PCCP

Unsupervised Machine Learning for Unbiased Chemical Classification in X-ray Absorption Spectroscopy and X-ray Emission Spectroscopy

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-06-2021-002903.R2
Article Type:	Paper
Date Submitted by the Author:	27-Sep-2021
Complete List of Authors:	Tetef, Samantha; University of Washington Govind, Niranjan; Pacific Northwest National Laboratory Seidler, Gerald; University of Washington

ARTICLE

Unsupervised Machine Learning for Unbiased Chemical Classification in X-ray Absorption Spectroscopy and X-ray Emission Spectroscopy

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Samantha Tetef[a], Niranjan Govind[b], Gerald T. Seidler[a,†]

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.

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 [1, 2]. Large databases, such as the Materials Project [3], Inorganic Crystal Structure Database [4, 5], and QM9 [6], along with open access packages for ML, have all contributed to this rise in popularity and reliability of machine learning analysis of data [7]. Recent work includes the use of ML to develop a way to represent molecular structures [8, 9], to study charge transport at the nanoscale level [10], or to automate chemical predictions from atomistic simulations [11].

X-ray absorption spectroscopy (XAS), an important chemical speciation technique, has seen impressive recent developments using ML [12-32]. 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 [33-35]. Furthermore, recent developments of reliable lab-based spectrometers in multiple energy ranges have facilitated an increase in accessibility of both XAS and XES measurements [36-40].

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 [41], electrical energy storage [42, 43], protein structure and function [44], terrestrial and lunar basalts [45], chemical catalysis [46] in biomolecules [47], and photochemical dynamics [48]. 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 [33, 49, 50]. 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 [51-54], prior knowledge is more commonly implicit via the user interaction with the standard tools for interpretation of EXAFS [55, 56] or XES spectra [57]. However, the analysis of XAS – and of XES, as seen here – is

[a] Department of Physics, University of Washington, Seattle, WA 98195, USA.
[b] Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA
† Corresponding author: seidler@uw.edu.
Electronic Supplementary Information (ESI) available: (details of any supplementary information available should be included here). See DOI: 10.1039/x0xx00000x

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. [27] 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. [31] 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 [12, 20, 29]. Moreover, supervised ML has seen use in classifying coordination and local chemical environments [14, 16] and the oxidation state [19] of 3d transition metals, and used to extract geometric properties [30], especially during high-throughput experiments [17] in real-time [26]. As another example with a pragmatic application, ML has recently been implemented for fitting XANES spectra [18]. Further work utilizing artificial intelligence for fitting EXAFS data is also actively being developed [24, 25]. 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 [22]. 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. [13] and Kiyohara, et al. [15], 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 regressional 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. [23], 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 [58], here quantitatively addressing whether the *chemically relevant* information in XANES and VtC-XES is indeed complementary or is instead highly coincident [59-61].

Based on our results, we propose that chemical classification problems are best addressed with unsupervised ML methods at least as a precursor analysis method [11], an approach that may enrich or suggest refinement of prior structure-specific inferential work in XAS [14, 16, 17, 26, 27, 31] and similar work in a wide and rapidly growing range of other spectroscopies in chemical sciences [62-64]. 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 [23] 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 [60, 65, 66]; and (3) the prior demonstration of good agreement between experiment and time-dependent density functional theory (TD-DFT) [66] calculation of XANES [67] and VtC-XES [66, 68-70].

For chemical context, the five "Types" of molecules used in our study are shown in Fig. 1. They are: (1) sulphides, (2) thiocarbonyls, (3) thiols, (4) sulphoxides, 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 sulphoxides, 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 [65] and Holden, et al. [66]. 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.

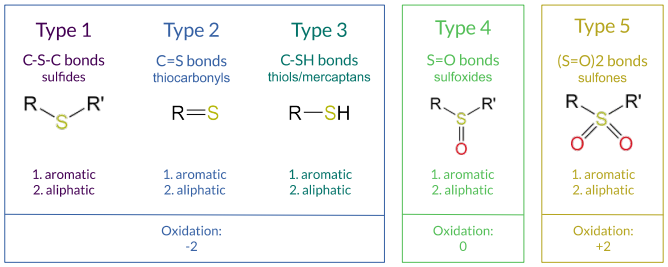

Fig 1. Schematic representation of the five types of sulphorganics investigated, along with sub-categories.

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.

2. Methods

2.1 Electronic Structure Calculations

Our data generation pipeline is shown schematically in Fig. 2. A list of sulphorganic compounds was created from a wide variety of sources, starting with the compounds in Yasuda and Kakiyama [65] and Holden et al.[66], 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 [71] via the `MolView` API [72]. All ground state structures, XANES [73], and VtC-XES [74] computations were performed with the open-source NWChem computational chemistry program [75, 76]. In total, 769 molecules are included in this work.

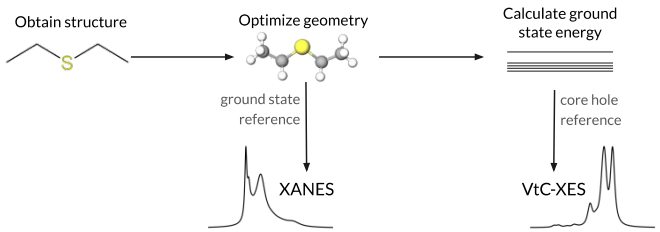

Fig. 2. Schematic depiction of the data generation pipeline.

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. 2."

The geometry optimizations utilized the 6-31G* basis sets [73, 74, 77, 78] and the B3LYP exchange correlation functional [79]. The XANES and VtC-XES spectra were then computed using the Sapporo QZP-2012 and Sapporo TZP-2012 basis set [80], respectively, for S, while the remaining atoms were represented using 6-31G* basis set, and PBE0 exchange correlation functional [81]. 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 [82].

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. [83]. 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 [83]. In this case, the energy-dependent broadening values of the transitions were chosen arbitrarily such that they most accurately depicted experimental features [60, 84]. Finally, the spectra were individually normalized by dividing their total Kα 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 [65, 66]. 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 Kα intensity to achieve a common intensity scale per S atom.

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` [85] package with a `Tensorflow` backend [86]. 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. S3 (Scheme 1: Oxidation), Fig. S3 (Scheme 2: Type), and Figs. S4 and 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.

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. S1. The trained VAE models, analysis code, and datasets are available on GitHub [87].

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` [88] 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.

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 [89]. 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). [90] Nonlinear dimensionality reduction can be achieved via unsupervised machine learning, specifically here, via the VAE neural network model [91]. Given that there is very scarce prior work using VAE's in spectroscopies, e.g., optical-wavelength spectroscopy in an astrophysical study [92], 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. [23]. 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. 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.

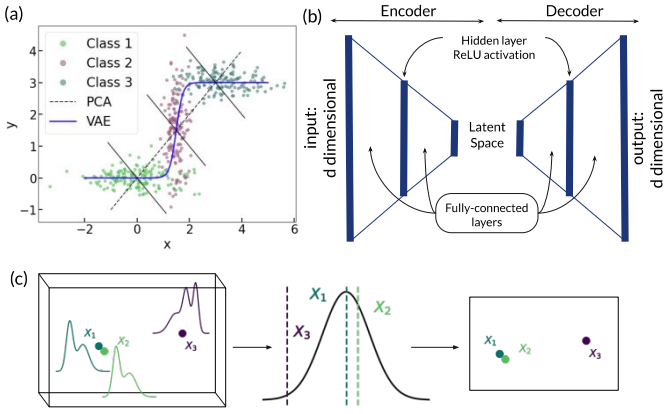

Fig. 3. (a) Clusters where nonlinear dimension reduction routines, such as from a neural network, might yield better clustering than a linear dimension reduction like PCA. (b) Architecture of a simple autoencoder (AE) with one hidden layer, demonstrating the dimension reduction utility of the AE via 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. 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 [92-95].

In Fig. 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) [91]. 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. 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. 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) [96] 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. 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.

4. Results and Discussion

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. 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.

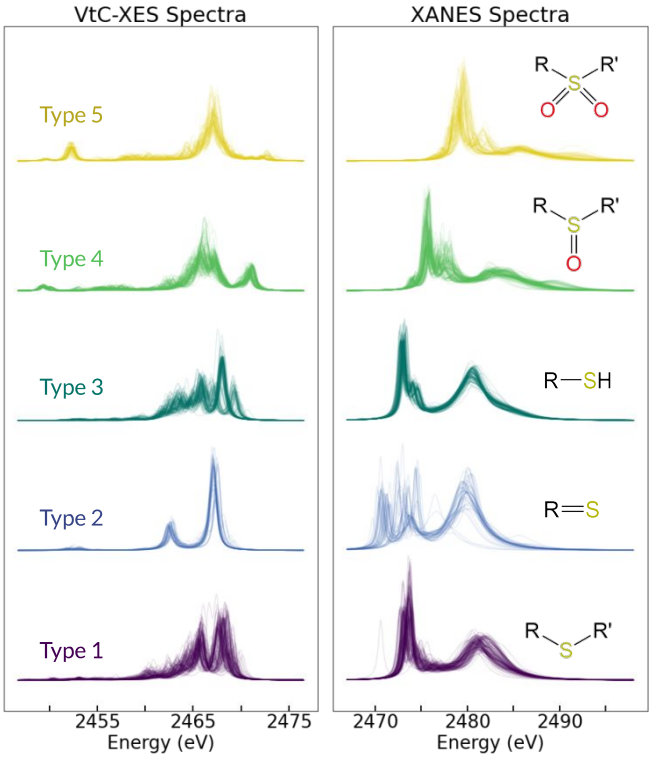

Fig. 4. VtC-XES (left) and XANES (right) spectra for all organosulphur compounds, displayed by compound type. Some spectra have been arbitrarily scaled or randomly removed for display purposes.

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. 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 ~ 6 – 8 principal components.

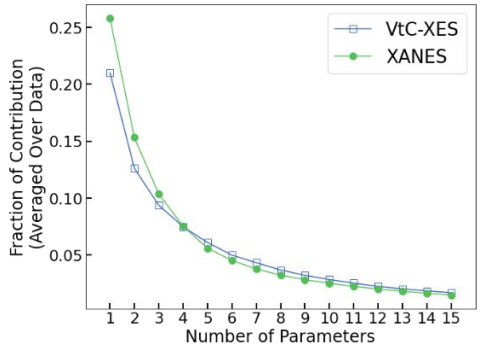

Fig. 5. Scree plot of PCA effectiveness for both VtC-XES and XANES. The vertical axis is the fraction of variance explained by each PC, e.g., the 10th PC.

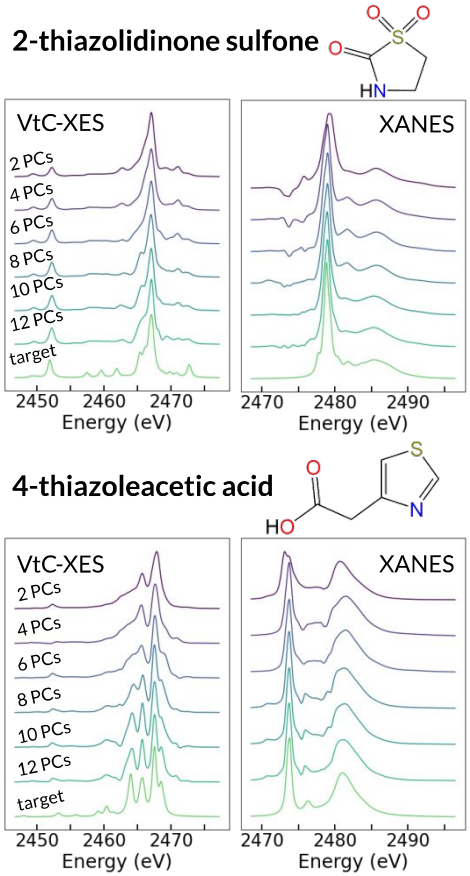

Fig. 6. Spectra reconstructed with increasing number of principal components (PCs) kept, for both VtC-XES and XANES of 2-thiazolidinone sulphone (Type 5) (top two panels) and 4-thiazoleaceticacid (Type 1) (bottom two panels).

To illustrate this fact, we show in Fig. 6 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. 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.

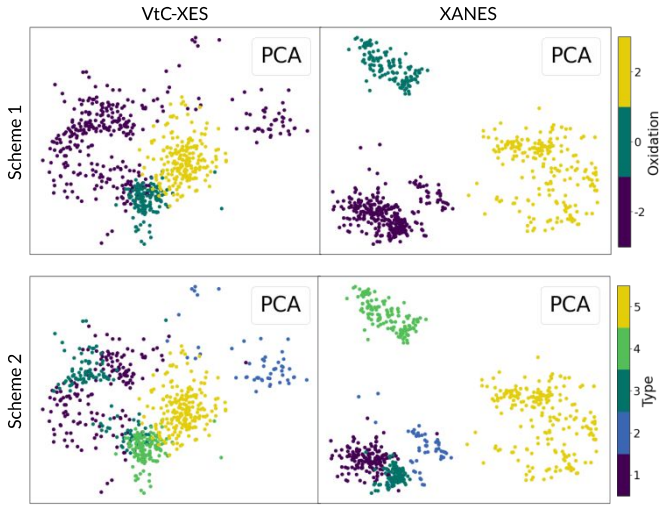

Fig. 7. Principal Component Analysis (PCA) projection for two dimensions, color-coded by the two different property classification schemes: Scheme 1 is by oxidation and Scheme 2 is by sulphur bond type.

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)[97], Factor Analysis (FA)[98, 99], and Non-negative Matrix Factorization (NMF)[100], as shown in Fig. S6. These three methods are other common linear dimensionality reduction routines and have been compared to PCA in other systems[101]. 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).

4.1.2 Variational Autoencoder

We again present in Fig. 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. 7, it is useful to establish some basic properties of the VAE training and resulting latent space.

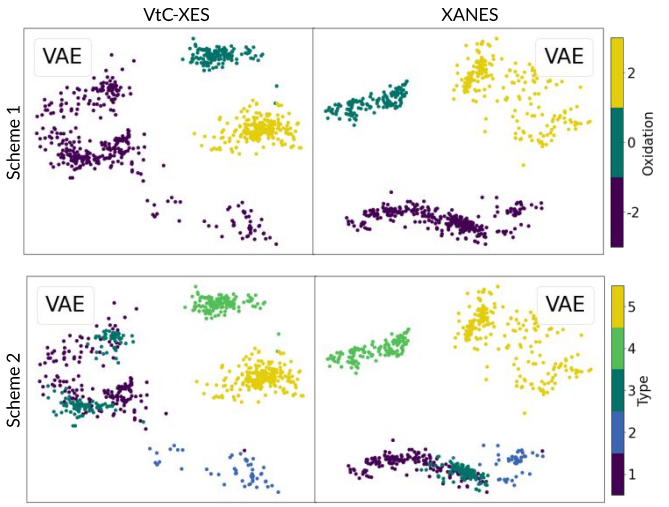

Fig. 8. Latent space representation in two dimensions via a Variational Autoencoder (VAE), color-coded by the two different property classification schemes: Scheme 1 is by oxidation and Scheme 2 is by sulphur bond type.

First, in Fig. 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. 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 protionamide 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.

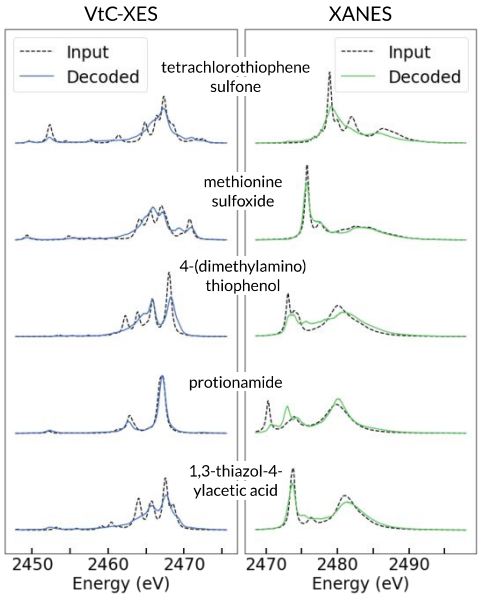

Fig. 9. Reconstruction of XES (left) and XANES (right) spectra from a two-dimensional latent space via a VAE. From bottom to top, the compounds are from Type 1, 2, 3, 4, and 5. The black dashed line represents the original inputted spectra, and the solid-colored line is the decoded spectra after it has 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. 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. 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.

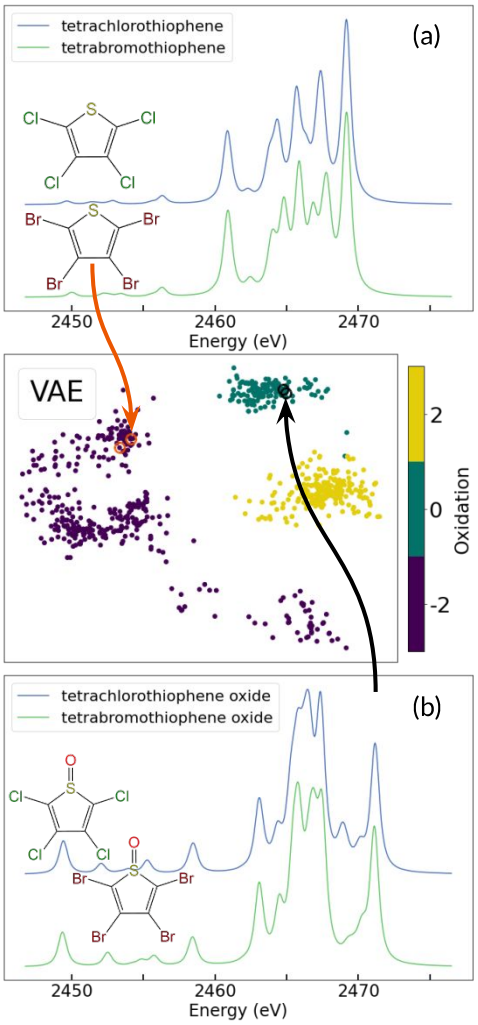

Fig. 10. Chemically similar compounds are nearby in the latent space. (a) The latent space location of tetrabromothiophene and tetrachlorothiophene, with the corresponding XES spectra on the right. (b) The same structures but oxidized to form tetrabromothiophene oxide and tetrachlorothiophene oxide.

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. 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.

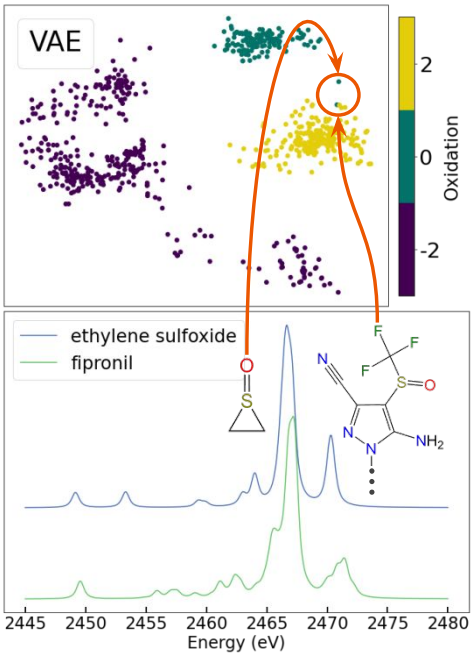

Fig. 11. A closer look at the outliers: the two "neutrally oxidized" compounds distinctly in the sulphone (+2 oxidation) cluster.

Moving now to the relative merits of the two-dimensional PCA representation (Fig. 7) and the VAE latent space (Fig. 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 sulphoxides 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. 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. 12. Here, we have color-coded spectra within types to indicate aromaticity, following Yasuda and Kakiyama [65], 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 [102], and have also observed sensitivity to aromaticity in a ligand, e.g. the sulphur K edge of sulphides [60, 65].

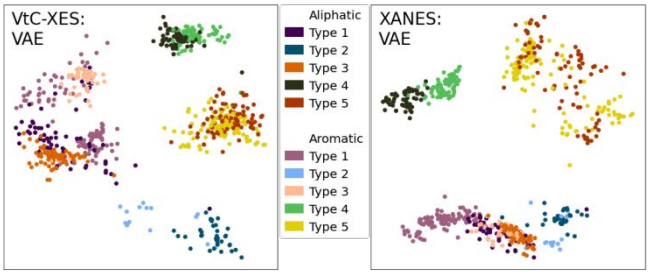

Fig. 12. Compounds with aromatic sulphur versus aliphatic sulphur, in the latent space (VAE) for both VtC-XES (left) and XANES (right).

As shown in Fig. 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 [65], which notes the aromaticity of the compound increases the energy but lowers the intensity of that peak, likely due to the presence of the π 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.

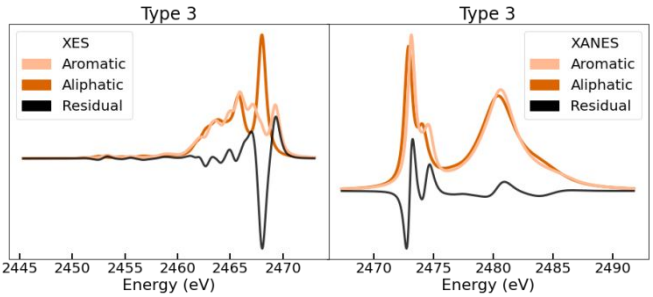

Fig. 13. Residuals between the average of the aromatic and aliphatic spectra of Type 3 (thiols).

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.

4.1.3 t-SNE, Clustering Without Mapping

In Fig. 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. 7 and Fig. 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. 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 isothiocynates, which are distinct from the other Type 2 thioketones.

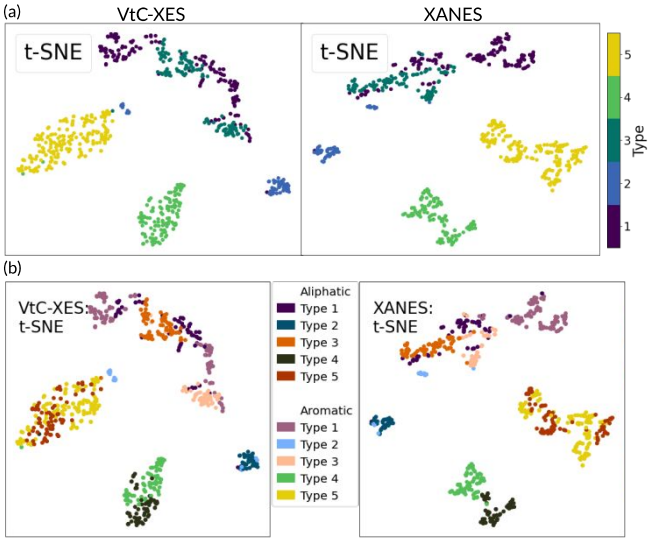

Fig. 14. t-SNE for VtC-XES (left) and XANES (right). (a) is color-coded by Type, while (b) is color-coded by aromaticity within each Type.

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 [65] on VtC-XES and by Qureshi et al. [60] 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. 15 for XANES. First, we examine the further splitting of the Type 1 aromatic compounds as shown in Fig. 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. 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. 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.

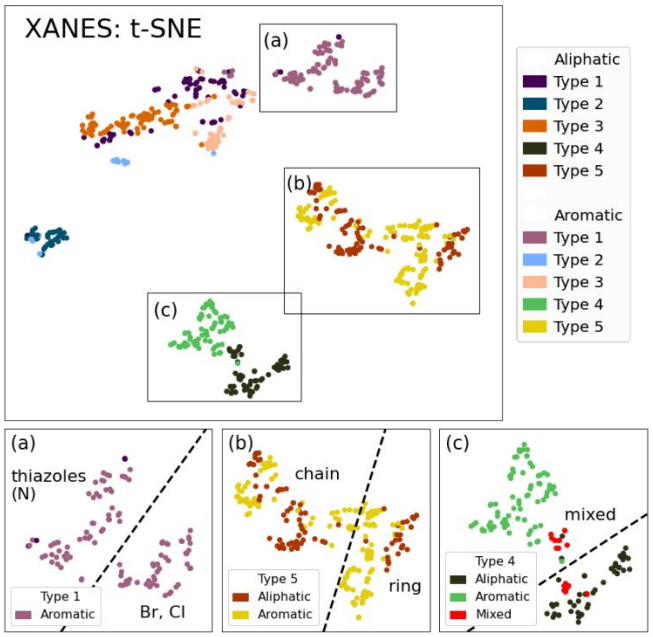

Fig. 15 (Main) A closer look the the subclusternig in the XANES t-SNE plot. (a) Separation of Type 1 aromatic compounds based on inclusion of chlorine or bromine in the aromatic system. (b) Separation of Type 5 aliphatic compounds based on bond strain via the inclusion of sulphur in a ring versus a chain. (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.

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. 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.

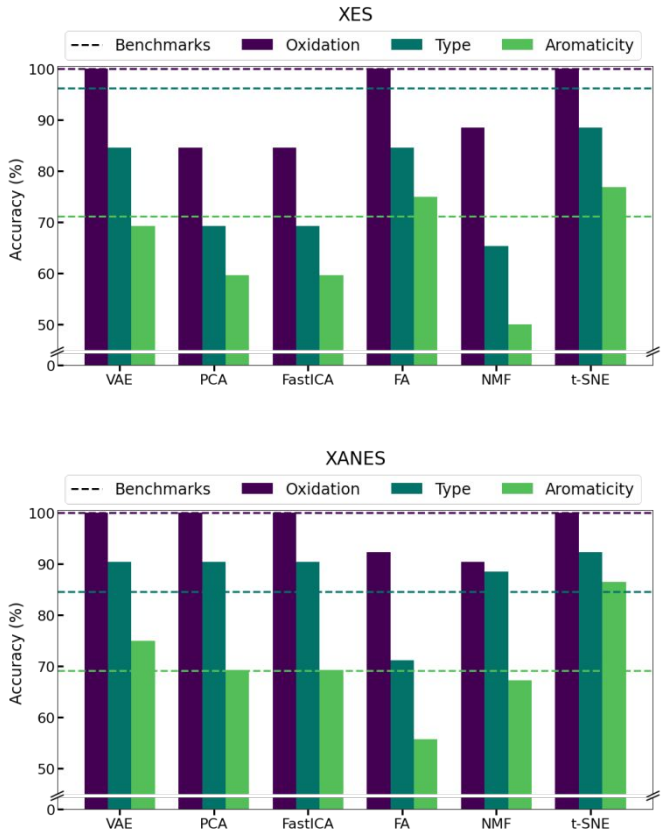

Fig. 16. Accuracy of KNN classification schemes on all dimensionally reduced spaces for both VtC-XES (top) and XANES (bottom).

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. 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. 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.

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. [20], 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 [59]. 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. 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. 17a we show the evolution from goitrin (oxidation state -2) to thiophene oxide (oxidation state 0). In Fig. 17b, we have the decoded spectra from the points in Fig. 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. 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.

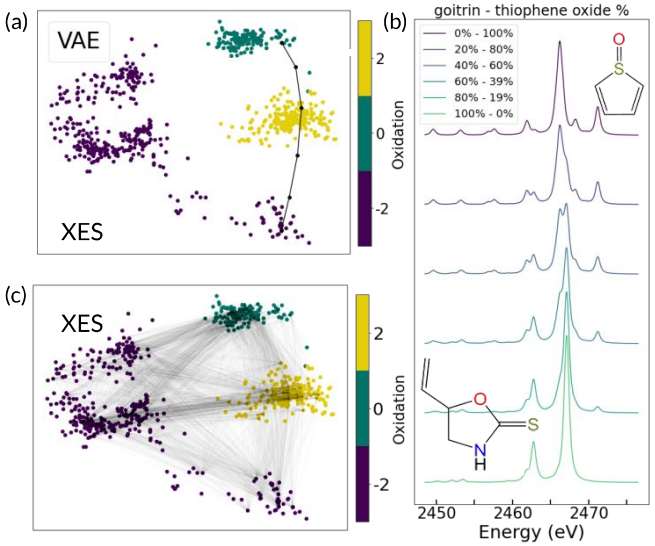

Fig. 17. As shown in (a), the evolution from goitrin (oxidation -2) to thiophene oxide (oxidation 0). (b) The linear combination of the spectra of thiophene oxide (top) and goitrin (bottom) that correspond to the points along the track in (a). (c) Tracks of 3000 different species evolutions.

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. [27] 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.

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. 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" [96] 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. 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.

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, DOI: 10.1146/annurev-chembioeng-101220-102232.
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. 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.
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. Brounce, J. Boyce, F. M. McCubbin, J. Humphreys, J. Reppart, 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. Zatsepin, 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, 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. Francl, 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. 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.
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. G. V. Fabian Pedregosa, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard 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, 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.