Informed Chemical Classification of Organophosphorus Compounds via Unsupervised Machine Learning of X-ray Absorption Spectroscopy and X-ray Emission Spectroscopy

Samantha Tetef$^+$, Vikram Kashyap$^+$, Alexandra Velian$^2$, Niranjan Govind$^3$, Gerald T. Seidler$^1$*

$^+$Co-first authors

$^1$Department of Physics, University of Washington, Seattle WA 98195, USA

$^2$Department of Chemistry, University of Washington, Seattle WA 98195, USA

$^3$Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

**Corresponding Author**

*seidler@uw.edu
ABSTRACT
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 which 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.

TOC GRAPHICS

KEYWORDS X-ray absorption fine structure, valence-to-core X-ray emission spectroscopy, Gaussian Process, UMAP, unsupervised machine learning.
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 inherent 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 \(^1-^5\). These data-driven models can frequently disentangle and infer patterns from lossy measurements 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 \(^6\). This necessarily assumes that sufficient information is, in fact, encoded in spectra; otherwise, ML models will correlate spurious features to requested properties. This detail of encoded information is often addressed by hand-selecting a targeted training domain which depends heavily on prior knowledge \(^7\). However, issues arise if the training domain is too small or 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 \(^8\), unwanted bias, especially from unrepresentative data, blindly undermines reliability of inferences and has led to contemporary ethical concerns \(^9-^{12}\).

In the effort to combat unwanted bias as well as provide generalizability to complex datasets, this study demonstrates the value of the pipeline exemplified in Figure 1, which 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 pipeline removes implicit biases and spurious correlations by adding steps (3) and (4) to a typical ML pipeline, which validate spectral sensitivity to properties requested during supervised predictions.

**Figure 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.

We utilize this pipeline for a spectroscopy method that has seen an ongoing exploration of ML applications: X-ray absorption spectroscopy (XAS)\(^{13-31}\). 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\(^ {32,33}\), catalysis\(^ {34}\), and photochemical dynamics\(^ {35}\). 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 X-ray emission spectroscopy (XES) for insight into the occupied electronic states. 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. 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, including in studies of sufficient scale for data science methods.

In the first study to utilize ML in XAS, Timoshenko, et al. predicted coordination from XANES spectra using a neural network, while Zheng, et al. also predicted coordination, except using a random forest model. Notably, Torrisi, et al. used a random forest model, except to correlate polynomial fitting parameters of spectra to properties like bond distance. Other works utilizing machine learning in XAS include a XANES matching algorithm, hierarchical clustering on spectra, and use of an autoencoder to correlate coordination to a reduced dimensional representation of spectra. Most of these studies assumed desired information was in fact encoded in spectra, largely because of hand-crafting relevant training datasets. However, our pipeline, via the unsupervised machine learning precursor, allows for explorative and unbiased refinement of chemical descriptors – a step that we propose is both necessary, and likely sufficient, when addressing much more complex datasets.

The present study is prompted by our recent work that compared the variance and information content of sulfur K-edge XANES to VtC-XES Kβ spectra for sulforganics. We found that nonlinear dimensionality reduction algorithms, a subset of unsupervised ML, provided
an effective way to extract 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.

Here, we investigate the information content and sensitivity of phosphorus K-edge XANES and VtC-XES Kβ 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 dataset of spectra we analyze is calculated from molecular structures gathered from the PubChem database using moldl, a Python module we have written to aid in collecting and managing molecular structure datasets. moldl is open-source and freely available to anyone. See the SI for more details. For the rest of this paper, we will refer to the phosphorus K-edge XANES and VtC-XES Kβ 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 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 utilize Uniform Manifold Approximation and Projection (UMAP) 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), which was used in our recent work to...
UMAP has additional benefits compared to t-SNE, such as being parametric and preserving global structure, which allows for future data compression as well as interpretation of overall global similarities. These advantages have led to its recent popularity, such as in single cell RNA sequencing (scRNA-seq) data analysis \(^4_8\), but has not yet seen use in XAS analysis.

To begin, heuristically one expects coordination to yield the strongest distinguishing feature 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 to further vary the effective charge on the phosphorus. The spectral averages for both the VtC-XES and XANES spectra for each tricoordinate phosphorus and tetracoordinate phosphorus class are shown in Fig. S1. 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.

Individual classes within each coordination are shown in columns A and B. Additionally, all R groups are constrained to 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). As expected, coordination distinguishes most of the groupings of the compounds, with a handful of outliers.
Figure 2 UMAP representation of VtC-XES (top) and XANES (bottom), color-coded by coordination. R₁, R₂, and R₃ are defined to be carbon-based aryl or alkyl chains, with only phosphates allowed to have R₁ and R₂ as H atoms.

It follows that there are chemically relevant sub-groupings within each coordination.

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. We expected 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 Fig. S1) will increase in both energy and intensity with an increase in phosphorus oxidation. From a molecular orbital perspective, this trend is from a larger overlap between the ligand valence orbital and the phosphorus 3p orbital.
(valence shell). In general, this feature (which also changes with different ligand symmetries and orientation) is why VtC-XES is so strongly sensitive to ligand identity\textsuperscript{49}. 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 by the average spectra in Fig. S1.

**Figure 3** UMAP representation of VtC-XES (top) and XANES (bottom) for tricoordinate phosphorus (A) and tetracoordinate phosphorus (B) compounds, color-coded by number of oxygens bonded to the phosphorus within each coordination.

Note that the phosphates are segregated from the other tetracoordinate phosphorus compounds and seem to sub-cluster as well. This observation brings 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, observed by changes in the ligand peak feature. 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 tetracooordinate phosphorus compounds (phosphates) were evaluated in which the oxygen substituents were replaced with one or two sulfur atoms. Compared to oxygen, sulfur is significantly less electronegative, with a Pauling electronegativity value near that of carbon and phosphorus. Thus, these oxygen-to-sulfur ligand substitutions likely cause the biggest spectral change by adjusting the effective charge on the phosphorous. The resulting clusters are shown in Figure 4.

![Figure 4 UMAP representation of VtC-XES (left) and XANES (right) for compounds with sulfur ligands, color-coded by number of sulfurs.](image)

As expected, the different ligand identities are contributing to cluster separations. The VtC-XES also clearly has an outlier – the orange phosphorothioate in the red dithiophosphate cluster at the bottom right of that figure. Chemically, that compound (PubChem CID 104781) is
structurally different from others because the oxygens form one edge of a carbon tetrahedrane. Thus, UMAP clearly identifies chemical outliers.

We then analyzed 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. In general, this distinction seems to be better illuminated by the VtC-XES spectra than the XANES (which is shown in Fig. S5), 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.

![Figure 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.](image)
In the above discussion, we have 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. Applying UMAP to just phosphates, we achieve the embedding shown in Figure 6, which has labeled the phosphates into four clusters determined by the dbscan 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. All compounds in Clusters I to IV can be viewed in Figs. S10 to S13. For some example compounds in each cluster along with their spectra and structure, see Figs. S6 to S9. 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 Fig. S14, and found no strong correlation. This further supports the nonlinear nature of spectra and the idea that spectral fingerprints in complex datasets do not correlate solely to a single high-variant property like the absorption edge, but rather a combination of properties.

**Figure 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 – independent of the specific dimensionality reduction algorithm used – 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 and thus has the potential to show evolutions through the domain space, similar to the latent space of a variational autoencoder (VAE) ⁵². This property facilitates real-time analysis of high-throughput experiments. 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.

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. 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 ⁴⁴ as well as greatly reduce both the computational cost and the effect of spectral noise.

Furthermore, we used a Gaussian Process (GP) in order to incorporate prior knowledge into our models and generate an informed predictor ⁵³. 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.

![Figure 7](image)

**Figure 7** Gaussian Process Classifier prediction accuracies with corresponding average probability (“confidence”) for all chemically driven and cluster-driven classification schemes.

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.

By utilizing UMAP and analyzing the resulting clustering in a two-dimensional
embedding of VtC-XES and XANES spectra of an ensemble of organophosphorus compounds,
we noticed 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 (which resulted from an unexpected, unintuitive
fingerprint). However, all these results culminated 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,
visualized in Figure 1, is broadly applicable – it can easily be expanded to both other systems
and other one-dimensional spectroscopies – providing a way to validate predictions instead of
relying solely on the initial construction of an appropriate training dataset.
ASSOCIATED CONTENT

Supporting Information.

The following files are available free of charge:

**Computational Methods** (docx)

**Figure S1** Class averages of spectra with different coordination (png)

**Figure S2** Scree plot of VtC-XES and XANES data (png)

**Figure S3** PCA reconstruction of VtC-XES spectra (png)

**Figure S4** PCA reconstruction of XANES spectra (png)

**Figure S5** UMAP representation of XANES with H atom substitutions (png)

**Figure S6** Phosphate sub-cluster I example spectra (png)

**Figure S7** Phosphate sub-cluster II example spectra (png)

**Figure S8** Phosphate sub-cluster III example spectra (png)

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

**Figure S10** Phosphate sub-cluster I structures (png)

**Figure S11** Phosphate sub-cluster II structures (png)

**Figure S12** Phosphate sub-cluster III structures (png)

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

**Figure S14** Phosphate sub-clusters correlation (png)

**Figure S15** 3D UMAP visualizations (png)

**Table S1** Classification table (docx)

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.
REFERENCES


