THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA

Article

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

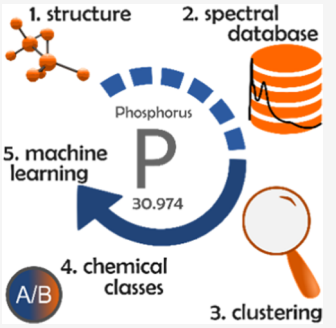
Samantha Tetef, Vikram Kashyap, William M. Holden, Alexandra Velian, Niranjan Govind, and Gerald T. Seidler*

Cite This: *J. Phys. Chem. A* 2022, 126, 4862–4872

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 that match our initial qualitative assessments of clustering. The results further support the benefit of utilizing unsupervised machine learning as a precursor to supervised machine learning, which we term Unsupervised Validation of Classes (UVC), a result that goes beyond the present case of X-ray spectroscopies.



I. INTRODUCTION

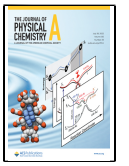
The information content in any spectroscopy method is constrained by the lossiness of the underlying quantum mechanics that connects an atomic-scale structure and dynamics to experimental observables. Further limitations to the sensitivity of spectroscopy techniques often include the inherently nonlinear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

In general, supervised ML studies across a wide range of spectroscopies target either predicting properties from spectra or correlating specific properties of interest to spectral features.[6] This necessarily assumes that sufficient information is, in fact, encoded in spectra; otherwise, supervised ML models will correlate spurious features to requested properties. This detail of encoded information is often addressed by hand-selecting a targeted training domain, an approach that is deeply contingent on the accuracy and completeness of prior knowledge.[7] Clearly, issues will 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 an effort to combat unwanted bias as well as provide generalizability to complex datasets, this study demonstrates the value of the discovery cycle exemplified in Figure 1. This process validates encoded information via unsupervised machine learning, i.e., cluster analysis on a reduced-dimensional embedding of the spectra, before passing either the embedding or the original spectra—selected as an unbiased training (sub)set—to a supervised ML model. This approach decreases the risk of implicit biases and spurious correlations introduced by supervised ML by adding steps (3) and (4) to validate spectral sensitivity of the training dataset to properties requested during supervised predictions. The continuation of inferences from supervised ML back to experimental design and (primarily) data simulation is obviously informed by the resulting errors achieved by the supervised ML model.

This cycle touches on other related ways people have used unsupervised ML as a precursor to or in a cycle with supervised ML. These approaches have included semi-supervised machine learning,[13] pretraining a neural network,[14]

Received: May 26, 2022
Revised: July 1, 2022
Published: July 15, 2022



https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

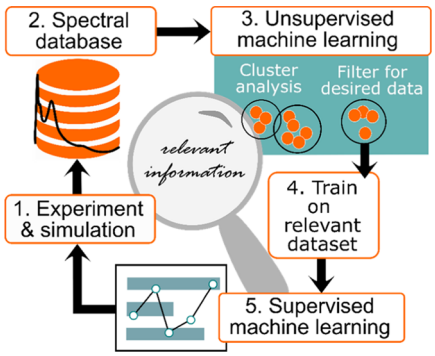


Figure 1. Flowchart of an analysis framework that uses unsupervised machine learning (such as cluster analysis) as a precursor to predictions on spectra via supervised machine learning, which can then inform experimental design and data creation.

feature selection or generation,[15] and human-in-the-loop learning,[16] which have been used in a multitude of fields, such as gene expression[17] and marketing.[18] Given the ubiquity of concerns about the scope and bias when constructing training datasets in supervised ML, we propose that our approach, which we term Unsupervised Validation of Classes (UVC), has relevance beyond the present case of X-ray spectroscopies as well as contributes to efforts to close the loop between artificial intelligence and scientific understanding.[19]

Here, we apply the framework of Figure 1 to both X-ray absorption spectroscopy (XAS) and X-ray emission spectroscopy (XES). XAS has seen an explosion of ML applications.[20−41] 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,[42,43] catalysis,[44] and photochemical dynamics.[45] XAS, which includes both the X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS), probes the unoccupied electronic states of the excited state of a chosen atomic species.

Conversely, relaxation to fill the core hole results in either nonradiative (Auger) or radiative processes. The latter results in the emission of X-ray fluorescence that can be finely characterized by XES for insight into the occupied electronic states.[46−48] Often discussed as complementary to XANES in information content, valence-to-core XES (VtC-XES) is produced when electrons deexcite from the valence shell to fill the core hole, giving direct information about the occupied electronic states involved in bonding.[49,50] 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 laboratory-based spectrometers,[51−53] including in studies of sufficient scale for data science methods.[54]

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

The present study is prompted by our recent work[55] 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 spectral features and thus important chemical information encoded in spectra. Moreover, our results exemplified the benefits of utilizing unsupervised ML to mold and understand the full potential of supervised ML analysis.[56]

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 proximity of phosphorus to sulfur in the periodic table allows for the same theoretical parameters to generate spectra (and thus obtain similar experimental agreement) as our previous study and also leverages the more diverse bonding environment of phosphorus. The dataset of spectra is calculated from molecular structures gathered from the PubChem[57] database using `moldl`, a new open-source tool that we have developed for this purpose.[58] 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 chemically relevant information that may be extracted using dimensionality reduction algorithms, especially when confined to very limited dimensions. These explorations allow for full utilization of real spectral information during supervised ML predictions.

To this end, we use the Uniform Manifold Approximation and Projection (UMAP)[59] 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),[60] which was used in our recent work[55] to extract chemical classes. Like t-SNE, UMAP constructs a graph-based representation of the data in the high-dimensional space to generate a similarity comparison, and then it tries to match the similarity comparison in a low-dimensional representation of the data. However, UMAP utilizes a different cost function, namely, cross-entropy instead of KL divergence, which further enables the global structure to be preserved, albeit at the cost of the "crowding problem".[60]

Moreover, given the proper choice in hyperparameters, UMAP can retain global similarity such that distances between clusters can be interpreted (given the manifold remains connected). This contrasts t-SNE, where its cost function,

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

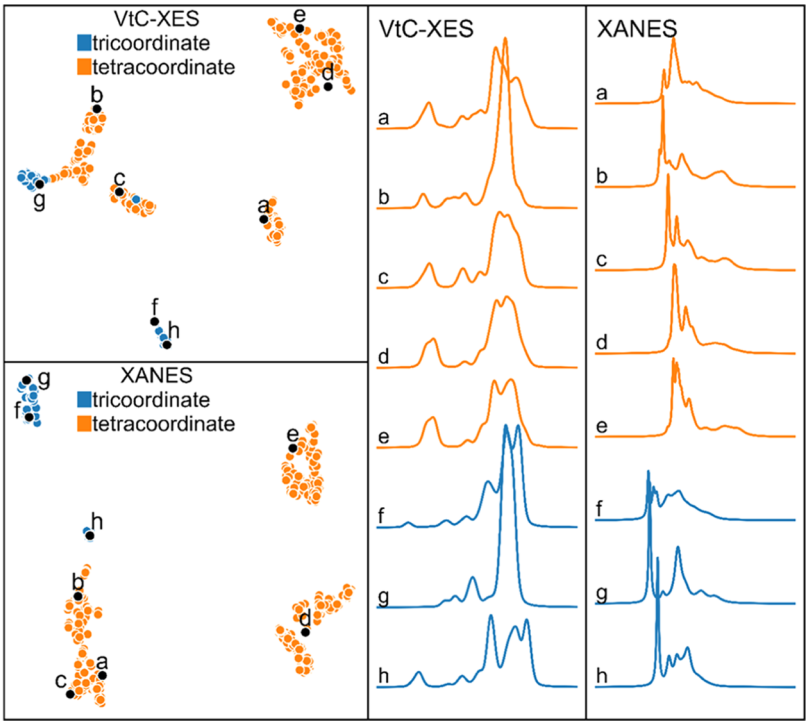


Figure 2. UMAP representation of VtC-XES (top) and XANES (bottom), color-coded by coordination, with some example spectra (as calculated by `NWChem`) shown to the right.

the KL divergence, goes to zero at large distances. The result is that t-SNE is not penalized for putting unalike data either far or very far away, and thus interpretation of similarity is only valid on a relatively local (intracluster) scale. These properties of UMAP allow it to generate a mapping function that can then be used to map subsequent data, which is why UMAP is called a "parametric embedding" and contrasts t-SNE's requirement that the entire training dataset must be used to predict new data. Thus, UMAP can be used for future data compression and has the potential for better interpretation of overall global similarities. These advantages have led to its recent popularity, such as in single-cell RNA sequencing (scRNA-seq) data analysis,[61] but UMAP has not yet seen use in XAS analysis.

II. METHODS

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

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

For VtC-XES, the calculated transitions were all shifted by −19 eV to align to the experiment.[68] A FWHM Lorentz broadening of 0.5 eV and a FWHM Gaussian broadening of 1.5 eV were added to each transition to agree with experimental data.[68] Because `NWChem` calculates a self-consistent field density functional theory (DFT) solution for both XANES[69] and VtC-XES,[70] this solution serves as a reference for the time-dependent DFT (TDDFT)-based X-ray spectroscopy calculations and thus only one internally consistent energy shift is required for each system.

Finally, both XANES and VtC-XES were individually normalized by their total Kα intensities. The Kα transitions scale in intensity proportional to the compound size (like VtC-XES and XANES calculations) but are very nearly independent of all environmental effects, thus providing an absolute scale to maintain relative intensities across the entire ensemble.

The sulforganics study of Holden et al.[54] for the experimental VtC-XES and `NWChem` calculations showed excellent agreement, as did additional calculations and comparison to XANES in Tetef et al.[55] Here, in Figures S1 and S2, we more modestly validate the performance of `NWChem` against several VtC-XES taken with the same instrument and methodology as Holden et al.,[54] and also

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

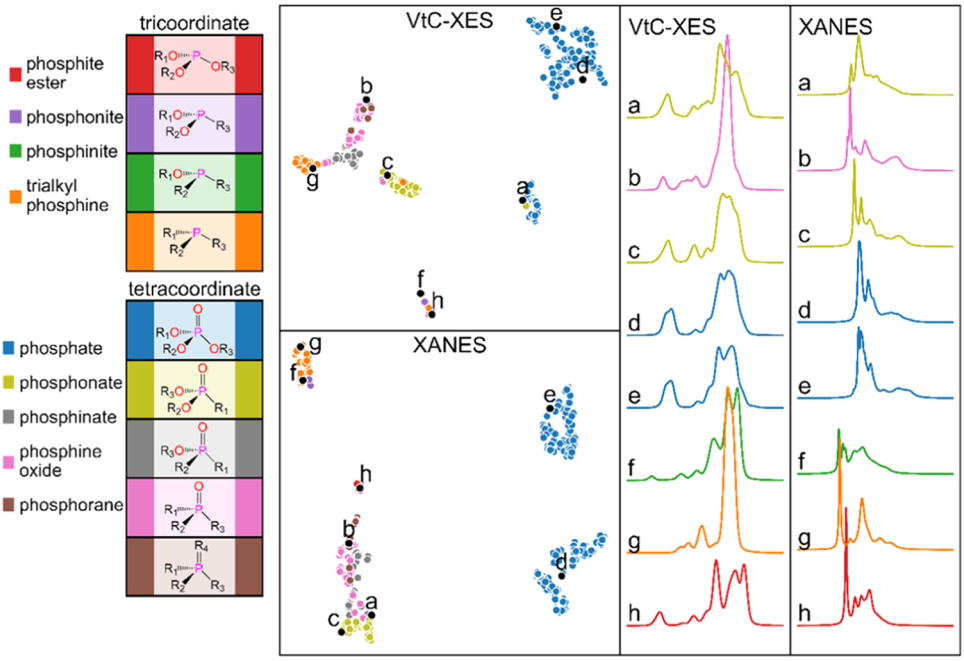


Figure 3. UMAP representation of VtC-XES (top) and XANES (bottom) for tricoordinate phosphorus and tetracoordinate phosphorus compounds, color-coded by number of oxygens bonded to phosphorus within each coordination. The same example spectra as before are shown to the right, as calculated by `NWChem`.

validate the performance against several XANES spectra from Persson et al.[67]

Briefly, Principal Component Analysis (PCA) was implemented using the `scikit-learn`[71] package in Python and was applied to the original spectra before UMAP to speed up computation and decrease noise. The number of principal components kept from the PCA was the number of components necessary to explain at least 95% of the variance in the dataset. For example, the number of retained principal components was 7 and 14 components for VtC-XES and XANES spectra, respectively, for the dataset consisting of all tricoordinate and tetracoordinate compounds, as shown in Figure S3. Some reconstructed spectra using the 95% variance cutoff are shown in Figures S4 and S5. The difference in the number of principal components required for VtC-XES and XANES suggests that XANES spectra have more variation and thus more nonlinear features, which is unsurprising.

UMAP was implemented using the `umap-learn` module[72] with default hyperparameters. Again, as mentioned above, to accelerate computing, the UMAP algorithm was applied to the PCA coefficients at the 95% variance level, thus decreasing the dimensionality of the training space from 1000 to either 7 or 14. For Figures 2−6, the number of UMAP output components was constrained to two for visualization purposes, while for Figure 7, the output dimensionality was set to five (found through the hyperparameter optimization discussed below).

Finally, to help illustrate the value of unsupervised ML as a precursor to supervised ML, we applied supervised ML in the form of a Gaussian process[73] classifier to the UMAP representation for all five classification schemes determined by the two-dimensional cluster analysis. The Gaussian processes were implemented using `scikit-learn`,[71] which utilizes the Laplace approximation as detailed by Rasmussen and Williams.[73] A separate classifier was trained for each of the five classification schemes, shown in Table S1, for both VtC-XES and XANES data.

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

III. RESULTS AND DISCUSSION

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

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

III.I. Unbiased Verification of Heuristic Classes. To begin, heuristically, one expects phosphorus coordination to yield the strongest distinguishing features between spectra, specifically the distinction between tricoordinate phosphorus

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

and tetracoordinate phosphorus. Not only do these coordination geometries have different hybridized orbital characters but they are also often a proxy for the oxidation state. In organophosphorus compounds with tricoordinate phosphorus centers, phosphorus is typically in a 3+ oxidation state, whereas compounds with tetracoordinate phosphorus centers usually have phosphorus in a 5+ oxidation state.

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

As expected, coordination distinguishes most of the groupings of the compounds, with a handful of outliers. We have further labeled some example compounds **a**−**h** (right panels) in each cluster with their corresponding VtC-XES and XANES spectra. (The identity of compounds **a**−**h** is defined in Table S2 in the supplementary section, but it is sufficient to say that they span a wide range of local coordination and oxidation states.) Note how some compounds that are in the same cluster in the VtC-XES embedding are in different clusters in the XANES embedding, and vice versa. For example, compounds **a**, **b**, and **c** are together in the XANES embedding, but they are in three different clusters in the VtC-XES embedding, which we will discuss later as being due to the number of oxygen ligands. These observations clearly indicate that VtC-XES and XANES encode information differently and that there are chemically relevant subgroupings within each coordination.

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

First, we expect effective charge of phosphorus to have the biggest impact on both VtC-XES and XANES spectra. For VtC-XES, the ligand peaks (the small low-energy peak in Figure S6) will increase in both energy and intensity with an increase in phosphorus oxidation. From a molecular orbital perspective, this trend is from both a larger overlap between the ligand valence orbital and the phosphorus 3p orbital (valence shell) and the increased number of oxygen ligands. In general, this feature (which also changes with different ligand symmetries and orientation) is why VtC-XES is strongly sensitive to ligand identity.[75] For XANES spectra, an increase in the oxidation of phosphorus, i.e., the number of oxygen ligands within a coordination, will cause a blue shift of the absorption edge, also demonstrated again by the average spectra in Figure S6.

Second, in terms of successful information encoding, we see that the number of oxygen ligands supplies much more information to explain the groupings in the UMAP representation than just coordination. For example, the highest oxidation compounds—the phosphates (blue)—are separated from all other compounds in both VtC-XES and XANES embeddings and are even subdivided into two clusters for both (this is due to a combination of chemical properties, which we will explore later in (Section III.II) and is the reason compounds **e** and **d** are separated in the XANES embedding but not the VtC-XES embedding).

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

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

Finally, moving to apparent outliers, one clear example is the location of compound **a**, diethyl (chloromethyl) phosphonate in the VtC-XES embedding. Compound **a** is a phosphonate but has a chlorinated carbon ligand, which effectively pulls more charge from phosphorus, thus making the carbon act more like an oxygen and the phosphorus having a higher oxidation. Likewise, both phosphonates, like compound **a**, in the phosphate cluster having a chlorinated R_1 ligand are thus grouped with the nominally "higher oxidation" compounds instead of the cluster with compound **c** (diethyl methanephosphonate).

As for further outliers, note that although compound **f** is a phosphinite with nominal P(III) oxidation from its tricoordinate P, it has a distinct number of oxygen ligands (one) compared to **g** (trialkyl phosphine, tricoordinate, no O ligands) and **h** (phosphite ester, tricoordinate, three O ligands). In these UMAP embeddings, compound **f** is more

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

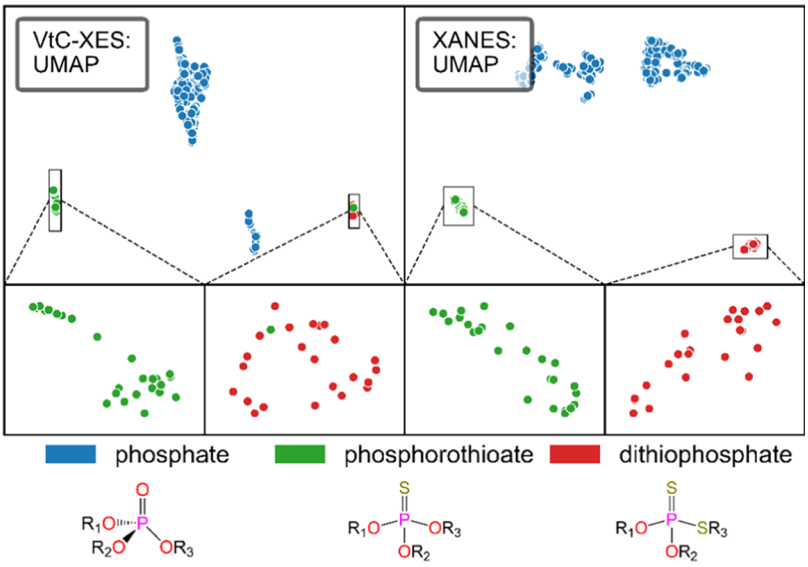


Figure 4. UMAP representation of VtC-XES (left) and XANES (right) for compounds with sulfur ligands, color-coded by the number of sulfurs. The pair of bottom insets on each panel are enlargements of the shown subregions to make it easier to see violations of cluster chemical classes, i.e., outlier compounds.

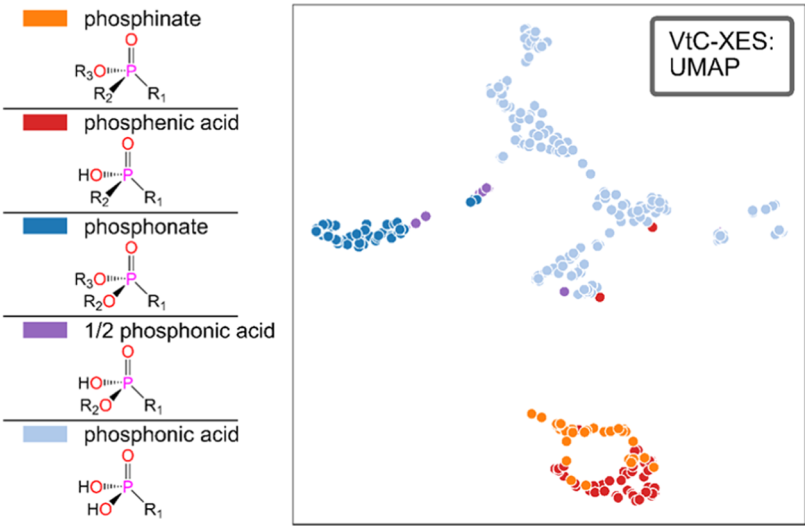


Figure 5. UMAP representation of the VtC-XES spectra of compounds with consecutively more R groups (if bonded to an oxygen) replaced with an H atom (to create hydroxyl groups), color-coded by chemical class.

similar to the higher oxidation compounds in VtC-XES compared to XANES spectra. Upon further examination, the other trialkyl phosphines in the cluster with compound **f** in the VtC-XES embedding are anomalous—they all have nitrile functional groups bonded to the phosphorus atom. Thus, in this case, VtC-XES seems to determine outliers more definitively than XANES, where the distinction falls on the second nearest-neighbor identity.

These observations bring us to our next hypothesis that VtC-XES and XANES are both sensitive to ligand identity. As stated earlier, VtC-XES is highly sensitive to ligand identity via changes in the ligand peak feature.[47] 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 XANES spectra.

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

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

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

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

In general, this distinction seems to be better illuminated by VtC-XES than the XANES (which is shown in Figure S8), as the clustering in 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 phosphorus, likely cause the biggest spectral changes and thus are the biggest contributing factor to clustering, which is consistent with our earlier observations.

III.II. Emergent Chemical Fingerprints from Clusters. Above, we motivated our classes by important chemical properties that we heuristically expected to yield the biggest spectral differences. However, even within this chemically driven framework, there are subclusters within our heuristic chemical classes that are instead emergent from UMAP. For example, we found that subclustering of the phosphate chemical class (exemplified by multiple separate subclusters 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 are sensitive to even more subtle details than anticipated.

Let us examine this subdivision of the phosphates, specifically in the UMAP embedding of their XANES spectra. For just phosphates, we achieve the embedding shown in Figure 6, which has labeled the phosphates into four clusters determined by the `dbscan`[77] 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 subcluster and 94% of the compounds have two hydroxyl groups bonded to phosphorus and the last R group an alkyl chain. These two clusters are the most distinct.

On the other hand, Clusters **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 phosphorus. Cluster **IV** compounds are structurally very similar to Cluster **III** compounds, even though their spectra are distinct. However, 54% of Cluster **IV** compounds have their third R group as aromatic rings. For some example compounds in each cluster along with their spectra and structure, see Figures S9−S12. All compounds in Clusters **I**−**IV** can also be viewed in Figures S13−S16. Additionally, given the linear nature of Clusters **I**, **III**, and **IV** in the UMAP embedding, we tested the correlation between the embedding location and the energy of the absorption edge, as shown in Figure S17, and found no strong correlation.

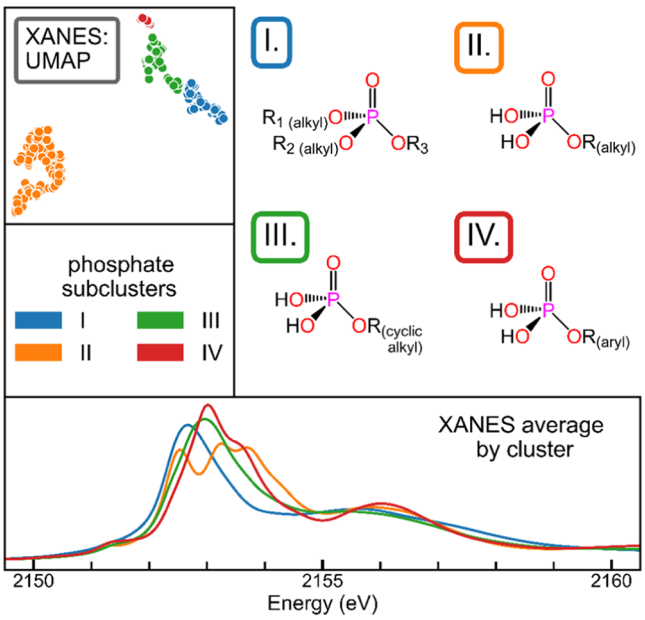


Figure 6. UMAP representation of the XANES spectra of phosphates, color-coded by subclusters. Cluster-averaged spectra and a summary of structural motifs for each cluster are also shown.

Furthermore, color-coding the phosphates based on a 10-dimensional clustering and then visualizing them in two dimensions yields very nearly the same classifications, as shown in Figure S18. Thus, the two-dimensional embedding is retaining enough information to categorize the phosphates appropriately. Even expanding the embedding space to three dimensions instead of two for all previous embeddings yields very nearly the same clustering, as shown in Figure S19. This retention in information—yet complex clustering of compounds—further supports the nonlinear nature of spectra and the idea that properties are complexly encoded in spectra and conversely, spectral features do not correlate solely to a single, high-variant attribute but rather a combination of electronic or chemical properties.

Taken en masse, these results show the extent to which chemically relevant information is, or is not, encoded by the quantum mechanics involved in XANES and VtC-XES. As to the specific algorithm, UMAP can be used iteratively as more data is collected. Thus, it has the potential to shown evolutions through the domain space, similar to the latent space of a variational autoencoder (VAE),[78] given proper tuning of its two hyperparameters: the number of expected neighbors in a cluster and the minimum distance between points. For an overview of the effect of those two hyperparameters on the

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

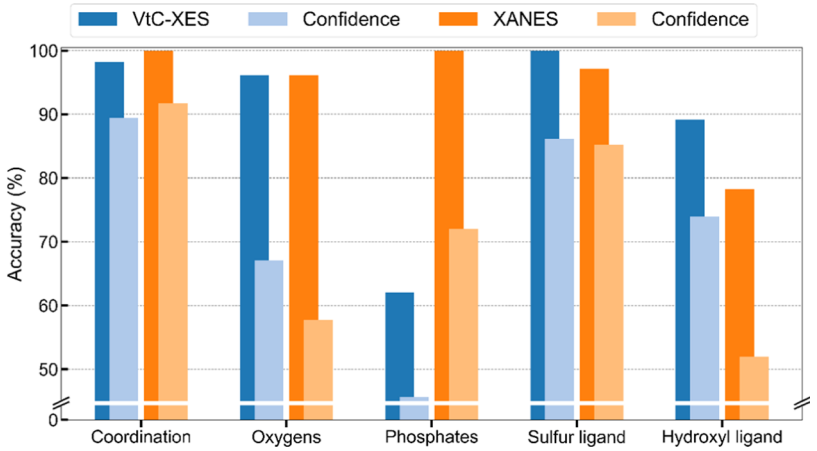


Figure 7. Gaussian process classifier prediction accuracies with corresponding average probability ("confidence") for all chemically driven and cluster-driven classification schemes.

UMAP embedding, see Figures S20 and S21. Finally, and of key importance here, UMAP can generate embeddings of spectra that can be used for unbiased refinement of the training dataset in addition to a preprocessing step before supervised ML predictions.

III.III. Validation of Chemical Fingerprints from Cluster Analysis. In our prior work on sulforganics[55] and in the present above work on the more complex case of organophosphorus compounds, we have demonstrated a convincing utility of advanced, nonlinear unsupervised ML tools for evaluating the chemically relevant information in VtC-XES and XANES spectra. We now return to our hypothesis presented in the introduction and illustrated in Figure 1, where we propose that such an unsupervised ML method can productively inform the use of supervised ML tasks.

The most common use of supervised ML in X-ray spectroscopy is to predict numerical properties, such as bond length or coordination, from XANES spectra.[20−22,24,31] 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[56] as well as greatly reduce both the computational cost and the effect of spectral noise.

Furthermore, we use a Gaussian process (GP) to incorporate prior knowledge into our models and generate an informed predictor.[73] A GP is a nonparametric kernel method that formally incorporates Bayes rule into the model, which not only allows for priors to be specified during training but also allows for a probabilistic interpretation of the results. This probability gives uncertainty estimates, or conversely confidence, of the predictions. We note that one of the biggest downsides of a GP is that it scales poorly, which is another reason why applying a nonlinear dimensionality reduction routine like UMAP beforehand can transform this problem into a computationally tractable one.

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

Finally, the accuracies and confidence of each prediction across VtC-XES and XANES data match what we observed in our two-dimensional UMAP figures. This correlation 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.

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

IV. CONCLUSIONS

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

These results culminate in a valuable analysis framework: (1) applying nonlinear dimensionality reduction routines and cluster analysis to check for both heuristic chemical sensitivities and emergent ones present in the spectra, (2) applying dimensionality reduction methods like UMAP before querying supervised ML models, and (3) utilizing models that incorporate prior knowledge, such as a Gaussian process, to estimate uncertainty or confidence of these predictions on the clustering-informed classes. Furthermore, this framework, which we call Unsupervised Validation of Classes (UVC) and illustrate in Figure 1, is broadly applicable—it can easily be expanded to both other systems and other spectroscopies—

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872

providing a way to inform methodology and validate predictions instead of relying solely on the scientist's knowledge and, possibly, bias in the initial construction of an appropriate training dataset.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.2c03635.

Theory versus experiment: VtC-XES (Figure S1); theory versus experiment: XANES (Figure S2); scree plot of VtC-XES and XANES data (Figure S3); PCA reconstruction of VtC-XES (Figure S4); PCA reconstruction of XANES spectra (Figure S5); class averages of spectra with different coordinations (Figure S6); cluster averages of spectra with different coordinations (Figure S7); UMAP representation of XANES with H atom substitutions (Figure S8); phosphate subcluster I example spectra (Figure S9); phosphate subcluster II example spectra (Figure S10); phosphate subcluster III example spectra (Figure S11); phosphate subcluster IV example spectra (Figure S12); phosphate subcluster I structures (Figure S13); phosphate subcluster II structures (Figure S14); phosphate subcluster III structures (Figure S15); phosphate subcluster IV structures (Figure S16); phosphate subcluster correlations (Figure S17); phosphate subclusters: 10-dimensional clustering (Figure S18); 3D UMAP visualizations (Figure S19); changing UMAP hyperparameters: number of neighbors (Figure S20); changing UMAP hyperparameters: minimum distance (Figure S21); classification table (Table S1); and table for compounds **a**–**h** (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Gerald T. Seidler − *Department of Physics, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0001-6738-7930; Email: seidler@uw.edu

Authors

[‖]**Samantha Tetef** − *Department of Physics, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-3098-8198

[‖]**Vikram Kashyap** − *Department of Physics, University of Washington, Seattle, Washington 98195, United States*

William M. Holden − *Department of Physics, University of Washington, Seattle, Washington 98195, United States*

Alexandra Velian − *Department of Chemistry, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0002-6782-7139

Niranjan Govind − *Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0003-3625-366X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.2c03635

Notes

The authors declare no competing financial interest.

[‖]S.T. and V.K. co-first authors.

■ ACKNOWLEDGMENTS

S.T. acknowledges funding from NRT-DESE: Data Intensive Research Enabling Clean Technologies (DIRECT) under grant no. NSF #1633216 and from NSF CHE-1904437. V.K. acknowledges support from the Washington NASA Space Grant from the Washington NASA Space Grant Consortium (WSGC). N.G. acknowledges support from the 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 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

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(2) Zhou, Z. Q.; He, Q. F.; Liu, X. D.; Wang, Q.; Luan, J. H.; Liu, C. T.; Yang, Y. Rational design of chemically complex metallic glasses by hybrid modeling guided machine learning. *npj Comput. Mater.* **2021**, *7*, No. 138.

(3) Liu, Y.; Zhao, T. L.; Ju, W. W.; Shi, S. Q. Materials discovery and design using machine learning. *J. Materiomics* **2017**, *3*, 159−177.

(4) Liu, Y.; Guo, B. R.; Zou, X. X.; Li, Y. J.; Shi, S. Q. Machine learning assisted materials design and discovery for rechargeable batteries. *Energy Storage Mater.* **2020**, *31*, 434−450.

(5) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501−1509.

(6) Ramirez, C. A. M.; Greenop, M.; Ashton, L.; Rehman, Iu. Applications of machine learning in spectroscopy. *Appl. Spectrosc. Rev.* **2021**, *56*, 733−763.

(7) Gordon, D. F.; Desjardins, M. Evaluation and selection of biases in machine learning. *Mach. Learn.* **1995**, *20*, 5−22.

(8) Wolpert, D. H.; Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67−82.

(9) Alelyani, S. Detection and evaluation of machine learning bias. *Appl. Sci.* **2021**, *11*, No. 6271.

(10) Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 1−35.

(11) Pot, M.; Kieusseyan, N.; Prainsack, B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights Imaging* **2021**, *12*, No. 13.

(12) Hiemstra, A. M. F.; Cassel, T.; Born, M. P.; Liem, C. C. S. The promises and perils of machine learning algorithms to reduce bias and discrimination in personnel selection procedures. *Gedrag Organ.* **2020**, *33*, 279−299.

(13) Belkin, M.; Niyogi, P.; Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **2006**, *7*, 2399−2434.

(14) Erhan, D.; Bengio, Y.; Courville, A.; Manzagol, P. A.; Vincent, P.; Bengio, S. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.* **2010**, *11*, 625−660.

(15) Liu, H.; Yu, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 491−502.

(16) Monarch, R. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*; Manning: Shelter Island, 2021.
(17) Omta, W. A.; van Heesbeen, R. G.; Shen, I.; de Nobel, J.; Robers, D.; van der Velden, L. M.; Medema, R. H.; Siebes, A. P. J. M.; Feelders, A. J.; Brinkkemper, S.; et al. Combining supervised and unsupervised machine learning methods for phenotypic functional genomics screening. *SLAS Discovery* **2020**, *25*, 655−664.
(18) Mathivanan, N. M. N.; Md Ghani, N. A.; Janor, R. M. Improving classification accuracy using clustering technique. *Bull. Electr. Eng. Inform.* **2018**, *7*, 465−470.
(19) Krenn, M.; Pollice, R.; Guo, S. Y.; Aldeghi, M.; Cervera-Lierta, A.; Friederich, P.; dos Passos Gomes, G.; Häse, F.; Jinich, A.; Nigam, A.et al.On scientific understanding with artificial intelligence, arXiv:2204.01467. arXiv.org e-Print archive. https://arXiv.2204.01467, 2022.
(20) Timoshenko, J.; Lu, D. Y.; Lin, Y. W.; Frenkel, A. I. Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles. *J. Phys. Chem. Lett.* **2017**, *8*, 5091−5098.
(21) Timoshenko, J.; Frenkel, A. I. "Inverting" X-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *ACS Catal.* **2019**, *9*, 10192−10211.
(22) Timoshenko, J.; Anspoks, A.; Cintins, A.; Kuzmin, A.; Purans, J.; Frenkel, A. I. Neural network approach for characterizing structural transformations by X-ray absorption fine structure spectroscopy. *Phys. Rev. Lett.* **2018**, *120*, No. 225502.
(23) Timoshenko, J.; Wrasman, C. J.; Luneau, M.; Shirman, T.; Cargnello, M.; Bare, S. R.; Aizenberg, J.; Friend, C. M.; Frenkel, A. I. Probing atomic distributions in mono- and bimetallic nanoparticles by supervised machine learning. *Nano Lett.* **2019**, *19*, 520−529.
(24) Zheng, C.; Chen, C.; Chen, Y.; Ong, S. P. Random forest models for accurate identification of coordination environments from X-Ray absorption near-edge structure. *Patterns* **2020**, *1*, No. 100013.
(25) Zheng, C.; Mathew, K.; Chen, C.; Chen, Y. M.; Tang, H. M.; Dozier, A.; Kas, J. J.; Vila, F. D.; Rehr, J. J.; Piper, L. F. J.; et al. Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Comput. Mater.* **2018**, *4*, No. 12.
(26) Kiyohara, S.; Miyata, T.; Tsuda, K.; Mizoguchi, T. Data-driven approach for the prediction and interpretation of core-electron loss spectroscopy. *Sci. Rep.* **2018**, *8*, No. 13548.
(27) Routh, P. K.; Liu, Y.; Marcella, N.; Kozinsky, B.; Frenkel, A. I. Latent representation learning for structural characterization of catalysts. *J. Phys. Chem. Lett.* **2021**, *12*, 2086−2094.
(28) Aarva, A.; Deringer, V. L.; Sainio, S.; Laurila, T.; Caro, M. A. Understanding X-ray spectroscopy of carbonaceous materials by combining experiments, density functional theory, and machine learning. Part I: Fingerprint spectra. *Chem. Mater.* **2019**, *31*, 9243−9255.
(29) Carbone, M. R.; Yoo, S.; Topsakal, M.; Lu, D. Classification of local chemical environments from x-ray absorption spectra using supervised machine learning. *Phys. Rev. Mater.* **2019**, *3*, No. 033604.
(30) Carbone, M. R.; Topsakal, M.; Lu, D.; Yoo, S. Machine-learning X-ray absorption spectra to quantitative accuracy. *Phys. Rev. Lett.* **2020**, *124*, No. 156401.
(31) Liu, Y.; Marcella, N.; Timoshenko, J.; Halder, A.; Yang, B.; Kolipaka, L.; Pellin, M. J.; Seifert, S.; Vajda, S.; Liu, P.; Frenkel, A. I. Mapping XANES spectra on structural descriptors of copper oxide clusters using supervised machine learning. *J. Chem. Phys.* **2019**, *151*, No. 164201.
(32) Martini, A.; Guda, S. A.; Guda, A. A.; Smolentsev, G.; Algasov, A.; Usoltsev, O.; Soldatov, M. A.; Bugaev, A.; Rusalev, Y.; Lamberti, C.; Soldatov, A. PyFitit: The software for quantitative analysis of XANES spectra using machine-learning algorithms. *Comput. Phys. Commun.* **2020**, *250*, No. 107064.
(33) Miyazato, I.; Takahashi, L.; Takahashi, K. Automatic oxidation threshold recognition of XAFS data using supervised machine learning. *Mol. Syst. Des. Eng.* **2019**, *4*, 1014−1018.
(34) Guda, A. A.; Guda, S. A.; Martini, A.; Kravtsova, A. N.; Algasov, A.; Bugaev, A.; Kubrin, S. P.; Guda, L. V.; Šot, P.; van Bokhoven, J. A.; et al. Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms. *npj Comput. Mater.* **2021**, *7*, No. 203.
(35) Fang, Z.; Hu, W.; Wang, M.; Wang, R.; Zhong, S.; Chen, S. X-ray absorption spectroscopy combined with machine learning for diagnosis of schistosomiasis cirrhosis. *Biomed. Signal Process. Control* **2020**, *60*, No. 101944.
(36) Torrisi, S. B.; Carbone, M. R.; Rohr, B. A.; Montoya, J. H.; Ha, Y.; Yano, J.; Suram, S. K.; Hung, L. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comput. Mater.* **2020**, *6*, No. 109.
(37) Trejo, O.; Dadlani, A. L.; De La Paz, F.; Acharya, S.; Kravec, R.; Nordlund, D.; Sarangi, R.; Prinz, F. B.; Torgersen, J.; Dasgupta, N. P. Elucidating the evolving atomic structure in atomic layer deposition reactions with in situ XANES and machine learning. *Chem. Mater.* **2019**, *31*, 8937−8947.
(38) Rankine, C. D.; Madkhali, M. M. M.; Penfold, T. J. A deep neural network for the rapid prediction of X-ray absorption spectra. *J. Phys. Chem. A* **2020**, *124*, 4263−4270.
(39) Rankine, C. D.; Penfold, T. J. Progress in the theory of X-ray spectroscopy: From quantum chemistry to machine learning and ultrafast dynamics. *J. Phys. Chem. A* **2021**, *125*, 4276−4293.
(40) Kiyohara, S.; Tsubaki, M.; Mizoguchi, T. Learning excited states from ground states by using an artificial neural network. *npj Comput. Mater.* **2020**, *6*, No. 68.
(41) Usoltsev, O. A.; Bugaev, A. L.; Guda, A. A.; Guda, S. A.; Soldatov, A. V. How much structural information could be extracted from XANES spectra for palladium hydride and carbide nanoparticles. *J. Phys. Chem. C* **2022**, *126*, 4921−4928.
(42) Cuisinier, M.; Cabelguen, P.-E.; Evers, S.; He, G.; Kolbeck, M.; Garsuch, A.; Bolin, T.; Balasubramanian, M.; Nazar, L. F. Sulfur speciation in Li−S batteries determined by operando X-ray absorption spectroscopy. *J. Phys. Chem. Lett.* **2013**, *4*, 3227−3232.
(43) Asakura, D.; Hosono, E.; Niwa, H.; Kiuchi, H.; Miyawaki, J.; Nanba, Y.; Okubo, M.; Matsuda, H.; Zhou, H.; Oshima, M.; et al. Operando soft x-ray emission spectroscopy of LiMn2O4 thin film involving Li−ion extraction/insertion reaction. *Electrochem. Commun.* **2015**, *50*, 93−96.
(44) Zhou, Y.; Doronkin, D. E.; Zhao, Z.; Plessow, P. N.; Jelic, J.; Detlefs, B.; Pruessmann, T.; Studt, F.; Grunwaldt, J. Photothermal catalysis over nonplasmonic Pt/TiO2 studied by operando HERFD-XANES, resonant XES, and DRIFTS. *ACS Catal.* **2018**, *8*, 11398−11406.
(45) Maiuri, M.; Garavelli, M.; Cerullo, G. Ultrafast spectroscopy: State of the art and open challenges. *J. Am. Chem. Soc.* **2020**, *142*, 3−15.
(46) Bunker, G. *Introduction to XAFS: A Practical Guide to X-ray Absorption Fine Structure Spectroscopy*; Cambridge University Press: Cambridge, 2010.
(47) Glatzel, P.; Bergmann, U. High resolution 1s core hole X-ray spectroscopy in 3d transition metal complexes—electronic and structural information. *Coord. Chem. Rev.* **2005**, *249*, 65−95.
(48) de Groot, F. High-resolution X-ray emission and X-ray absorption spectroscopy. *Chem. Rev.* **2001**, *101*, 1779−1808.
(49) Lee, N.; Petrenko, T.; Bergmann, U.; Neese, F.; DeBeer, S. Probing valence orbital composition with iron K beta X-ray emission spectroscopy. *J. Am. Chem. Soc.* **2010**, *132*, 9715−9727.
(50) Pollock, C. J.; DeBeer, S. Insights into the geometric and electronic structure of transition metal centers from valence-to-core X-ray emission spectroscopy. *Acc. Chem. Res.* **2015**, *48*, 2967−2975.
(51) Seidler, G. T.; Mortensen, D. R.; Remesnik, A. J.; Pacold, J. I.; Ball, N. A.; Barry, N.; Styczinski, M.; Hoidn, O. R. A laboratory-based hard x-ray monochromator for high-resolution x-ray emission spectroscopy and x-ray absorption near edge structure measurements. *Rev. Sci. Instrum.* **2014**, *85*, No. 113906.
(52) Malzer, W.; Schlesiger, C.; Kanngießer, B. A century of laboratory X-ray absorption spectroscopy − A review and an optimistic outlook. *Spectrochim. Acta, Part B* **2021**, *177*, No. 106101.

(53) Zimmermann, P.; Peredkov, S.; Abdala, P. M.; DeBeer, S.; Tromp, M.; Müller, C.; van Bokhoven, J. A. Modern X-ray spectroscopy: XAS and XES in the laboratory. *Coord. Chem. Rev.* **2020**, *423*, No. 213466.
(54) Holden, W. M.; Jahrman, E. P.; Govind, N.; Seidler, G. T. Probing sulfur chemical and electronic structure with experimental observation and quantitative theoretical prediction of Kα and valence-to-core Kβ X-ray emission spectroscopy. *J. Phys. Chem. A* **2020**, *124*, 5415−5434.
(55) Tetef, S.; Govind, N.; Seidler, G. T. Unsupervised machine learning for unbiased chemical classification in X-ray absorption spectroscopy and X-ray emission spectroscopy. *Phys. Chem. Chem. Phys.* **2021**, *23*, 23586−23601.
(56) Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* **2019**, *150*, No. 150901.
(57) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.
(58) github.com/vikramkashyap/moldl.
(59) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction, arXiv:1802.03426. arXiv.org e-Print archive. https://arXiv.1802.03426, 2020.
(60) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(61) Pont, F.; Tosolini, M.; Fournie, J. J. Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic Acids Res.* **2019**, *47*, No. e133.
(62) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.
(63) Aprà, E.; Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; van Dam, H. J. J.; Alexeev, Y.; Anchell, J.; et al. NWChem: Past, present, and future. *J. Chem. Phys.* **2020**, *152*, No. 184102.
(64) Noro, T.; Sekiya, M.; Koga, T. Segmented contracted basis sets for atoms H through Xe: Sapporo-(DK)-nZP sets (n = D, T, Q). *Theor. Chem. Acc.* **2012**, *131*, No. 1124.
(65) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(66) Bergner, A.; Dolg, M.; Küchle, W.; Stoll, H.; Preuß, H. Ab initio energy-adjusted pseudopotentials for elements of groups 13−17. *Mol. Phys.* **1993**, *80*, 1431−1441.
(67) Persson, I.; Klysubun, W.; Lundberg, D. A K-edge P XANES study of phosphorus compounds in solution. *J. Mol. Struct.* **2019**, *1179*, 608−611.
(68) Yasuda, S. Chemical effects on the X-ray K emission spectra of phosphorus in organic compounds. *Bull. Chem. Soc. Jpn.* **1984**, *57*, 3122−3124.
(69) Lopata, K.; Van Kuiken, B. E.; Khalil, M.; Govind, N. Linear-response and real-time time-dependent density functional theory studies of core-level near-edge X-ray absorption. *J. Chem. Theory Comput.* **2012**, *8*, 3284−3292.
(70) Zhang, Y.; Mukamel, S.; Khalil, M.; Govind, N. Simulating valence-to-core X-ray emission spectroscopy of transition metal complexes with time-dependent density functional theory. *J. Chem. Theory Comput.* **2015**, *11*, 5804−5809.
(71) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(72) Sainburg, T.; McInnes, L.; Gentner, T. Q. Parametric UMAP embeddings for representation and semisupervised learning. *Neural Comput.* **2021**, *33*, 1−27.
(73) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2006.
(74) github.com/Seidler-Lab/Phosphorus-ML-Project.
(75) Rovezzi, M.; Glatzel, P. Hard X-ray emission spectroscopy: a powerful tool for the characterization of magnetic semiconductors. *Semicond. Sci. Technol.* **2014**, *29*, No. 023002.
(76) Murphy, L. R.; Meek, T. L.; Allred, A. L.; Allen, L. C. Evaluation and test of Pauling's electronegativity scale. *J. Phys. Chem. A* **2000**, *104*, 5867−5871.
(77) Hahsler, M.; Piekenbrock, M.; Doran, D. dbscan: Fast density-based clustering with R. *J. Stat. Software* **2019**, *91*, 1−30.
(78) Shrestha, A.; Mahmood, A. Review of deep learning algorithms and architectures. *IEEE Access* **2019**, *7*, 53040−53065.

https://doi.org/10.1021/acs.jpca.2c03635
J. Phys. Chem. A 2022, 126, 4862−4872