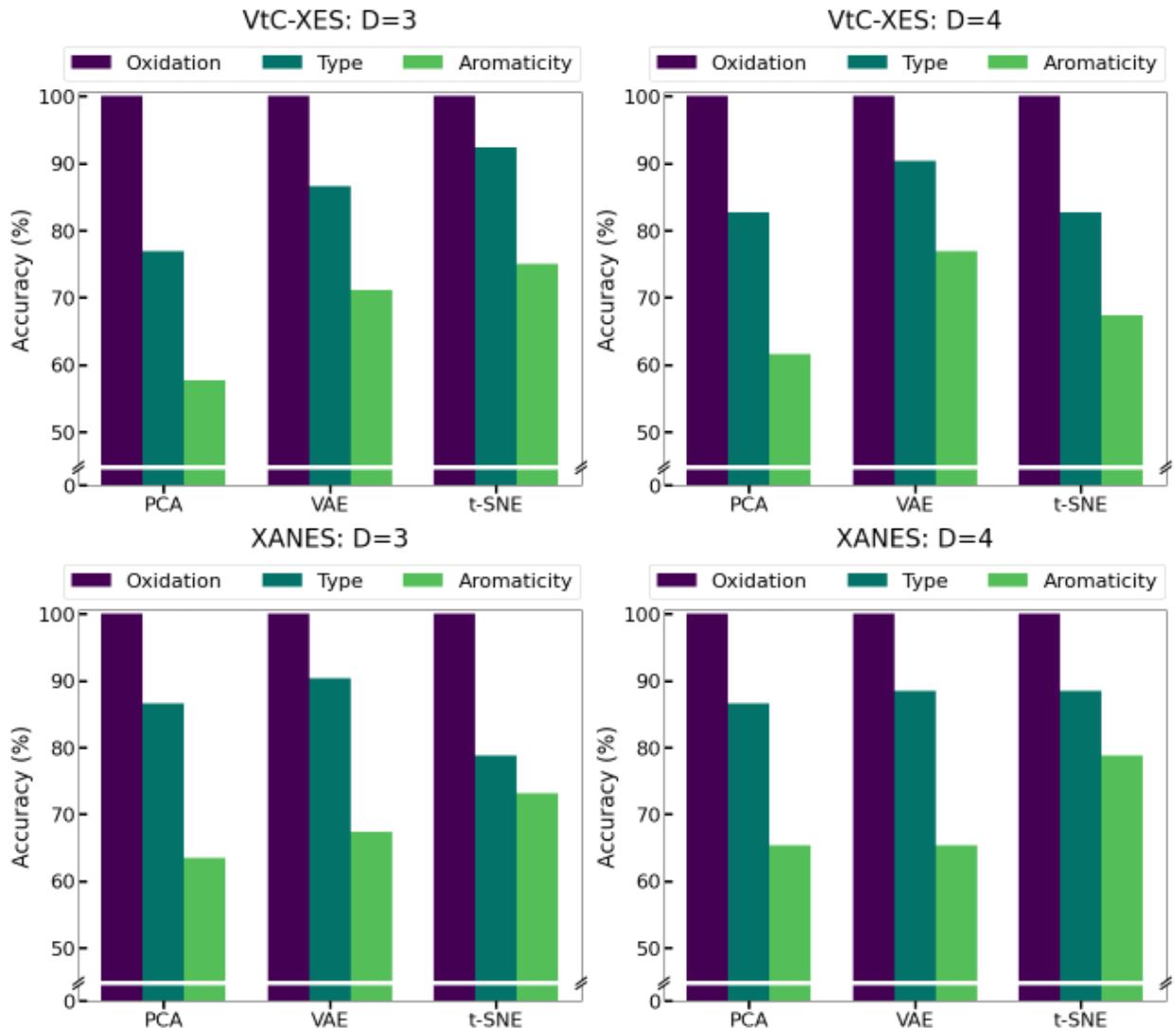
Aromaticity KNN: XANES

XANES



**Figure 6.S12** KNN classification for Aromaticity for XANES.

### Accuracy for increasing latent dimension



**Figure 6.S13** Accuracy of KNN classification schemes on the PCA, VAE, and t-SNE reduced spaces for VtC-XES (top) and XANES (bottom) while increasing the latent or embedding dimension, D.

### 6.7.5 References

1. A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo and S. Severini, *npj Quantum Information*, 2018, **4**, 28.
2. A. Hyvärinen and E. Oja, *Neural Networks*, 2000, **13**, 411-430.
3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
4. D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.
5. D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788-791.
6. S. C. Snook and R. L. Gorsuch, *Psychological Bulletin*, 1989, **106**, 148–154.

## 7 Chapter 7 – Clustering and Classification of Organophosphates

Originally published as: S. Tetef, V. Kashyap, W. M. Holden, A. Velian, N. Govind and G. T. Seidler. *The Journal of Physical Chemistry A* 2022 Vol. 126 Issue 29 Pages 4862-4872. S. Tetef and V. Kashyap contributed equally to this work. V. Kashyap calculated the majority of the spectra and handled data management and acquisition. S. Tetef wrote the text and conducted the machine learning analysis.

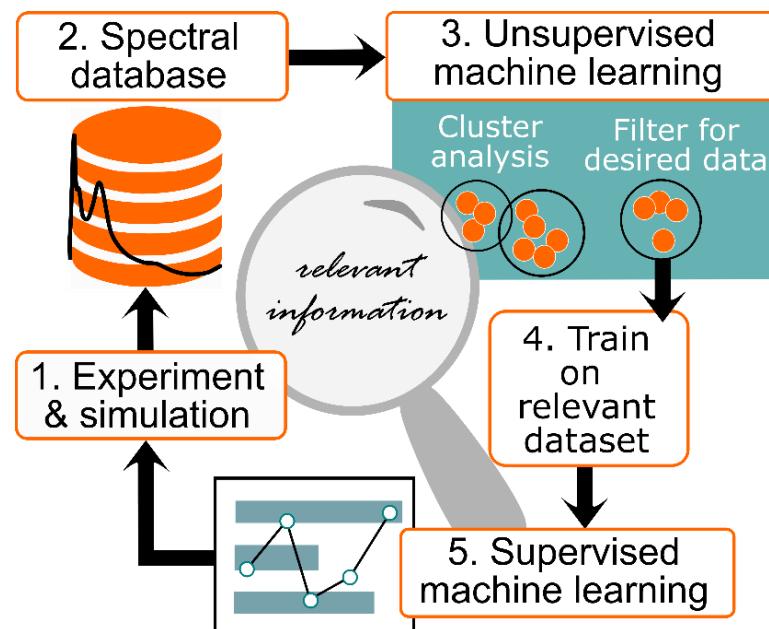
*We analyze an ensemble of organophosphorus compounds to form an unbiased characterization of the information encoded in their X-ray absorption near-edge structure (XANES) and valence-to-core X-ray emission spectra (VtC-XES). Data-driven emergence of chemical classes via unsupervised machine learning, specifically cluster analysis in the Uniform Manifold Approximation and Projection (UMAP) embedding, finds spectral sensitivity to coordination, oxidation, aromaticity, intramolecular hydrogen bonding, and ligand identity. Subsequently, we implement supervised machine learning via Gaussian process classifiers to identify confidence in predictions that match our initial qualitative assessments of clustering. The results further support the benefit of utilizing unsupervised machine learning as a precursor to supervised machine learning, which we term Unsupervised Validation of Classes (UVC), a result that goes beyond the present case of X-ray spectroscopies.*

## 7.1 Introduction

The information content in any spectroscopy method is constrained by the lossiness of the underlying quantum mechanics that connects atomic-scale structure and dynamics to experimental observables. Further limitations to the sensitivity of spectroscopy techniques often include the inherently nonlinear or stochastic responses of the experimental probe. These facts constrain our ability to correlate physical measurements, e.g., spectral features, to desired microscopic properties. Thus, the emergence of data science and machine learning (ML) in spectroscopy, with applications in all fields in the physical sciences, has exploded<sup>1-5</sup>. These data-driven models can frequently disentangle and infer patterns from inherently lossy observables as well as provide insight into the information encoded in spectra.

In general, supervised ML studies across a wide range of spectroscopies target either predicting properties from spectra or correlating specific properties of interest to spectral features<sup>6</sup>. This necessarily assumes that sufficient information is, in fact, encoded in spectra; otherwise, supervised ML models will correlate spurious features to requested properties. This detail of encoded information is often addressed by hand-selecting a targeted training domain, an approach that is deeply contingent on the accuracy and completeness of prior knowledge<sup>7</sup>. Clearly, issues will arise if the training domain is too small or is biased. First, if the training domain is too small, the model will be unable to generalize well beyond its specialized scope, which violates the essential assumption that the training and test data are sampled from the same distribution. Second, although some bias is essential for any machine learning model<sup>8</sup>, unwanted bias, especially from unrepresentative data, blindly undermines reliability of inferences and has led to contemporary ethical concerns<sup>9-12</sup>.

In the effort to combat unwanted bias as well as provide generalizability to complex datasets, this study demonstrates the value of the discovery cycle exemplified in Figure 1. This process validates encoded information via unsupervised machine learning, i.e., cluster analysis on a reduced-dimensional embedding of the spectra, before passing either the embedding or the original spectra – selected as an unbiased training (sub)set – to a supervised machine learning model. This approach decreases risk of implicit biases and spurious correlations introduced by supervised ML by adding steps (3) and (4) to validate spectral sensitivity of the training data set to properties requested during supervised predictions. The continuation of inferences from supervised ML back to experimental design and (primarily) data simulation is obviously informed by the resulting errors achieved by the supervised ML model.



**Figure 7.1** Flowchart of an analysis framework that uses unsupervised machine learning (such as cluster analysis) as a precursor to predictions on spectra via supervised machine learning, which can then inform experimental design and data creation.

This cycle touches on other related ways people have used unsupervised ML either as a precursor to or in a cycle with supervised ML. These approaches have included semi-supervised machine learning<sup>13</sup>, pre-training a neural network<sup>14</sup>, feature selection or generation<sup>15</sup>, and human-in-the-loop learning<sup>16</sup> which have been used in a multitude of fields, such as gene expression<sup>17</sup> and marketing<sup>18</sup>. Given the ubiquity of concerns about scope and bias when constructing training data sets in supervised ML, we propose that our approach, which we term Unsupervised Validation of Classes (UVC), has relevance beyond the present case of X-ray spectroscopies as well as contributes to efforts to close the loop between artificial intelligence and scientific understanding<sup>19</sup>.

Here, we apply the framework of Figure 1 to both X-ray absorption spectroscopy (XAS) and X-ray emission spectroscopy (XES). XAS has seen an explosion of ML applications<sup>20-41</sup>. XAS is most commonly used in chemistry, biology, and materials science to investigate the element-specific local coordination environment and electronic structure, with applications including energy storage<sup>42, 43</sup>, catalysis<sup>44</sup>, and photochemical dynamics<sup>45</sup>. XAS, which includes both X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS), probes the unoccupied electronic states of the excited state of a chosen atomic species.

Conversely, relaxation to fill the core hole results in either nonradiative (Auger) or radiative processes. The latter results in the emission of X-ray fluorescence that can be finely characterized by XES for insight into the occupied electronic states<sup>46-48</sup>. Often discussed as complementary to XANES in information content, valence-to-core XES (VtC-XES) is produced when electrons de-excite from the valence shell to fill the core hole, giving direct information about occupied electronic states involved in bonding<sup>49, 50</sup>. While XAS and XES have traditionally

been synchrotron-based methods, we note that their access, including for VtC-XES, is now being steadily augmented with a renaissance of lab-based spectrometers<sup>51-53</sup>, including in studies of sufficient scale for data science methods<sup>54</sup>.

In the first study to use supervised ML in XAS, Timoshenko, et al.<sup>20</sup> successfully inferred coordination numbers of Pt nanoclusters from XANES spectra using a neural network, a result that would otherwise require (human) expert analysis of EXAFS. Zheng, et al.<sup>24</sup> also predicted coordination, except using a random forest model. Notably, Torrisi, et al.<sup>36</sup> likewise used a random forest model to correlate polynomial fitting parameters of spectra to properties like bond distance. Other works utilizing both supervised and unsupervised machine learning in XAS include a XANES matching algorithm<sup>25</sup>, hierarchical clustering on spectra<sup>26</sup>, and use of an autoencoder to correlate coordination to a reduced dimensional representation of spectra<sup>27</sup>. Most of these studies assumed desired information was in fact encoded in spectra, largely because of hand-crafting relevant training datasets. However, our approach (Figure 1), via the unsupervised machine learning precursor, allows for explorative and unbiased refinement of chemical descriptors – a step that we propose is necessary, and likely sufficient, when addressing much more complex datasets.

The present study is prompted by our recent work<sup>55</sup> that compared the variance and information content of sulfur K-edge XANES to VtC-XES K $\beta$  spectra for sulfororganics. We found that nonlinear dimensionality reduction algorithms, a subset of unsupervised ML, provided an effective way to extract spectral features and thus important chemical information encoded in spectra. Moreover, our results exemplified the benefits of utilizing unsupervised ML to mold and understand the full potential of supervised ML analysis<sup>56</sup>.

Here, we investigate the information content and sensitivity of phosphorus K-edge XANES and VtC-XES K $\beta$  in a more complex chemical system, organophosphorus compounds, and indeed

find sensitivity to a wider range of chemical properties, including coordination, oxidation, aromaticity, intramolecular hydrogen bonding, and ligand identity. The proximity of phosphorus to sulfur in the periodic table allows for the same theoretical parameters to generate spectra (and thus obtain similar experimental agreement) as our previous study, but also leverages the more diverse bonding environment of phosphorus. The dataset of spectra is calculated from molecular structures gathered from the PubChem<sup>57</sup> database using moldl, a new open-source tool that we have developed for this purpose<sup>58</sup>. For the rest of this paper, we will refer to the phosphorus K-edge XANES and VtC-XES K $\beta$  as just XANES and VtC-XES, respectively, for brevity.

Organophosphorus compounds have much higher total variance than sulforganics, as well as higher variance within the same bonding geometry. We can therefore tune the input domain to account for these highly variant structures, allowing us to understand the sensitivity of these spectra to a wider range of properties. In addition, we can find, in an unbiased way, the extent of the chemically relevant information that may be extracted using dimensionality reduction algorithms, especially when confined to very limited dimensions. These explorations allow for full utilization of real spectral information during supervised ML predictions.

To this end, we use the Uniform Manifold Approximation and Projection (UMAP)<sup>59</sup> for dimensionality reduction, which allows us to develop chemical classes by examining clustering of spectra in a two-dimensional embedding. UMAP is a nonlinear embedding similar to t-distributed Stochastic Neighbor Embedding (t-SNE)<sup>60</sup>, which was used in our recent work<sup>55</sup> to extract chemical classes. Like t-SNE, UMAP constructs a graph-based representation of the data in the high-dimensional space to generate a similarity comparison, and then it tries to match that similarity comparison in a low-dimensional representation of the data. However, UMAP utilizes a

different cost function, namely cross entropy instead of KL-divergence, which further enables global structure to be preserved, albeit at the cost of the “crowding problem”<sup>60</sup>.

Moreover, given the proper choice in hyperparameters, UMAP can retain global similarity such that distances between clusters can be interpreted (given the manifold remains connected). This contrasts t-SNE, where its cost function, the KL-divergence, goes to zero at large distances. The result is that t-SNE is not penalized for putting unlike data either far or *very* far away, and thus interpretation of similarity is only valid on a relatively local (intra-cluster) scale. These properties of UMAP allow it to generate a mapping function that can then be used to map subsequent data, which is why UMAP is called a “parametric embedding” and contrasts t-SNE’s requirement that the entire training dataset must be used to predict new data. Thus, UMAP can be used for future data compression and has the potential for better interpretation of overall global similarities. These advantages have led to its recent popularity, such as in single cell RNA sequencing (scRNA-seq) data analysis<sup>61</sup>, but UMAP has not yet seen use in XAS analysis.

## 7.2 Methods

Our methods for the electronic structure calculations closely follow that of Tetef et al.<sup>55</sup>. Molecular structures were downloaded from the PubChem database using our open-source Python module called moldl<sup>58</sup> that allows for users to easily write scripts that can search the PubChem database and store the resulting structures, with metadata, in a local database indexed by PubChem Compound IDentification (CID) numbers. The downloaded structures can then be sorted using customizable filters, and selected molecules can be exported in multiple formats (SDF, MOL, and XYZ). This tool is accessible to any researcher for use in projects that require the collection and management of molecular structure datasets. A total of 1196 compounds were downloaded and managed in this study, while 756 of them were structurally viable for our desired analysis.

Both the XANES and VtC-XES spectra were calculated with the open-source NWChem computational chemistry software package<sup>62, 63</sup> via the same pipeline as specified in Tetef et al<sup>55</sup>. To summarize, both spectra were computed using the Sapporo QZP-2012 basis set<sup>64</sup> for P, while the remaining atoms were represented using the 6-31G\* basis set, and the PBE0 exchange correlation functional<sup>65</sup>. Additionally, the Stuttgart RLC ECP<sup>66</sup> was substituted for atoms heavier than phosphorus. As in Tetef et al.<sup>55</sup>, a post-processing energy-dependent linear broadening scheme was applied to the XANES transitions, starting with a full-width half-maximum (FWHM) Lorentz broadening of 0.5 eV at the whiteline, and then linearly increasing to 4.0 eV FWHM at 20 eV past the whiteline. An energy shift of 50 eV was applied to all XANES transitions to align with experimental data<sup>67</sup>.

For the VtC-XES, the calculated transitions were all shifted by -19 eV to align to experiment<sup>68</sup>. A FWHM Lorentz broadening of 0.5 eV and a FWHM Gaussian broadening of 1.5 eV was added to each transition to agree with experimental data<sup>68</sup>. Because NWChem calculates a self-consistent field density functional theory (DFT) solution for both the XANES<sup>69</sup> and VtC-XES<sup>70</sup>, this solution serves as a reference for the time-dependent DFT (TDDFT)-based X-ray spectroscopy calculations and thus only one internally consistent energy shift is required for each system. Finally, both the XANES and VtC-XES spectra were individually normalized by their total K $\alpha$  intensities. The K $\alpha$  transitions scale in intensity proportional to the compound size (like the VtC-XES and XANES calculations) but are very nearly independent of all environment effects, thus providing an absolute scale to maintain relative intensities across the entire ensemble.

The sulforganics study of Holden, et al.<sup>54</sup> for the experimental VtC-XES and NWChem calculation showed excellent agreement, as did additional calculations and comparison to XANES in Tetef, et al.<sup>55</sup> Here, in Figures S1 and S2, we more modestly validate the performance of

NWChem against several VtC-XES spectra taken with the same instrument and methodology as Holden, et al.<sup>54</sup>, and also validate performance against several XANES spectra from Persson, et al.<sup>67</sup>.

Briefly, PCA was implemented using the scikit-learn<sup>71</sup> package in Python and was applied to the original spectra before UMAP to speed up computation and decrease noise. The number of principal components kept from the PCA was the number of components necessary to explain at least 95% of the variance in the dataset. For example, the number of retained principal components was 7 and 14 components for the VtC-XES and XANES spectra respectively for the dataset consisting of all tricoordinate and tetracoordinate compounds, as shown in Figure S3. Some reconstructed spectra using the 95% variance cutoff are shown in Figures S4 and S5.

The difference in the number of principal components required for the VtC-XES and XANES suggests that the XANES spectra have more variation and thus more nonlinear features, which is unsurprising. UMAP was implemented using the umap-learn module<sup>72</sup> with default hyperparameters. Again, as mentioned above, to accelerate computing the UMAP algorithm was applied to the PCA coefficients at the 95% variance level, thus decreasing the dimensionality of the training space from 1000 to either 7 or 14. For Figures 2 to 6, the number of UMAP output components was constrained to two for visualization purposes, while for Figure 7, the output dimensionality was set to five (found through the hyperparameter optimization discussed below).

Finally, to help illustrate the value of unsupervised ML as a precursor to supervised ML, we applied supervised ML in the form of a Gaussian Process<sup>73</sup> Classifier to the UMAP representation for all five classification schemes determined by the two-dimensional cluster analysis. The Gaussian Process was implemented using scikit-learn<sup>71</sup>, which utilizes the Laplace

approximation as detailed in Rasmussen and Williams<sup>73</sup>. A separate classifier was trained for each of the five classification schemes, shown in Table S1, for both the VtC-XES and XANES data.

A test set was specified for each classifier, which comprised of a random selection of 15% within each class, with the rest of the data specified as training. A validation set was then randomly selected within that training set to optimize model hyperparameters. These hyperparameters were found to be five dimensions for the UMAP embedding, with the optimal kernels for the Gaussian Process selected as Rational Quadratic for both the VtC-XES spectra and XANES spectra. All data and analysis code for this study is publicly available<sup>74</sup>.

### 7.3 Results and Discussion

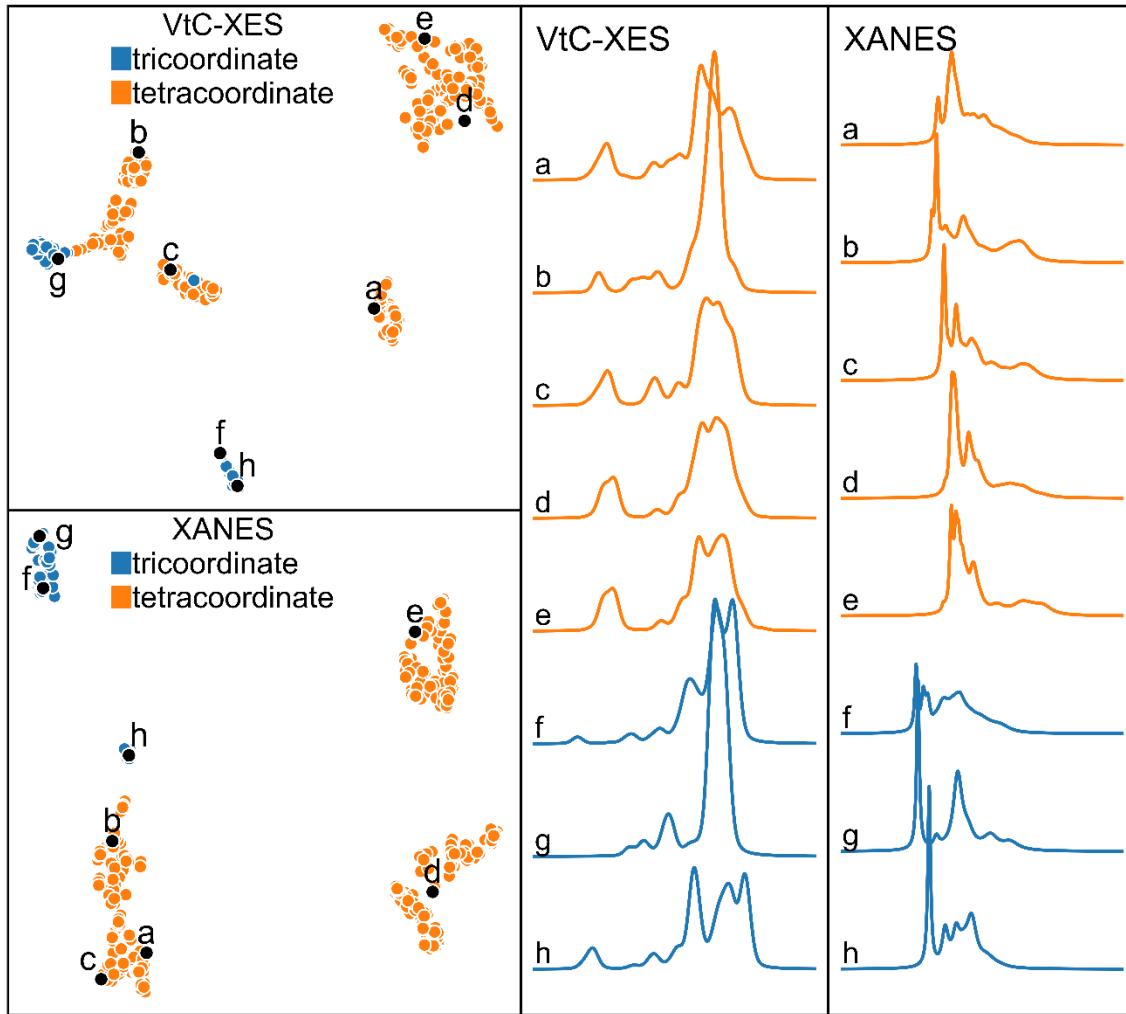
The first two sections below follow the general approach in Tetef, et al.<sup>55</sup>, wherein we investigate the heuristically expected chemical sensitivities in VtC-XES and XANES (IIA) and then, when subclusters are observed within an expectedly dominant chemical classification, we investigate unexpected sensitivities to further structure or electronic refinements (IIB). This has several important results, including delineation of both similar and different sensitivities of VtC-XES and XANES to chemical classifications, as well as the emergence of spectral sensitivity to second-shell coordination for phosphates.

The final section (IIIC), on the other hand, seeks to address the motivating hypothesis illustrated in Figure 1, i.e., that unsupervised ML can usefully inform supervised ML. We demonstrate that confidence of predictions directly correlates to our qualitative cluster analysis, thus validating that the strength of information encoded in VtC-XES and XANES can vary between spectroscopies depending on the system and property of interest.

### 7.3.1 Unbiased verification of heuristic classes

To begin, heuristically one expects the phosphorus coordination to yield the strongest distinguishing features between spectra, specifically the distinction between tricoordinate phosphorus and tetracoordinate phosphorus. Not only do these coordination geometries have different hybridized orbital character, but they are often a proxy for oxidation state. In organophosphorus compounds with tricoordinate phosphorus centers, the phosphorus is typically in a 3+ oxidation state, whereas compounds with tetracoordinate phosphorus centers usually have the phosphorus in a 5+ oxidation state.

We chose compounds with a diverse number of oxygens bonded to phosphorus within these two coordination configurations (with all other bonding atoms as carbon) to further vary the effective charge on the phosphorus. We then applied UMAP to the VtC-XES and XANES spectra to create a two-dimensional embedding of the ensemble. The results are color-coded based on whether the compound includes tricoordinate phosphorus or tetracoordinate phosphorus, as shown in Figure 2. All R groups bonded to the phosphorus (or bonded to the oxygens bonded to the phosphorus) are constrained to be exclusively carbons (e.g., alkyl or aryl chains), and sometimes hydrogens (when bound to the oxygen) to achieve hydroxyl groups, but only for phosphates (which we will explore later).



**Figure 7.2** UMAP representation of VtC-XES (top) and XANES (bottom), color-coded by coordination, with some example spectra (as calculated by NWChem) shown to the right.

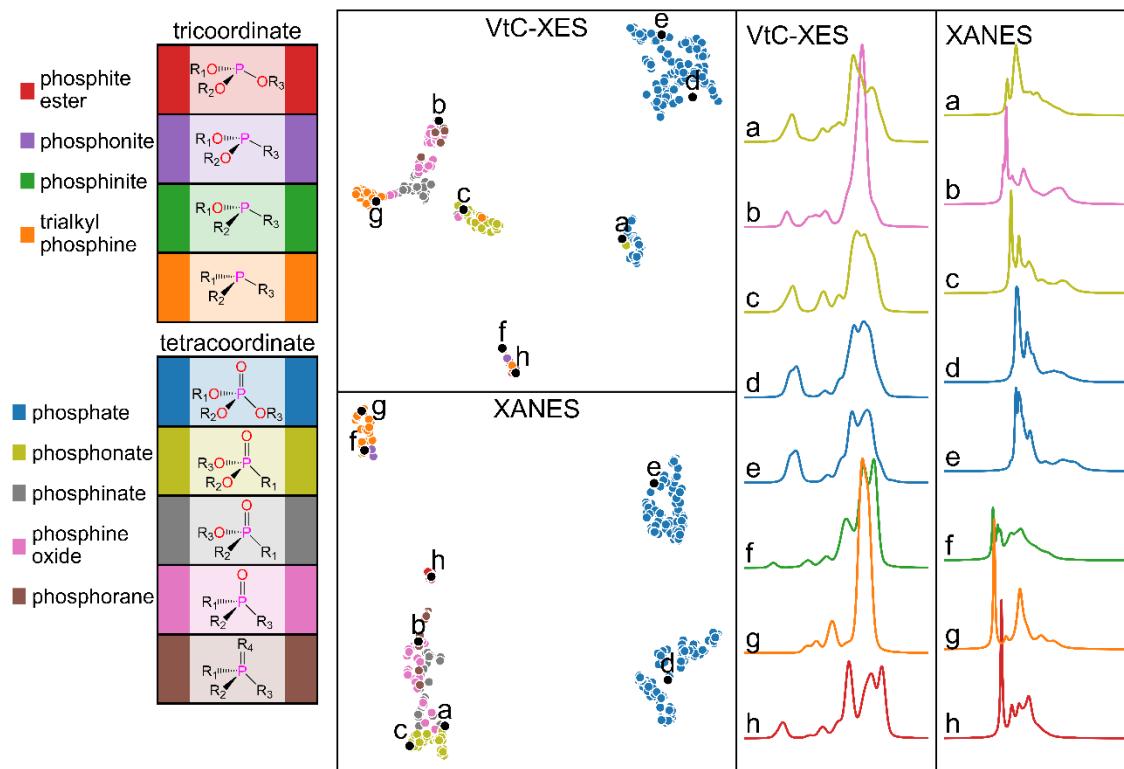
As expected, coordination distinguishes most of the groupings of the compounds, with a handful of outliers. We have further labelled some example compounds **a** to **h** (right panels) in each cluster with their corresponding VtC-XES and XANES spectra. (The identity of compounds **a** to **h** are defined in Table S2 in the supplementary section, but it is sufficient to say that they span a wide range of local coordination and oxidation states.) Note how some compounds which are in

the same cluster in the VtC-XES embedding are in different clusters in the XANES embedding, and vice versa. For example, compounds **a**, **b**, and **c** are together in the XANES embedding, but they are in three different clusters in the VtC-XES embedding, which we will discuss later as being due to the number of oxygens ligands. These observations clearly indicate that VtC-XES and XANES encode information differently, and that there are chemically relevant sub-groupings within each coordination.

Seeking to elucidate the chemical sub-grouping, Figure 3 shows the embedding color-coded within each of the tri- and tetra-coordinate classes based on the number of oxygens bonded to the phosphorus and the corresponding named chemical classifications. The spectral averages for both the VtC-XES and XANES spectra for each *class* are shown in Figure S6, while the spectral averages for each *cluster* are shown in Figure S7. Figure 3 shows very clear retention of chemically relevant information, with some similarities and differences between the VtC-XES and XANES. We will now discuss the expected change in spectra based on the chemical signatures in this ensemble, the resulting successes in information encoding, the differences between the two spectroscopies, and (importantly) the occurrence of outliers in the UMAP embedding; specifically, if the outliers correspond to molecules whose electronic structure is somehow strongly anomalous with respect to their general chemical class.

First, we expect effective charge of the phosphorus to have the biggest impact on both the VtC-XES and XANES spectra. For the VtC-XES, the ligand peaks (the small low-energy peak in Figure S6) will increase in both energy and intensity with an increase in phosphorus oxidation. From a molecular orbital perspective, this trend is from both a larger overlap between the ligand valence orbital and the phosphorus 3p orbital (valence shell) and the increased number of oxygen ligands. In general, this feature (which also changes with different ligand symmetries and

orientation) is why VtC-XES is strongly sensitive to ligand identity<sup>75</sup>. For the XANES spectra, an increase in the oxidation of the phosphorus, i.e., the number of oxygen ligands within a coordination, will cause a blueshift of the absorption edge, also demonstrated again by the average spectra in Figure S6.



**Figure 7.3** UMAP representation of VtC-XES (top) and XANES (bottom) for tricoordinate phosphorus and tetracoordinate phosphorus compounds, color-coded by number of oxygens bonded to the phosphorus within each coordination. The same example spectra as before are shown to the right, as calculated by NWChem.

Second, in terms of successful information encoding, we see that the number of oxygen ligands supplies much more information to explain the groupings in the UMAP representation than just coordination. For example, the highest oxidation compounds – the phosphates (blue) – are

separated from all other compounds in both the VtC-XES and XANES embeddings and are even sub-divided into two clusters for both (this is due to a combination of chemical properties which we will explore later in section IIIB and is the reason compounds **e** and **d** are separated in the XANES embedding but not the VtC-XES).

Third, we consider the similarities and differences of information encoding by XANES and VtC-XES in Figure 3. In terms of differences, the VtC-XES segregated the ytetracoordinate phosphonates (yellow) from other compounds, whereas the XANES segregated the tricoordinate trialkyl phosphines (orange) from the rest of the ensemble. Additionally, the VtC-XES separated the phosphine oxides (pink) into two subclusters, not seen in the XANES embedding, while the tricoordinate phosphite esters (red) get their own cluster in the XANES but not in the VtC-XES embedding.

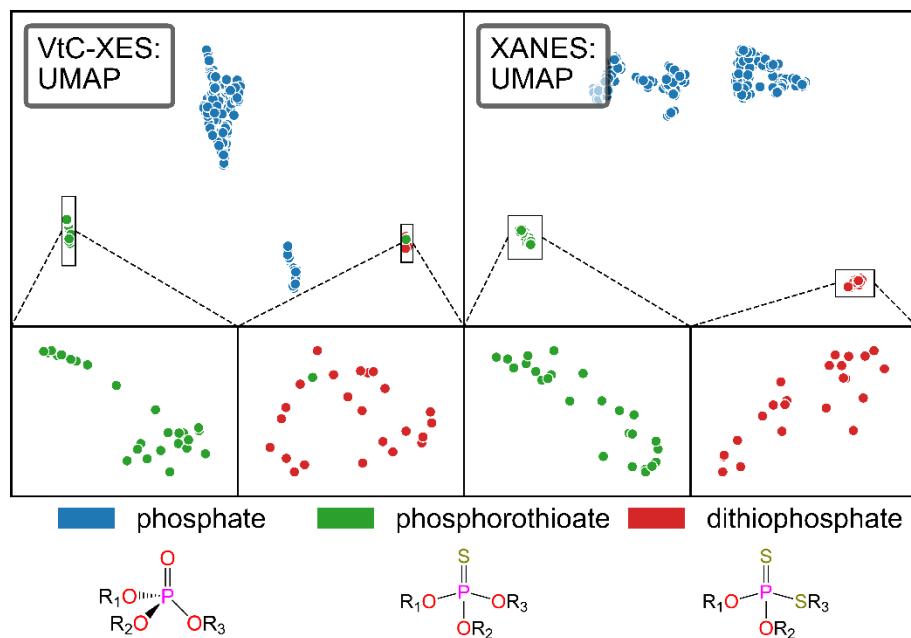
A closer look at these differences in the UMAP embeddings for VtC-XES and XANES is exemplified by the example compounds **a**, **b**, and **c**. In this case, although compound **b** (tetracoordinate, phosphine oxide, one oxygen ligand) has a more reduced P atom compared to compounds **a** and **c** (both tetracoordinate, phosphates, four oxygen ligands), it is in a different cluster in the VtC-XES embedding but in the same cluster, albeit at the opposite end, as compounds **a** and **c** in the XANES embedding. We see that the VtC-XES spectrum for compound **b** is in fact vastly different than the spectrums of **a** and **c**, but its XANES counterpart is more similar to the others. This difference in grouping is likely indicative of the variation within the two spectroscopies. Because UMAP compares both local and global similarities between spectra, this trend might indicate that VtC-XES spectra have more discrete spectral features (especially regarding charge on the phosphorus) compared to a continuous variation in XANES spectral features (for example, a continuous shift in the absorption edge).

Finally, moving to apparent outliers, one clear example is the location of compound **a**, diethyl (chloromethyl) phosphonate in the VtC-XES embedding. Compound **a** is a phosphonate but has a chlorinated carbon ligand, which effectively pulls more charge from the phosphorus, thus making the carbon act more like an oxygen and the phosphorus have a higher oxidation. Likewise, both phosphonates, like compound **a**, in the phosphate cluster have a chlorinated R<sub>1</sub> ligand are thus grouped with the nominally “higher oxidation” compounds instead of the cluster with compound **c** (diethyl methanephosphonate).

As for further outliers, note that although compound **f** is a phosphinite with nominal P(III) oxidation from its tricoordinate P, it has a distinct number of oxygen ligands (one) compared to **g** (trialkyl phosphine, tricoordinate, no O ligands) and **h** (phosphite ester, tricoordinate, three O ligands). In these UMAP embeddings, compound **f** is more similar to the higher oxidation compounds in the VtC-XES compared to the XANES spectra. Upon further examination, the other trialkyl phosphines in the cluster with compound **f** in the VtC-XES embedding are anomalous – they all have nitrile functional groups bonded to the phosphorus atom. Thus, in this case, the VtC-XES seems to determine outliers more definitively than XANES, where the distinction falls on the second nearest neighbor identity.

These observations bring us to our next hypothesis that VtC-XES and XANES are both sensitive to ligand identity. As stated earlier, VtC-XES is highly sensitive to ligand identity via changes in the ligand peak feature <sup>47</sup>. Again, because the absorption edge of a XANES spectrum shifts with oxidation, the electronegativity of ligands will cause the biggest spectral change. However, even for ligands with approximately the same electronegativity, different phase shifts and cross sections cause finer changes to the XANES spectra.

To systematically probe the effect of ligand identity, a series of tetracoordinate phosphorus compounds (phosphates) were evaluated in which the oxygen substituents were replaced with one or two sulfur atoms with the local bonding environment around the phosphorus otherwise unchanged. Compared to oxygen, sulfur is significantly less electronegative, with a Pauling electronegativity value near that of carbon and phosphorus<sup>76</sup>. Thus, while differences in photoelectron scattering can influence the XANES, we generally expect that these oxygen-to-sulfur ligand substitutions cause the biggest spectral change by adjusting the effective charge on the phosphorous. The resulting clusters are shown in Figure 4. Note that the phosphates are the same compounds that were used in the ensemble appearing in Figures 2 and 3, but that we have added additional chemical classes – phosphorothioates and dithiophosphates – to create the ensemble appearing in Figure 4.

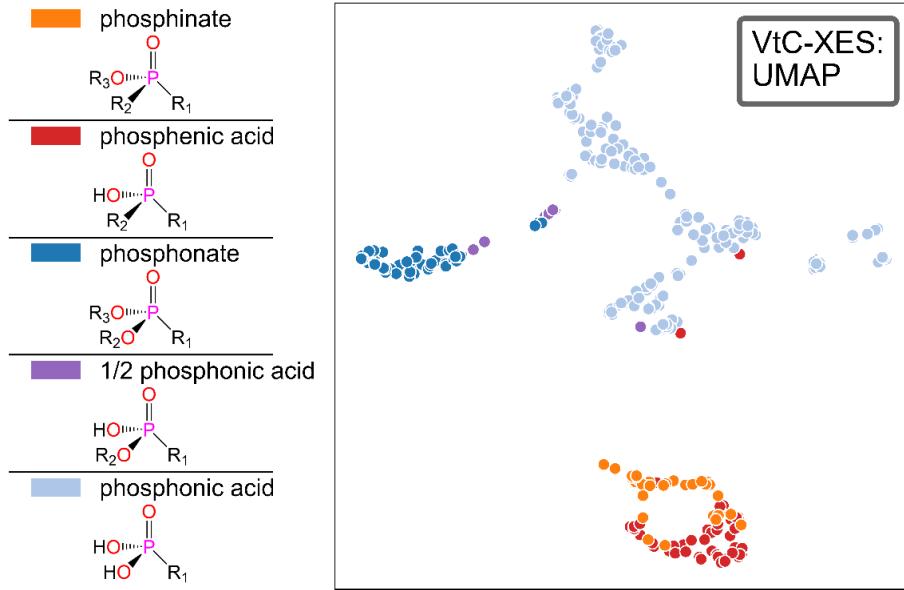


**Figure 7.4** UMAP representation of VtC-XES (left) and XANES (right) for compounds with sulfur ligands, color-coded by number of sulfurs. The pair of bottom insets on each panel are

enlargements of the shown sub-regions to make it easier to see violations of cluster chemical classes, i.e., outlier compounds.

The different ligand identities drive cluster separations in Figure 4, but do not exhaust the refinement of chemical classification – we return below to the question of further classification within phosphates. However, in Figure 4, the VtC-XES has a clear outlier – the phosphorothioate (green) in the dithiophosphate cluster (red) in the second inset of that figure. Chemically, that compound (PubChem CID 104781, tert-Butylbicycolphosphorothionate) is structurally different from others because the oxygens form one edge of a carbon tetrahedrane. Thus, a clear chemical outlier, in terms of electronic structure, is also flagged as an outlier in the UMAP embedding because it grouped this compound with dithiophosphates instead of with phosphorothioates.

We next analyze whether the spectra would be sensitive to substitutions of R groups (if bonded to an oxygen) with a hydrogen atom, thus forming hydroxyl groups, as shown in Figure 5. Here, we have taken phosphinate and phosphonate as starting points, and consecutively replaced O-R groups with OH groups. Note that the phosphinates and phosphonates are the same compounds that were used in the ensemble appearing in Figures 2 and 3, but that we have added additional chemical classes – phosphenic acids, half phosphonic acids, and phosphonic acids – to create the ensemble appearing in Figure 5.



**Figure 7.5** UMAP representation of the VtC-XES of compounds with consecutively more R groups (if bonded to an oxygen) replaced with an H atom (to create hydroxyl groups), color-coded by chemical class.

In general, this distinction seems to be better illuminated by the VtC-XES spectra than the XANES (which is shown in Figure S8), as the clustering in the VtC-XES is suggestive of a sensitivity to hydroxyl groups. However, Figure 5 also exemplifies that first-nearest neighbors, e.g., the oxygen ligands directly bonded to the phosphorus, likely cause the biggest spectral changes and thus are the biggest contributing factor to clustering, which is consistent with our earlier observations.

### 7.3.2 Emergent chemical fingerprints from clusters

Above, we motivated our classes by important chemical properties that we heuristically expected to yield the biggest spectral differences. However, even within this chemically driven framework, there are sub-clusters within our heuristic chemical classes which are instead emergent

from UMAP. For example, we found that sub-clustering of the phosphate chemical class (exemplified by the multiple separate sub-clusters in Figures 3 and 4) was caused by unexpected variations in the secondary substituent (atoms bound to oxygens, not directly to phosphorus), indicating that XANES spectra is sensitive to even more subtle details than anticipated.

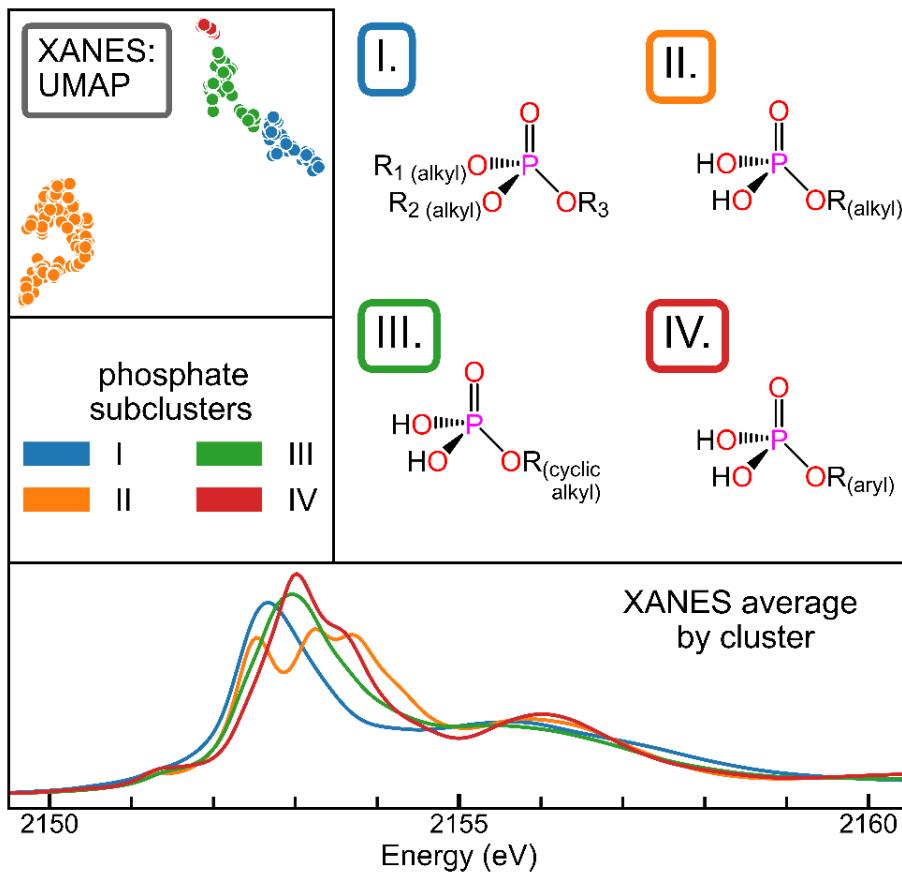
Let us examine this sub-division of the phosphates, specifically in the UMAP embedding of their XANES spectra. For just phosphates, we achieve the embedding shown in Figure 6, which has labeled the phosphates into four clusters determined by the dbSCAN<sup>77</sup> clustering algorithm: **I**, **II**, **III**, and **IV**. The average spectrum for each cluster is shown at the bottom and the common structural motifs for each cluster are shown to the right.

77% of Cluster **I** is comprised of compounds with two alkyl R groups and the third group either alkyl or aryl rings. This distinction is different from Clusters **II** to **IV** as they instead typically have two R groups as H atoms instead of carbon-based groups. Cluster **II** is the largest sub-cluster and 94% of the compounds have two hydroxyl groups bonded to the phosphorus and the last R group an alkyl chain. These two clusters are the most distinct.

On the other hand, Cluster **III** and **IV** are similar in composition. Cluster **III** is comprised of compounds with the third R group as: (a) alkyl rings, or cycloalkanes (36%), (b) aromatic rings (23%), or (c) take part in intramolecular hydrogen bonding with one of the hydroxyl groups bonding to the phosphorus. Cluster **IV** compounds are structurally very similar to Cluster **III** compounds, even though their spectra are distinct. However, 54% of Cluster **IV** compounds have their third R group as aromatic rings. For some example compounds in each cluster along with their spectra and structure, see Figures S9 to S12. All compounds in Clusters **I** to **IV** can also be viewed in Figures S13 to S16. Additionally, given the linear nature of Clusters **I**, **III**, and **IV** in

the UMAP embedding, we tested the correlation between the embedding location and the energy of the absorption edge, as demonstrated in Figure S17, and found no strong correlation.

Furthermore, color-coding the phosphates based on a ten-dimensional clustering and then visualizing them in two dimensions yields very nearly the same classifications, as shown in Figure S18. Thus, the two-dimensional embedding is retaining enough information to categorize the phosphates appropriately. Even expanding the embedding space to three dimensions instead of two for all previous embeddings yields very nearly the same clustering, as shown in Figure S19. This retention in information – yet complex clustering of compounds – further supports the nonlinear nature of spectra and the idea that properties are complexly encoded in spectra and conversely, spectral features do not correlate solely to a single, high-variant attribute but rather a combination of electronic or chemical properties.



**Figure 7.6** UMAP representation of XANES of phosphates, color-coded by sub-clusters. Cluster-averaged spectra and a summary structural motif for each cluster are also shown.

Taken *en masse*, these results show the extent to which chemically relevant information is, or is not, encoded by the quantum mechanics involved in XANES and VtC-XES. As to the specific algorithm, UMAP can be used iteratively as more data is collected. Thus, it has the potential to show evolutions through the domain space, similar to the latent space of a variational autoencoder (VAE)<sup>78</sup>, given proper tuning of its two hyperparameters: the number of expected neighbors in a cluster and the minimum distance between points. For an overview of the effect of those two hyperparameters on the UMAP embedding, see Figures S20 and S21. Finally, and of key importance here, UMAP can generate embeddings of spectra that can be used for unbiased

refinement of the training data set in addition to a preprocessing step before supervised ML predictions.

### 7.3.3 Validation of chemical fingerprints from cluster analysis

In our prior work on sulfororganics<sup>55</sup> and in the present above work on the more complex case of phosphorganics, we have demonstrated a convincing utility of advanced, nonlinear unsupervised ML tools for evaluating the chemically-relevant information in VtC-XES and XANES spectra. We now return to our hypothesis presented in the introduction and illustrated in Figure 1, where we propose that such an unsupervised ML method can productively inform the use of supervised ML tasks.

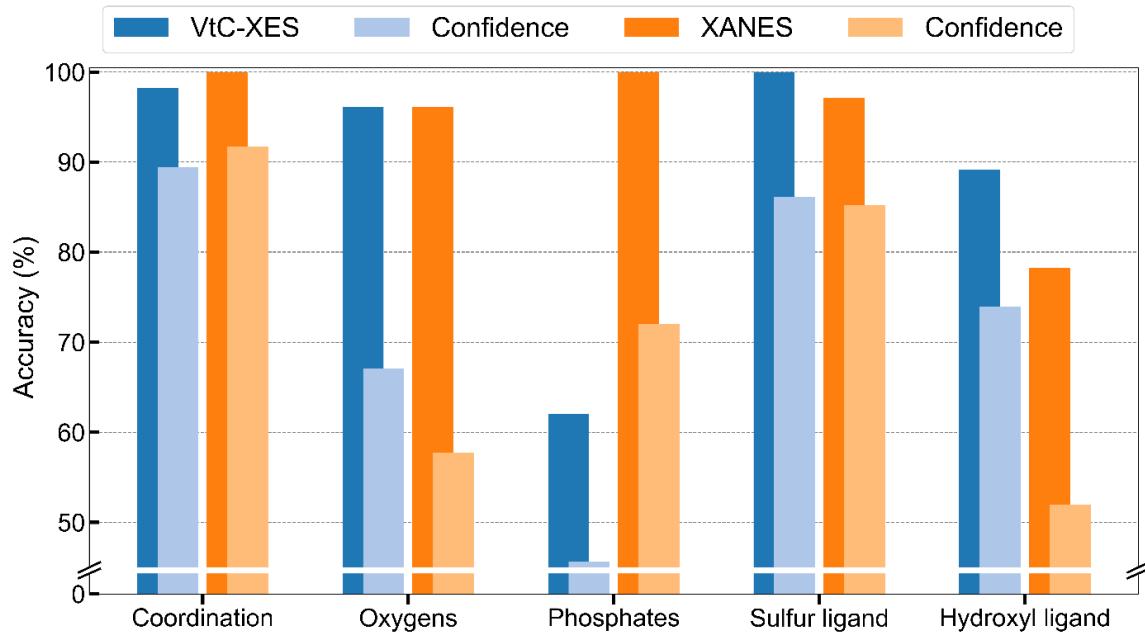
The most common use of supervised ML in X-ray spectroscopy is to predict numerical properties, such as bond length or coordination, from XANES spectra<sup>20-22, 24, 31</sup>. Here, we instead predict chemical classes from both VtC-XES and XANES spectra. Moreover, we predict these classes from a five-dimensional UMAP representation of the spectra instead of from the original spectra themselves. Such preprocessing through dimensionality reduction can help separate inherently correlated and nonlinear spectral features<sup>56</sup> as well as greatly reduce both the computational cost and the effect of spectral noise.

Furthermore, we use a Gaussian Process (GP) in order to incorporate prior knowledge into our models and generate an informed predictor<sup>73</sup>. A GP is a non-parametric kernel method that formally incorporates Bayes rule into the model, which not only allows for priors to be specified during training, but also allows for a probabilistic interpretation of the results. This probability gives uncertainty estimates, or conversely confidence, of the predictions. We note that one of the biggest downsides of a GP is that it scales poorly, which is another reason why applying a nonlinear

dimensionality reduction routine like UMAP beforehand can transform this problem into a computationally tractable one.

The results of training a GP on each of the five classification schemes (see Table S1) we developed – coordination, number of oxygen ligands, phosphate subcluster, number of sulfur ligands, and number of hydroxyl ligands – are shown in Figure 7, with the average accuracy score on the test set as well as the probability of that prediction, i.e., the confidence score, shown. There is a clear correlation between the average accuracy and confidence, indicating that the GP is, in fact, properly modeling uncertainty of predictions.

Finally, the accuracies and confidence of each prediction across the VtC-XES and XANES data matches what we observed in our two-dimensional UMAP figures. This is clearly demonstrated in the hydroxyl ligand and phosphate subcluster classification schemes, where the XANES and VtC-XES, respectively, poorly cluster by these schemes, and the low corresponding GP confidence reflects this. Overall, these results further validate that visualizing data via a dimensionality reduction algorithm like UMAP correlates to extractable information content and can properly inform classes to be used for supervised ML.



**Figure 7.7** Gaussian Process Classifier prediction accuracies with corresponding average probability (“confidence”) for all chemically driven and cluster-driven classification schemes.

However, we note that care must be taken to ensure transferability when training any supervised ML model on theoretical spectra to then make predictions on experimental data, the obvious next step of our GPs. Ensuring transferability might mean appropriately modeling for noise, the spectral line shape, or any systematics errors in the theoretical model.

## 7.4 Conclusions

By utilizing Uniform Manifold Approximation and Projection (UMAP) and analyzing the resulting clustering in a two-dimensional embedding of VtC-XES and XANES spectra of an ensemble of organophosphorus compounds, we find sensitivity to coordination and ligand identity (specifically by distinguishing number of oxygen ligands, sulfur ligands, and hydroxyl groups). Additionally, the XANES was clearly more sensitive to phosphate sub-groupings due to an

unexpected, unintuitive fingerprint that emerged from the clustering in the unsupervised machine learning tool, UMAP.

These results culminate in a valuable analysis framework: (1) applying nonlinear dimensionality reduction routines and cluster analysis to check for both heuristic chemical sensitivities and emergent ones present in spectra, (2) applying dimensionality reduction methods like UMAP before querying supervised ML models, and (3) utilizing models that incorporate prior knowledge, such as a Gaussian Process, to estimate uncertainty or confidence of these predictions on the clustering-informed classes. Furthermore, this framework, which we call Unsupervised Validation of Classes (UVC) and illustrate in Figure 1, is broadly applicable – it can easily be expanded to both other systems and other spectroscopies – providing a way to inform methodology and validate predictions instead of relying solely on the scientist’s knowledge and, possibly, bias in the initial construction of an appropriate training dataset.

## ASSOCIATED CONTENT

### **Supporting Information.**

The following files are available free of charge:

**Figure S1** Theory versus experiment: VtC-XES (png)

**Figure S2** Theory versus experiment: XANES (png)

**Figure S3** Scree plot of VtC-XES and XANES data (png)

**Figure S4** PCA reconstruction of VtC-XES spectra (png)

**Figure S5** PCA reconstruction of XANES spectra (png)

**Table S1** Classification table (docx)

**Table S2** Table for compounds a to h (docx)

**Figure S6** Class averages of spectra with different coordination (png)

**Figure S7** Cluster averages of spectra with different coordination (png)

**Figure S8** UMAP representation of XANES with H atom substitutions (png)

**Figure S9** Phosphate sub-cluster I example spectra (png)

**Figure S10** Phosphate sub-cluster II example spectra (png)

**Figure S11** Phosphate sub-cluster III example spectra (png)

**Figure S12** Phosphate sub-cluster IV example spectra (png)

**Figure S13** Phosphate sub-cluster I structures (png)

**Figure S14** Phosphate sub-cluster II structures (png)

**Figure S15** Phosphate sub-cluster III structures (png)

**Figure S16** Phosphate sub-cluster IV structures (png)

**Figure S17** Phosphate sub-clusters correlation (png)

**Figure S18** Phosphate subclusters: 10-dim clustering (png)

**Figure S19** 3D UMAP visualizations (png)

**Figure S20** Changing UMAP hyperparameters: number of neighbors (png)

**Figure S21** Changing UMAP hyperparameters: minimum distance (png)

## AUTHOR INFORMATION

The authors declare no competing financial interests.

## ACKNOWLEDGMENT

ST acknowledges funding from NRT-DESE: Data Intensive Research Enabling Clean Technologies (DIRECT) under grant no. NSF #1633216 and acknowledge funding from NSF CHE-1904437. VK acknowledges support from the Washington NASA Space Grant from the Washington NASA Space Grant Consortium (WSGC). NG acknowledges support from the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences under Award No. KC-030105172685. AV acknowledges support from the Research Corporation for Science Advancement through a Cottrell Scholars Award. This research benefited from computational resources provided by the Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at PNNL. PNNL is operated by Battelle Memorial Institute for the United States Department of Energy under DOE Contract No. DE-AC05-76RL1830. Additionally, this work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

## 7.5 References

1. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547-555.
2. Z. Q. Zhou, Q. F. He, X. D. Liu, Q. Wang, J. H. Luan, C. T. Liu and Y. Yang, *npj Computational Materials*, 2021, **7**, 138.
3. Y. Liu, T. L. Zhao, W. W. Ju and S. Q. Shi, *Journal of Materomics*, 2017, **3**, 159-177.
4. Y. Liu, B. R. Guo, X. X. Zou, Y. J. Li and S. Q. Shi, *Energy Storage Materials*, 2020, **31**, 434-450.
5. J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *JOM*, 2013, **65**, 1501-1509.
6. C. A. Meza Ramirez, M. Greenop, L. Ashton and I. u. Rehman, *Applied Spectroscopy Reviews*, 2021, **56**, 733-763.
7. D. F. Gordon and M. Desjardins, *Machine Learning*, 1995, **20**, 5-22.
8. D. H. Wolpert and W. G. Macready, *IEEE Transactions on Evolutionary Computation*, 1997, **1**, 67-82.
9. S. Alelyani, *Applied Sciences*, 2021, **11**, 6271.
10. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, *ACM Computing Surveys*, 2021, **54**, 1-35.
11. M. Pot, N. Kieusseyan and B. Prainsack, *Insights into Imaging*, 2021, **12**, 13.
12. A. M. F. Hiemstra, T. Cassel, M. P. Born and C. C. S. Liem, *Gedrag en Organisatie*, 2020, **33**, 279-299.
13. M. Belkin, P. Niyogi and V. Sindhwani, *Journal of Machine Learning Research*, 2006, **7**, 2399-2434.
14. D. Erhan, Y. Bengio, A. Courville, P. A. Manzagol, P. Vincent and S. Bengio, *Journal of Machine Learning Research*, 2010, **11**, 625-660.
15. H. Liu and L. Yu, *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17**, 491-502.
16. R. Monarch, *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*, Manning, Shelter Island, 2021.
17. W. A. Omta, R. G. van Heesbeen, I. Shen, J. de Nobel, D. Robers, L. M. van der Velden, R. H. Medema, A. P. J. M. Siebes, A. J. Feelders, S. Brinkkemper, J. S. Klumperman, M. R. Spruit, M. J. S. Brinkhuis and D. A. Egan, *SLAS Discovery*, 2020, **25**, 655-664.
18. N. Muhammad Noor Mathivanan, N. Azura Md.Ghani and R. Mohd Janor, *Bulletin of Electrical Engineering and Informatics; Vol 7, No 3: September 2018 DOI - 10.11591/eei.v7i3.1272*, 2018.
19. M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, Z. Yao and A. Aspuru-Guzik, *arXiv*, 2022.
20. J. Timoshenko, D. Y. Lu, Y. W. Lin and A. I. Frenkel, *Journal of Physical Chemistry Letters*, 2017, **8**, 5091-5098.
21. J. Timoshenko and A. I. Frenkel, *Acs Catalysis*, 2019, **9**, 10192-10211.
22. J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans and A. I. Frenkel, *Physical Review Letters*, 2018, **120**, 225502.
23. J. Timoshenko, C. J. Wrasman, M. Luneau, T. Shirman, M. Cargnello, S. R. Bare, J. Aizenberg, C. M. Friend and A. I. Frenkel, *Nano Letters*, 2019, **19**, 520-529.
24. C. Zheng, C. Chen, Y. Chen and S. P. Ong, *Patterns*, 2020, **1**, 100013.

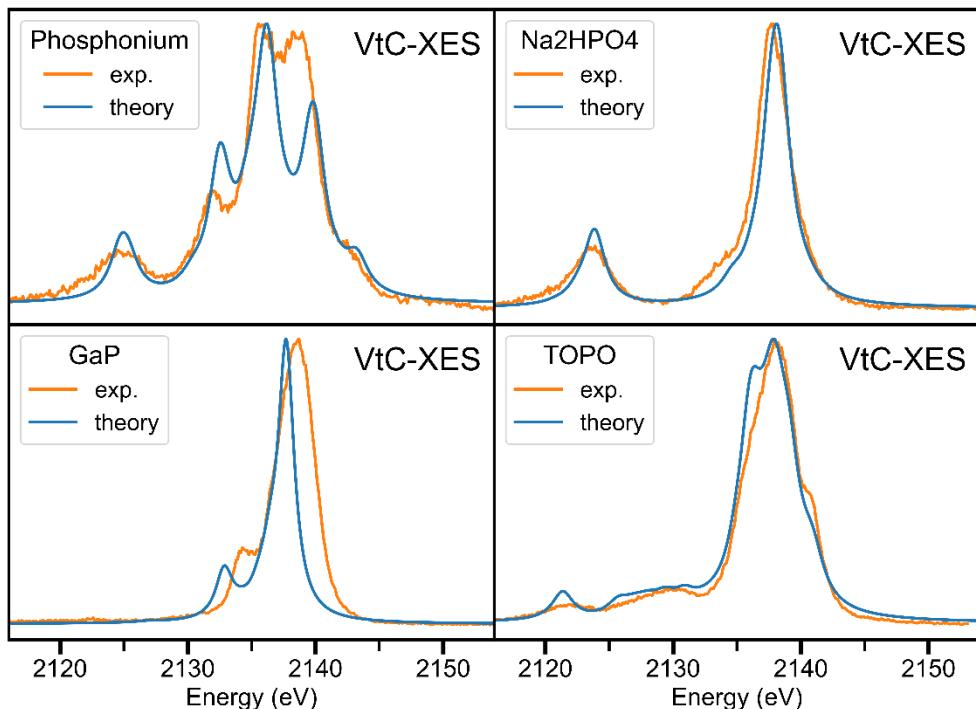
25. C. Zheng, K. Mathew, C. Chen, Y. M. Chen, H. M. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson and S. P. Ong, *Npj Computational Materials*, 2018, **4**, 12.
26. S. Kiyohara, T. Miyata, K. Tsuda and T. Mizoguchi, *Scientific Reports*, 2018, **8**, 13548.
27. P. K. Routh, Y. Liu, N. Marcella, B. Kozinsky and A. I. Frenkel, *The Journal of Physical Chemistry Letters*, 2021, **12**, 2086-2094.
28. A. Aarva, V. L. Deringer, S. Sainio, T. Laurila and M. A. Caro, *Chemistry of Materials*, 2019, **31**, 9243-9255.
29. M. R. Carbone, S. Yoo, M. Topsakal and D. Lu, *Physical Review Materials*, 2019, **3**, 033604.
30. M. R. Carbone, M. Topsakal, D. Lu and S. Yoo, *Physical Review Letters*, 2020, **124**, 156401(156406).
31. Y. Liu, N. Marcella, J. Timoshenko, A. Halder, B. Yang, L. Kolipaka, M. J. Pellin, S. Seifert, S. Vajda, P. Liu and A. I. Frenkel, *The Journal of Chemical Physics*, 2019, **151**, 164201.
32. A. Martini, S. A. Guda, A. A. Guda, G. Smolentsev, A. Algasov, O. Usoltsev, M. A. Soldatov, A. Bugaev, Y. Rusalev, C. Lamberti and A. V. Soldatov, *Computer Physics Communications*, 2020, **250**, 107064.
33. I. Miyazato, L. Takahashi and K. Takahashi, *Molecular Systems Design & Engineering*, 2019, **4**, 1014-1018.
34. A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot, J. A. van Bokhoven, C. Copéret and A. V. Soldatov, *npj Computational Materials*, 2021, **7**, 203.
35. Z. Fang, W. Hu, M. Wang, R. Wang, S. Zhong and S. Chen, *Biomedical Signal Processing and Control*, 2020, **60**, 101944.
36. S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram and L. Hung, *npj Computational Materials*, 2020, **6**, 109.
37. O. Trejo, A. L. Dadlani, F. De La Paz, S. Acharya, R. Kravec, D. Nordlund, R. Sarangi, F. B. Prinz, J. Torgersen and N. P. Dasgupta, *Chemistry of Materials*, 2019, **31**, 8937-8947.
38. C. D. Rankine, M. M. M. Madkhali and T. J. Penfold, *The Journal of Physical Chemistry A*, 2020, **124**, 4263-4270.
39. C. D. Rankine and T. J. Penfold, *The Journal of Physical Chemistry A*, 2021, **125**, 4276-4293.
40. S. Kiyohara, M. Tsubaki and T. Mizoguchi, *Npj Computational Materials*, 2020, **6**, 68.
41. O. A. Usoltsev, A. L. Bugaev, A. A. Guda, S. A. Guda and A. V. Soldatov, *The Journal of Physical Chemistry C*, 2022, **126**, 4921-4928.
42. M. Cuisinier, P.-E. Cabelguen, S. Evers, G. He, M. Kolbeck, A. Garsuch, T. Bolin, M. Balasubramanian and L. F. Nazar, *The Journal of Physical Chemistry Letters*, 2013, **4**, 3227-3232.
43. D. Asakura, E. Hosono, H. Niwa, H. Kiuchi, J. Miyawaki, Y. Nanba, M. Okubo, H. Matsuda, H. Zhou, M. Oshima and Y. Harada, *Electrochemistry Communications*, 2015, **50**, 93-96.
44. Y. Zhou, D. E. Doronkin, Z. Zhao, P. N. Plessow, J. Jelic, B. Detlefs, T. Pruessmann, F. Studt and J.-D. Grunwaldt, *ACS Catalysis*, 2018, **8**, 11398-11406.

45. M. Maiuri, M. Garavelli and G. Cerullo, *Journal of the American Chemical Society*, 2020, **142**, 3-15.
46. G. Bunker, *Introduction to XAFS: A Practical Guide to X-ray Absorption Fine Structure Spectroscopy*, Cambridge University Press, Cambridge, 2010.
47. P. Glatzel and U. Bergmann, *Coordination Chemistry Reviews*, 2005, **249**, 65-95.
48. F. de Groot, *Chemical Reviews*, 2001, **101**, 1779-1808.
49. N. Lee, T. Petrenko, U. Bergmann, F. Neese and S. DeBeer, *Journal of the American Chemical Society*, 2010, **132**, 9715-9727.
50. C. J. Pollock and S. DeBeer, *Accounts of Chemical Research*, 2015, **48**, 2967-2975.
51. G. T. Seidler, D. R. Mortensen, A. J. Remesnik, J. I. Pacold, N. A. Ball, N. Barry, M. Styczinski and O. R. Hoidn, *Review of Scientific Instruments*, 2014, **85**, 113906.
52. W. Malzer, C. Schlesiger and B. Kanngießer, *Spectrochimica Acta Part B: Atomic Spectroscopy*, 2021, **177**, 106101.
53. P. Zimmermann, S. Peredkov, P. M. Abdala, S. DeBeer, M. Tromp, C. Müller and J. A. van Bokhoven, *Coordination Chemistry Reviews*, 2020, **423**, 213466.
54. W. M. Holden, E. P. Jahrman, N. Govind and G. T. Seidler, *The Journal of Physical Chemistry A*, 2020, **124 (26)**, 5415-5434.
55. S. Tetef, N. Govind and G. T. Seidler, *Phys. Chem. Chem. Phys.*, 2021, **23**, 23586-23601.
56. M. Ceriotti, *The Journal of Chemical Physics*, 2019, **150**, 150901.
57. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Research*, 2020, **47**.
58. [github.com/vikramkashyap/moldl](https://github.com/vikramkashyap/moldl).
59. L. McInnes, J. Healy and J. Melville, *arXiv*, 2020.
60. L. van der Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579-2605.
61. F. Pont, M. Tosolini and J. J. Fournie, *Nucleic Acids Research*, 2019, **47**.
62. M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. De Jong, *Computer Physics Communications*, 2010, **181**, 1477-1489.
63. E. Apra, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Atta-Fynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauet, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Fruchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Gotz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jonsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. M. del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, A. Vazquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D.

- Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Wolinski, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison, *J. Chem. Phys.*, 2020, **152**, 26.
- 64. T. Noro, M. Sekiya and T. Koga, *Theoretical Chemistry Accounts*, 2012, **131**, 1124.
  - 65. C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158-6170.
  - 66. A. Bergner, M. Dolg, W. Küchle, H. Stoll and H. Preuß, *Molecular Physics*, 1993, **80**, 1431-1441.
  - 67. I. Persson, W. Klysubun and D. Lundberg, *Journal of Molecular Structure*, 2019, **1179**, 608-611.
  - 68. S. Yasuda, *Bulletin of the Chemical Society of Japan*, 1984, **57**, 3122-3124.
  - 69. K. Lopata, B. E. Van Kuiken, M. Khalil and N. Govind, *Journal of Chemical Theory and Computation*, 2012, **8**, 3284-3292.
  - 70. Y. Zhang, S. Mukamel, M. Khalil and N. Govind, *Journal of Chemical Theory and Computation*, 2015, **11**, 5804-5809.
  - 71. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825-2830.
  - 72. T. Sainburg, L. McInnes and T. Q. Gentner, *Neural Computation*, 2021, **33**, 2881-2907.
  - 73. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.
  - 74. [github.com/Seidler-Lab/Phosphorus-ML-Project](https://github.com/Seidler-Lab/Phosphorus-ML-Project).
  - 75. M. Rovezzi and P. Glatzel, *Semicond. Sci. Technol.*, 2014, **29**.
  - 76. L. R. Murphy, T. L. Meek, A. L. Allred and L. C. Allen, *The Journal of Physical Chemistry A*, 2000, **104**, 5867-5871.
  - 77. M. Hahsler, M. Piekenbrock and D. Doran, *Journal of Statistical Software*, 2019, **91**, 1 - 30.
  - 78. A. Shrestha and A. Mahmood, *IEEE Access*, 2019, **7**, 53040-53065.

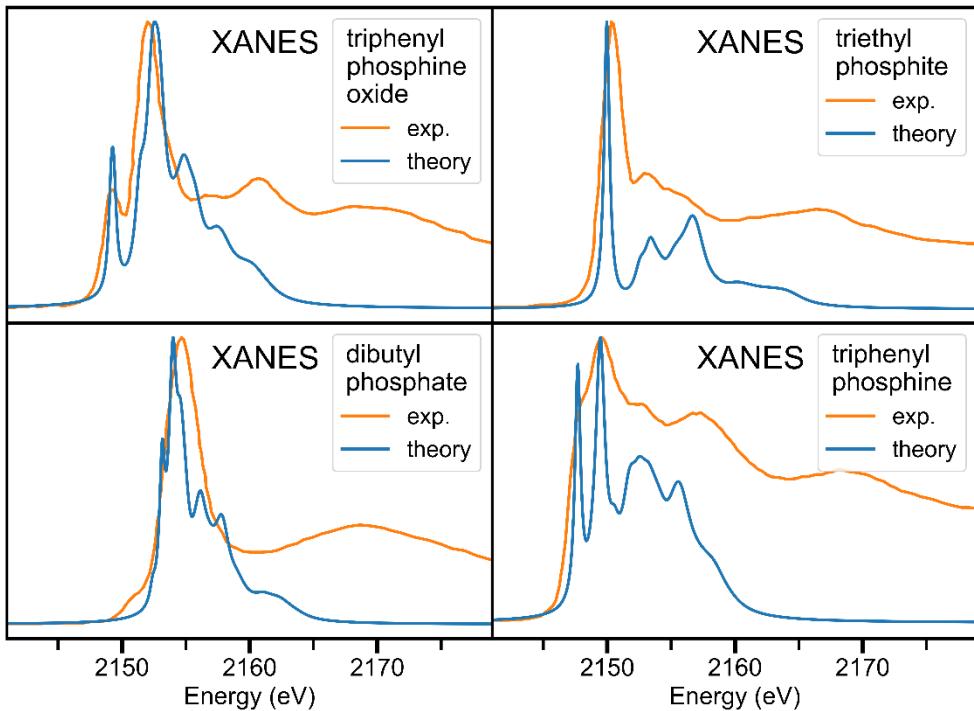
## 7.6 Supplementary Information

Theory versus experiment: VtC-XES



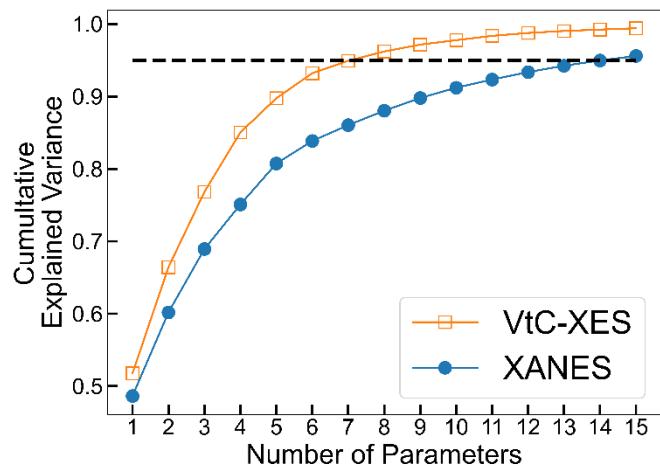
**Figure 7.S1** Experimental spectra versus theoretically calculated VtC-XES spectra using NWChem<sup>1</sup>. The experimental procedure follows the same protocol as Holden et al.<sup>2</sup> There is relatively good agreement in the existence and location of resonances, except a modest edge shift for GaP (bottom left).

### Theory versus experiment: XANES



**Figure 7.S2** Experimental spectra versus theoretically calculated XANES spectra using NWChem<sup>1</sup>. Experimental data is from Persson et al.<sup>3</sup> There is relatively good qualitative agreement in the existence and energy of near-edge features.

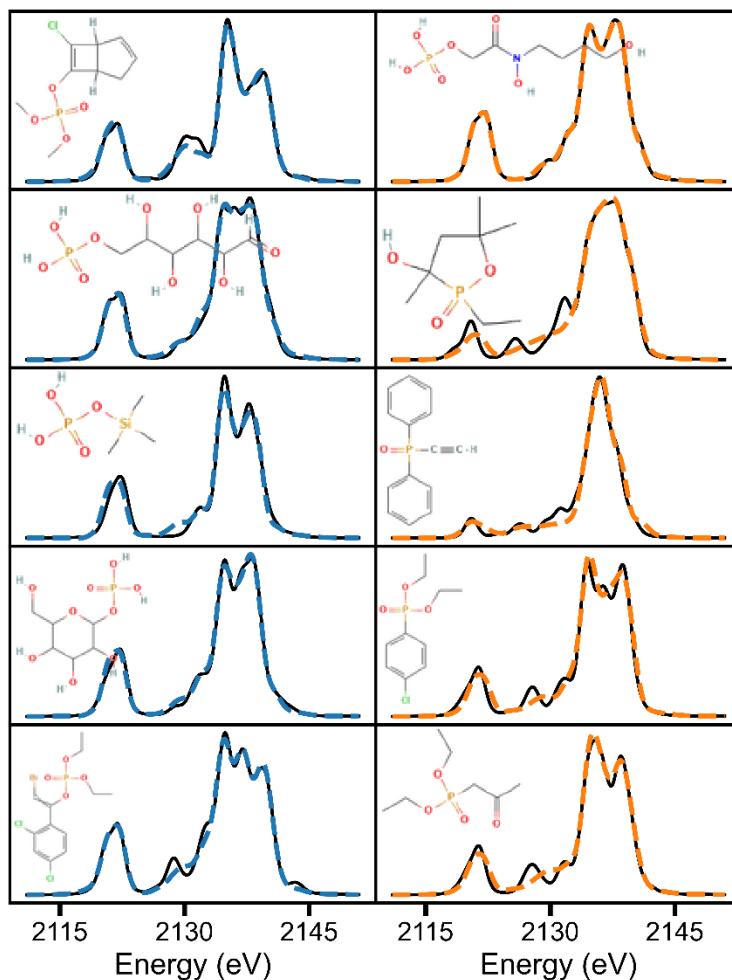
Scree plot of VtC-XES and XANES data



**Figure 7.S3** PCA preprocessing step to keep only 95% of the variance of the dataset.

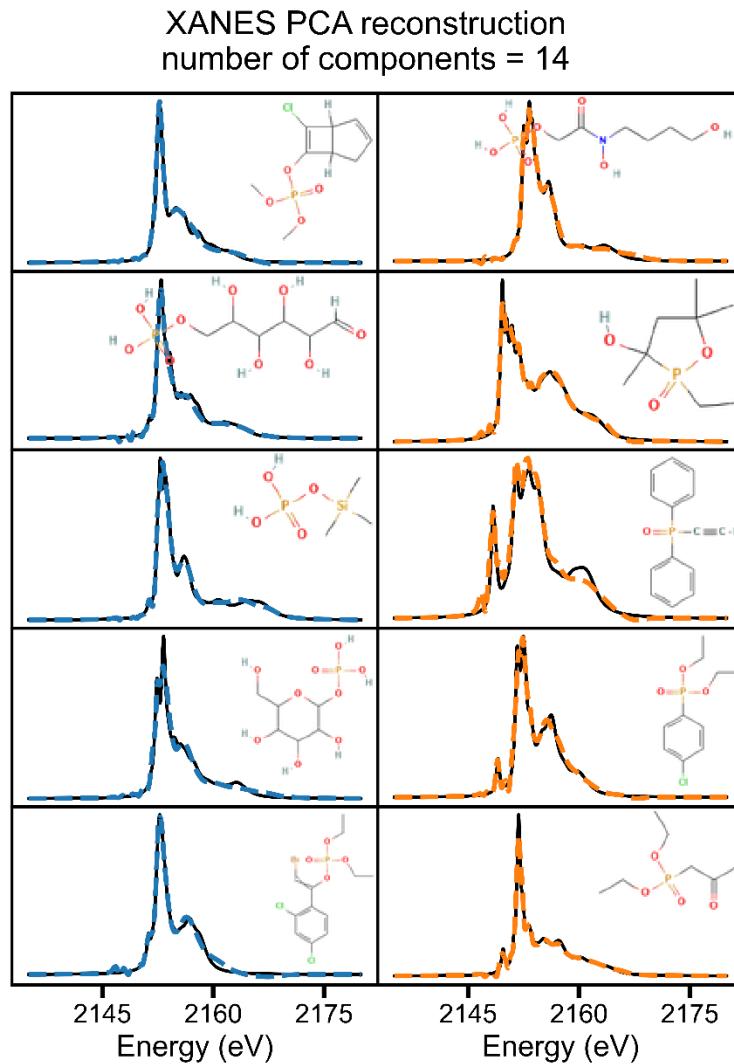
### PCA reconstruction of VtC-XES spectra

VtC-XES PCA reconstruction  
number of components = 7



**Figure 7.S4** Reconstructed VtC-XES spectra of randomly selected compounds (PubChem CIDS shown in top left for each) after being passed through the PCA pre-processing step to keep 95% of the variance of the dataset. The black lines are the theoretically calculated spectra, while the dashed colored lines are the reconstructed spectra after 95% variance, according to PCA, is retained.

### PCA reconstruction of XANES spectra



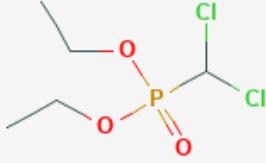
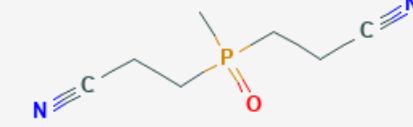
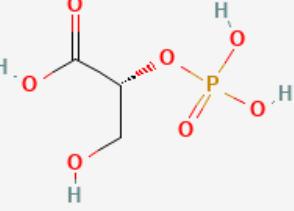
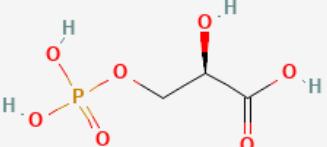
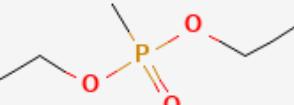
**Figure 7.S5** Reconstructed XANES spectra of randomly selected compounds (PubChem CIDS shown in top left for each) after being passed through the PCA pre-processing step to keep 95% of the variance of the dataset. The black lines are the theoretically calculated spectra, while the dashed colored lines are the reconstructed spectra after 95% variance, according to PCA, is retained.

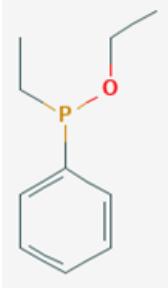
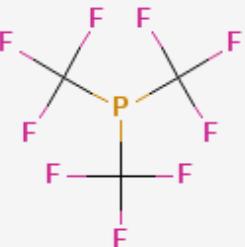
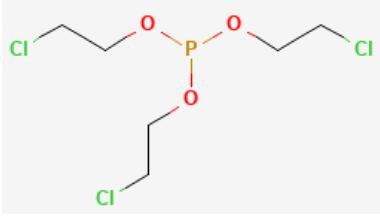
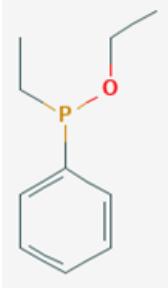
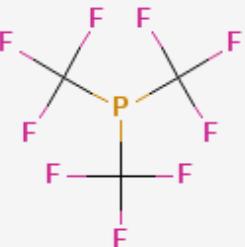
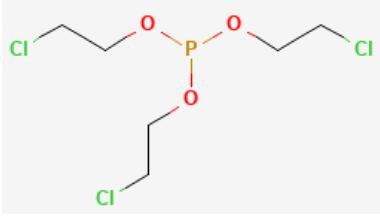
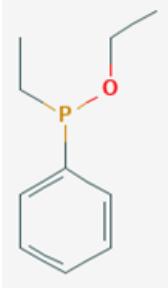
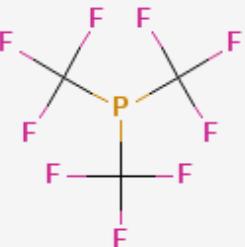
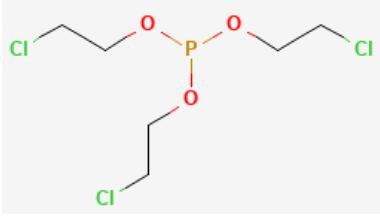
### Classification Table

| <i>Classification Scheme</i>     | <i>Number of classes</i> | <i>Corresponding figure</i> |
|----------------------------------|--------------------------|-----------------------------|
| <i>Coordination</i>              | 2                        | 2                           |
| <i>Number of oxygen ligands</i>  | 9                        | 3                           |
| <i>Number of sulfur ligands</i>  | 3                        | 4                           |
| <i>Number of hydroxyl groups</i> | 5                        | 5                           |
| <i>Phosphate subclusters</i>     | 4                        | 6                           |

**Table 7.S1** Summary of all classification schemes developed with UMAP, with references to the figures in the main text where they are first displayed. These classifications are equivalent to the ones predicted in Figure 7 in the main text.

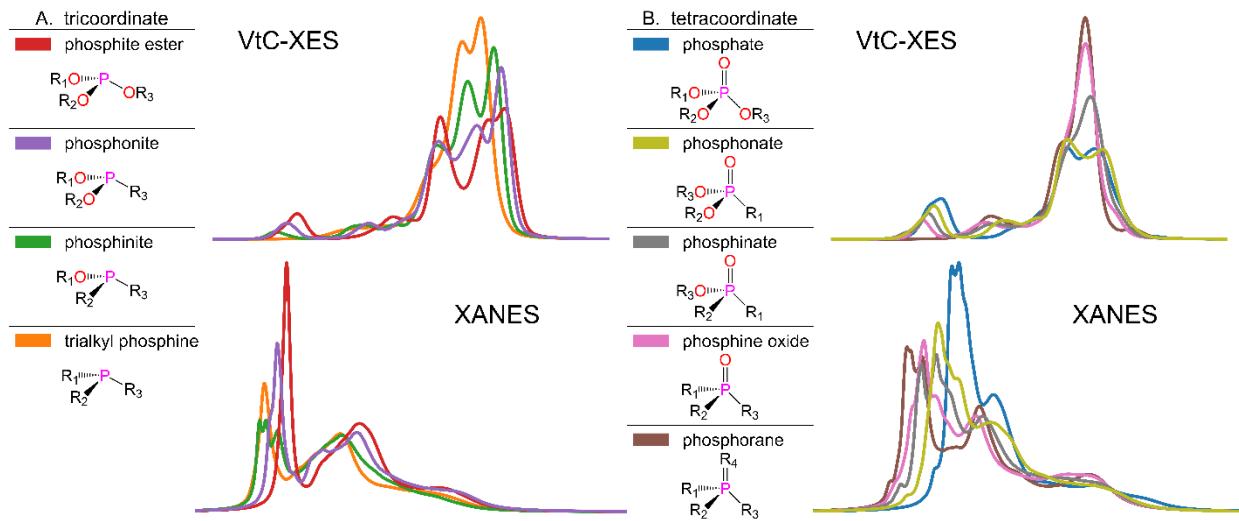
Table for compounds a to h

| Compound label | CID/<br>Class               | Structure  |
|----------------|-----------------------------|--|
| <b>a</b>       | 11085347<br>phosphonate     |    |
| <b>b</b>       | 54042304<br>phosphine oxide |    |
| <b>c</b>       | 439278<br>phosphonate       |   |
| <b>d</b>       | 439183<br>phosphate         |  |
| <b>e</b>       | 12685<br>phosphate          |  |

|  |   |          |  |             |  |        |  |                    |  |      |   |                 |  |
|--|---|----------|--|-------------|--|--------|--|--------------------|--|------|---|-----------------|--|
| <b>f</b><br><br><b>g</b><br><br><b>h</b> | <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%;">12496115</td><td style="text-align: center; width: 70%;">  </td></tr> <tr> <td>phosphinite</td><td></td></tr> </table><br><table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%;">136280</td><td style="text-align: center; width: 70%;">  </td></tr> <tr> <td>trialkyl phosphine</td><td></td></tr> </table><br><table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%;">8783</td><td style="text-align: center; width: 70%;">  </td></tr> <tr> <td>phosphite ester</td><td></td></tr> </table> | 12496115 |  | phosphinite |  | 136280 |  | trialkyl phosphine |  | 8783 |  | phosphite ester |  |
| 12496115                                 |   |          |  |             |  |        |  |                    |  |      |   |                 |  |
| phosphinite                              |   |          |  |             |  |        |  |                    |  |      |   |                 |  |
| 136280                                   |   |          |  |             |  |        |  |                    |  |      |   |                 |  |
| trialkyl phosphine                       |   |          |  |             |  |        |  |                    |  |      |   |                 |  |
| 8783                                     |    |          |  |             |  |        |  |                    |  |      |   |                 |  |
| phosphite ester                          |   |          |  |             |  |        |  |                    |  |      |   |                 |  |

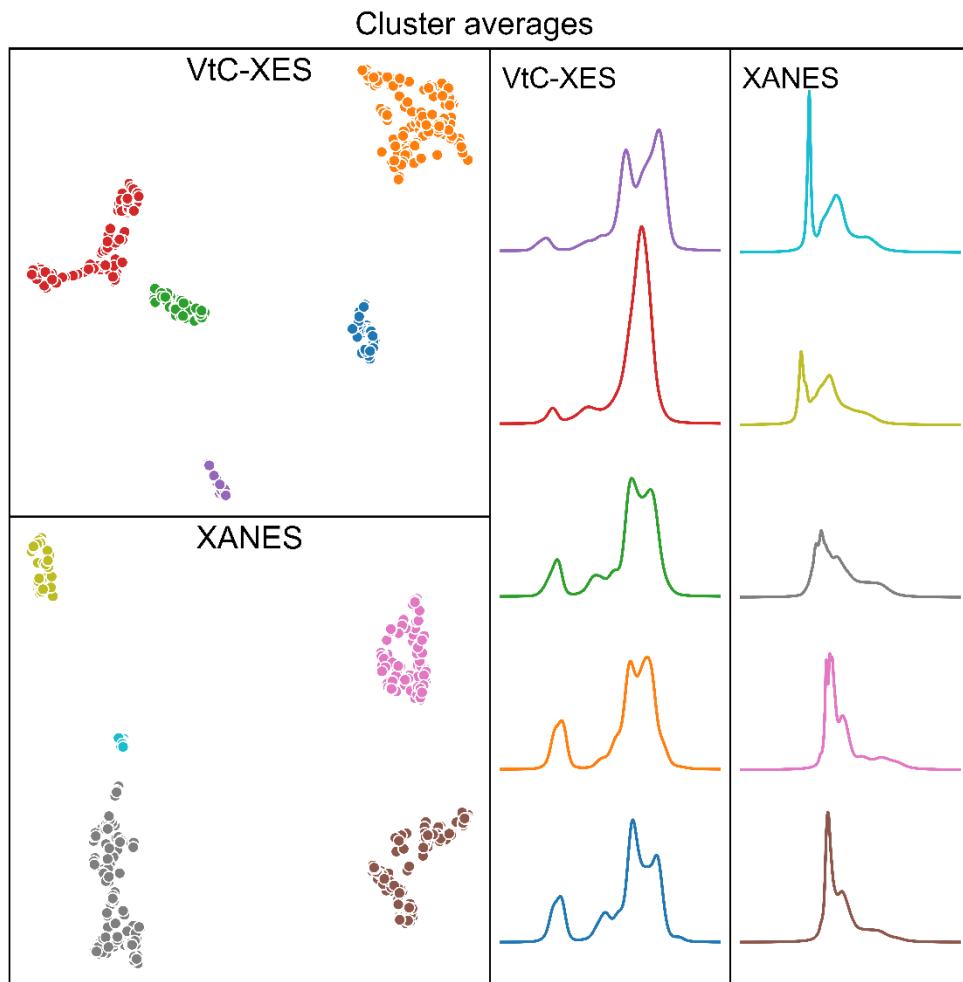
**Table 7.S2** Structure, CID, and chemical class for each compound **a** to **h** labeled in Figures 2 and 3 in the main text.

### Class averages of spectra with different coordination



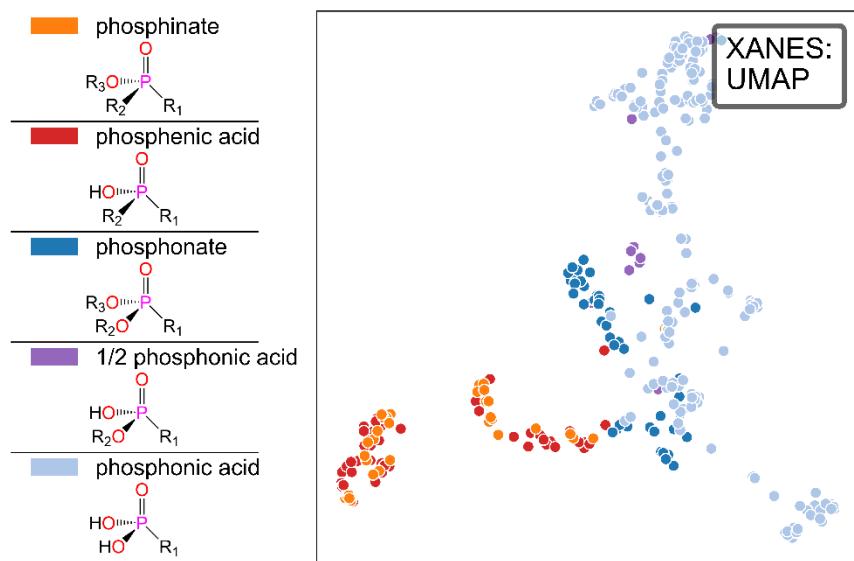
**Figure 7.S6** Average spectra for each chemical class within the two coordination geometries.

Cluster averages of spectra with different coordination



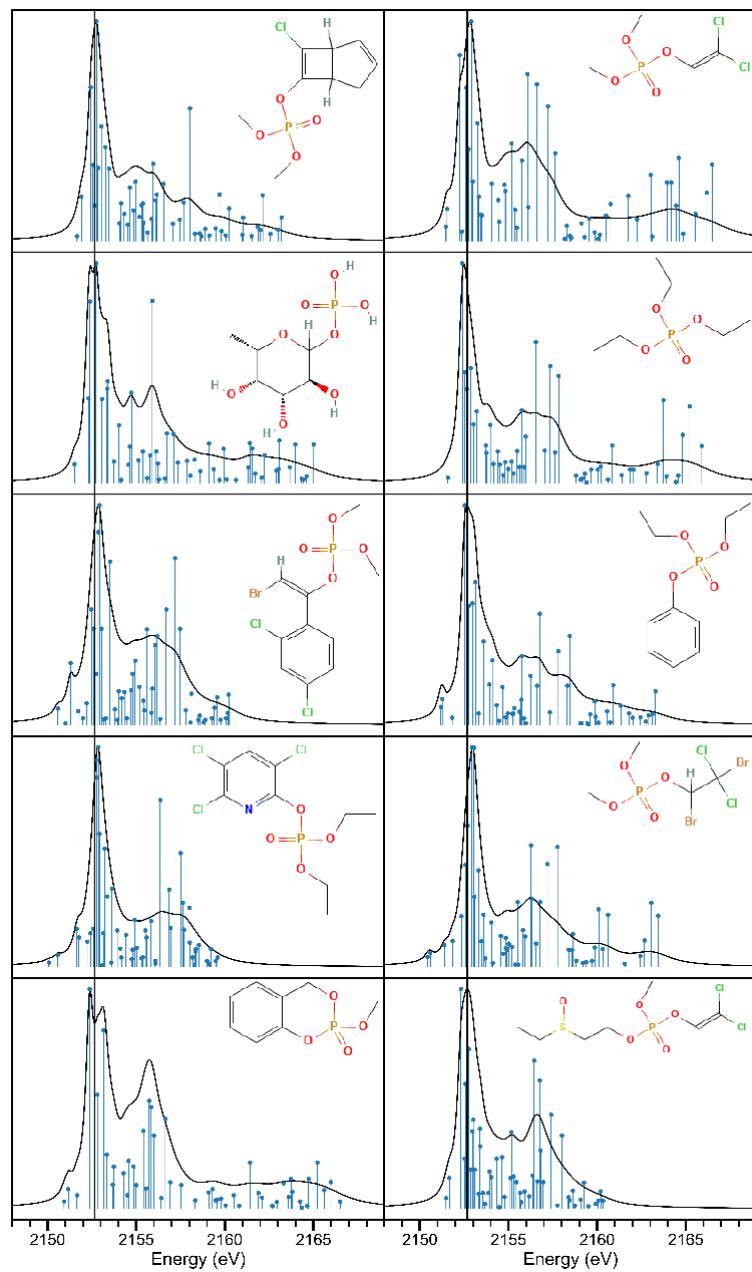
**Figure 7.S7** Cluster averages of compounds as they appear in the embeddings in Figures 2 and 3 in the main text.

UMAP representation of XANES with H atom substitutions



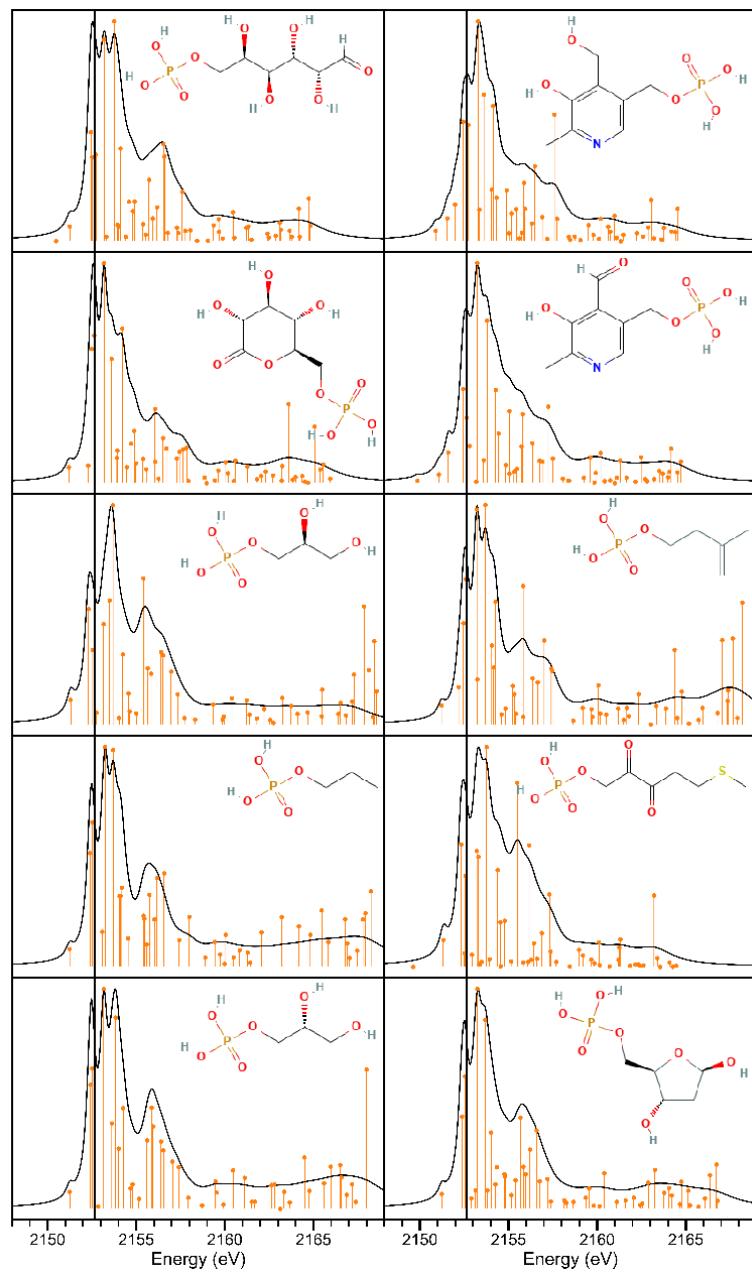
**Figure 7.88** The XANES embedding corresponding to Figure 7.5, i.e., substitution of O-R with hydroxyl groups. The XANES does not cluster as well as the VtC-XES for this classification scheme.

### Phosphate sub-cluster I example spectra



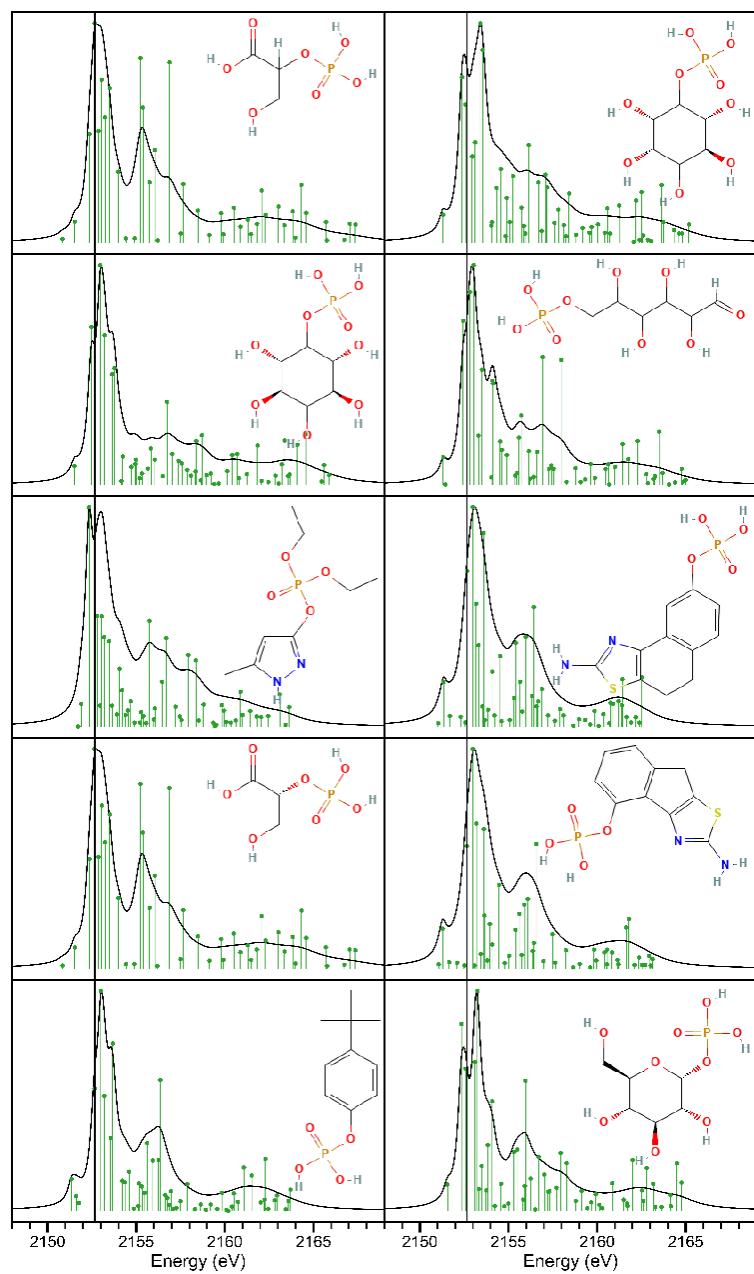
**Figure 7.S9** Example compounds and their corresponding spectra and transitions in phosphate sub-cluster I.

### Phosphate sub-cluster II example spectra



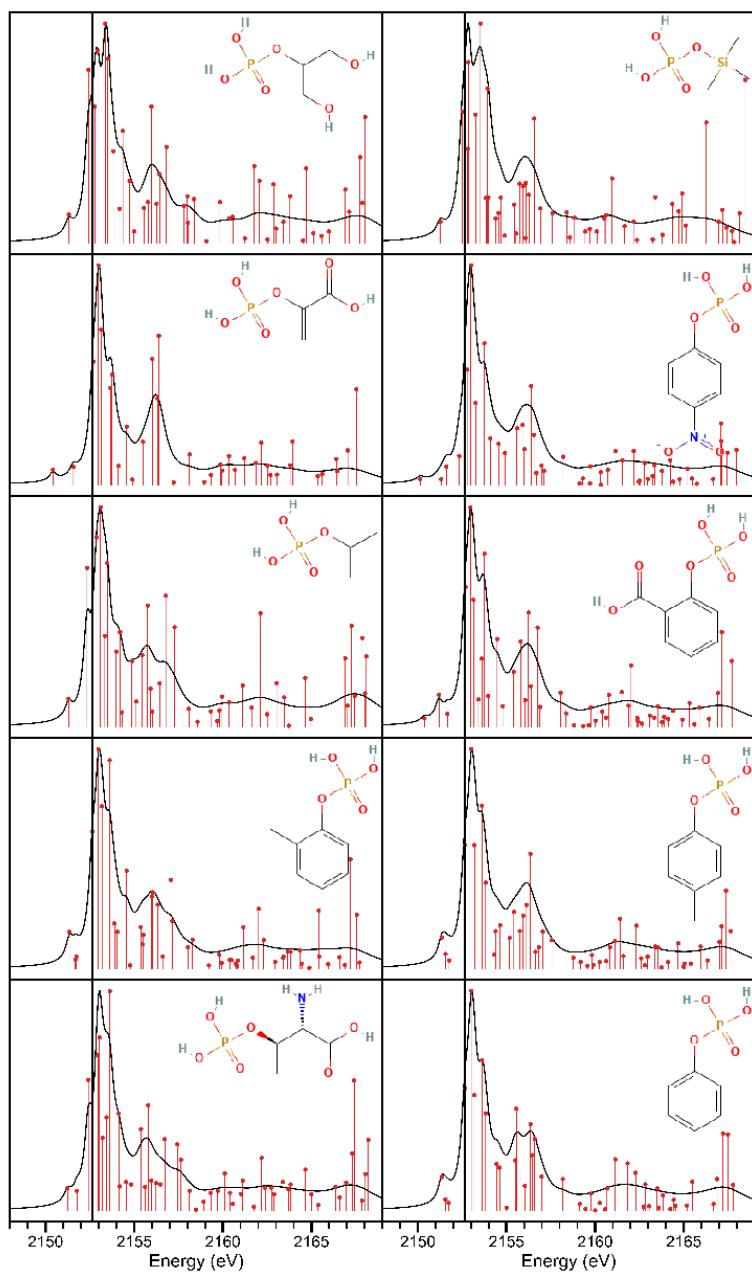
**Figure 7.S10** Example compounds and their corresponding spectra and transitions in phosphate sub-cluster II.

### Phosphate sub-cluster III example spectra



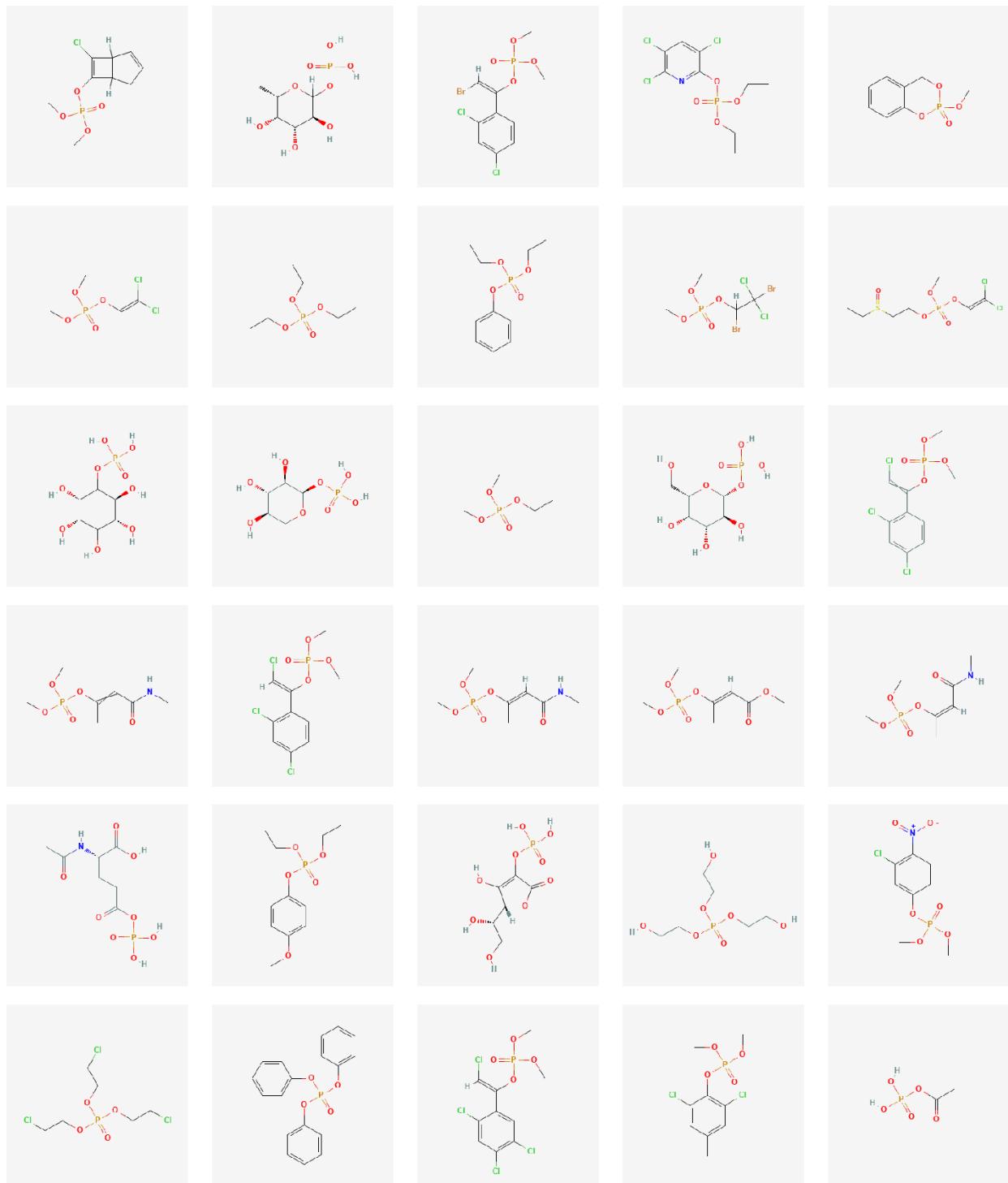
**Figure 7.S11** Example compounds and their corresponding spectra and transitions in phosphate sub-cluster III.

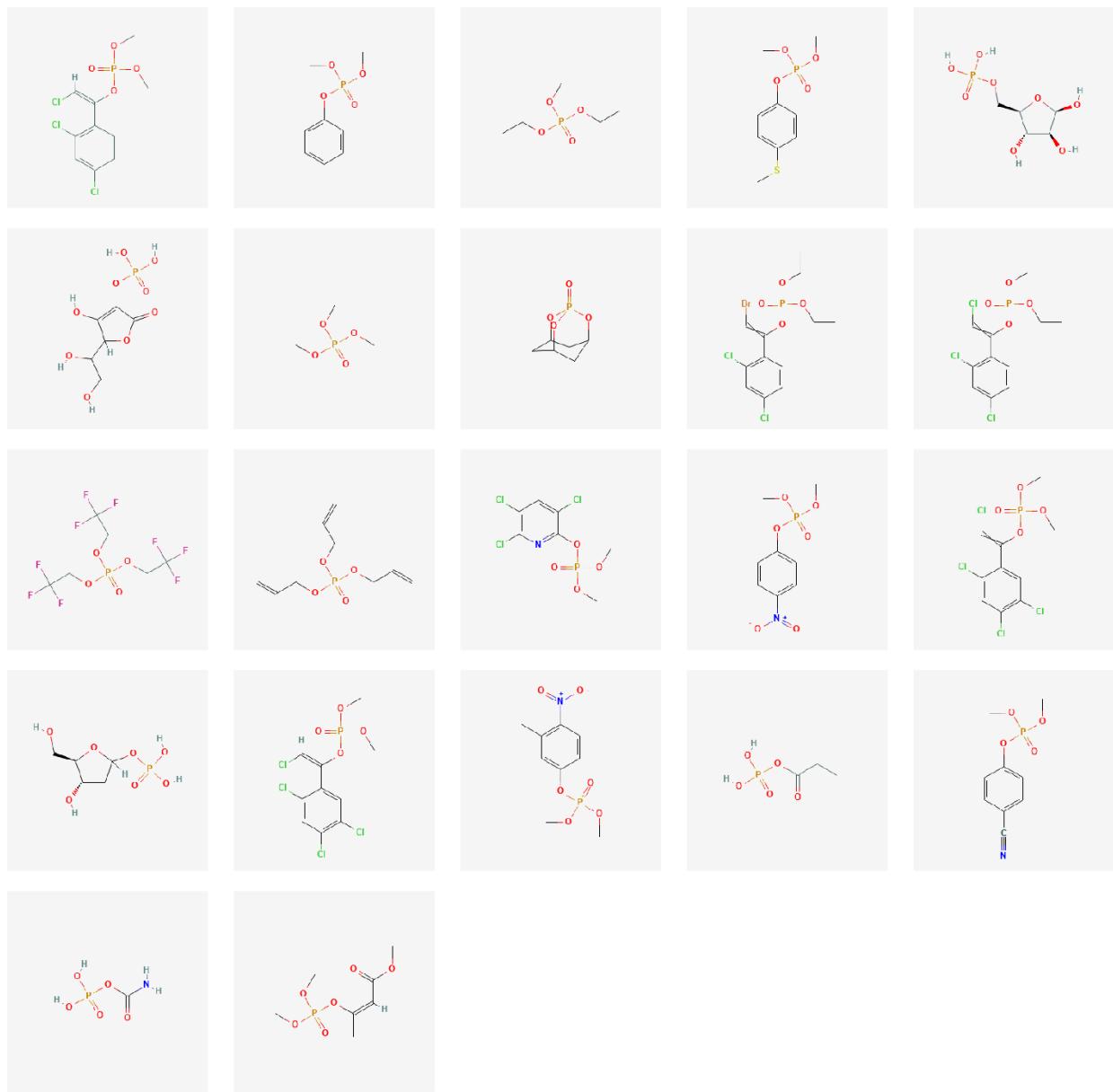
Phosphate sub-cluster IV example spectra



**Figure 7.S12** Example compounds and their corresponding spectra and transitions in phosphate sub-cluster IV.

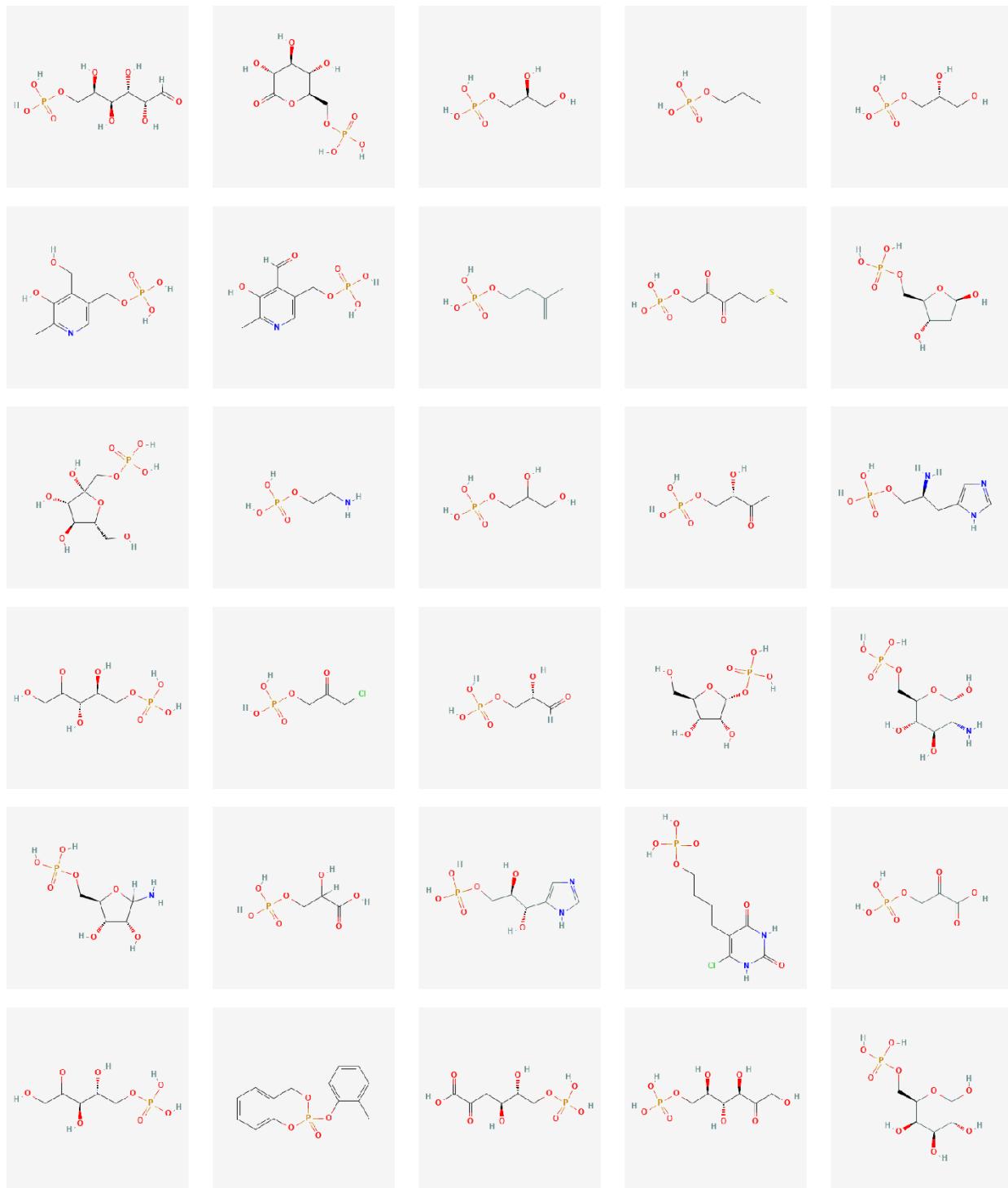
## Phosphate sub-cluster I structures

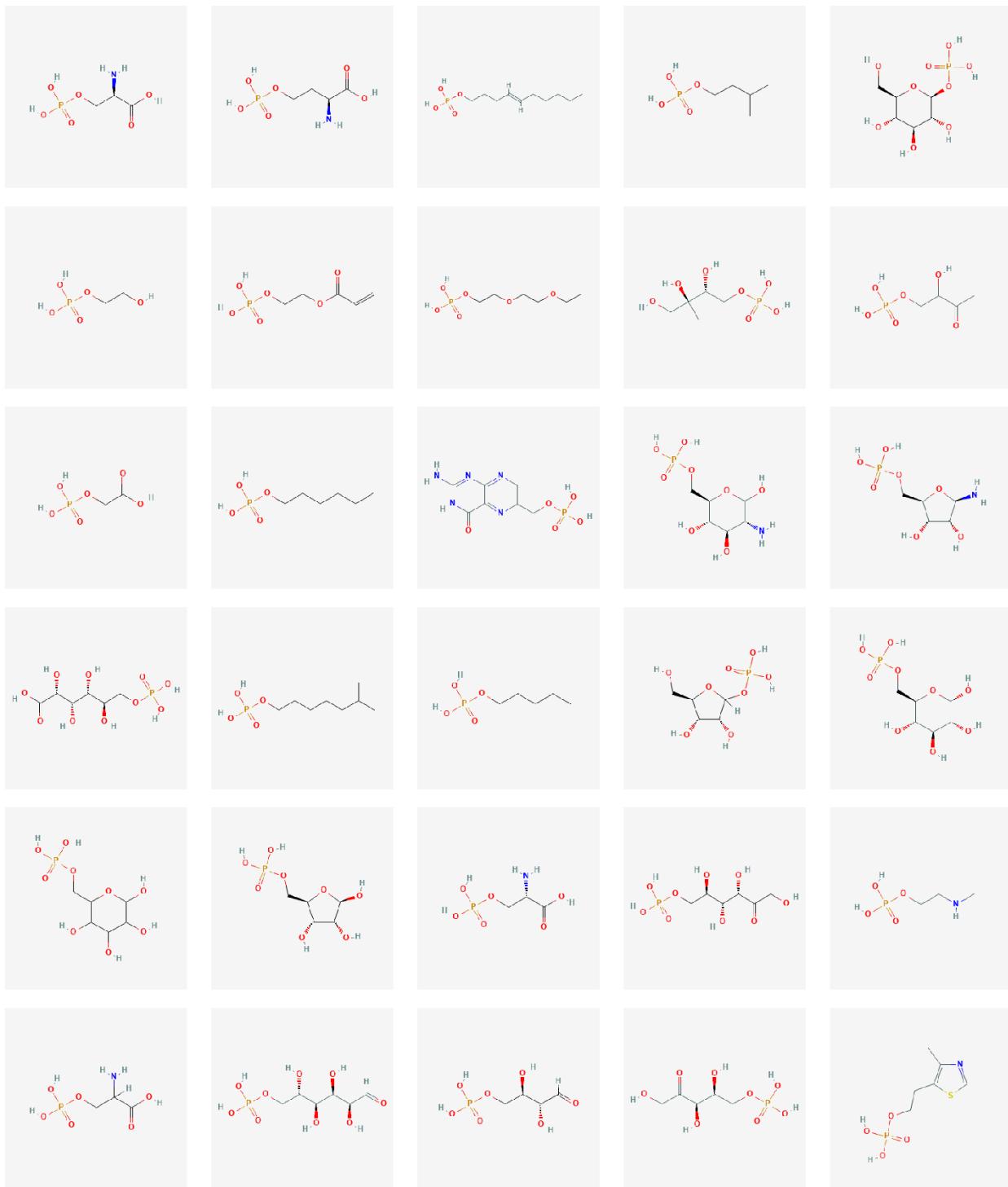


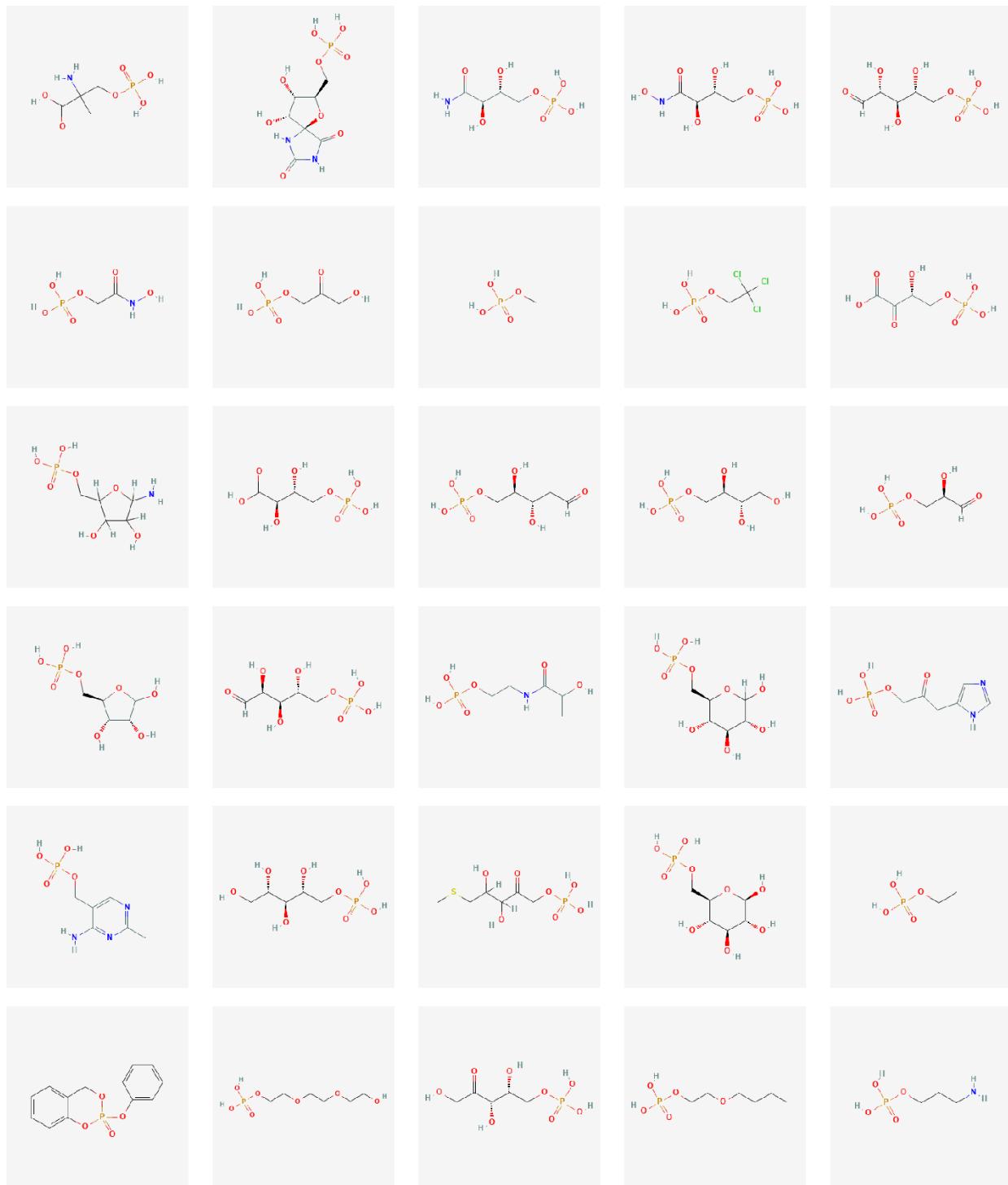


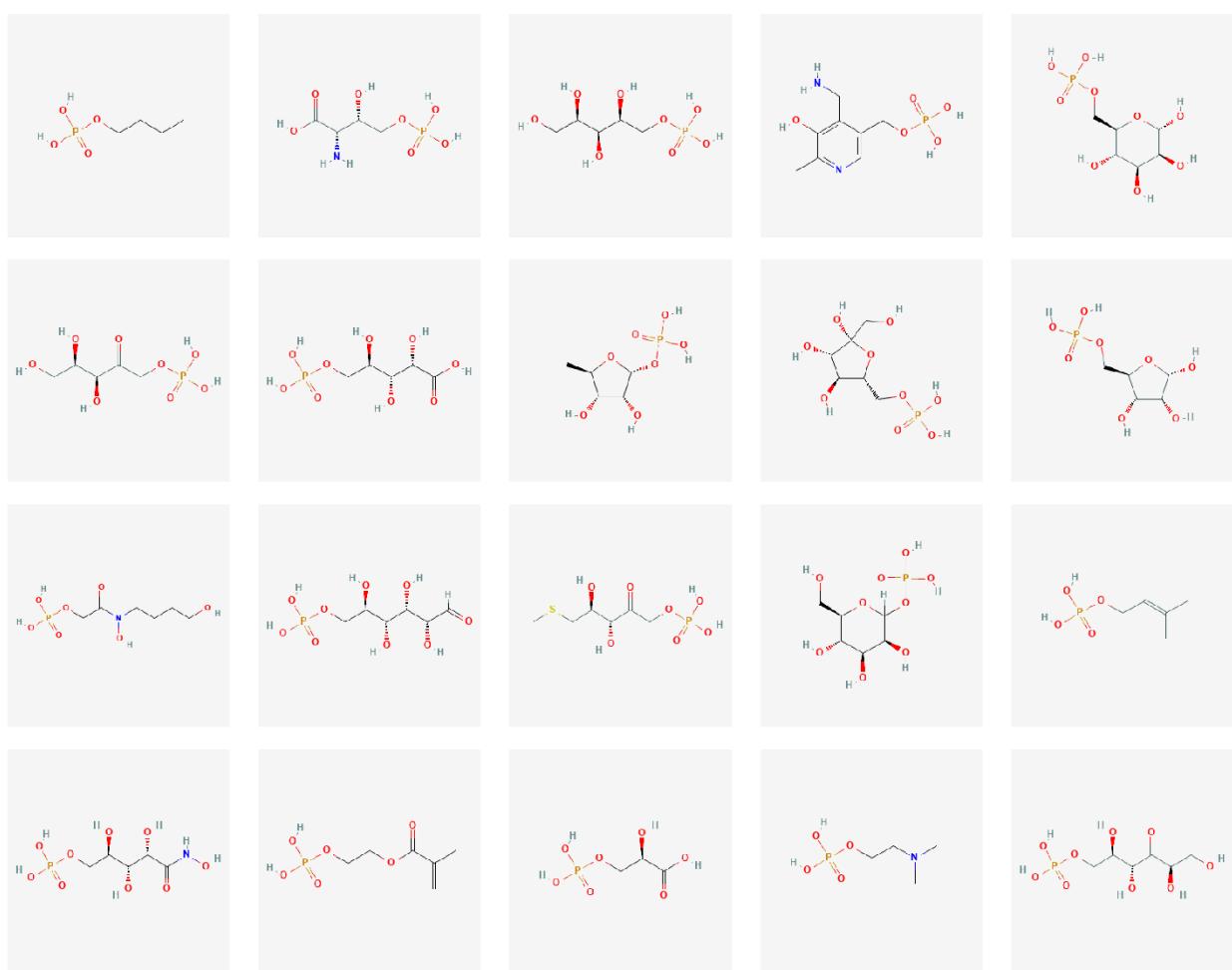
**Figure 7.S13** Compounds belonging phosphate sub-cluster I.

## Phosphate sub-cluster II structures



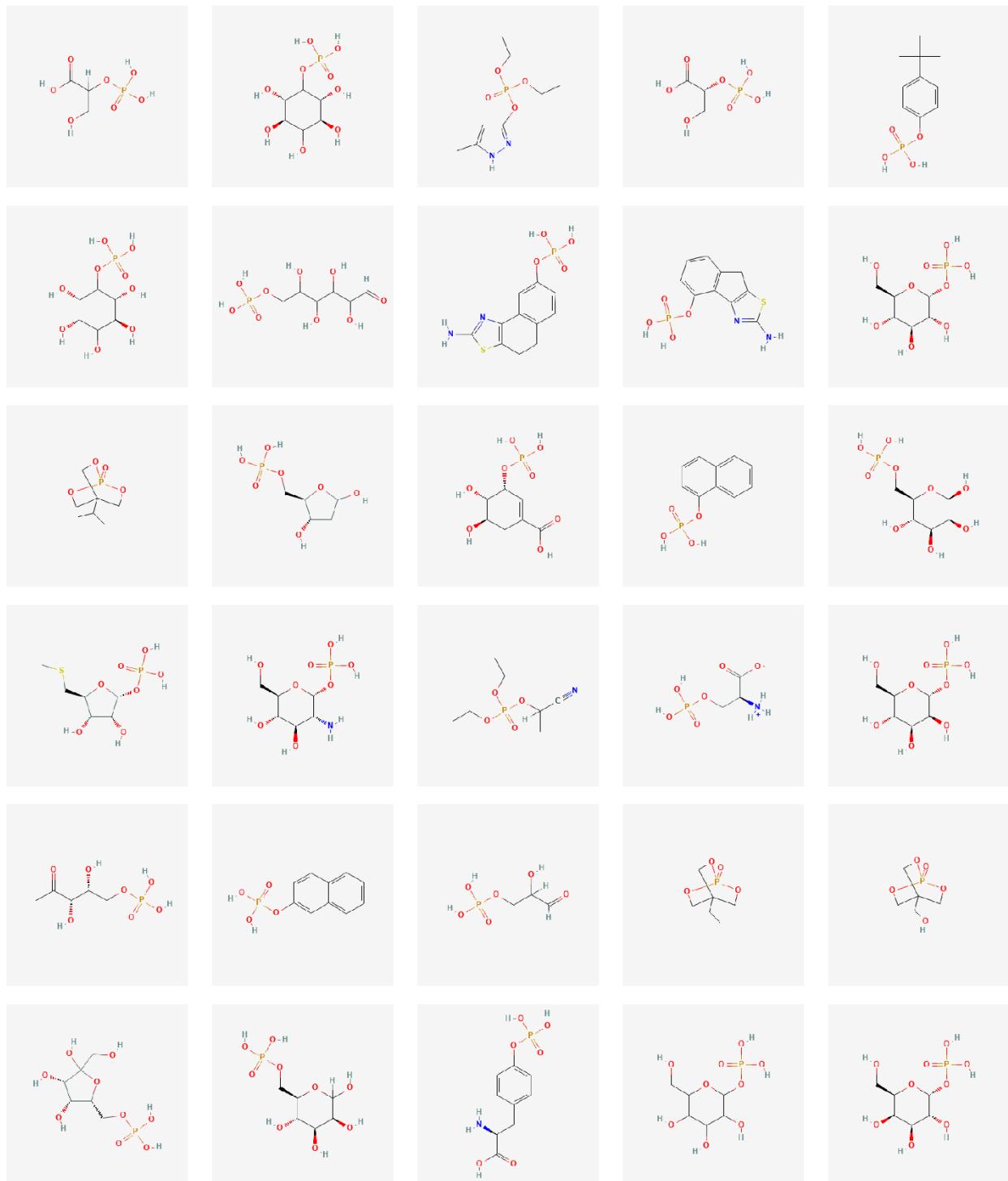


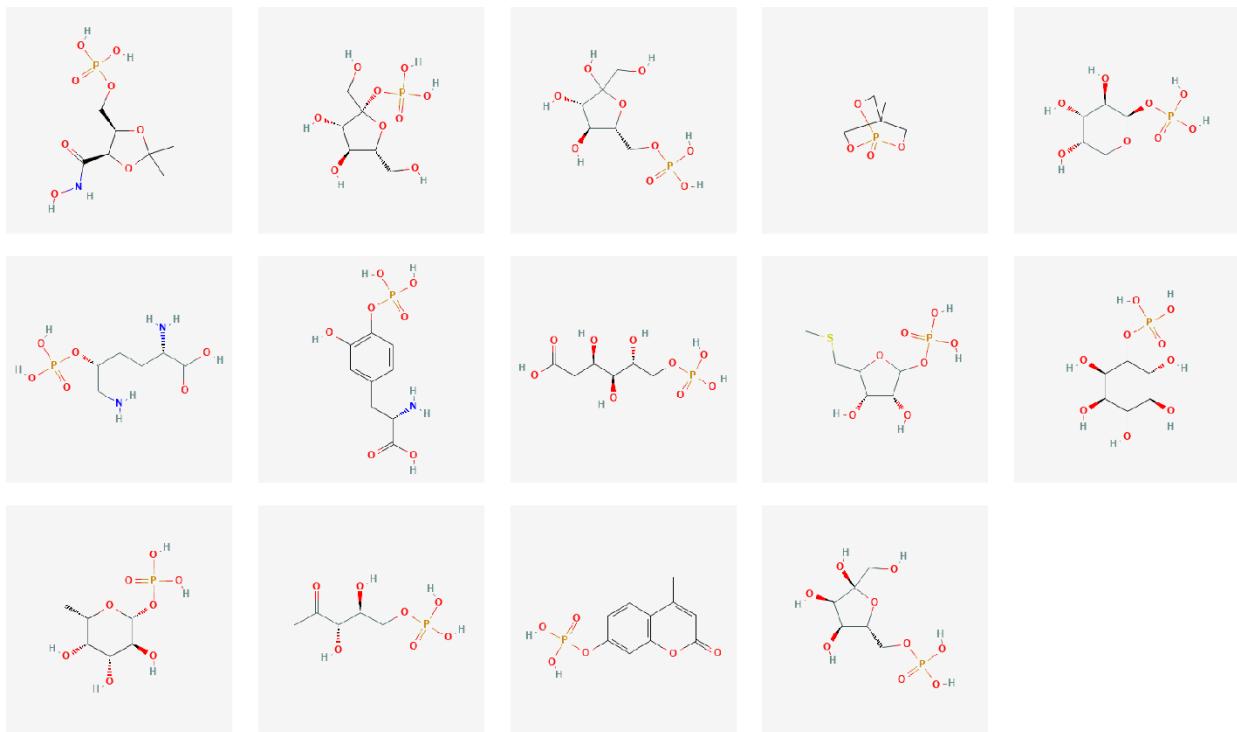




**Figure 7.S14** Compounds belonging phosphate sub-cluster II.

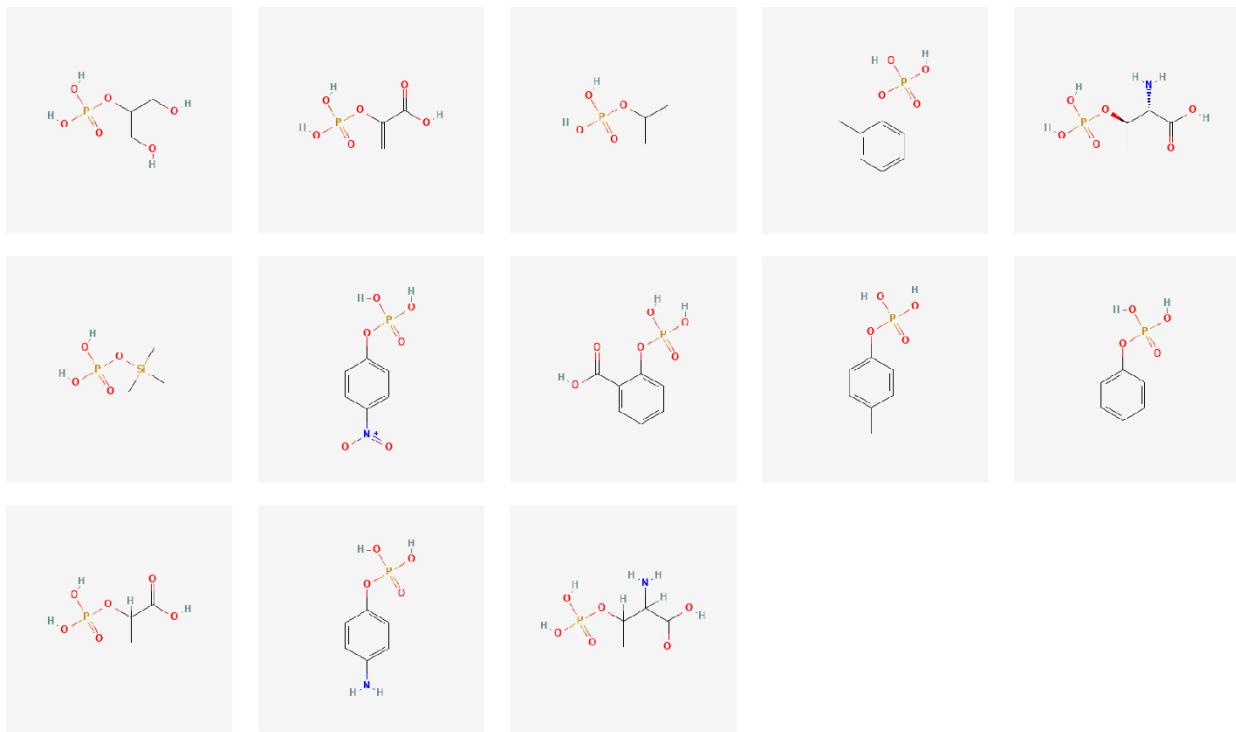
### Phosphate sub-cluster III structures





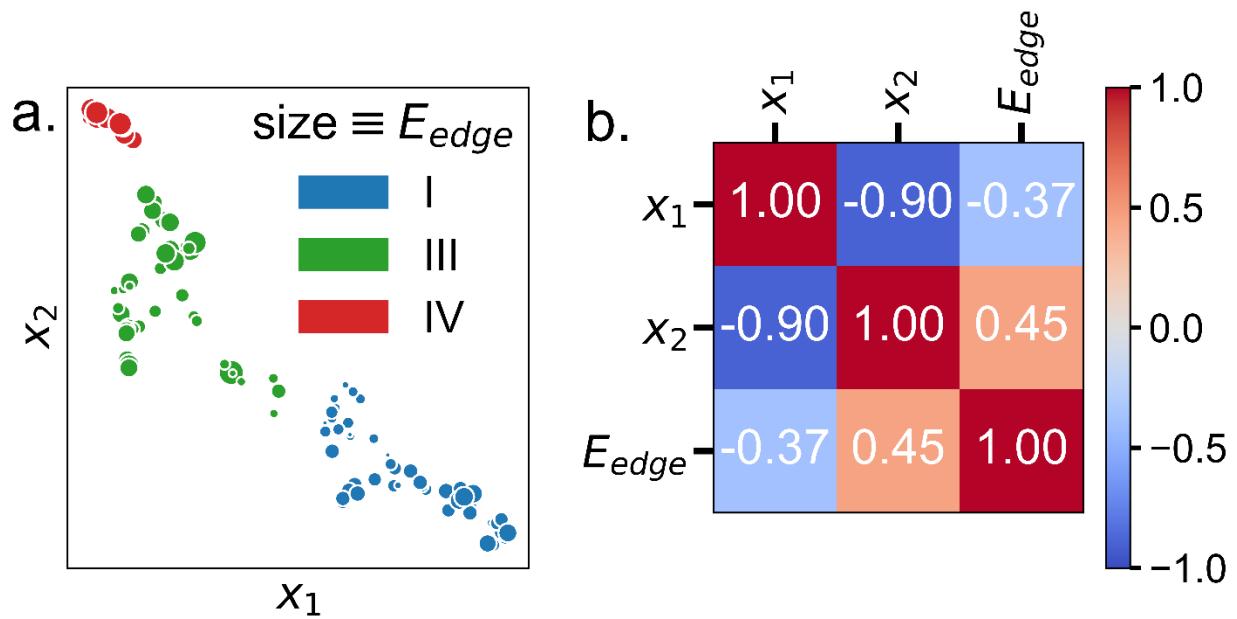
**Figure 7.S15** Compounds belonging phosphate sub-cluster III.

### Phosphate sub-cluster IV structures



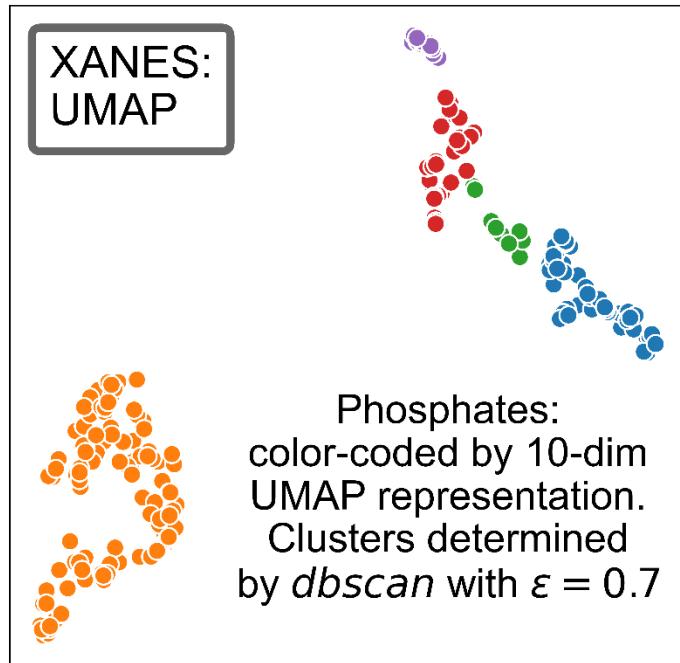
**Figure 7.S16** Compounds belonging phosphate sub-cluster IV.

Phosphate sub-clusters correlation



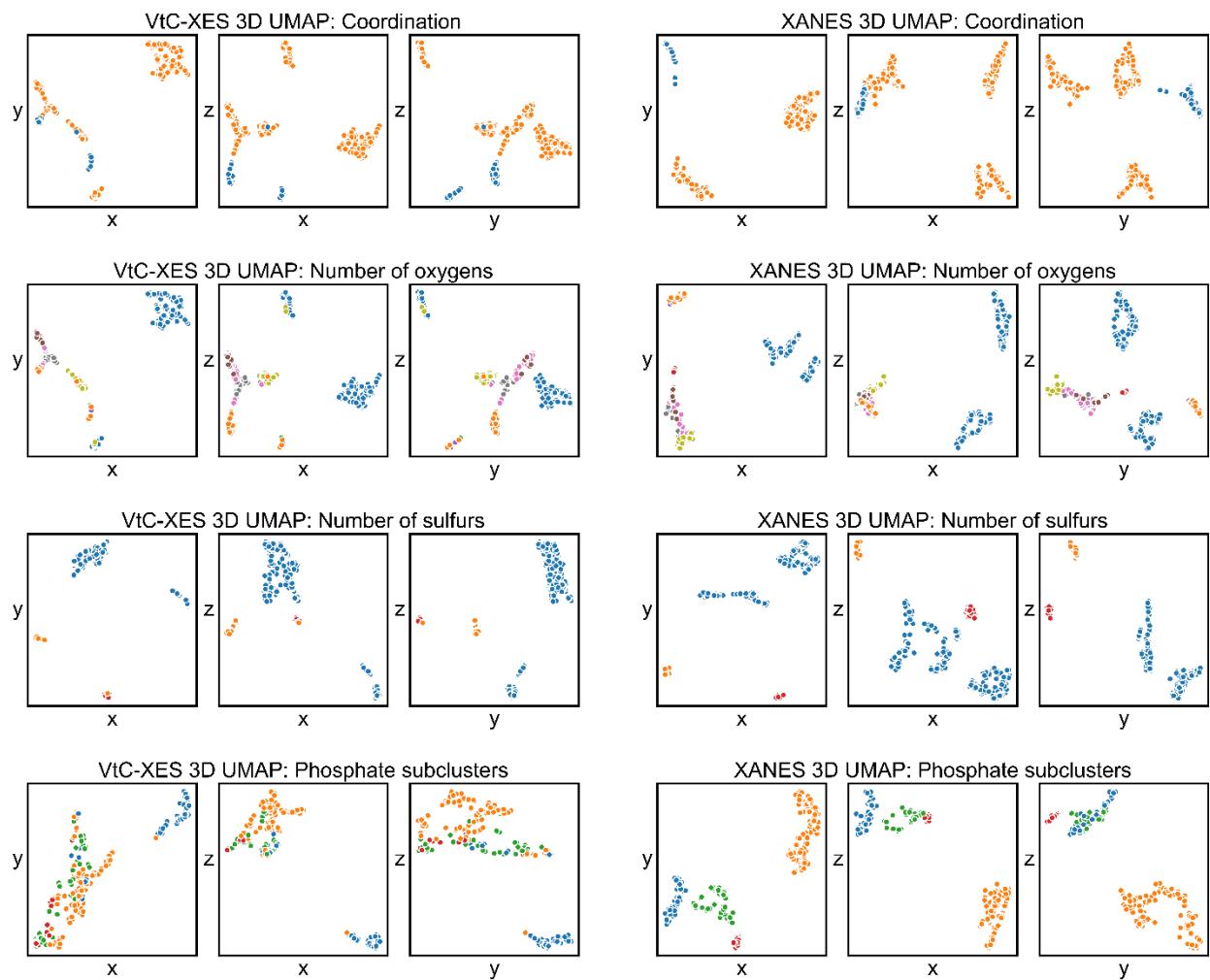
**Figure 7.S17** Phosphate sub-clusters **I**, **III**, and **IV** and their weak correlation to the energy of the absorption edge, defined as the spectral point with the greatest first derivative. (a) UMAP representation with points scaled to be different sizes based on the location of the edge, i.e., a higher energy edge yields a bigger data point. (b) Correlation matrix between the two UMAP axes and the edge energy.

Phosphates subclusters: 10-dim clustering



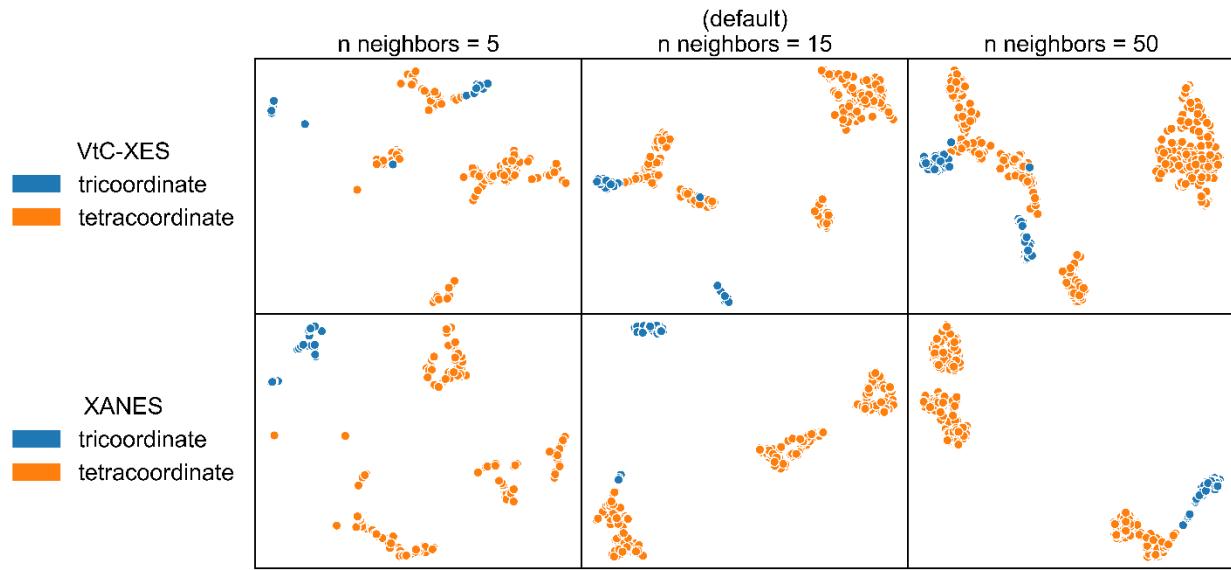
**Figure 7.S18** Two-dimensional visualization of phosphate clustering in 10-dimensions. The dbSCAN clustering algorithm generally clusters the phosphates in the same groups as clustering applied directly to the 2D representation (as shown in Fig. 6). Instead, here the Cluster **III** compounds are divided into two separate groups. Thus, the 2D representation of the phosphates is retaining the great majority of the information in the spectra, except likely a detail in Cluster **III** compounds that gets thrown away when reduced to so few dimensions. However, the overall classes are very similar and robust against UMAP dimension.

### 3D UMAP visualizations



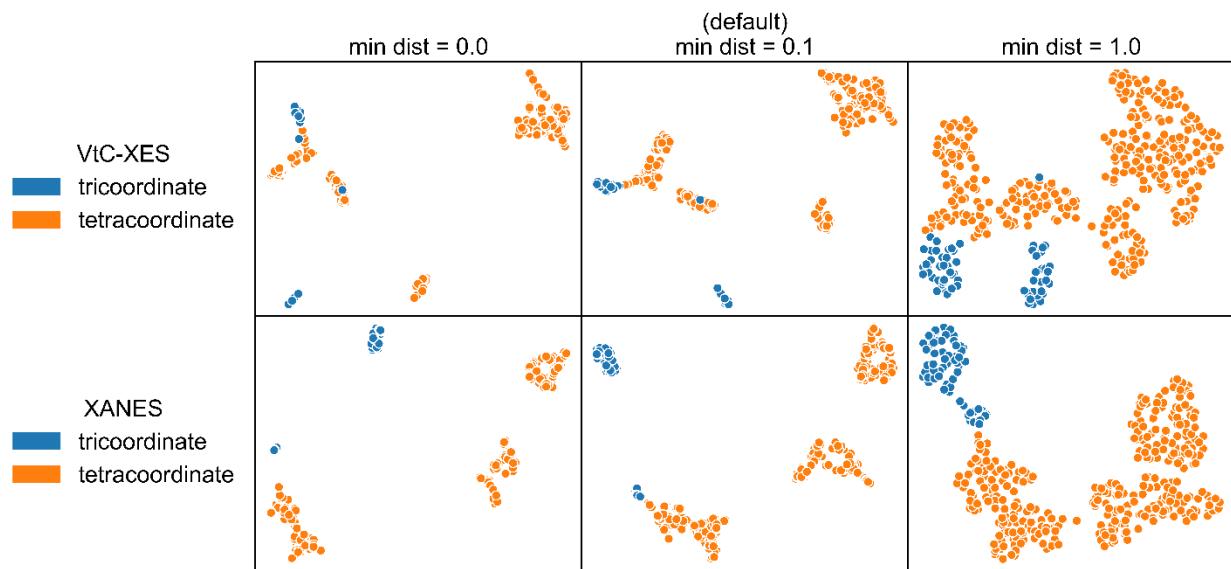
**Figure 7.S19** Three-dimensional UMAP projections for various classification schemes for both the VtC-XES (left three panels) and XANES (right three panels). Most clustering in 3D seems to be the same as the 2D embeddings in the main text, indicating that a two-dimensional embedding captures most of the useful cluster information as three dimensions.

### Changing UMAP hyperparameters: number of neighbors



**Figure 7.S20** Two-dimensional UMAP projections for both VtC-XES (top) and XANES (bottom) of tricoordinate P and tetracoordinate P compounds when varying the hyperparameter number of expected neighbors in a cluster. The number of neighbors balances the global versus local structure preserved by UMAP, and it is similar to the perplexity hyperparameter in t-SNE. For small `n_neighbors` values, local similarities are stressed, while large values stress global similarities (at the cost of losing fine details)<sup>4</sup>.

### Changing UMAP hyperparameters: minimum distance



**Figure 7.S21** Two-dimensional UMAP projections for both VtC-XES (top) and XANES (bottom) of tricoordinate P and tetracoordinate P compounds when varying the hyperparameter minimum distance between points. The minimum distance hyperparameter controls how tightly packed points in the reduced space can be. Generally low minimum distance values focus on more detailed topological structure, while large values stress broad topological structure. Smaller values of minimum distance are thus better for clear clusters for our analysis<sup>4</sup>.

### 7.6.1 References

1. E. Apra, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Attafynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauet, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Fruchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Gotz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jonsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. M. del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, A. Vazquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Wolinski, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison, *J. Chem. Phys.*, 2020, **152**, 26.
2. W. M. Holden, E. P. Jahrman, N. Govind and G. T. Seidler, *The Journal of Physical Chemistry A*, 2020, **124** (26), 5415-5434.
3. I. Persson, W. Klysubun and D. Lundberg, *Journal of Molecular Structure*, 2019, **1179**, 608-611.
4. L. McInnes, J. Healy and J. Melville, *arXiv*, 2020.

## 8 Chapter 8 – Manifold Projection Image Segmentation for Nano-XANES Imaging

S. Tetef, A. Pattammattel, Y. S. Chu, M. K. Y. Chan, G. T. Seidler. Manifold Projection Image Segmentation for Nano-XANES Imaging. *In preparation.* A. Pattammattel conducted the experiment and collected all data. S. Tetef wrote the text and conducted all data analysis.

*As spectral imaging techniques are becoming more prominent in science, advanced image segmentation algorithms are required to identify appropriate domains in these images. We present a version of image segmentation called manifold projection image segmentation (MPIS) that is generally applicable to a broad range of systems without the need for training because MPIS uses unsupervised machine learning with a few physically motivated hyperparameters. We apply MPIS to nano-XANES imaging, where X-ray Absorption Near Edge Structure (XANES) spectra are collected with nanometer spatial resolution. We show the superiority of manifold projection over linear transformations, such as the commonly used Principal Component Analysis (PCA). Moreover, MPIS maintains accuracy while reducing computation time and sensitivity to noise compared to the standard nano-XANES imaging analysis procedure. Finally, we demonstrate how multimodal information, such as X-ray Fluorescence (XRF) data and spatial location of pixels, can be incorporated into the MPIS framework. We propose that MPIS is adaptable for any spectral imaging technique where the length scale of domains is larger than the resolution of the experiment.*

## 8.1 Introduction

The increased popularity in various scientific fields of utilizing high-throughput imaging techniques, especially spectral imaging experiments, has benefited from advanced image segmentation algorithms so that researchers can identify regions in the image belonging to the same domain, object, phase, etc. Image segmentation methods that utilize multimodal characterization measurements as input, which potentially could be high-dimensional, are especially beneficial for the scientific community <sup>1-4</sup>. However, most common image segmentation algorithms utilize either hand-crafted rules or convolutional neural networks, both of which can suffer from lack of generalizability. Moreover, not enough training data or unreliable data simulations may make transfer learning unrealistic when using neural networks. Instead, an alternative is to utilize manifold projection and clustering based on spectral similarity rather than deep learning, effectively performing *semantic* image segmentation. This manifold projection image segmentation, which we will refer to as MPIS, has seen success when applied to mass spectroscopy images <sup>5</sup> and flow cytometry data <sup>6</sup>.

Here, we apply MPIS to a hyperspectral imaging technique called nanoscale X-ray absorption near edge structure (nano-XANES) <sup>7-12</sup>. XANES is a common experimental technique in materials science, chemistry, and biology as it is sensitive to local electronic structure around a chosen atomic species <sup>13</sup>. Nano-XANES imaging is a fitting application for MPIS. The goal of nano-XANES imaging is often to generate a compositional map of the local coordination, or phase, of the element of interest. The most common practice to make these maps is to perform linear combination fitting (LCF) to a reference library of spectra at every pixel, treating the image as an ensemble of independent XANES spectra and ignoring the spatial location of each spectrum. The analysis of XANES spectra using LCF is highly constrained by the prior knowledge of the system

as well as the limited information encoded in the spectra. Uncertainty in the system can lead to an overly large library with poor linear independence. Only after LCF are the fit results used to construct the spatial phase maps. This approach is slow by requiring many fits, and any errors in the LCF fitting process are propagated when creating the spatial maps.

We show that MPIS has three major benefits compared to the above standard practice. First, by implementing MPIS, we flip the order of generating spatial domains and identifying the compositions of those domains. By switching this order and decoupling the image segmentation from the LCF, one can substitute improved or specialized classification or regression techniques as needed while maintaining persistent image segmentation, or domain identification, via MPIS. Therefore, phase maps are independent of the selection of a reference library and any errors in the LCF results. Second, MPIS can cluster the reference library in the context of the experimental spectra. Because the reference library can be a large set of numerically similar spectra, researchers often report LCF fits by grouping reference spectra by chemical class. MPIS instead provides a data-driven way to group references together.

Furthermore, MPIS is adaptable for encoding multimodal data into the image segmentation pipeline. In almost all nano-XANES imaging studies, XRF maps are simultaneously acquired for every XANES spectrum, producing a higher dimensional dataset enabling both spectroscopic and elemental analysis. We show that MPIS applied to an augmented encoding of both XANES and XRF spectra can better separate low signal-to-noise from high signal-to-noise data. Furthermore, encoding the position of the pixel into MPIS can generate smaller domain regions – divided by spatial location rather than global groupings – that is more akin to instance image segmentation, for example separating out each physical particle in the same phase. Finally, to force sparsity in the fits, the standard LCF practice uses stepwise regression. We instead substitute stepwise

regression with LASSO regression, as presented in Jahrman et al.<sup>14</sup>, to speed up computations. We additionally perform LCF on the cluster-averaged spectra specified by MPIS. By having a data-informed way to average spectra together without losing spatial resolution, our LCF is more robust to noise.

We propose MPIS can be broadly applied to a wide range of spectroscopy techniques, including multimodal experiments. While we demonstrate MPIS on a nano-XANES image, MPIS can be used to cluster any ensemble-based measurement because the pixel location in the image is encoded as optional multimodal information. Furthermore, MPIS decreases the chances of overfitting by requiring fewer, and physically meaningful, hyperparameters compared to deep learning. Finally, MPIS increases the reliability and efficiency of high-throughput analysis by speeding up computations and reducing sensitivity to noise in subsequent analysis.

## 8.2 Methods

### 8.2.1 Experimental Methods

Our sample is composed of Lithium iron phosphate (LFP), pyrite, hematite, and stainless steel. The Lithium iron phosphate (LFP) and pyrite (Pyr) were purchased from Sigma Aldrich, St Louis, MO. Hematite (Hem) and stainless-steel (SS) nanoparticles were obtained from US Nano Research (Houston, TX, USA). A heterogeneous mixture of the above-mentioned particles was created by physically mixing in acetone, followed by ultrasonication for 5 minutes. About 5 mL of the dispersed mixture was drop-casted onto a silicon nitride membrane (Norcia, Edmonton, Canada) and the solvent was dried in air. All data was collected at the Hard X-ray Nanoprobe (HXN) Beamline at National Synchrotron Light Source II (NSLS-II) at Brookhaven National

Laboratory<sup>15, 16</sup>. A detailed methodology for nano-XANES acquisition was previously published.

12

We measured Fe K-edge XANES using nano-XANES imaging, where our data consists of a 3D image with 155 x 160 spatial components and photon energy in the Z axis. In other words, every pixel in the image has a XANES spectrum with just over 70 energies measured between 7.08 to 7.20 keV. The nano-XANES data of the present sample is comprised of about 25,000 spectra. Further processing of the stack was performed via the XMIDAS program<sup>17</sup>. The energy stack was first spatially aligned using the image registration tool in XMIDAS that uses the PyStackReg package<sup>18</sup>. Spectra are preprocessed via normalization and alignment using the standard procedures<sup>19</sup>). Finally, the spectra were assembled to create a 3D array (Fe energy stack) for XANES analysis. For each scan point, an energy X-ray Fluorescence spectrum was collected with a three-element silicon drift detector (Vortex, Hitachi Inc) positioned at 90 deg to the sample. The XRF spectra were processed using the PyXRF software<sup>20</sup> to compute elemental maps.

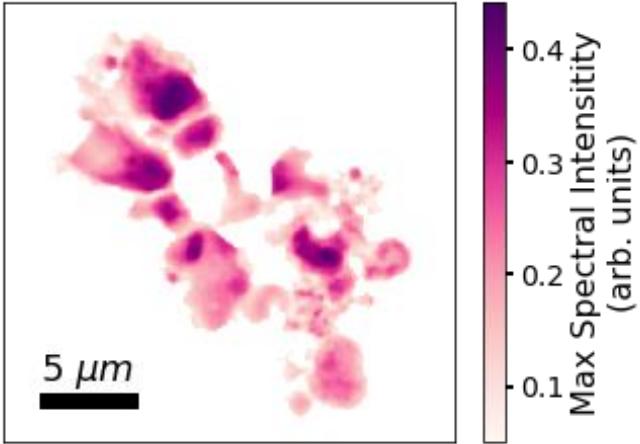
### 8.2.2 Computational Methods

Uniform Manifold Approximation and Projection (UMAP)<sup>21</sup> was implemented using the umap-learn Python package. UMAP requires two main hyperparameters – the number of neighbors (to control cluster sizes and thus global versus local similarity) and the minimum distance between points in the cluster (to control how tightly packed the clusters are). For all UMAP spaces, we set the minimum distance to zero (for the tightest-packed clusters possible). We set the number of neighbors to be between 20 and 80. Changes with this hyperparameter within a reasonable range (20 to 80) did not change the clustering results.

Principal Component Analysis (PCA)<sup>22</sup>, k-means clustering, and dbSCAN were implemented using `sklearn`. Although PCA does not require any hyperparameters, a scree plot was used to determine the number of principal components to keep given a specified threshold of explained variance. The value of  $k$  (the number of clusters) in k-means was determined to be between 3 and 6 such that it appeared to qualitatively distinguish the original nano-XANES image while reasonably explaining the reduced space. The clustering approach dbSCAN uses the epsilon hyperparameter, which we set to be one for all UMAP spaces. We qualitatively checked that this epsilon value appropriately labelled the UMAP clusters by visualizing the UMAP space color-coded by the dbSCAN labels.

### 8.3 Results and Discussion

A two-dimensional display of our sample, colored by maximum XANES intensity (and thus identifying regions with the highest photon counts) is shown in Fig. 8.1, where background spectra are filtered out such that only the sample region is examined. This sample – and thus dataset – is the same as the one found in Pattammattel, et al.<sup>17</sup> Each “pixel” (150 nm wide) represents a processed XANES spectrum.



**Figure 8.1** Nano-XANES map, color-coded by the maximum spectral intensity of the Fe K-edge XANES spectra (to indicate the most likely places with sample due to the high photon counts). Each pixel is 150 nm. Note that background spectra are filtered out.

Fig. 8.2 demonstrates our MPIS procedure in relation to both the standard nano-XANES analysis and LCF procedures. To start the MPIS procedure, we first apply Principal Component Analysis (PCA)<sup>22</sup> to the pre-processed ensemble of XANES spectra. Next, we have the option to encode multimodal data. Either we exclusively take the coefficients of the six highest principal components (determined by a scree plot, see Methods), or we use the joint encoding of those principal component coefficients with multimodal information.

To be specific, the multimodal encoding starts with an array of the principal component coefficients, i.e.,

$$\vec{S}(x, y) = (PC_1, \dots, PC_6) \quad (8.1)$$

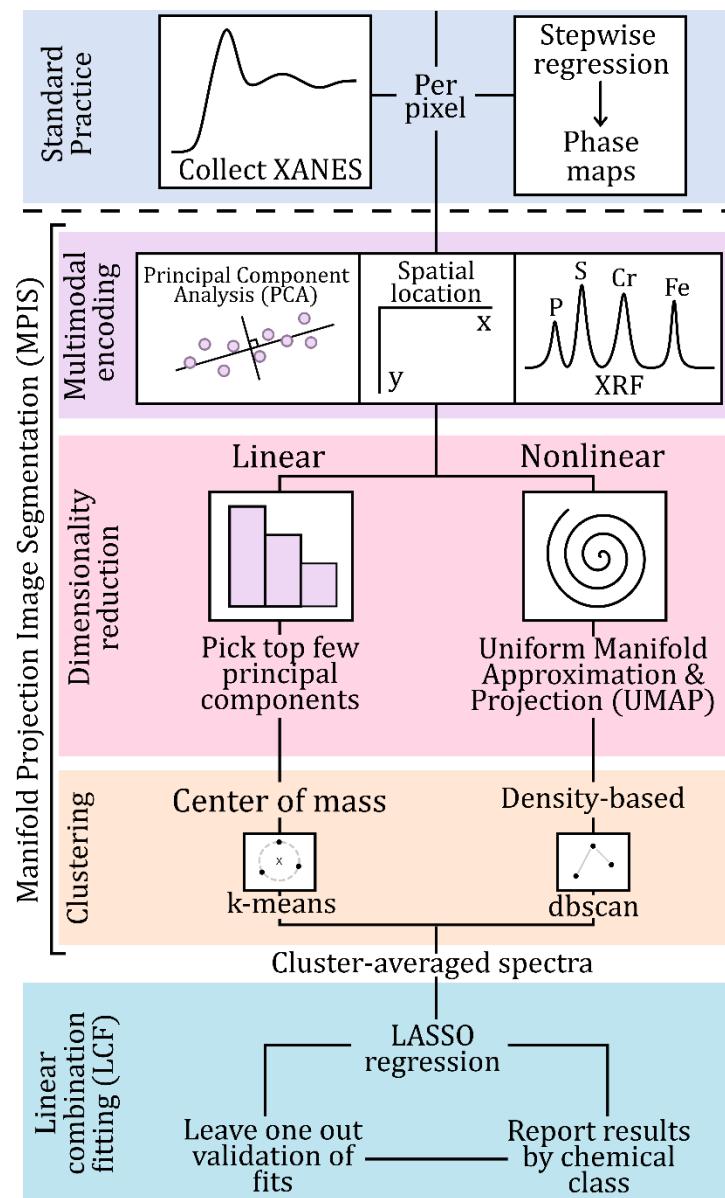
Then, the spatial location and/or XRF of the four elements are appended to that array. In its most complex case, where both spatial location and XRF are jointly encoded, the encoding takes the form of the following vector:

$$\vec{S}(x, y) = (PC_1, \dots, PC_6, I_P^{XRF}, I_S^{XRF}, I_{Cr}^{XRF}, I_{Fe}^{XRF}, x, y) \quad (8.2)$$

where the first six components belong to the coefficients of the first size principal components, the next four components are the normalized XRF data (each of the P, S, and Cr XRF maps are divided by the Fe XRF map), and the last two components belong to the x and y positions (which are scaled to be between 0 and 1). The relative importance of the different components is then tuned by two new hyperparameters  $\alpha$  and  $\beta$  dictating the informational strength or the importance of the XRF and spatial location, respectively. Thus, the above encoding is implemented as follows:

$$\vec{S}(x, y) = (PC_1, \dots, PC_6, \alpha I_P^{XRF}, \alpha I_S^{XRF}, \alpha I_{Cr}^{XRF}, \alpha I_{Fe}^{XRF}, \beta x, \beta y) \quad (8.3)$$

where  $\alpha$  and  $\beta$  represent an independent scaling of importance for each distinct multimodal measurement. This procedure can be easily extended to encode other types of multimodal information.

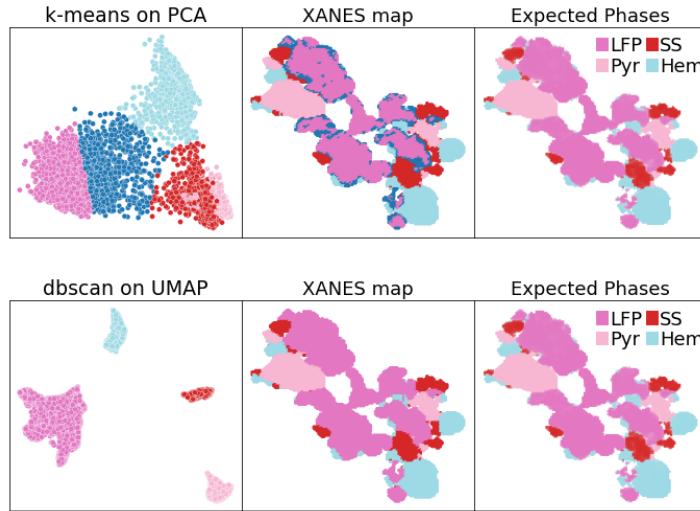


**Figure 8.2** Our manifold projection image segmentation (MPIS) and linear combination fitting (LCF) pipeline for analyzing our nano-XANES image.

We then take the (multimodal) encoding and pass it to a nonlinear dimensionality reduction routine to identify spectral clusters. In general, applying dimensionality reduction before clustering increases the reliability of the clustering labels by combating the “curse of dimensionality” (as

opposed to clustering applied directly to the spectra). We compared using a linear routine – namely PCA – to a nonlinear routine – namely Uniform Manifold Approximation and Projection (UMAP)<sup>21</sup> – when performing the dimensionality reduction step of MPIS. Prior work has shown that nonlinear dimensionality reduction, compared to linear, does better at disentangling the inherently nonlinear spectral features in X-ray absorption spectroscopy<sup>23, 24</sup>, albeit linear routines are often sufficient<sup>25, 26</sup>. However, maintaining PCA as a preparation step for UMAP speeds up UMAP and filters out unimportant noise in the spectra, as shown in Fig. 8.S1

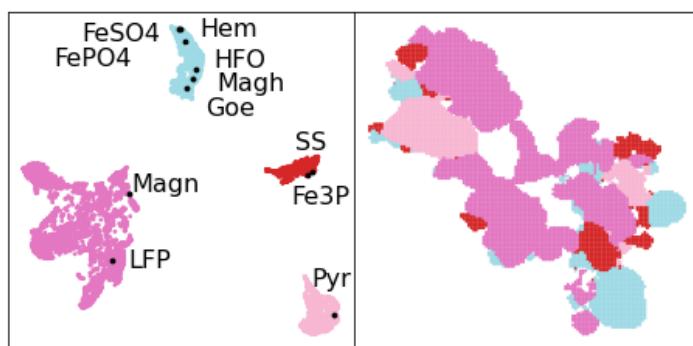
While a center-of-mass-based clustering algorithm such as k-means<sup>27</sup> pairs well with PCA, we opted for a density-based cluster algorithm called dbSCAN<sup>28</sup>) for the nonlinear embedding via UMAP. To see the effectiveness of UMAP and dbSCAN for clustering as opposed to PCA and k-means, see Fig. 8.3. The left panels in Fig. 8.3 show distinct and well separated clusters when UMAP and dbSCAN are used as opposed to overlapping or non-separated clusters using PCA and k-means. Moreover, k-means needed five clusters to appropriately group the data in the PCA space, which is larger than the expected four known phases. Although Fig. 8.3 shows a two-dimensional projection of the data, we used six principal components in our MPIS pipeline as six principal components explained 97% of the variance of the data. See Figs. 8.S2-5 for the triangle plots that visualize the PCA and UMAP hypercubes we used as well as other supplementary figures relating to MPIS.



**Figure 8.3** (top) k-means clustering on the first two principal components. (bottom) dbSCAN clustering on a two-dimensional UMAP embedding. The clusters and labeling in the two-dimensional UMAP representation not only match expectations, but they are easier to see and thus interpret than the k-means clusters on the top two PCA components.

Finally, we identified the composition of each cluster by performing linear combination fitting (LCF) to a reference library using the MPIS cluster-averaged spectra. To do so, we utilized the procedure presented in Jahrman, et al.<sup>14</sup>, which utilizes Least Absolute Selection and Shrinkage Operator (LASSO) regression instead of stepwise regression. However, instead of bootstrapping our data to generate estimates in uncertainty, we utilized leave-one-out validation. Specifically, we refit each spectrum with one reference spectrum in the library removed at a time and noted the changes in the fit results. In addition to speeding up computation time by avoiding stepwise regression, we reduced the total number of fits by performing LCF on the average spectrum for each cluster rather than the spectrum at every pixel individually. See the Supplementary Information for details on LASSO regression. In brief, hyperparameters were chosen via 5-fold cross validation.

Our reference library for LCF was composed of both the known phases – LiFe(II)PO<sub>4</sub> (LFP), pyrite, stainless steel, and hematite – and additional mineral phases to model a typical experiment, namely HFO (hydrinous ferric oxyhydroxide), goethite, maghemite, magnetite, Fe<sub>3</sub>P, Fe(III)PO<sub>4</sub>, and Fe(III)SO<sub>4</sub>. Specifically, hematite, goethite, maghemite, and HFO are all oxides and have very similar spectra, while Fe<sub>3</sub>P has the same oxidation state as elemental Fe, which is the same as stainless steel. The selection of this library was based on a quick XRF measurement and the availability of experimental reference spectra. Moreover, this reference library represents a realistic uncertainty for chemical speciation of Fe-phases in heterogeneous samples with *a priori* knowledge.



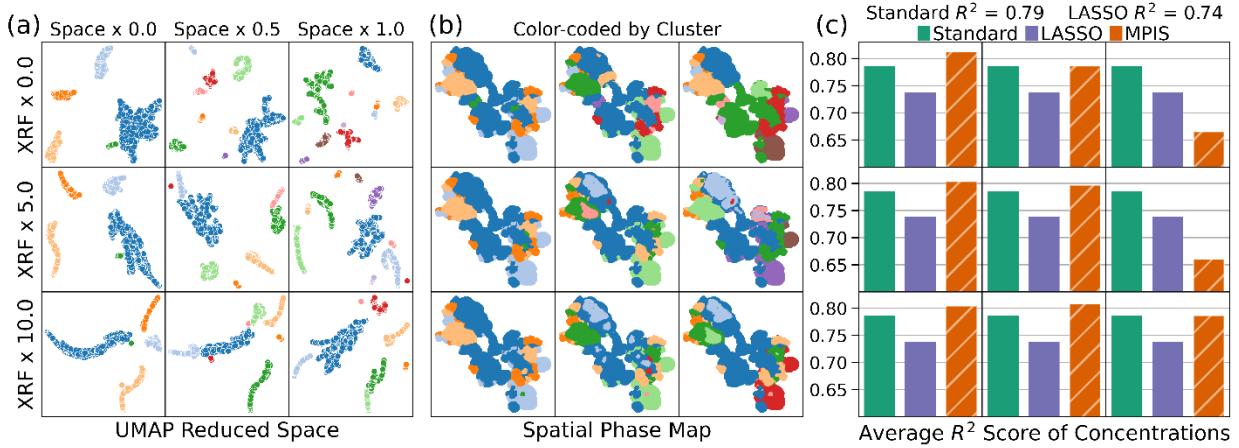
**Figure 8.4** Reference chemical classes. Often, LCF results are reported using the chemical class of the references. These classes are usually created using chemical knowledge of the system. Instead, we offer a completely data-driven way one can generate these classes, specifically by projecting references onto the UMAP space determined by the experimental spectra.

Finally, we reported LCF fit results by the chemical classes for the references, which we developed by projecting the reference spectra onto the UMAP space of the experimental data, as shown in Fig. 8.4. We divided the references from the same cluster if the XRF would be able to distinguish between references. For example, while Fe<sub>3</sub>P and stainless steel appeared in the same

cluster, they are theoretically distinguishable using both the P and Cr XRF data. Following this procedure, all references were split into their own class besides the “oxides” – hematite, maghemite, HFO, and goethite – which were grouped into one combined class. To see the correlation matrix for references, see Fig. 8.S6.

We hypothesized that applying MPIS and LASSO regression rather than pixel-by-pixel stepwise regression would speed up computation time while maintaining accuracy. We ran both procedures and found MPIS took about 30 seconds compared to the standard pixel-by-pixel stepwise regression procedure of enumerating all quaternary combinations of 11 references, which took 4 minutes (using 8 GB RAM on an Intel i5 CPU). However, the time complexity of stepwise regression fits grows as  $O(n^k)$  given the reference library size  $n$  and the combination size  $k$ , so a larger reference library will greatly increase computation time.

We then compared the effect of encoding the XRF and spatial location of every spectrum into MPIS on the LCF results, as shown in Fig 5. The uppermost left panel shows the LCF results with no multimodal encoding in MPIS. We compared predicted coefficients using the “standard” approach (non-negative least squares for every pixel), likewise using “LASSO” for every pixel, and then using “MPIS” and predicting concentrations from cluster averages. These coefficients were scored against the “true” concentrations, which were obtained by non-negative least squares per pixel using only the four known phases in the reference library (rather than all 11 as in the “standard” case).

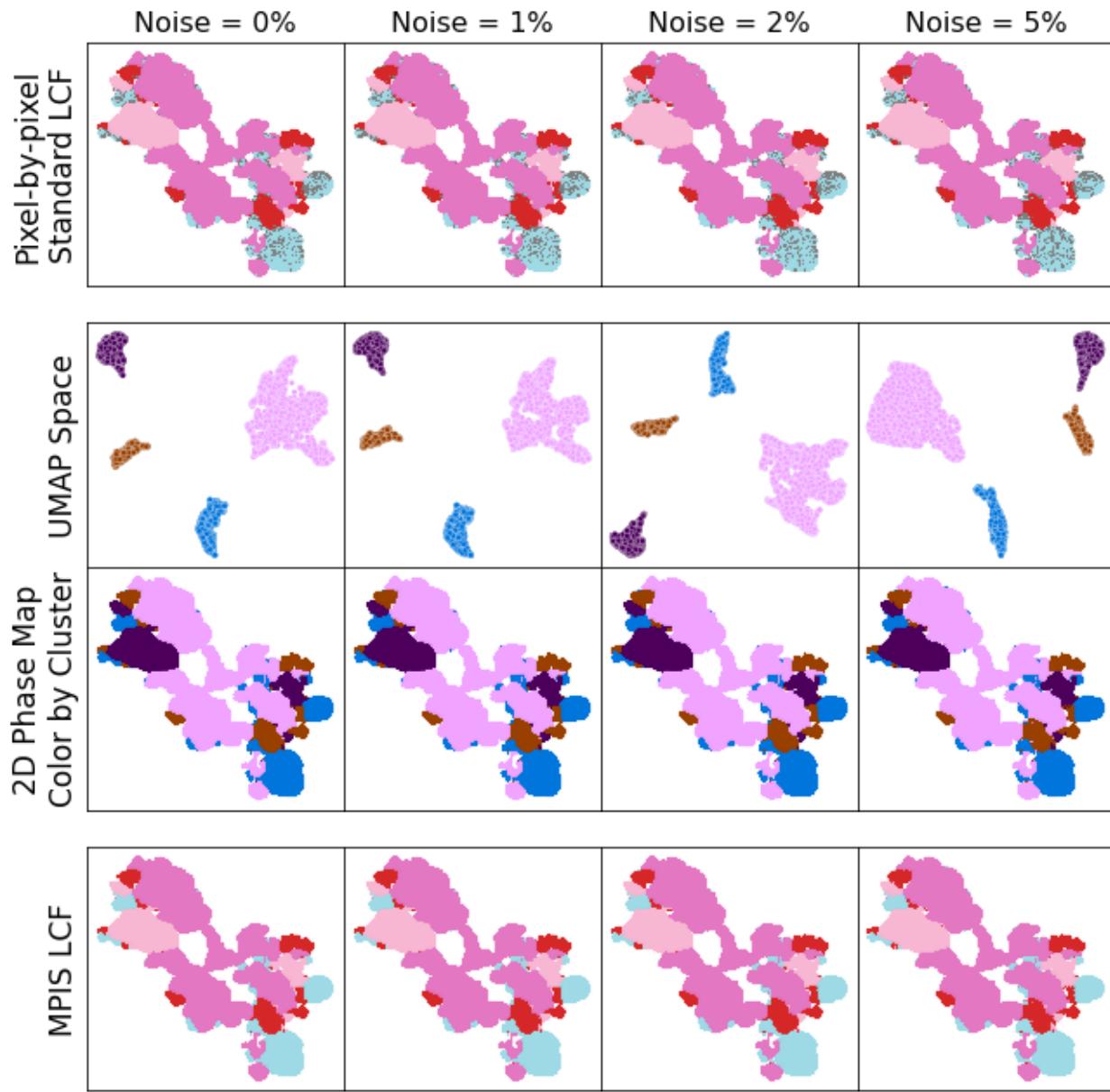


**Figure 8.5** (a) Effects on the clusters when encoding XRF data and spatial location into the MPIS pipeline. (b) The resulting 2D phase maps, colored by cluster. (c) Score of linear combination fitting (LCF) predictions via the standard pixel-by-pixel analysis (“Standard”), pixel-by-pixel LASSO regression (“LASSO”), and LASSO regression via MPIS (“MPIS”). The upper leftmost panel shows no joint information encoding.

In general, MPIS scored just as well as the standard approach (if the spatial strength is not too large) but in less time. Moreover, by identifying domains, fits are less sensitive to uncorrelated noise in the dataset. To demonstrate this effect, we augmented the experimental spectra with additional uncorrelated Gaussian noise with increasing intensity and compared domain identification using the standard pixel-by-pixel analysis with domains identified by MPIS, as shown in Fig. 8.6. Because the standard procedure constructs phase maps after LCF, there are spurious single-pixel phases when the noise is large. However, MPIS is more robust against these unphysical single-pixel fluctuations. Moreover, generating cluster-averaged spectra via MPIS is an informed way to average noisy spectra together for LCF fits without needing to lose resolution by Gaussian blurring the image. Furthermore, when noise is so high that MPIS on just the spectra

fails, encoding XRF and spatial location into MPIS recovers the analysis, as shown in Fig. 8.S7.

Fig. 8.S8 compares the average error in predicting concentrations.



**Figure 8.6** Adding noise (as a percentage of spectral intensity) to the experimental spectra causes pixel-by-pixel analysis to have small unphysical fluctuations in the phase maps, resulting from

uncertain LCF fits (top row). By applying MPIS (second row), the phase maps (third row) are more robust to noise, demonstrated by the consistent LCF results (bottom row).

## 8.4 Conclusions

The standard procedure for analyzing nano-XANES imaging is performing linear combination fitting (LCF) via stepwise regression for every pixel independently and then constructing phase maps using the fit results. Instead of stepwise regression, we encourage sparsity by performing LASSO regression to LCF onto our reference library. LASSO regression reduces computation time by decreasing the required number of fits. We also implement manifold projection image segmentation (MPIS) to cluster experimental spectra first before performing LCF, enabling LCF to only identify the composition of the phases rather than informing the spatial maps. By identifying domains first, we decouple the reliance on correct LCF fit results for appropriate phase maps, which can be greatly impacted by noise. The other benefit to using MPIS rather than the traditional deep learning approaches is that it requires fewer hyperparameters (which avoids overfitting), all of which are physically motivated. Moreover, MPIS is adaptable to include multimodal data, which we demonstrated by encoding X-ray Fluorescence and spatial location of pixels in addition to the XANES spectra. Because the spatial location of pixels is encoded as additional information, the basic procedure of MPIS can be applied to any ensemble-based spectroscopy measurements where clustering is important. Furthermore, we propose that MPIS can be applied to any spectral imaging measurement where the experimental resolution (pixel size) is smaller than the intrinsic length scale of domains.

## Acknowledgements

This work is supported by the U.S. Department of Energy (DOE) Office of Science Scientific User Facilities project titled “Integrated Platform for Multimodal Data Capture, Exploration and Discovery Driven by AI Tools.” M.C. acknowledges the support from the BES SUFD Early Career award. This research used Hard X-ray Nanoprobe beamline of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. Work performed at the Center for Nanoscale Materials, a U.S. Department of Energy Office of Science User Facility, was supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. All code is available through GitHub at [github.com/stetef/nano-XANES-microscopy-of-Fe<sup>29</sup>](https://github.com/stetef/nano-XANES-microscopy-of-Fe).

## 8.5 References

1. C. Zhang, J. Zhou, H. Wang, T. Tan, M. Cui, Z. Huang, P. Wang and L. Zhang, *Journal*, 2022, **14**.
2. M. K. Jakubowski, W. Li, Q. Guo and M. Kelly, *Journal*, 2013, **5**, 4163-4186.
3. C. Yapp, E. Novikov, W.-D. Jang, T. Vallius, Y.-A. Chen, M. Cicconet, Z. Maliga, C. A. Jacobson, D. Wei, S. Santagata, H. Pfister and P. K. Sorger, *Communications Biology*, 2022, **5**, 1263.
4. W. C. Schwartzkopf, A. C. Bovik and B. L. Evans, *IEEE Transactions on Medical Imaging*, 2005, **24**, 1593-1610.
5. H. Hu, R. Yin, H. M. Brown and J. Laskin, *Analytical Chemistry*, 2021, **93**, 3477-3485.
6. I. Stolarek, A. Samelak-Czajka, M. Figlerowicz and P. Jackowiak, *iScience*, 2022, **25**.
7. I. Nakai, C. Numako, S. Hayakawa and A. Tsuchiyama, *Journal of Trace and Microprobe Techniques*, 1998, **16**, 87-98.
8. R. Belissont, M. Munoz, M. C. Boiron, B. Luais and O. Mathon, *Minerals*, 2019, **9**.
9. M. Cusack, Y. Dauphin, J. P. Cuif, M. Salome, A. Freer and H. Yin, *Chemical Geology*, 2008, **253**, 172-179.
10. M. Bonnin-Mosbah, N. Métrich, J. Susini, M. Salomé, D. Massare and B. Menez, *Spectrochimica Acta Part B: Atomic Spectroscopy*, 2002, **57**, 711-725.
11. L. Mino, E. Borfecchia, C. Groppo, D. Castelli, G. Martinez-Criado, R. Spiess and C. Lamberti, *Catalysis Today*, 2014, **229**, 72-79.

12. A. Pattammattel, R. Tappero, M. Ge, Y. S. Chu, X. Huang, Y. Gao and H. Yan, *Science Advances*, 2020, **6**, eabb3615.
13. G. Bunker, *Introduction to XAFS: A Practical Guide to X-ray Absorption Fine Structure Spectroscopy*, Cambridge University Press, Cambridge, 2010.
14. E. P. Jahrman, L. L. Yu, W. P. Krekelberg, D. A. Sheen, T. C. Allison and J. L. Molloy, *Journal of Analytical Atomic Spectrometry*, 2022, **37**, 1247-1258.
15. E. Nazaretski, H. Yan, K. Lauer, N. Bouet, X. Huang, W. Xu, J. Zhou, D. Shu, Y. Hwu and Y. S. Chu, *Journal of Synchrotron Radiation*, 2017, **24**, 1113-1119.
16. H. Yan, N. Bouet, J. Zhou, X. Huang, E. Nazaretski, W. Xu, A. P. Cocco, W. K. S. Chiu, K. S. Brinkman and Y. S. Chu, *Nano Futures*, 2018, **2**, 011001.
17. A. Pattammattel, R. Tappero, D. Gavrilov, H. Zhang, P. Aronstein, H. J. Forman, P. A. O'Day, H. Yan and Y. S. Chu, *Metallomics*, 2022, **14**, mfac078.
18. P. Thevenaz, U. E. Ruttimann and M. Unser, *IEEE Transactions on Image Processing*, 1998, **7**, 27-41.
19. B. Ravel and M. Newville, *Journal of Synchrotron Radiation*, 2005, **12**, 537-541.
20. L. Li, Y. Hanfei, X. Wei, Y. Dantong, H. Annie, L. Wah-Keat, I. C. Stuart and S. C. Yong, 2017.
21. L. McInnes, J. Healy and J. Melville, *arXiv*, 2020.
22. S. Wold, K. Esbensen and P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37-52.
23. S. Tetef, N. Govind and G. T. Seidler, *Phys. Chem. Chem. Phys.*, 2021, **23**, 23586-23601.
24. S. Tetef, V. Kashyap, W. M. Holden, A. Velian, N. Govind and G. T. Seidler, *The Journal of Physical Chemistry A*, 2022, **126**, 4862-4872.
25. M. Lerotic, C. Jacobsen, J. B. Gillow, A. J. Francis, S. Wirick, S. Vogt and J. Maser, *Journal of Electron Spectroscopy and Related Phenomena*, 2005, **144-147**, 1137-1143.
26. M. A. Marcus, *Journal of Electron Spectroscopy and Related Phenomena*, 2023, **264**, 147310.
27. K. P. Sinaga and M. S. Yang, *IEEE Access*, 2020, **8**, 80716-80727.
28. M. Hahsler, M. Piekenbrock and D. Doran, *Journal of Statistical Software*, 2019, **91**, 1 - 30.
29. [github.com/stetef/nano-XANES-microscopy-of-Fe](https://github.com/stetef/nano-XANES-microscopy-of-Fe).

## 8.6 Supplementary Information

### 8.6.1 Linear combination fitting objective function

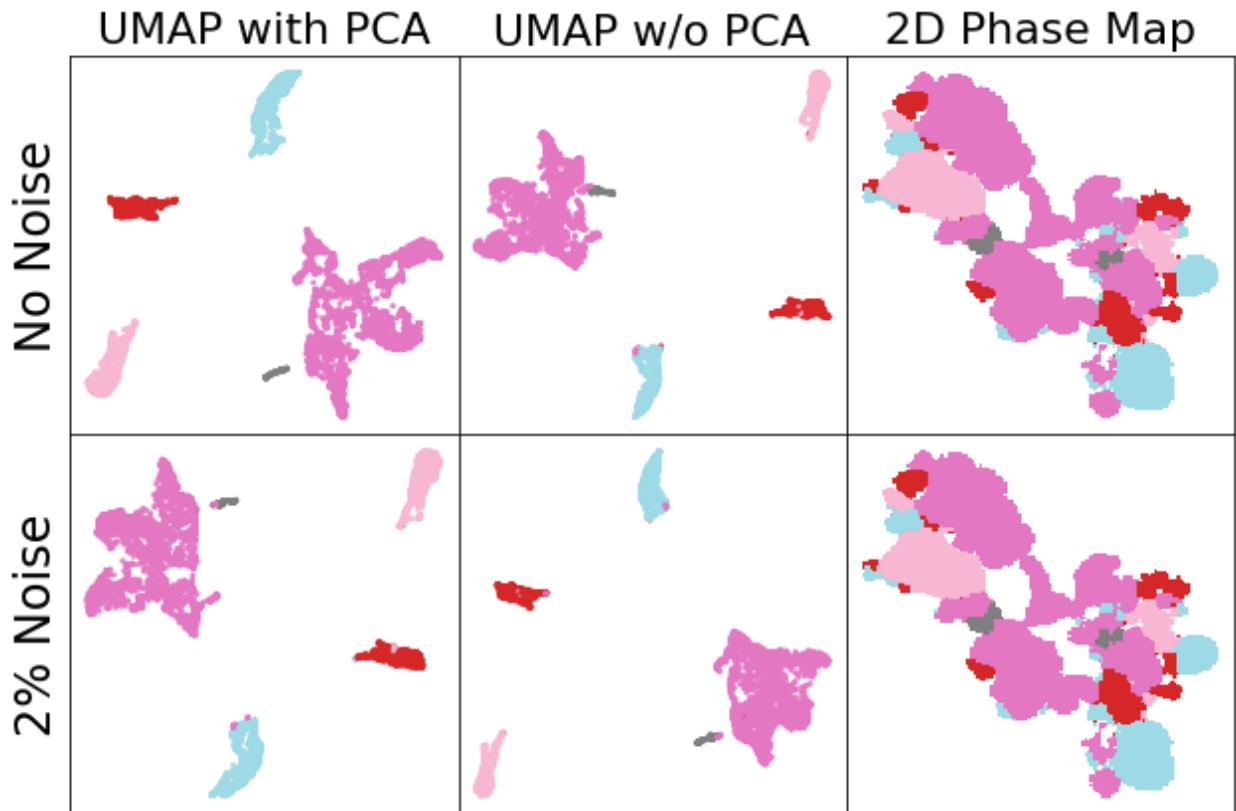
Our linear combination fitting (LCF) objective function is:

$$\widehat{\vec{c}_j} = \operatorname{argmin}_{\vec{c}_j} \left[ \frac{1}{2} \left\| \vec{y}_j - R^T \cdot \vec{c}_j \right\|_2^2 + \lambda_1 \left\| \vec{c}_j \right\|_1 + \lambda_2 \left\| 1 - \Sigma_i c_{ij} \right\|_2^2 \right] \quad (8.S1)$$

where  $\mathbf{y}$  is the unknown experimental spectra,  $R$  is the matrix composed of reference spectra, and  $\mathbf{c}$  is the coefficients contributing to the spectra. The first term represents reconstruction error (via a  $L_2$  norm, which is equivalent to a Euclidean distance metric), and the second term is the regularizer, modified by a Lagrange multiplier. The regularization was set to be the  $L_1$  norm to encourage sparsity. These terms effectively constitute LASSO regression. However, the input for each spectrum is the same – specifically the reference set  $R$  – so each spectrum is fit independently of the others.

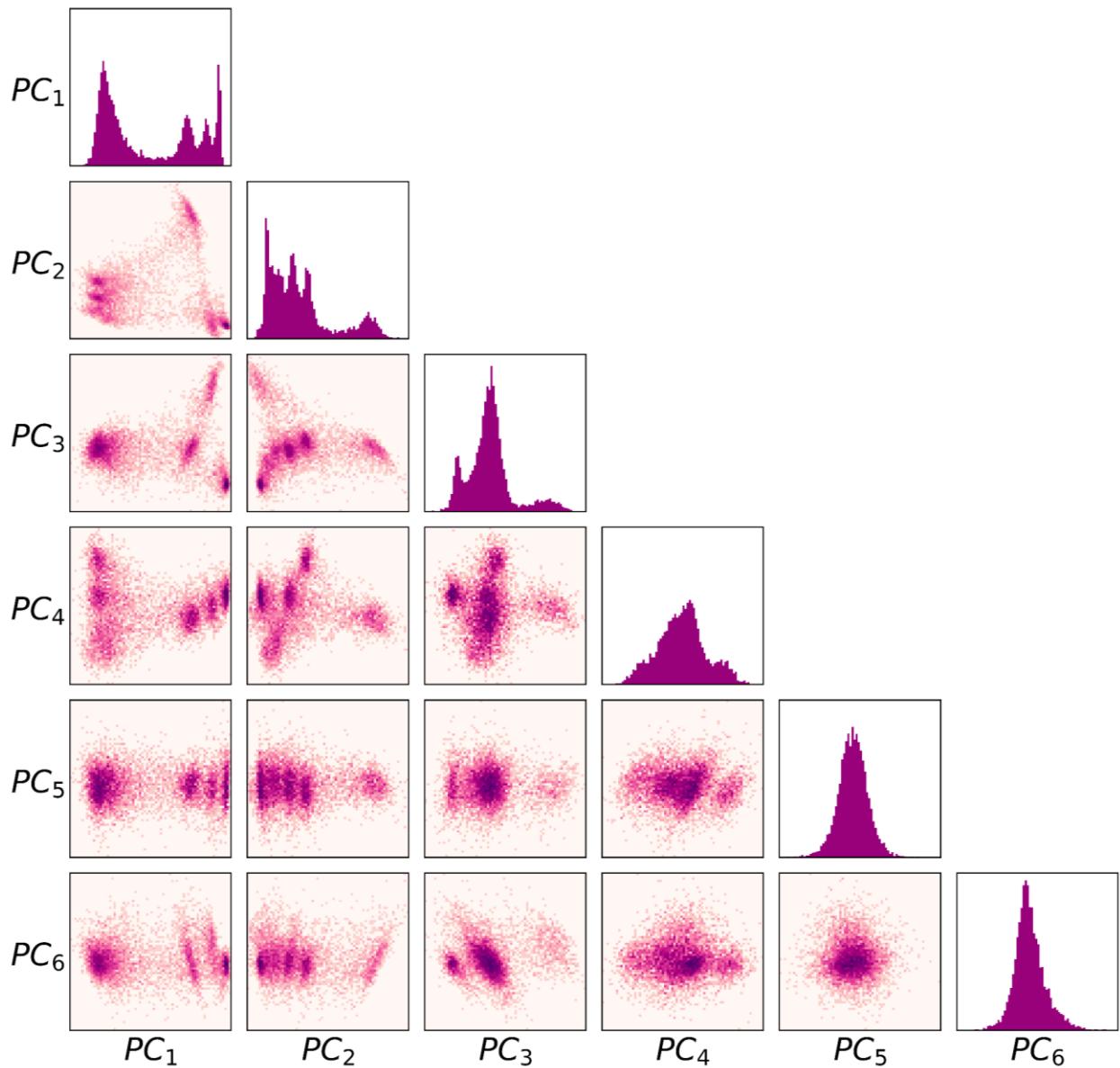
The hyperparameters for the fits were found via 5-fold cross-validation on a dataset composed of linear combinations of reference spectra (with forced sparsity and various levels of noise introduced). Specifically, we found a  $\lambda_1$  value of 0.0006 and  $\lambda_2$  value of 10 to be best, with consistent convergence onto a solution. Again, note that this objective function is minimized for every spectrum (or data point) and is therefore not trained on any training dataset as the input (the reference set  $R$ ) is the same for every fit. To minimize the above objective function, we used `scipy`'s `minimize` function with bounds on the weights to be in the range [0, 1]. Moreover, we used `scipy`'s built-in Sequential Least Squares Programming (SLSQP) optimization method, which is a quasi-Newton method.

UMAP spaces with and without PCA processing

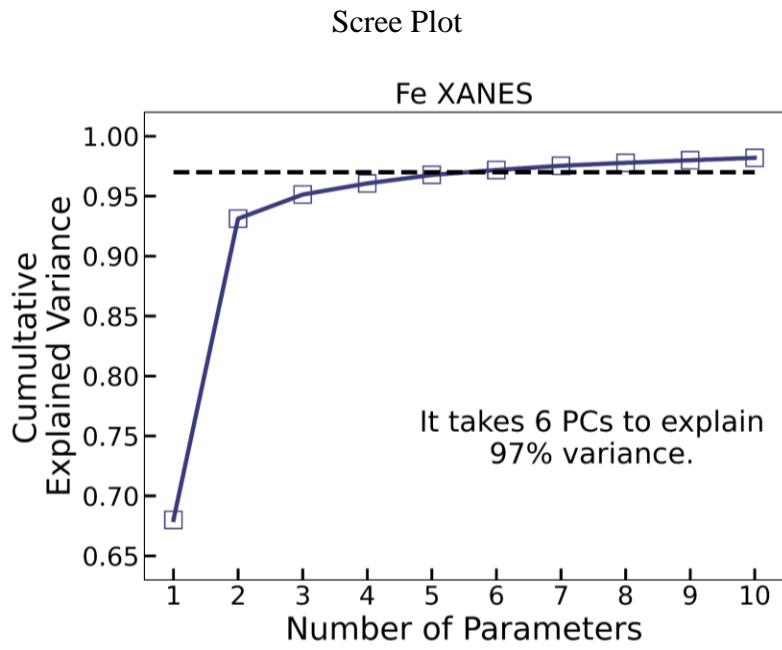


**Figure 8.S1** UMAP spaces with and without PCA processing. UMAP applied to the first 6 principal components produces clusters that are very similar to the clusters made when UMAP is applied directly to the spectra, both on the raw experimental data and with augmented noise added to the spectra.

PCA triangle plot

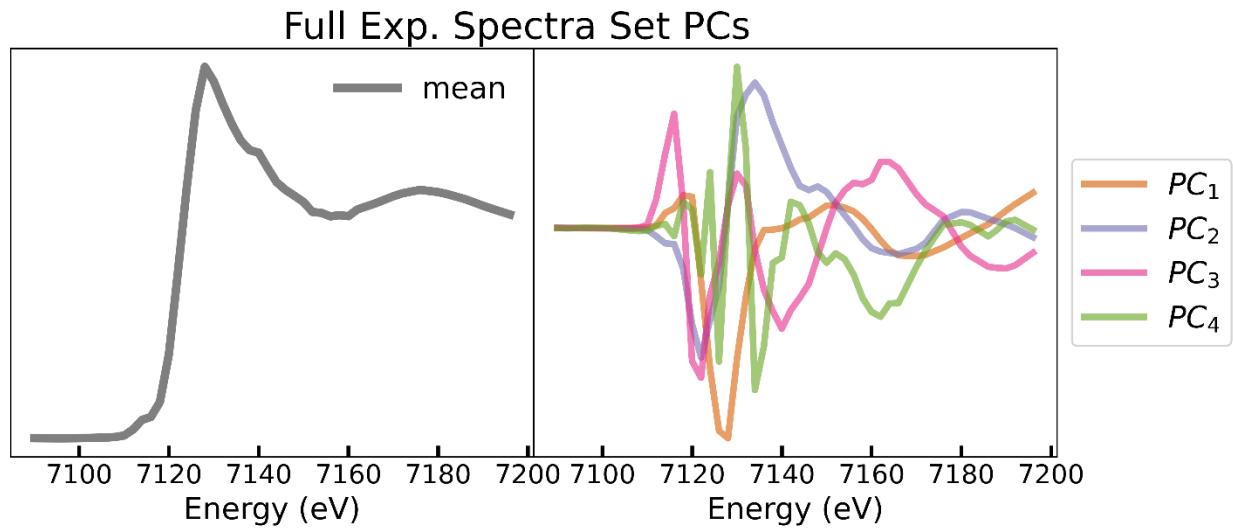


**Figure 8.S2** PCA triangle plot of all two-dimensional projects of the six-dimensional hypercube of the top principal components of the spectral dataset. Six dimensions were chosen because it takes the top six principal components (PCs) to explain 97% of the variance.



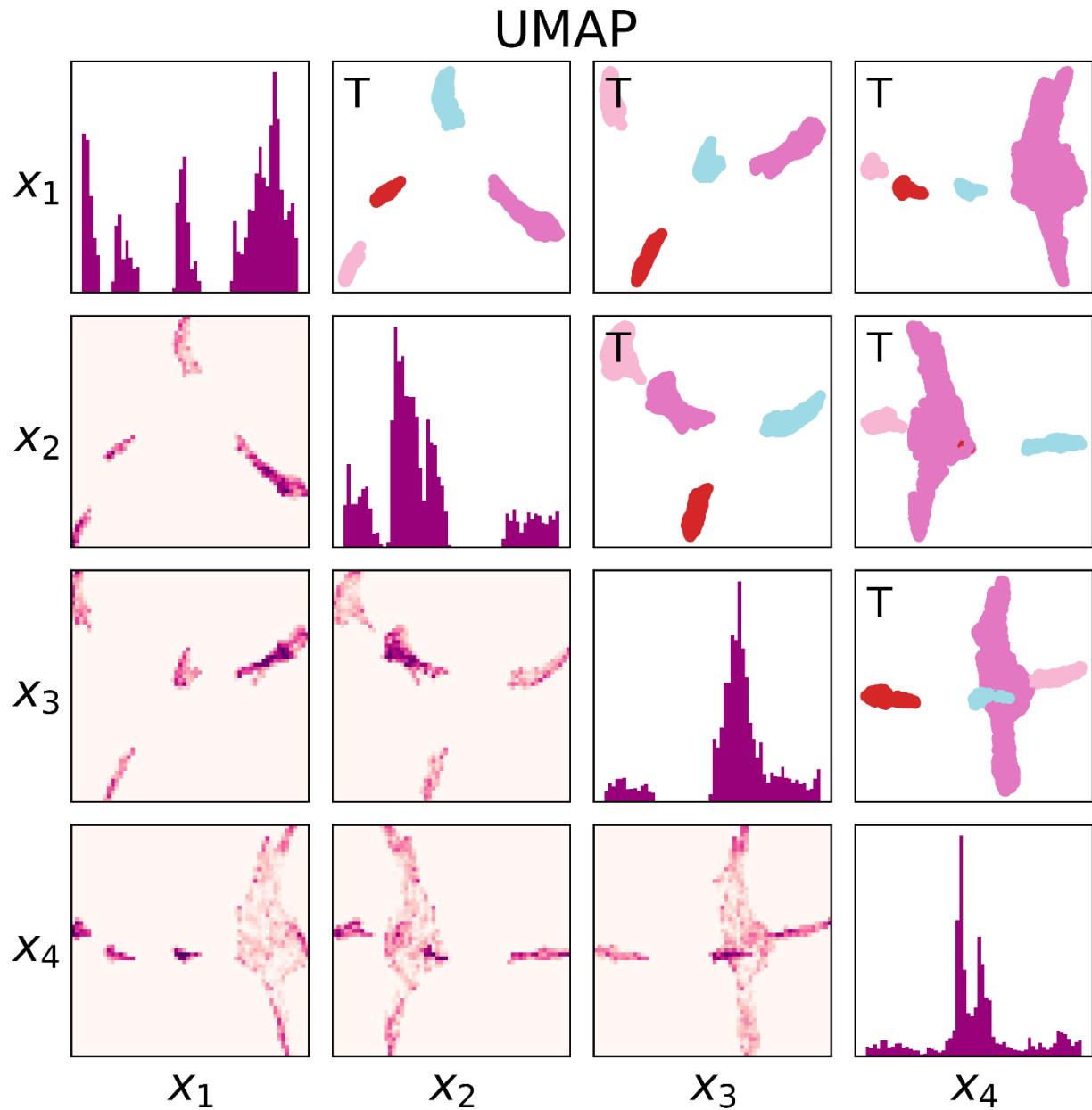
**Figure 8.S3** Scree plot of experimental spectra. It takes 6 PCs to explain 97% variance (dashed line).

First four principal components



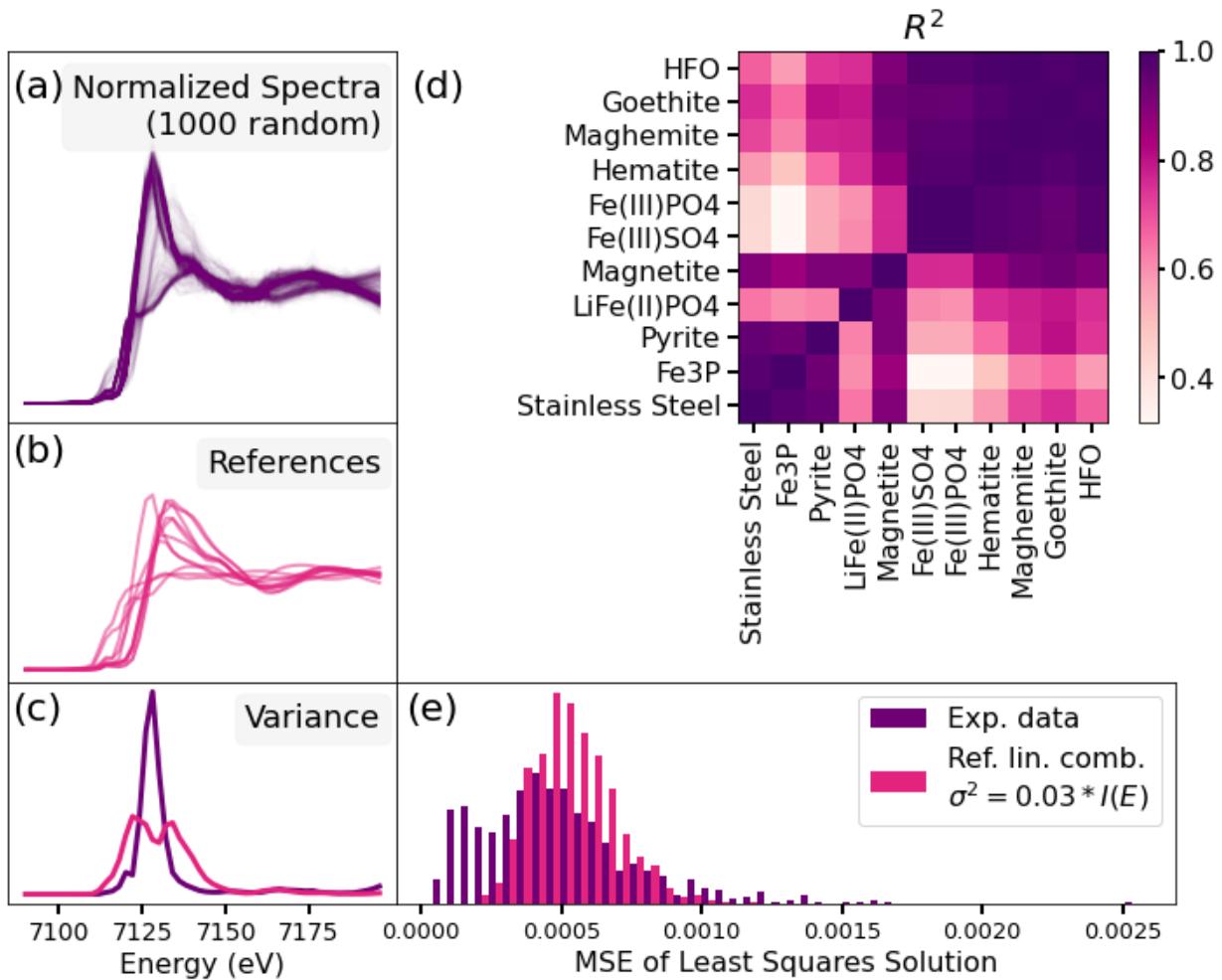
**Figure 8.S4** First four principal components of the measured spectra.

### Four-dimensional UMAP hypercube



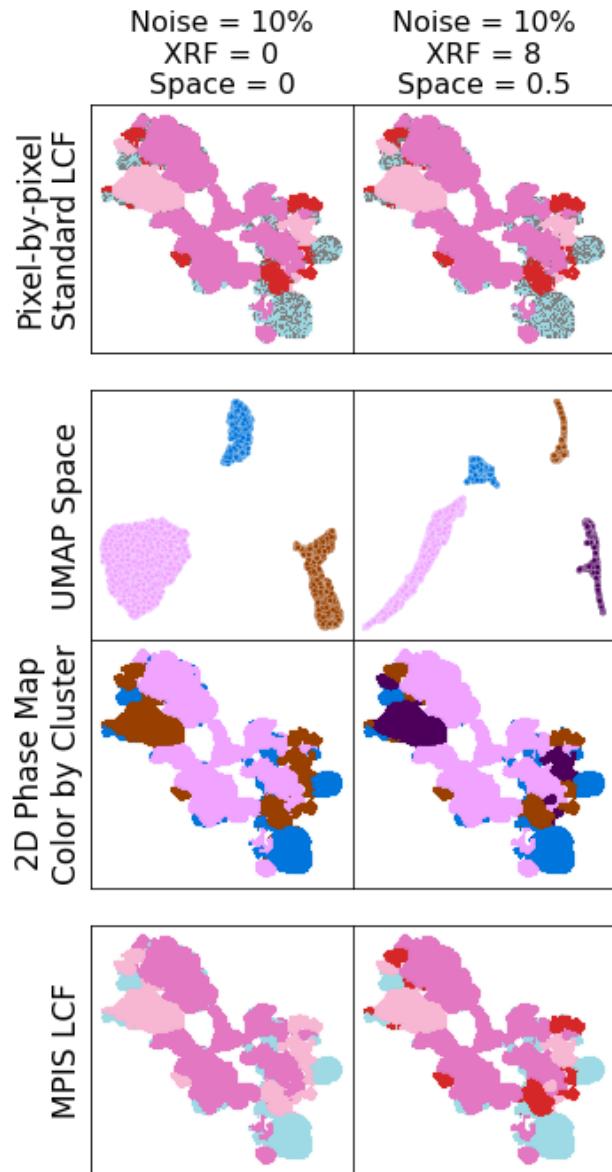
**Figure 8.S5** Four-dimensional UMAP hypercube (applied to the top PCA components), with two-dimensional projections (color-coded by density) shown in the bottom left corner. The upper right corner is composed of the same projections (transposed so that the upper and lower triangles match), except instead color-coded by the dbSCAN clustering labels.

### Correlation of reference spectra



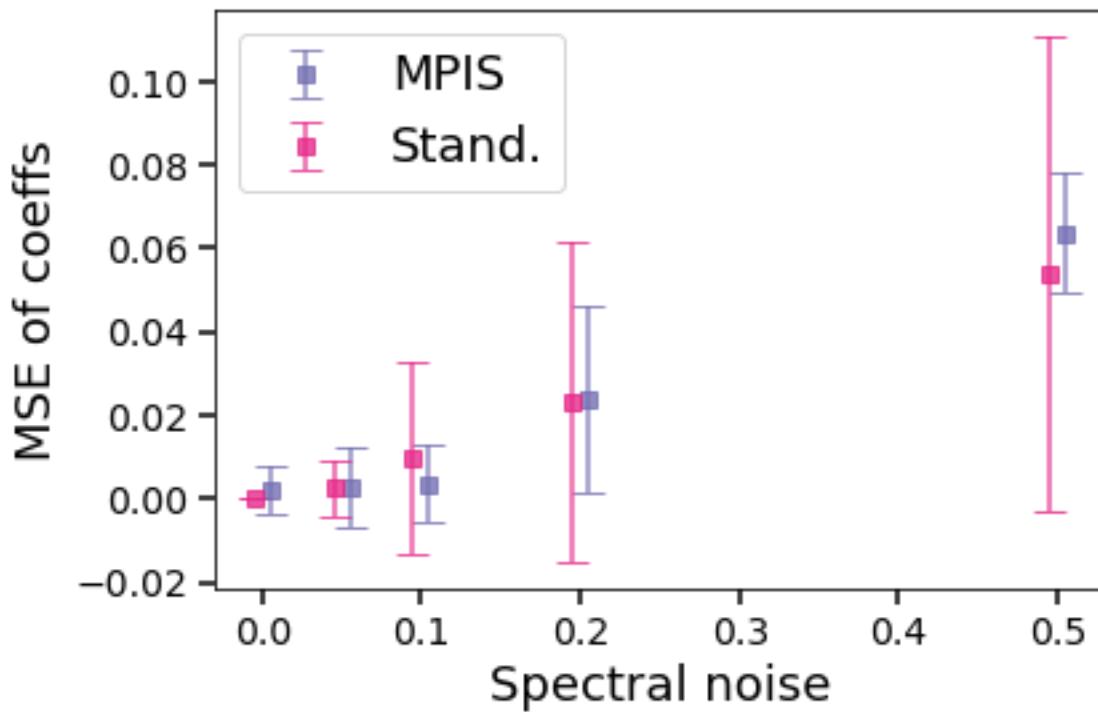
**Figure 8.S6** Correlation of reference spectra. (a) 1000 randomly sampled experimental spectra (which passed the background filter) normalized following the standard procedure in Athena<sup>19</sup>. (b) Reference spectra used in this study (11 total). (c) Variation of both the 1000 experimental spectra and the reference spectra. (e) Mean squared error between the original spectra and the fitted (via least squares) spectra of both the experimental data and true linear combinations of references (with random normal noise with a variance of 3% of the spectral intensity to model true experimental noise).

### Augmented MPIS versus noise



**Figure 8.S7** Augmented MPIS versus noise. When noise is set to 10%, only three clusters appear (left column). However, adding XRF and spatial encoding recovers the lost cluster (right column).

Error in predicted concentrations versus noise



**Figure 8.S8** Error in predicted concentrations versus noise (on a dataset composed of true linear combinations of references) for both our MPIS cluster-averaged pipeline and the standard individual spectrum analysis. The variance in predictions decreases when spectra are averaged together in an informed way via MPIS.

## 9 Chapter 9 – Recursive Feature Elimination for Nano-XANES Imaging

S. Tetef, A. Pattamattel, Y. S. Chu, M. K. Y. Chan, G. T. Seidler. Using Recursive Feature Elimination to Reduce Experimental Time of High-Throughput Nano-XANES Imaging. *In preparation.* A. Pattamattel conducted the experiment and collected all data. S. Tetef wrote the text and conducted all data analysis.

*We utilized a type of feature selection method called Recursive Feature Elimination (RFE) to reduce experimental time by decreasing the required number of measurements to perform. By applying this feature selection technique to nano-XANES imaging, a high-throughput experiment that collects tens of thousands of high-dimensional X-ray Absorption Near-Edge Structure (XANES) spectra with nanometer precision, we can reduce the total experimental time from around 24 hours to approximately three hours, about a 90% improvement in efficiency. Moreover, we can capitalize on the most common analysis procedure – linear combination fitting onto a reference library – to train the RFE algorithm and learn the most optimal measurements within this context. We compared predictions using both the full energy point spectra and the reduced energy point (sub)spectra, with energies chosen by the RFE algorithm, and found that we maintained sufficient accuracy in inferences using the highly constrained (sub)spectra. While RFE suggested measurements that produced better when compared to other feature selection methods, we further explored recommendations to maintain reliable RFE results, especially when there is larger uncertainty in the system and thus a more expansive reference library is constructed. In general, while feature selection can be highly beneficial for any high-throughput experiment that*

*produces high-dimensional spectra, not just nano-XANES imaging, careful evaluation that reliable results can be maintained is required before performing the constrained experiment.*

## 9.1 Introduction

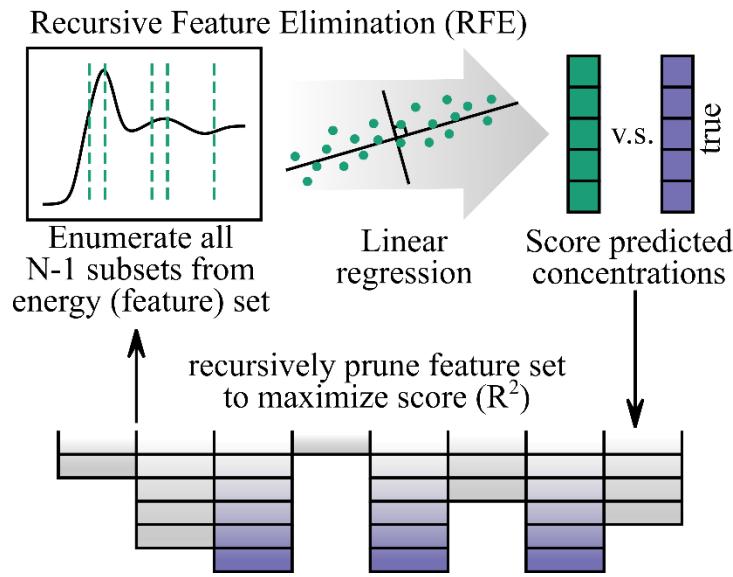
High-throughput experiments and multimodal characterization techniques are rising in popularity in all fields of science due to their efficiency and comprehensive scope. While the large amounts of data collected in these studies allow more opportunity for insight, it is not true that more data is *always* beneficial. For example, if the experiment measures correlated parameters, meaning multiple data points rise or fall together such that more than one data point is unnecessary, later inferences may be unreliable. One way to disentangle measurements is to determine which ones strongly contribute to desired effects by performing feature selection. Feature selection is a machine learning technique that takes a set of input measurements, also called features, and determines which ones are important, often in the context of some goal or target. Feature selection, in addition to reducing the susceptibility of learning from correlation rather than causation, can minimize experimental time by requiring fewer measurements to be taken<sup>1</sup>.

Here, we will perform feature selection to choose the best measurements for a type of high-dimensional spectral imaging technique called nanoscale X-ray absorption near edge structure (nano-XANES)<sup>2-7</sup>. Nano-XANES measures a XANES spectrum at every pixel (with nanometer precision), where each spectrum is usually composed of 50 to 100 energy measurements. The total experimental time for nano-XANES imaging is about 24 hours (for a typical 75 energy point measurement for each XANES spectrum). This timeframe makes nano-XANES imaging too time consuming for more expansive nano-XANES experiments, such as time-resolved in-situ kinetics studies or tomography XANES (four-dimensional measurements rather than three). Thus, the

bottleneck for high throughput nano-XANES acquisition is largely due to the number of energy measurement in each spectrum.

Given this difficulty, we hypothesized that feature selection can help reduce the number of needed energy points and remove the redundant energy values in the spectra, but without compromising on post-experimental analysis. This approach differs from previous work<sup>8, 9</sup> that has instead selected spatial regions of interest to gather full spectra, thus compromising global spatial information rather than spectral information. Even though there have been recent advances in speeding up micro- and nano-XANES imaging<sup>10, 11</sup> from an implementation perspective, our design to reduce the number of required energies further reduces the total experimental time to only a few hours instead of 24, which is approximately a 90% improvement in efficiency.

The type of feature selection we implemented is called Recursive Feature Elimination (RFE)<sup>12</sup>, a wrapper-based supervised feature selection routine. Fig. 9.1 demonstrates the RFE algorithm. As the name suggests, the RFE model recursively decides which input features are the most important by recursively pruning the input space such that the least important features are removed first. The algorithm decides the importance of each feature with one of a few options. For example, the RFE can correlate a subset of features to the accuracy of predicted target labels by training a base machine learning estimator on that specific feature subset. Or, in the case of linear regression, the RFE can rank the model weights which correspond to each input feature such that the feature with the largest weight is deemed most important and the feature corresponding to the smallest weight is least important. The RFE algorithm will then recursively retrain the base machine learning model on smaller and smaller feature subsets until a desired number of features remains.



**Figure 9.1** Recursive Feature Elimination (RFE) optimizes the feature subset to measure, in this case energies, by training a base machine learning model, such as linear regression, to predict target variables from spectra.

## 9.2 Methods

The sample and experimental data is the same as it appeared in A. Pattammattel, et al.<sup>7</sup> and S. Tetef, et al.<sup>13</sup>. See those works for the experimental details. Training data for the feature selection and machine learning models was generated by linear combinations of reference spectra, where a random dropout was included such that there was enforced sparsity in the number of references contributing to any one spectrum. Our reference library is the same as in A. Pattammattel, et al.<sup>7</sup> and S. Tetef, et al.<sup>13</sup>, which includes the four known phases – stainless steel (SS), lithium iron phosphate (LFP), pyrite (Pyr), and hematite (Hem) – and seven additional ones –HFO (hydrous ferric oxyhydroxide), goethite, maghemite, magnetite, Fe<sub>3</sub>P, Fe(III)PO<sub>4</sub>, and Fe(III)SO<sub>4</sub>.

All feature selection methods, including recursive feature elimination (RFE), random forest (RF), decision tree (DT), and linear regression (LR), were implemented using the `sklearn`

python package. The RFE algorithm was trained on a dataset composed of 1000 linear combinations of references (without additional noise) and with linear regression as the base estimator. We then applied principal component analysis (PCA) to the reference library and projected the 1000 generated linear combinations using the principal component vectors obtained from the reference library. The PCA-projected spectra were similarly given as training input to the other feature selection models, with the PCA-projected coefficients as the target (or output) variables.

We kept the top 13 energies, which were selected as most important from a dataset of 50,000 linear combinations of reference spectra that were subsequently PCA-projected, where the RFE ranked *all* energies (it stopped when only one energy was left). We then *ad hoc* chose three additional energies to ensure proper normalization of spectra – two in the far pre-edge (maximally spaced) and one in the post-edge (highest energy available). Thus, we kept a total of 16 energies from the 54 possible. Of note, these 54 energies were interpolated from the original 74 energies experimentally measured in order to align energy scales with the reference library. To normalize (sub)spectra, we fit a line to the first two energies in each spectrum (energies which we added for that purpose) and subtracted that line from each spectrum. We then fit another background post-edge line to all energies above 7150 eV. We used the maximum of the (sub)spectra to determine edge location (rather than the maximum in the derivative, as is commonly done with full spectra) to generate “flattened” spectra by dividing by the post-edge line in the region past the edge (so that the post-edge features fall along the  $y = 1$  line on average).

As a baseline, we obtained “true” linear combination fitting (LCF) results using the full-energy experimental spectra by performing pixel-by-pixel non-negative least squares linear combination fitting (NNLS-LCF) onto a smaller reference library composed of only the four

known phases (SS, Hem, Pyr, LFP). Details of the alternative LCF approach – LASSO-LCF via Manifold Projection Image Segmentation (MPIS) – are in S. Tetef, et al.<sup>13</sup>. In short, we used Uniform Manifold Approximation and Projection (UMAP)<sup>14</sup> and dbSCAN clustering<sup>15</sup> to globally group spectra together and then performed LCF on the cluster-averaged spectra.

## 9.3 Results and Discussion

### 9.3.1 Recursive Feature Elimination (RFE) Training, Recommendations, and Validation

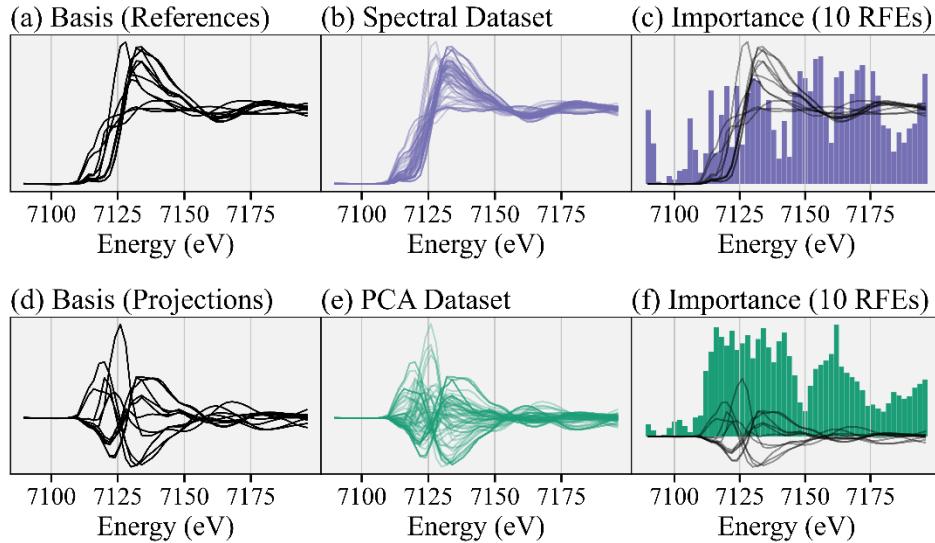
Because the RFE is a supervised routine, we synthesized a training dataset of linear combinations of reference spectra, where the reference spectra correspond to possible phases in our sample. Moreover, this training dataset incorporates prior knowledge of our system and mirrors post-experimental analysis, particularly by inverting the analysis process, which often entails linear combination fitting (LCF) onto a reference library. We knew our sample was made of stainless steel (SS), lithium iron phosphate (LFP), pyrite (Pyr), and hematite (Hem). However, we added other iron-containing phases to the reference library to represent a typical experiment for measuring Fe K-edge XANES, specifically HFO (hydrous ferric oxyhydroxide), goethite, maghemite, magnetite, Fe<sub>3</sub>P, Fe(III)PO<sub>4</sub>, and Fe(III)SO<sub>4</sub>. Note that our sample, dataset, and reference library are the same as it appeared in A. Pattammattel, et al.<sup>7</sup> and S. Tetef, et al.<sup>13</sup>. See Methods for details.

First, we notably observed that the choice of reference spectra, or generically the choice of basis vectors to generate linear combinations for a training dataset for the RFE, plays a critical role in the reliability of the RFE results. For example, randomly selecting 50 experimental spectra as a basis set to generate linear combinations for the training data was problematic. Specifically, the

experimental spectra contained so much linear dependence that linear combinations of them effectively had nonunique solutions. Thus, the base machine learning model at the center of the RFE learned unreliable solutions, so the RFE produced recommendations that were contrary to our intuition – recommending energies with low spectral variance (Fig. 9.S1). However, the RFE matched our intuition – identifying regions with high variance as important – when the basis vectors were chosen to be linearly independent. For example, the RFE matched our human intuition when we used distinct Gaussian distributions as the basis vectors for training the RFE (Fig. 9.S2). Because reference libraries for XANES studies are usually highly correlated, or in other words linearly dependent, we recommend applying Principal Component Analysis (PCA)<sup>16</sup> to the generated linear combination training dataset to force feature inputs and target outputs to be linearly independent and thus there exists unique solutions for the base model within the RFE to learn.

Fig. 9.2 compares the RFE recommendations using the linearly *dependent* reference spectra versus the linear *independent* principal components as training data. Specifically, the first row uses the reference library (Fig. 9.2a) to make linear combinations (Fig. 9.2b) to train an ensemble of 10 RFE models and obtain a collective importance of every energy (Fig. 9.2c). On the other hand, using the same spectra but projected onto the first few principal components for both the references (Fig. 9.2d) and training dataset (Fig. 9.2e), the RFE results are relatively similar (Fig. 9.2f) to the recommendations made by training directly on the spectra (Fig. 9.2c). Fig. 9.S3 quantitatively compares the results of training the RFE on the linear dependent versus independent pairs of features and target variables. While an appropriate choice in a small but comprehensive reference library might mitigate the effects of linear dependence of the basis set when training the RFE model, applying PCA first is a flexible procedure that allows for inclusion of a larger reference

library, thus providing robustness against uncertainty in the system.



**Figure 9.2** Comparing RFE results with forced linear independence in the basis set and thus the training dataset. (a) The spectral reference library. (b) Training dataset of linear combinations of references. (c) The RFE results, trained on the spectral linear combinations, where the basis set can have linear dependence. (d) Reference spectra projected onto the first six principal components from PCA. PCA forces the basis set to be linear independent and thus the linear combinations are unique. (e) The same training data as before, which are also projected using PCA. (f) The RFE results trained on the PCA projections.

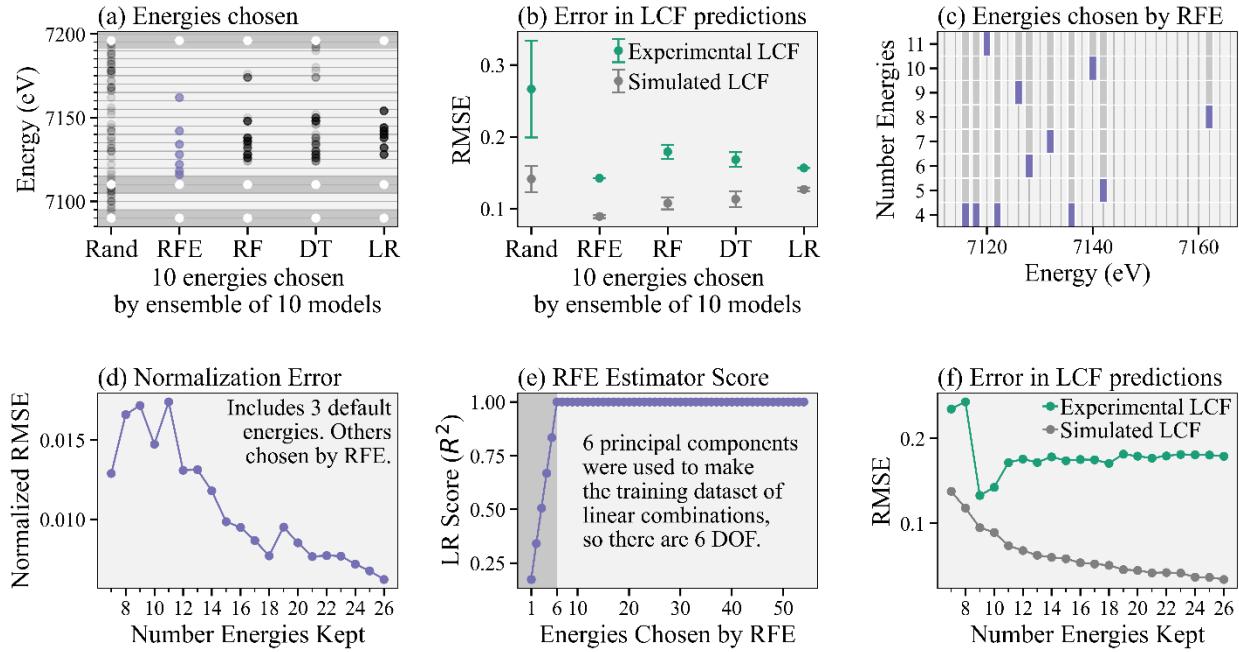
Further characterization and validation of the RFE is shown in Fig. 9.3. Note that all feature selection methods, including the RFE, are trained on a dataset composed of linear combinations of reference spectra which are then projected onto principal components. Fig 9.3a shows four different feature selection models, including the RFE – namely random forest (RF), decision tree (DT), and linear regression (LR) – and compares them to a random selection of energies. We show the combined results of an ensemble of 10 feature selection models (including 10 random draws),

where each model selects the top 10 (arbitrarily chosen number) energies. We then add the same three energies to ensure normalization, as indicated by white points in the dark gray regions, for a total of 13 energies kept. Fig. 9.3b shows the corresponding average and standard deviation of LASSO-LCF predictions given the energy selections for each feature selection model. We compare the errors on both the simulated LCF (using a generated test dataset of linear combinations of references) and the actual experimental spectra. For the experimental spectra, we determine the “true” coefficients by performing non-negative least squares (NNLS) onto the four known reference spectra using the full spectra. For each reduced energy point (sub)spectra, we perform LASSO regression onto the references (also reduced in energy points) to obtain the predicted coefficients. Unsurprisingly, errors in the experimental LCF are higher than the simulated LCF, although the RFE generally produced (sub)spectra with the lowest errors in predictions.

Fig. 9.3c shows the consecutive energies discarded by the RFE as fewer energies are kept. Of note, the same energies are kept during each retraining of the RFE, where each retraining the RFE picks fewer energies. This pattern is demonstrated by the purple stopping points and indicates that the RFE recommendations are consistent, regardless of the hyperparameter determining the number of features (or energies) to keep. Fig. 9.3d shows the error in normalizing XANES spectra using the reduced energy point (sub)spectra. Specifically, the normalized root mean squared error (RMSE / number of energies chosen) is shown, where the error is calculated between (sub)spectra that are *normalized after* energy cuts from the raw experimental spectra and the spectra *normalized first* using the full energy spectra and then sliced by energy to make the (sub)spectra. Because there is no spectral variation in the far pre-edge, the RFE does not choose energies in that region. However, normalization requires fitting a line in that region. Thus, we add two default energies in

the far pre-edge to ensure this line can be appropriately determined as well as another high energy point to similarly help with normalization. We see the error in normalization is reasonably small when more than 15 total energies are kept (12 chosen by the RFE plus the three default ones). However, we recommend taking further care to determine the number of energy points needed to ensure normalization.

Fig. 9.4e shows the score (coefficient of determination, or  $R^2$ ) of the base estimator inside the RFE, in this case linear regression (LR), as more and more energies are chosen by the RFE. Because each spectrum has six degrees of freedom (DOF) – one for each of the principal components the spectra are projected onto – the score for the base estimator is imperfect when fewer than six energies are kept, exactly because the system of equations is underdetermined in that regime. Thus, we recommend keeping enough energy points such that number of energies is greater than the number of principal components required to explain 99% variance of the reference set. To see how increasing uncertainty in the system, or including a larger reference library, affects the number of principal components, see Fig. 9.54. This large Fe K edge XANES reference library was taken from M. Marcus and P. Lam<sup>17</sup>. Finally, Fig. 9.3f compares errors in LCF predictions (on both the simulated linear combinations and experimental data) using different number of energies in the (sub)spectra. Again, we have added three default energies to ensure normalization, so the RFE algorithm is recommending between 4 and 23 energies for a total of 7 to 26 energies kept, as shown. We see that errors in LCF predictions on the experimental spectra converge once 11 energies total are kept, providing a lower bound on our (sub)spectra size. The slight drop in error at 9 and 10 energies kept is likely due to differences in normalization.

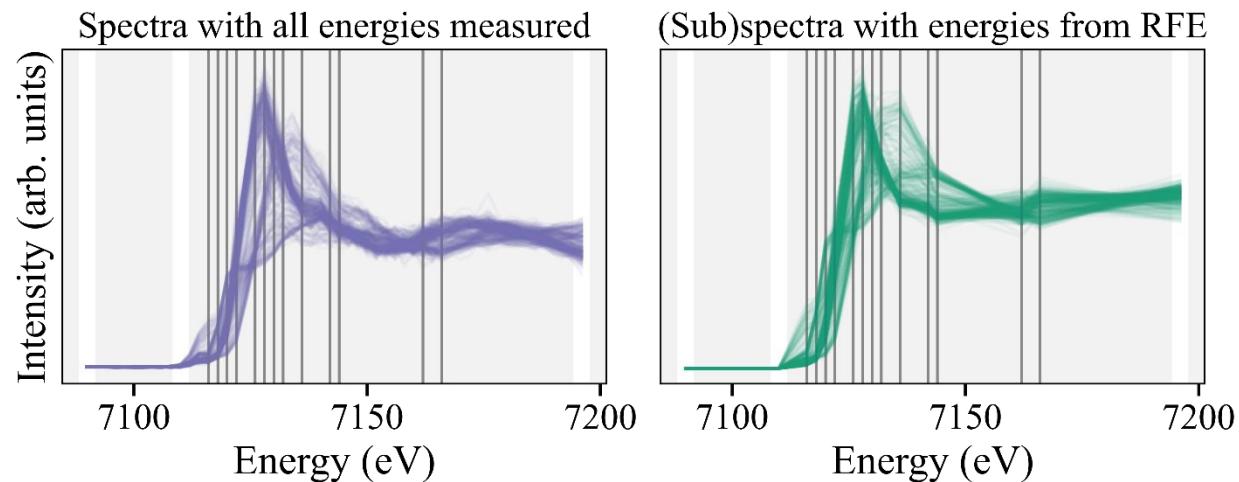


**Figure 9.3** Characterization and validation of RFE algorithm. (a) Collection of energies chosen by different feature selection algorithms: random selection (Rand), recursive feature elimination (RFE), random forest (RF), decision tree (DT), and linear regression (LR). The dark bars include the three default energies (white) to ensure normalization. (b) Corresponding errors in LCF predictions on both a generated test dataset and the experimental spectra for all models. (c) Energies consecutively removed by the RFE as fewer energy points are kept, which shows consistency in training. (d) Error in reconstructing spectra using normalization parameters from reduced energy point (sub)spectra of different sizes compared to normalized full energy spectra. (e)  $R^2$  score of the linear regression (LR) base estimator in the RFE. (f) Error in LCF predictions versus (sub)spectra size on both simulated test data and the experimental spectra.

### 9.3.2 Reliability of Inferences Using Measurements Chosen by RFE

Following the recommendations above, we chose the 13 most important energies as recommended by the RFE algorithm and then *ad hoc* added three energies to ensure normalization, thus keeping a total of 16 energy points in our (sub)spectra. We then took energy cuts of the

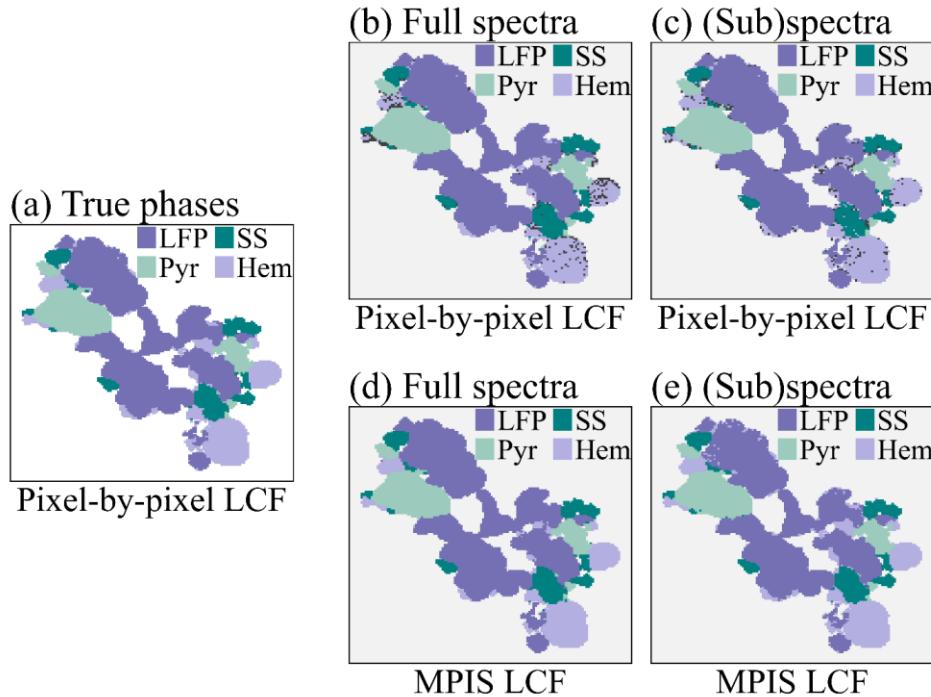
experimental and reference spectra using these 16 energies and renormalized all (sub)spectra independently. We attempted to combat any systematic errors in normalization in the experimental (sub)spectra by renormalizing the reference (sub)spectra as well. The full experimental spectra and 16-energy (sub)spectra are shown in Fig. 9.4, with the gray lines indicating the RFE recommended energies and the white lines indicating the energies we added for normalization. Fig. 9.S5 shows correlation matrices for both the full reference spectra and reference (sub)spectra and Fig. 9.S6 shows scree plots for the experimental dataset for the full and (sub)spectra. Both these figures show that most of the information in the full spectra is retained in our (sub)spectra.



**Figure 9.4** Fully measured experimental XANES spectra (left) compared to the reduced energy point (sub)spectra (right), with energies recommended by the RFE algorithm (vertical gray lines). The vertical white lines indicate energies we subsequently added for normalization purposes.

Next, we applied Manifold Projection Image Segmentation (MPIS) to cluster spectra in the nano-XANES image and then performed Linear Combination Fitting (LCF) via Least Absolute Selection and Shrinkage Operator regression, or LASSO-LCF, as detailed in S. Tetef, et al.<sup>13</sup> See

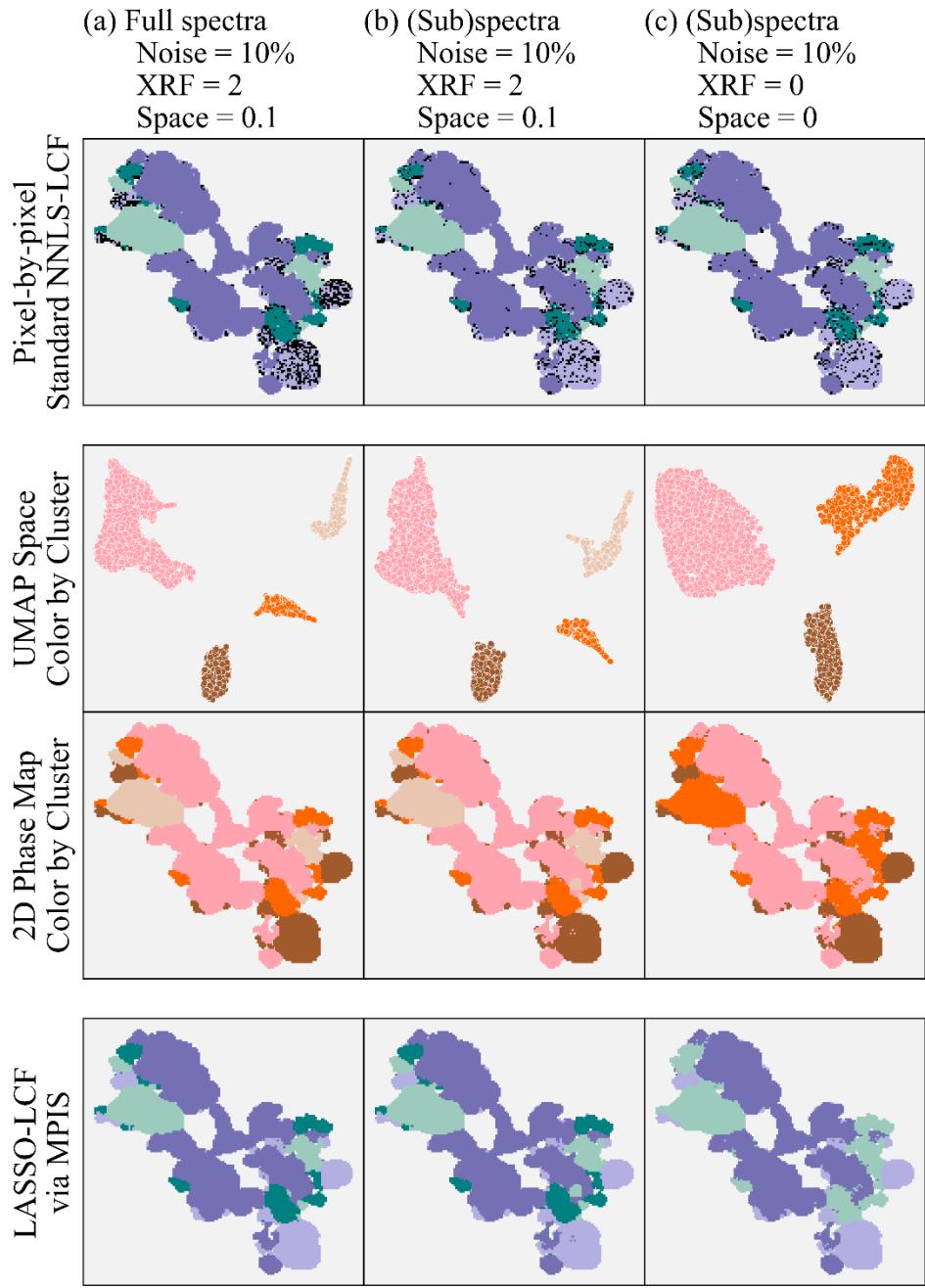
Figs S7-S9 for the corresponding MPIS figures. The end results for LASSO-LCF via MPIS are shown in Fig. 9.5. We calculated the “true results” (Fig. 9.5a) via pixel-by-pixel non-negative least squares linear combination fitting (NNLS-LCF) regression using just the four known phases as our reference library. We then compared the standard analysis procedure – pixel-by-pixel NNLS-LCF – using the *full reference library* on the full-energy spectra (Fig. 9.5b) versus the 16-energy (sub)spectra (Fig. 9.5c). The dark speckles in these images are pixels where the LCF reported phases that were not one of the true phases of LFP, stainless steel, pyrite, or hematite. Finally, we compared these results instead using LASSO-LCF via MPIS on the full-energy spectra (Fig. 9.5d) and the 16-energy point (sub)spectra (Fig. 9.5e). The results for the MPIS on the (sub)spectra (Fig. 9.5e) are almost identical to the full-spectra results (Fig. 9.5d), indicating the 16-energy (sub)spectra retained enough information to maintain accurate inferences. Moreover, using the MPIS before performing LASSO-LCF removed the spurious and incorrect LCF results, indicated by the lack of gray regions in the MPIS panels.



**Figure 9.5** Linear Combination Fitting (LCF) results via standard pixel-by-pixel analysis and Manifold Projection Image Segmentation (MPIS). (a) The “true” results, using the full energy spectra via pixel-by-pixel NNLS-LCF onto the four known reference phases. (b) Pixel-by-pixel NNLS-LCF applied to the full energy spectra. (c) Pixel-by-pixel NNLS-LCF applied to the reduced energy point (sub)spectra. (d) LASSO-LCF via MPIS applied to the full energy spectra. (e) LASSO-LCF via MPIS applied to the reduced energy point (sub)spectra.

While the above results are promising, we hypothesized that encoding two additional modes of information – the spatial location of each spectrum as well as the elemental composition of every pixel, specifically sulfur, phosphorus, and chromium using the X-ray fluorescence (XRF) intensities – would further increase the robustness and quality of MPIS results, especially for noisy spectra. We demonstrate this effect in Fig. 9.6, where Fig. 9.6a shows the results for noisy full energy spectra (Gaussian noise with a standard deviation of 10% of the spectral intensity at each

energy is added to the experimental spectra). We have included augmented information by tuning the strength of the encoding of the XRF data and spatial location of every pixel using the “XRF” and “Space” weighting hyperparameters, respectively. The detail of this encoding is explained in S. Tetef, et al. <sup>13</sup>. To view the overall effects of varying the two hyperparameters in MPIS that control spatial segregation, see Fig. 9.S10. Fig. 9.6b shows the results for the same augmented information except using the 16-energy point (sub)spectra. The MPIS generates similar phase maps for both the full spectra and (sub)spectra with the additional information encoded. However, performing MPIS on the (sub)spectra without the augmented information (Fig. 9.6c) fails to appropriately separate out two of the four phases, indicated by the UMAP space only containing three clusters rather than four. Thus, the extra information encoded into the MPIS pipeline helped to recover the extra cluster, distinguishing hematite from stainless steel when noise levels are high.



**Figure 9.6** MPIS and LASSO-LCF results using (a) the full spectra and multimodal encoding, (b) the (sub)spectra and multimodal encoding, and (c) the (sub)spectra by themselves *without* augmented information. Here, the total XRF intensity of sulfur, phosphorus, and chromium and the spatial location of pixels are multimodal information.

## 9.4 Conclusions

We have shown that Recursive Feature Elimination (RFE) can be used to select the most important measurements to perform, which can help speed up high-throughput, high-dimensional spectroscopy experiments. Nano-XANES imaging, in particular, is highly time constrained by the number of energies to measure in each XANES spectrum and thus greatly benefits from using feature selection to determine optimal energy measurements. We observed that there are two key contributors to determining the minimum number of energy points to measure. First, ensuring energies are chosen such that proper normalization can occur is critical in maintaining reliable linear combination fitting (LCF) results. Second, we recommend keeping the number of additional energies to measure at least equal to the degrees of freedom of the reference library, where Principal Component Analysis (PCA) can be utilized to parameterize the number of linearly independent components in the library and thus quantify the uncertainty in the system. Finally, when implementing RFE, we recommend processing the training dataset of linear combinations of references with PCA to ensure that input and output vectors are linearly independent and thus the learned solutions are unique. The PCA pre-processing step for the RFE creates more robust recommendations for larger reference libraries, which are inherently more prone to linear dependence within the set and can thus cause unreliable RFE results. While the RFE algorithm showed promising results, especially compared to other feature selection methods, in the future, careful validation is critical before performing the experiment using the recommended measurements, especially to ensure reliable inferences can be maintained within the constrained experiment.

## Acknowledgements

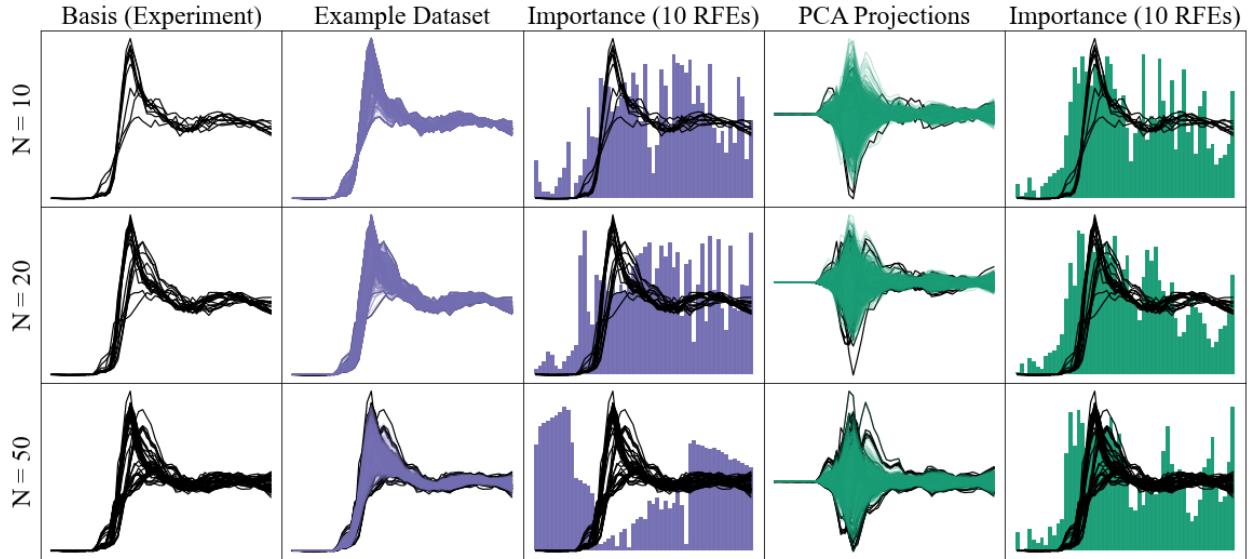
Thank you to P. Lam and M. Marcus for sharing their Fe K-edge XANES reference library.

## 9.5 References

1. J. D. Li, K. W. Cheng, S. H. Wang, F. Morstatter, R. P. Trevino, J. L. Tang and H. Liu, *ACM COMPUTING SURVEYS*, 2018, **50**.
2. I. Nakai, C. Numako, S. Hayakawa and A. Tsuchiyama, *Journal of Trace and Microprobe Techniques*, 1998, **16**, 87-98.
3. R. Belissont, M. Munoz, M. C. Boiron, B. Luais and O. Mathon, *Minerals*, 2019, **9**.
4. M. Cusack, Y. Dauphin, J. P. Cuif, M. Salome, A. Freer and H. Yin, *Chemical Geology*, 2008, **253**, 172-179.
5. M. Bonnin-Mosbah, N. Métrich, J. Susini, M. Salomé, D. Massare and B. Menez, *Spectrochimica Acta Part B: Atomic Spectroscopy*, 2002, **57**, 711-725.
6. L. Mino, E. Borfecchia, C. Groppo, D. Castelli, G. Martinez-Criado, R. Spiess and C. Lamberti, *Catalysis Today*, 2014, **229**, 72-79.
7. A. Pattammattel, R. Tappero, M. Ge, Y. S. Chu, X. Huang, Y. Gao and H. Yan, *Science Advances*, 2020, **6**, eabb3615.
8. N. Mölders, P. J. Schilling, J. Wong, J. W. Roos and I. L. Smith, *Environmental Science & Technology*, 2001, **35**, 3122-3129.
9. M. Grafe, E. Donner, R. N. Collins and E. Lombi, *ANALYTICA CHIMICA ACTA*, 2014, **822**, 1-22.
10. B. E. Etschmann, E. Donner, J. Brugger, D. L. Howard, M. D. de Jonge, D. Paterson, R. Naidu, K. G. Scheckel, C. G. Ryan and E. Lombi, *ENVIRONMENTAL CHEMISTRY*, 2014, **11**, 341-350.
11. U. Boesenborg, C. G. Ryan, R. Kirkham, A. Jahn, A. Madsen, G. Moorhead, G. Falkenberg and J. Garrevoet, *Journal of Synchrotron Radiation*, 2018, **25**, 892-898.
12. H. Jeon and S. Oh, *APPLIED SCIENCES-BASEL*, 2020, **10**.
13. S. Tetef, 2023, in prep.
14. L. McInnes, J. Healy and J. Melville, *arXiv*, 2020.
15. M. Hahsler, M. Piekenbrock and D. Doran, *Journal of Statistical Software*, 2019, **91**, 1 - 30.
16. S. Wold, K. Esbensen and P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37-52.
17. M. A. Marcus and P. J. Lam, *Environmental Chemistry*, 2014, **11**, 10-17.

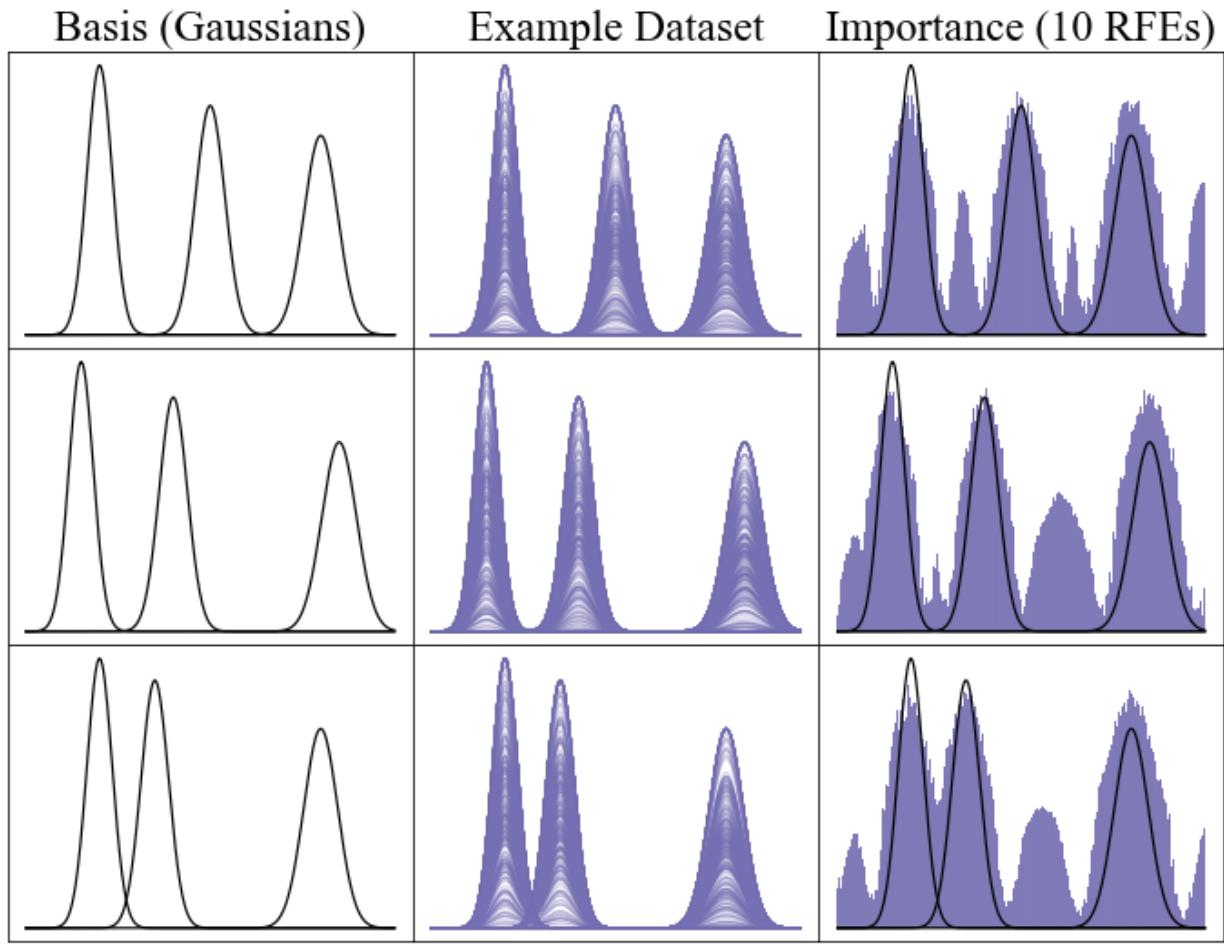
## 9.6 Supplementary Information

Figure S1 RFE results on linear combinations of experimental data



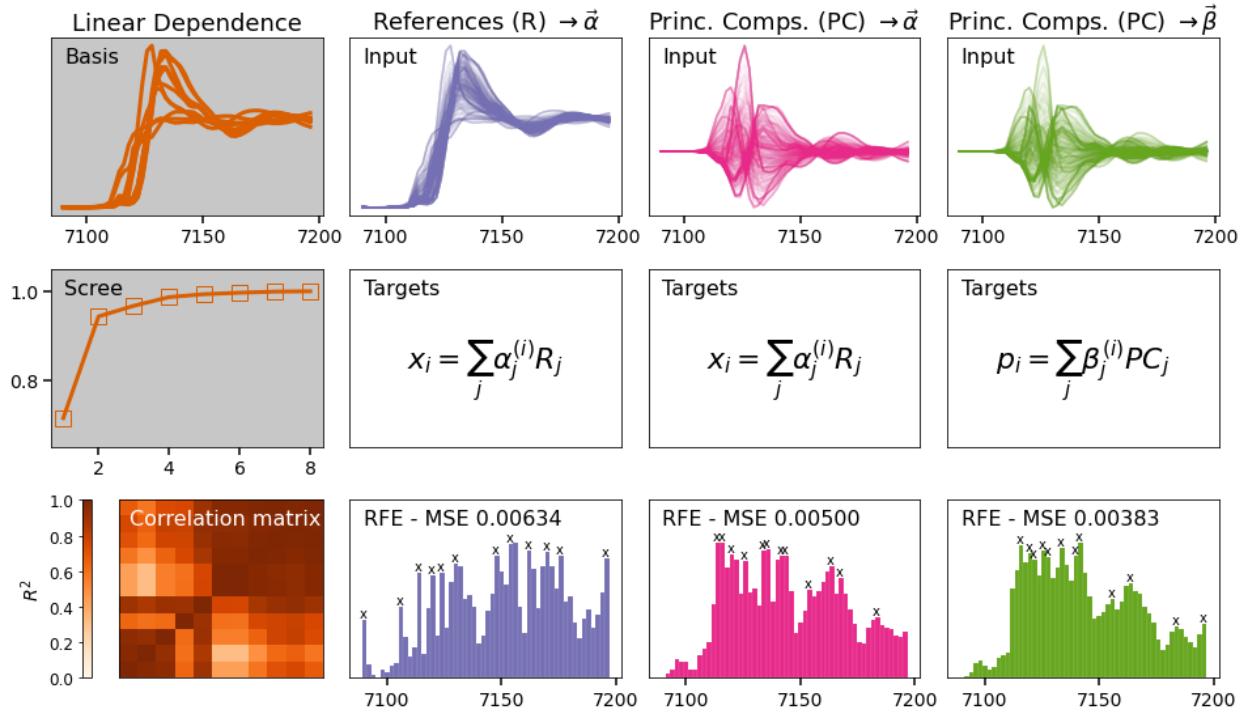
**Figure 9.S1** (a) Random sampling of experimental data to act as a basis for linear combinations of spectra. Note that there is no guarantee that the basis spectra span or equally sample the experimental domain. (b) 1000 linear combinations generated from the corresponding basis. (c) The compiled results of an ensemble of 10 RFEs trained on the spectra. (d) The equivalent dataset except projected onto the first six principal components. (e) The compiled results of an ensemble of 10 RFEs trained on the principal components. When  $N$  (the basis set size) is 50, there is so much linear dependence in the basis set that the RFE fails because it chooses points in the pre-edge, which has no variation.

Figure S2 RFE results on gaussian basis sets



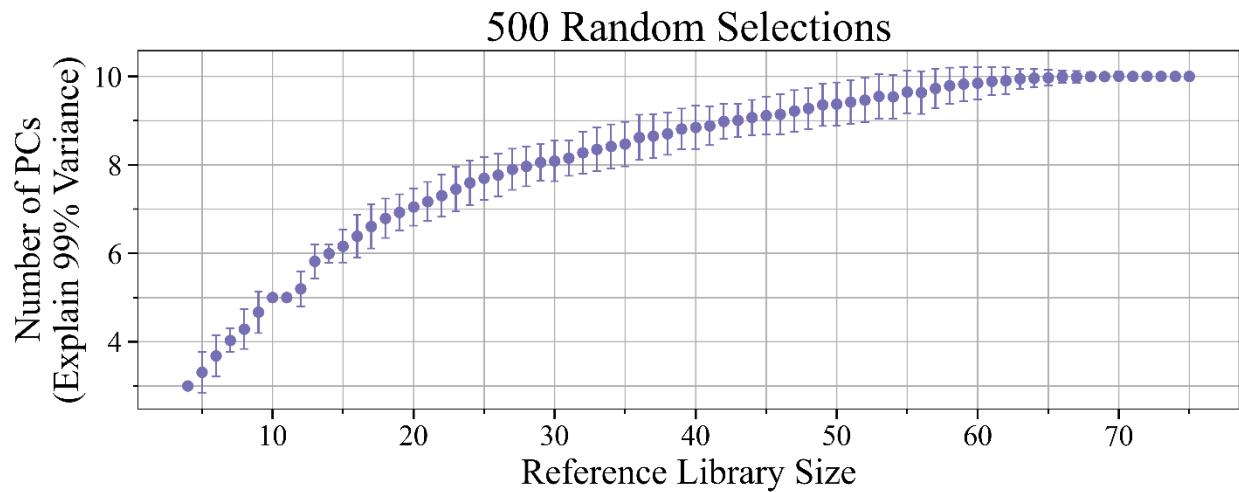
**Figure 9.S2** Test of the RFE by using three gaussians used as basis spectra (left-most column) to make linear combinations (middle column). The RFE clearly picks features (i.e., “energies”) that correspond to the highest variation. We used linear regression as our base estimator. However, after the regions corresponding to the three distributions are filled, the RFE must rank areas in between peaks where there is no signal. The peaks in importance between the Gaussians represent random selections in these regimes.

Figure S3 RFE results on both linear and nonlinear input/output pairs



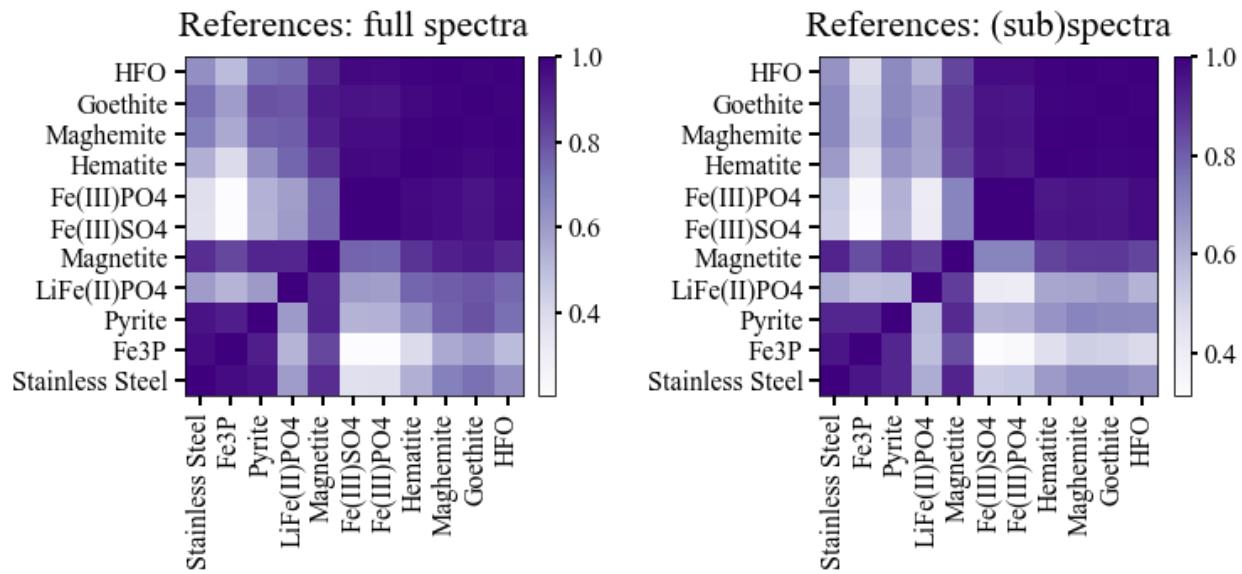
**Figure 9.S3** Comparing results of linear and nonlinear inputs to the RFE. The RFE results, where linear combinations of spectra are the input and the concentrations that created those spectra are output, is shown in purple. Instead, using projections onto the first few principal components as input (with the same output) is shown in pink. The green shows the RFE results using both linear input and outputs. The total results for an ensemble of 10 RFE algorithms for each are shown at the bottom, along with the mean squared error (MSE) of predictions using LASSO linear combination fitting (LASSO-LCF) on a generated dataset of linear combinations of reference spectra.

Figure S4 Scree plot showing PCs needed for increasing reference library size



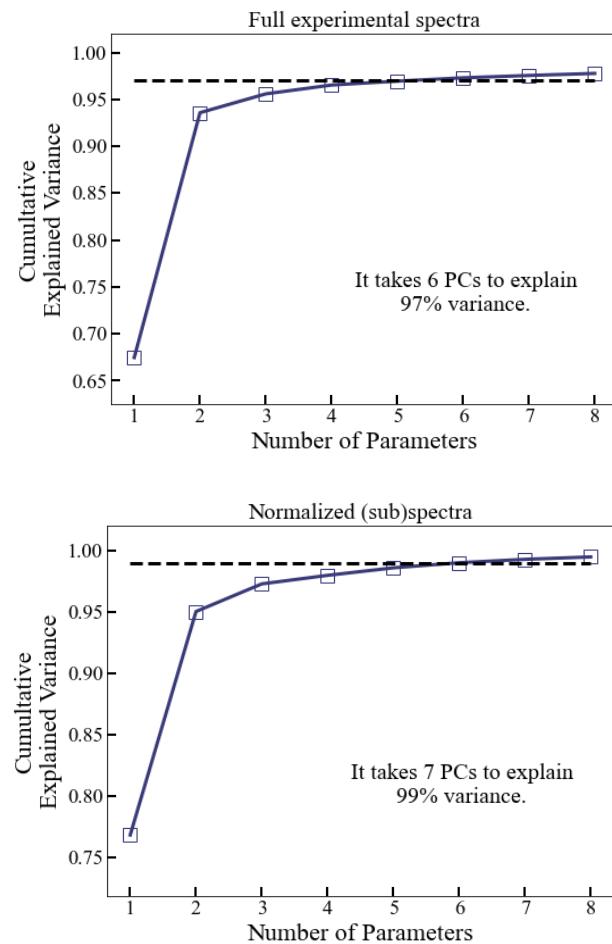
**Figure 9.S4** The number of principal components (PCs) needed to explain 99% variance of the reference set. Starting with the four known references, we randomly selected additional references from the set of 11 total references used in this study. After the 11 references were chosen, we randomly selected additional references from another larger set of 64 Fe K edge XANES to constitute the reference library. We reselected these random additions 50 times and show the average and standard deviation of the calculated number of principal components for that reference library size.

Figure S5 Correlation matrices of references



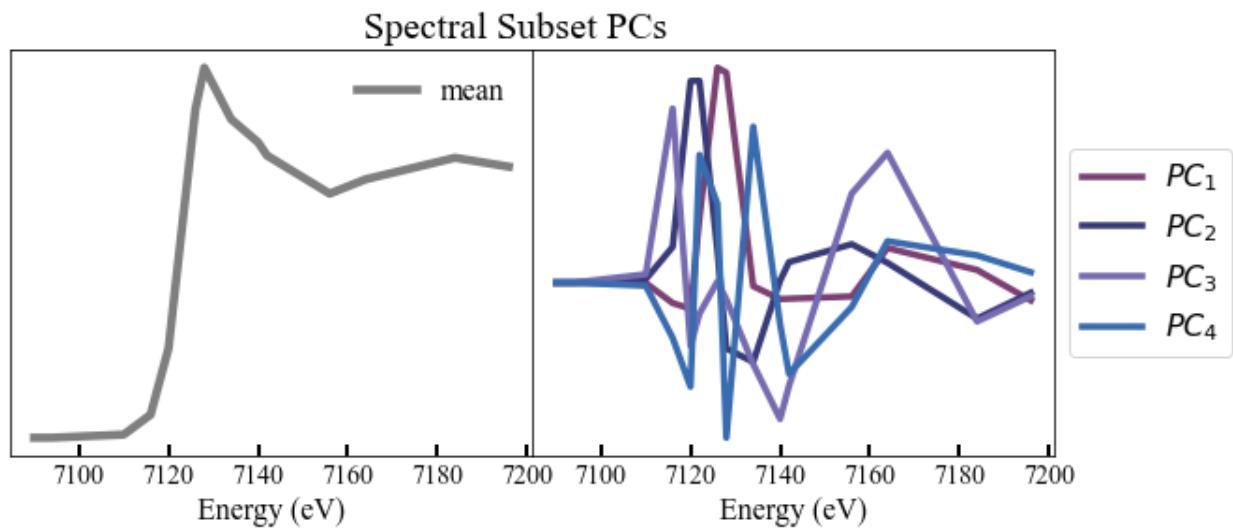
**Figure 9.S5** Correlation, or similarity matrices, of the reference set for both the entire spectra and the 14-energy (sub)spectra. The correlation coefficient ( $R^2$ ) qualitatively looks the same for both, although the quantitative range for the (sub)spectra is larger, indicating global correlations (and information) is retained.

Figure S6 Scree plot of experimental data



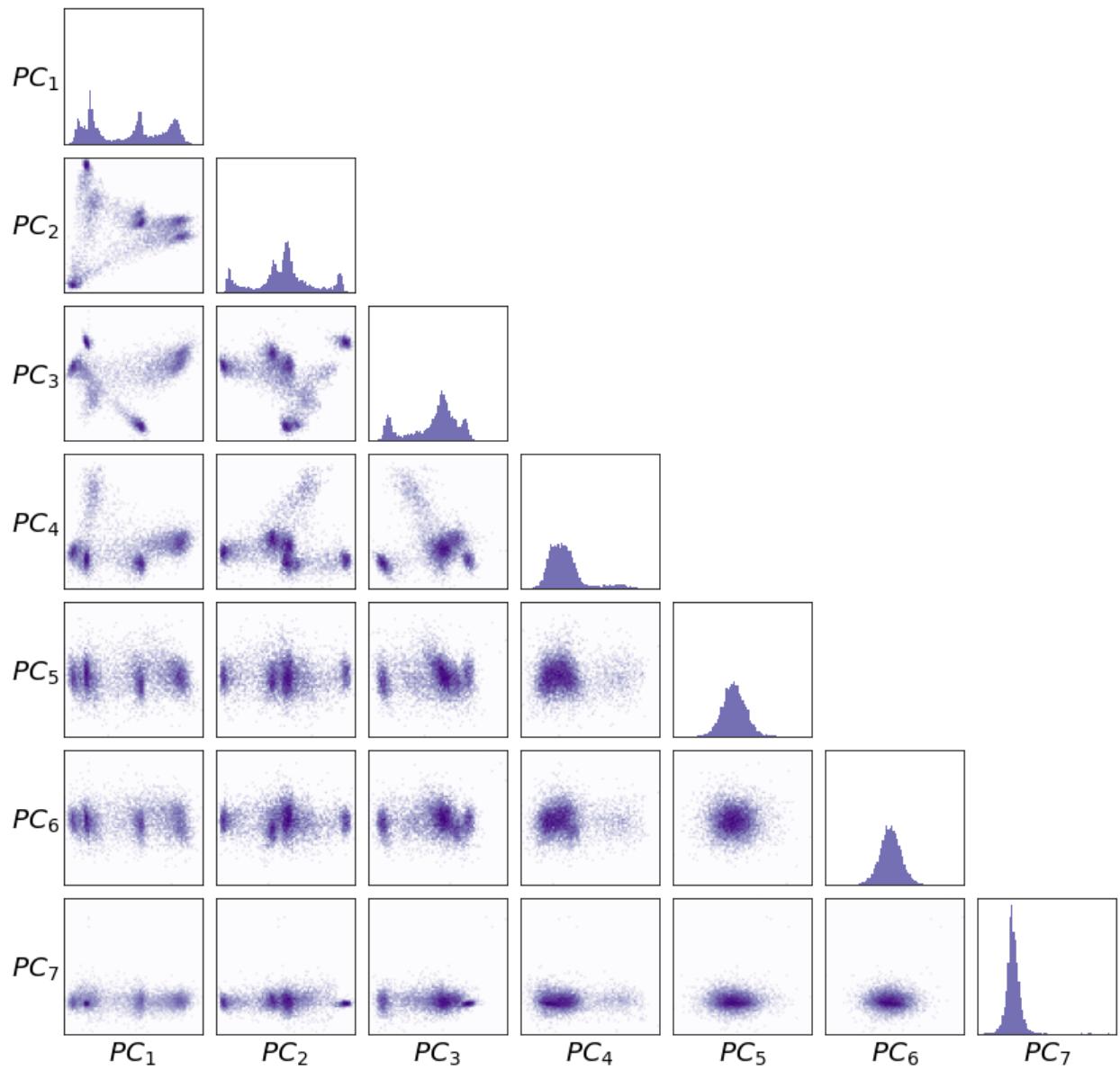
**Figure 9.S6** Scree plot of experimental data on full spectra (top) versus (sub)spectra (bottom).

Figure S7 First four PCs of sub(spectra)



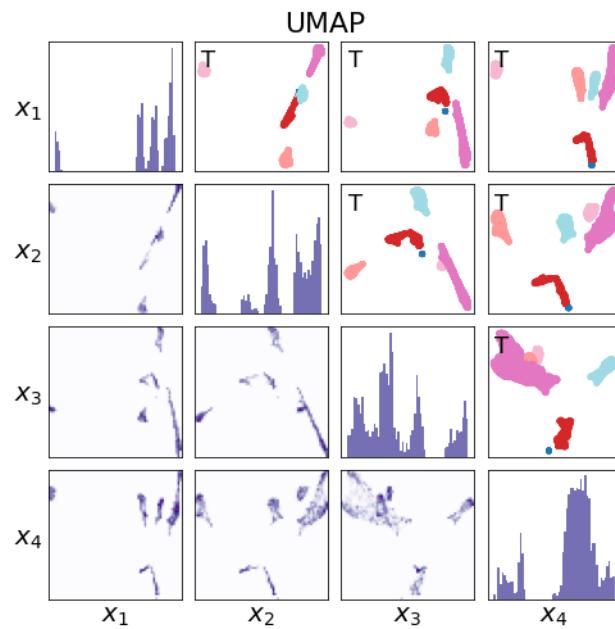
**Figure 9.S7** First four principal components of the spectral subset. These components, in theory, should match with the principal components from the full spectral dataset, if all information is retained.

Figure S8 PCA triangle plot on (sub)spectra



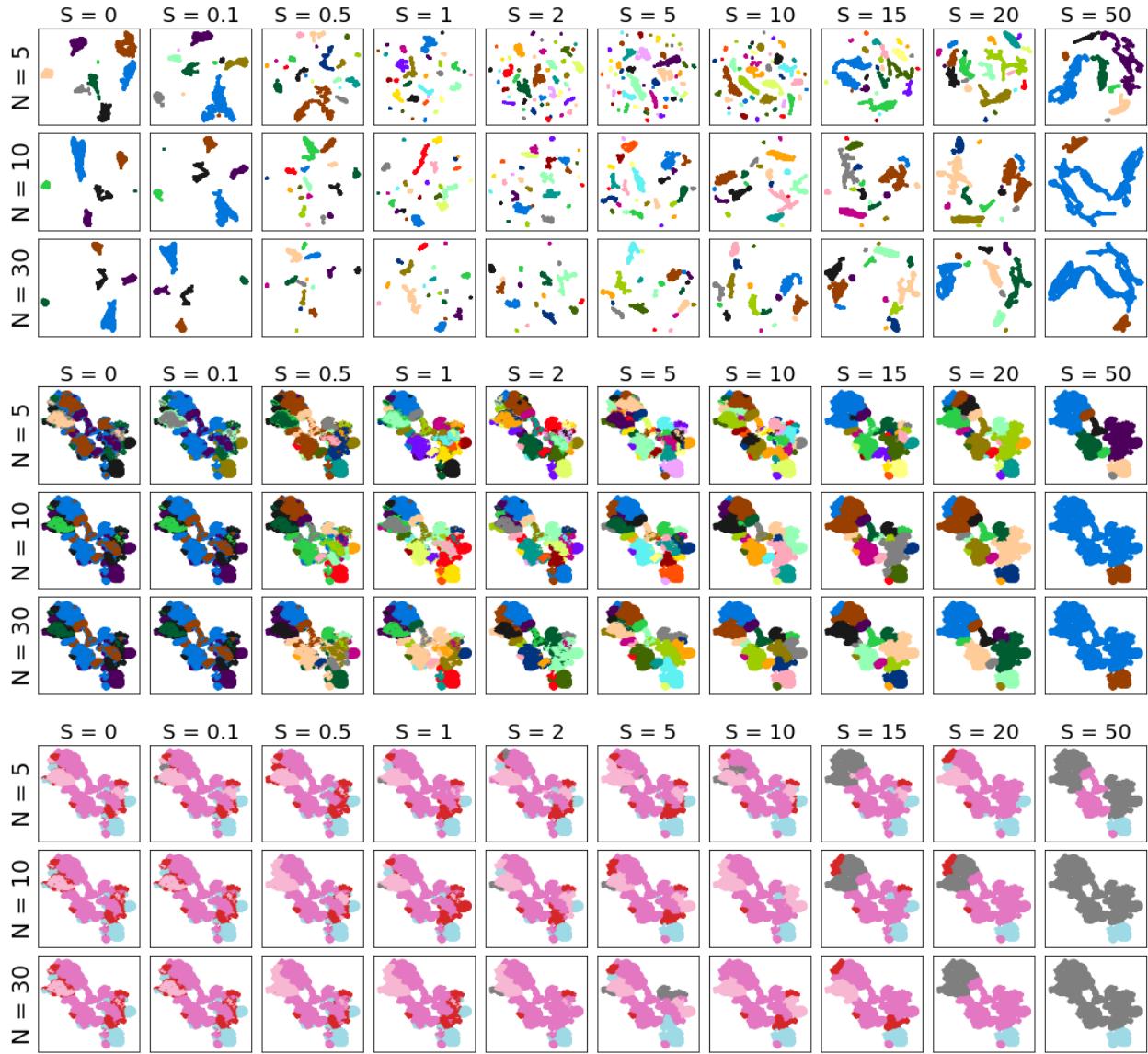
**Figure 9.S8** PCA triangle plot of experimental data on (sub)spectra.

Figure S9 UMAP and dbSCAN clustering on (sub)spectra



**Figure 9.S9** UMAP and dbSCAN on energy subset.

Figure S10 Effects of varying hyperparameters that control spatial grouping



**Figure 9.S10** Strength of spatial encoding ( $S$ ) versus UMAP's number of neighbors ( $N$ ). The minimum distance in UMAP is 0 and dbSCAN epsilon is 1 for all. The top section shows the UMAP space color-coded by dbSCAN clusters, the middle shows the same clusters but on the 2D map, and the bottom shows the max contributions from the LCF fits. Pink = Pyrite, magenta = LFP, blue = Hematite, red = SS, and gray = all other references.