npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships

Steven B. Torrisi[1,2]✉, Matthew R. Carbone[3], Brian A. Rohr[1], Joseph H. Montoya[1], Yang Ha[4], Junko Yano[5], Santosh K. Suram[1]✉ and Linda Hung[1]✉

X-ray absorption spectroscopy (XAS) produces a wealth of information about the local structure of materials, but interpretation of spectra often relies on easily accessible trends and prior assumptions about the structure. Recently, researchers have demonstrated that machine learning models can automate this process to predict the coordinating environments of absorbing atoms from their XAS spectra. However, machine learning models are often difficult to interpret, making it challenging to determine when they are valid and whether they are consistent with physical theories. In this work, we present three main advances to the data-driven analysis of XAS spectra: we demonstrate the efficacy of random forests in solving two new property determination tasks (predicting Bader charge and mean nearest neighbor distance), we address how choices in data representation affect model interpretability and accuracy, and we show that multiscale featurization can elucidate the regions and trends in spectra that encode various local properties. The multiscale featurization transforms the spectrum into a vector of polynomial-fit features, and is contrasted with the commonly-used "pointwise" featurization that directly uses the entire spectrum as input. We find that across thousands of transition metal oxide spectra, the relative importance of features describing the curvature of the spectrum can be localized to individual energy ranges, and we can separate the importance of constant, linear, quadratic, and cubic trends, as well as the white line energy. This work has the potential to assist rigorous theoretical interpretations, expedite experimental data collection, and automate analysis of XAS spectra, thus accelerating the discovery of new functional materials.

npj Computational Materials (2020)6:109 ; https://doi.org/10.1038/s41524-020-00376-6

INTRODUCTION

Rapid extraction of structure-property relationships is critical to the discovery of functional materials. One avenue to accelerate this process involves the use of machine learning (ML) models, which are becoming more reliable with the availability of libraries generated by high-throughput materials experiments and calculations[1–8]. Using these libraries, data-driven techniques are now powerful enough that bulk structure-property relationships can be extracted from experimental X-ray diffraction data using automated agents[9,10]. Data-driven probes of relevant local properties (such as those descriptive of electrochemical behavior[11]) could further help to accelerate scientific discovery, with the ultimate promise of *in operando* characterization and automated planning of experiments[2,12,13]. However, accessing local chemical properties of metal centers by spectroscopy can be challenging due to the contribution of multiple factors, such as ligand type, coordination numbers, and charge and spin states.

X-ray absorption spectroscopy (XAS)[14,15] is a characterization technique that is sensitive to local electronic and atomic structure, and has been important for discovering and understanding functional materials for a wide range of energy applications, such as CO_2 capture by metal oxide nanoparticles[16,17], solar water splitting[18], and catalysis[19–21]. It is particularly suitable as a local probe thanks to its general robustness, large signal-to-noise ratio[22], element specificity, and unique sensitivity to the chemical environments of absorbing atoms[23–27]. A given XAS spectrum is unique to the absorbing atom and the edge energy (from which a core electron is excited), and can be divided into the X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS) regions. Each region carries unique information about the environment of absorbing atoms. Among other things, XANES encodes symmetry and electronic structure information of the absorbing site, while EXAFS expresses the structure of excitations and back-scattering at photoelectron energies exceeding the threshold for ionization[26,27], thus containing information about neighboring atoms and excited state phenomena. Because XANES spectra have an inherently higher signal-to-noise ratio, they can be sampled in orders of magnitude less time compared to EXAFS. Targeted spectrum-property signatures accessible via XANES could thus enable high-throughput experiments by targeting maximally informative regions of interest. The impact of automation in analysis and extraction of new spectrum-property trends make advances in XANES characterization highly desirable to the community.

The XANES region is comprised of the pre-edge, rising-edge, and post-edge regions. Over the years, certain trends in each region have been identified which help characterize different local chemical and structural properties such as oxidation state[28,29] and coordination environment[30–34]. Some trends associated with electronic transitions can be explained by quantum mechanical symmetry arguments; for example, Farges and co-workers[30–32]

[1]Accelerated Materials Design and Discovery, Toyota Research Institute, Los Altos, CA 94022, USA. [2]Department of Physics, Harvard University, Cambridge, MA 02138, USA. [3]Department of Chemistry, Columbia University, New York, NY 10027, USA. [4]Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [5]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ✉email: torrisi@g.harvard.edu; santosh.suram@tri.global; linda.hung@tri.global

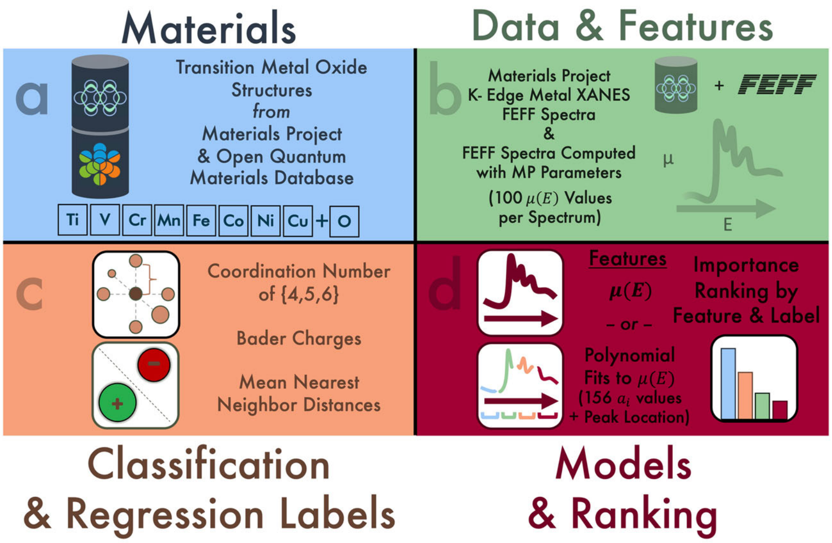
Materials
Data & Features
a
Transition Metal Oxide Structures from Materials Project & Open Quantum Materials Database
Ti V Cr Mn Fe Co Ni Cu + O
b
Materials Project K- Edge Metal XANES FEFF Spectra & FEFF Spectra Computed with MP Parameters (100 $\mu(E)$ Values per Spectrum)
+ FEFF
μ
E
c
Coordination Number of {4,5,6}
Bader Charges
Mean Nearest Neighbor Distances
d
Features
$\mu(E)$
– or –
Polynomial Fits to $\mu(E)$ (156 a_i values + Peak Location)
Importance Ranking by Feature & Label
Classification & Regression Labels
Models & Ranking

Fig. 1 Workflow summary. Visual description of our workflow. **a** The materials that we study consist of 3*d* transition metal oxide structures drawn from the Materials Project (MP) database[7] and Open Quantum Materials Database (OQMD)[8]. **b** The inputs to our ML models are XANES spectra computed using FEFF 9[40], either downloaded from the MP or computed using the same set of parameters. **c** The local properties to be predicted from spectra are the coordination number (limited to 4, 5, or 6), the mean nearest-neighbor distance, and the Bader charge. **d** The models we train are random forests, where features are either the entire spectra projected onto a uniformly spaced 100-point energy grid, or the coefficients of overlapping polynomials fit to partitions of the spectra. Feature rankings from the two different featurizations are compared to each other and to known trends.

showed that in Ti and Ni complexes, pre-edge peak intensity is suppressed as the coordination number increases from $4 \rightarrow 6$, since the number of locally coordinated metal ligands can allow certain electronic transitions (like ligand p-orbital mixing) in the pre-edge energy region[28,35,36]. Further, for some Ni materials, this pre-edge feature was found to correlate with the Ni-O distance[32]. Similarly useful, but limited fingerprints exist for other properties of interest. For instance, for some materials, the peak location clearly shifts with the oxidation state[27].

However, depending on the absorbing atom, ligand types, and property of interest, previously known heuristics may not provide sufficient information, and close study of a small group of compounds is required to discover and understand new trends. Farges pointed out that conventional heuristics fail for MnO and $MnCO_3$, which share the same oxidation state and coordination number, yet exhibit radically different Mn K-edge shapes and a large 4.5 eV shift between their peaks[37]. Known coordination number heuristics are also not as straightforward to apply to heavier transition metals like Fe[33,38,39]. Furthermore, oxidation number heuristics differ based on the type(s) of ligand anions due to differences in charge transfer processes.

In cases where heuristics fail, researchers typically must rely on their existing knowledge of specific materials' spectra, and also use software such as FEFF[40] to predict theoretical spectra from input crystal structures. These known or computational spectra can be used either through direct comparison or through specialized algorithms, such as those in MXAN[41–43] and Pyfitit[44,45]. This theoretical mapping of crystal structure to spectra can provide a thorough understanding of the materials under investigation, as long as the material's structure can be identified. However, identifying the structure can be computationally expensive if spectra for multiple geometrical configurations must be computed during the search. The experimental material may also have a structure that is significantly different from anything in the current library or list of structural candidates (the sample could even be amorphous) which would make it difficult to apply these techniques.

Due to the desire to rapidly characterize spectra for arbitrary local environments, data-driven methods for XAS are now enjoying great interest across various communities[46–51]; for a general overview, we recommend a review by Timoshenko and Frenkel[52]. These methods attempt to exploit all of the information contained within a spectrum, as opposed to the subset which a heuristic describes, and are enabled by the high availability of theoretical data and the promise of high-throughput experimental XAS data. ML models have been used to automate the analysis of experimental XANES and EXAFS data to gain insights into system properties and behavior[49,51,53,54]. Previous work has demonstrated the feasibility of classifying certain structural properties, such as oxidation number and coordination, from said spectra via ensemble learning[55,56]. Recent work has also used artificial neural networks[46] and random forests[47] to focus on coordination alone.

Comparatively less attention has been paid to data-assisted discovery of new spectrum-property trends with XANES[57], although feature ranking of input values to random forests have seen success in discovering heuristics for materials behavior[58,59]. Studies which focus on interpretable ML models are of particular interest to the field, as they could yield physical insights from automatically highlighted spectrum-property patterns. At a coarser level of XANES interpretation, some works have studied the importance of different spectral regions: the post-edge region of XANES spectra has been shown to be important to coordination classification from a marked drop in performance after training a model without it[46]. A split into to three energy regions before, at, and after the edge using random forests compared the relative information content in different energy regions[47].

In this work, we demonstrate three new capabilities of interpretable ML models for XAS. First, we demonstrate successful machine-learned analysis of XANES data for coordination, in agreement with prior work[46,47], and to our knowledge, the first models for mean nearest-neighbor distance and Bader charge (the partial charge distribution on atoms, which is correlated with the oxidation state—see "Methods" section and Supplementary Figs. 4–11). Secondly, we represent XANES spectra using two types of featurization and two choices of normalization, and find that models using post-edge normalized data and multiscale polynomial featurization are readily interpretable and conform with multiple known trends. Finally, we use the multiscale polynomial featurization to uncover high-level spectrum-property relationships for experimentally relevant material parameters, comparing the relative importance of white line energy, absorption magnitude, and linear/quadratic/cubic character of spectra across elements and properties. The results of this study highlight which energy regions are important, which could accelerate experimental structural characterization depending on individual elements and use cases. While most heuristics apply to a smaller subset of materials, these trends are based on thousands of spectra, and so may be applicable to a wide range of chemistries and structures.

Figure 1 gives an overview of our workflow. Thousands of computational XANES spectra are used to train random forest models for eight absorbing elements; models specific to each element take a spectrum as input to predict the absorbing atom's local coordination number, mean nearest-neighbor distance, and Bader charge. We demonstrate strong performance for coordination number and regression of the mean nearest-neighbor distances, and for some elements, show that Bader charges can be accurately reproduced. We then compare and contrast the interpretability based on property, featurization, and normalization.

The XANES spectra used as input to models in the main text are all post-edge normalized (we address the effect of normalization in the discussion section and SI), and are featurized in two different ways in order to coax enhanced interpretability from our models. We refer to featurization based on the full 100 equally spaced values of the spectrum as "pointwise" featurization. Pointwise feature ranking illuminates which energy regions matter, but it is unclear from the ranking alone if the slopes or peaks of the spectra are what drive the predictions, or if the spectral magnitudes are the primary contributors. This motivates the use of a multiscale polynomial featurization which is new to this work. Our featurization technique (depicted in Fig. 2) captures information about the entire spectrum on both a coarse and fine level, describing trends in subdivided domains ranging from 2.5 eV to 12.5 eV. Because the polynomial terms correspond to the local constant, linear, quadratic, and cubic character of the spectra, tracking the importance of each coefficient conveys if the magnitude (a_0) of $\vec{\mu}(E)$ is most important for the prediction task, or if the local derivatives as parameterized by polynomial fits (a_1, a_2, a_3) are what matter. This disambiguates which aspects of the spectrum are contributing to the decisions of the model. We additionally add to the vector the white line energy, or peak absorption energy. We find that trends in feature importance are highly dependent on the element in question, the prediction task at hand, and the normalization. Crucially, we find that featurizing the spectra with overlapping polynomial fits usually results in similar or improved model performance and improved interpretability compared to pointwise featurization.

RESULTS

We present the performance across our pointwise-featurized models in Fig. 3 and across our polynomially-featurized models in Fig. 4. These plots include the accuracies (the percentage of the test set classified correctly) and F1 scores (See Eq. 7 in SI) associated with specific coordination number classes, as well as coefficients of determination R^2 (See Eq. 5 in SI) for nearest-neighbor distance and Bader charge regression. Detailed metrics and plots comparing the two performances are available in the SI; we see only nominal changes in performance between the two featurizations, with few exceptions (See Supplementary Fig. 1). The baseline accuracy shown in the figures refers to a naive model which simply guesses the modal class of 4-fold, 5-fold, or 6-fold coordination depending on the element. The baseline accuracy is higher for elements with a large modal class, and comparing baseline accuracy to accuracy shows the resolving power that the model has 'learned'[47].

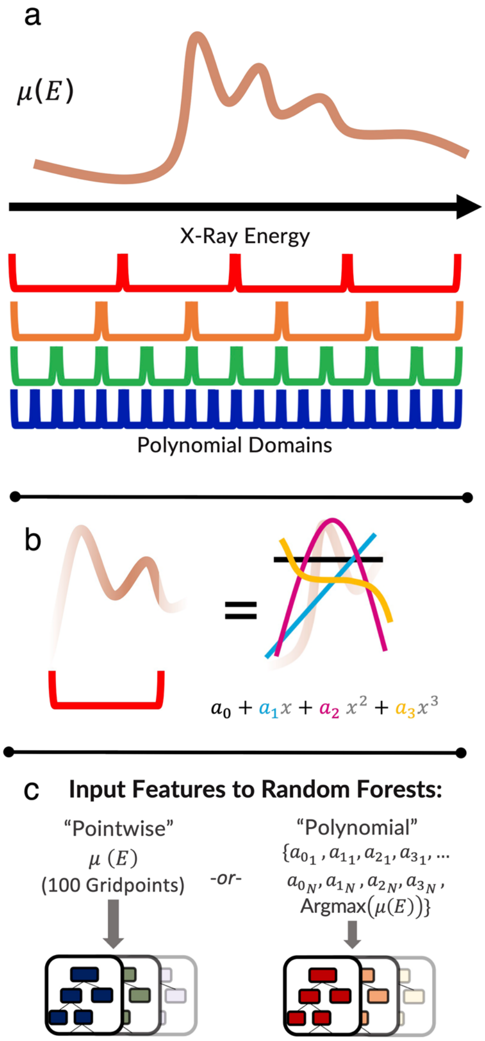
a
μ(E)
X-Ray Energy
Polynomial Domains
b
=
$a_0 + a_1x + a_2x^2 + a_3x^3$
c Input Features to Random Forests:
"Pointwise"
μ (E)
(100 Gridpoints)
-or-
"Polynomial"
{$a_{0_1}, a_{1_1}, a_{2_1}, a_{3_1}, \ldots$
$a_{0_N}, a_{1_N}, a_{2_N}, a_{3_N}$,
Argmax(μ(E))}

Fig. 2 Featurization summary. **a** We start with the pointwise representation of spectra, where values of $\vec{\mu}(E)$ are projected onto 100 uniformly-spaced values of E. For $N = 4$, 5, 10, and 20, we then partition the spectrum into N equally-sized regions. **b** Cubic polynomials $a_0 + a_1x + a_2x^2 + a_3x^3$ are fit to the spectrum within each partition (setting $x = 0$ at the center of the partition). **c** Each polynomial thus yields four coefficients (constant, linear, quadratic, and cubic), which are used as input features to the model, for $(4 + 5 + 10 + 20) \times 4 = 156$ total polynomial coefficients, and 157 features total when including the peak energy value (a.k.a. the white line energy).

In the following subsections, we comment on model performance and the interpretations which follow from the models for each property prediction task in turn. By comparing the feature rankings from pointwise and polynomial featurizations, we confirm which regions substantively inform model predictions, and identify relevant local spectral behavior in the regions of interest.

Coordination number

For models predicting coordination number, we find average accuracies of 85.3/85.4% and average F1 scores of 81.8/81.7% for the pointwise/polynomial fits, respectively. Our coordination number dataset has significant class imbalances for all metals except for V (see Table 1), and we find that the F1 scores tend to

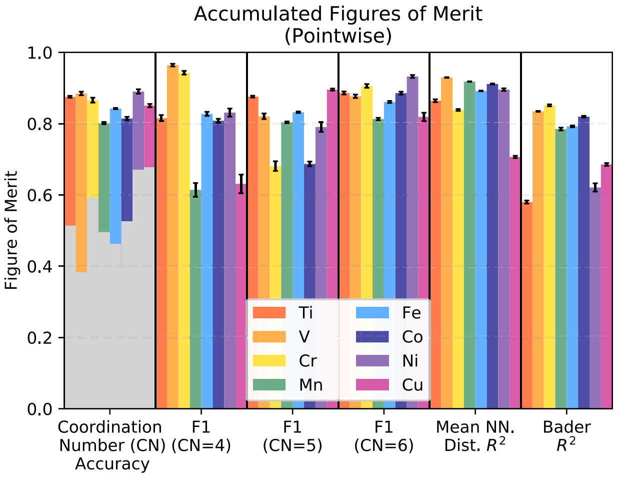

Fig. 3 Performance of random forest models using the pointwise featurization of the spectrum $\mu(E)$ as input. Baseline accuracy is shown in gray for the coordination number classification accuracy, and describes performance of a naive model which simply guesses the modal class. F1 scores for coordination number models are presented for each class. Error bars represent ±1 standard deviation obtained from 10 random forests trained on the same data with different random seeds.

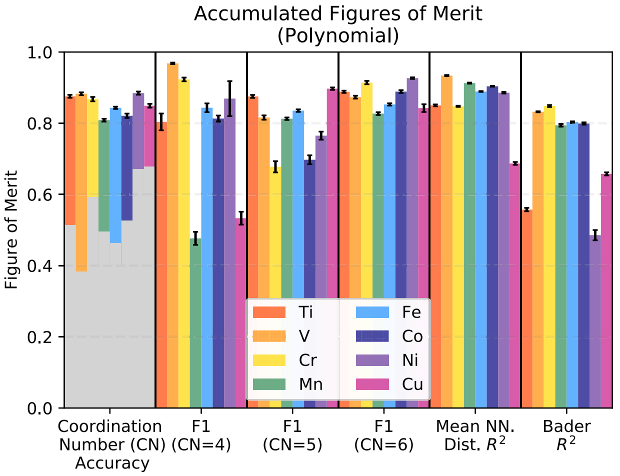

Fig. 4 Performance of random forest models using the polynomial featurization of the spectrum $\mu(E)$ as input. See Fig. 3 for a detailed description of the plot.

track the imbalances of each class. As seen in Figs 3 and 4, there is poorer average performance for Mn, and Cu—largely owing to difficulties in classifying fourfold coordinated spectra for each case, and for Cr and Co, classifying fivefold coordinated spectra. This may be due to a lack of distinguishing features for those elements, as it was shown for Cr and Co in classifying five-fold and six-fold coordinated structures (see subplots in Fig. 2 of ref. [46]). Despite the dataset class imbalances, however, we note that the overall performance of the coordination accuracy is clustered around an accuracy of 80–88% for every element, providing sufficient basis for interpreting feature importance.

The F1 scores for coordination number here (over a test set combining OQMD and MP structures) are slightly lower than those in previous work using artificial neural networks for classification on a smaller data set of only MP structures (see Table I of ref. [46]); the better performance of neural networks may also be explained by their larger parameter space. We note that a direct performance comparison with earlier work by Zheng and Chen et al. is difficult, due to differences in the targeted property—specific discrete coordination numbers here, compared to continuously weighted coordination motifs in ref. [47]. Their high accuracies of >0.9 for the same absorbing elements suggest that weighted coordinations may be a better choice of target for training future ML models.

For both pointwise and polynomial featurization in Fig. 5, the most informative parts of the spectra shift from the pre-edge region to other regions of the spectrum when moving from lighter to heavier 3*d* metals. The importance of the post-edge region accords with prior work[46,47], but our polynomial featurization provides the additional insight that coordination is largely determined by the constant terms of the polynomial in that region, indicating that the magnitude of the spectra is more important than local linear or quadratic behavior.

The most important non-constant features are located in the pre-edge region for the four lighter metals (Ti, V, Cr, Mn), in agreement with known qualitative trends[31,60]. The quadratic character of Ti and V pre-edges as the second-most important features perfectly lines up with domain knowledge that the intensity of a pre-edge peak is critical for coordination classification for lighter metals; Cr also has multiple higher order coefficients ranking as important in the pre-edge region[32]. We also see that pre-edge features remain important features for Mn, Fe, Co, Ni, and Cu, consistent with earlier studies[33,38,39]. Still, for the four heavier metals (Fe, Co, Ni, and Cu), the polynomial-featurized models primarily rely on absorption in the edge and post-edge regions. These findings have particular relevance in identifying the minimum energy range for an experimental scan in order to identify the coordination number an absorbing atom.

Table 1. Breakdown of spectra which were used to train, validate, or test the ML models.

Metal	Total spectra	4-fold coordination	5-fold coordination	6-fold coordination	Total coordination/ nearest-neighbor distance spectra	Bader charge spectra
Titanium	4793	334	2301	2074	4709	3201
Vanadium	6929	1954	2404	2504	6862	2863
Chromium	2395	436	580	1326	2342	1809
Manganese	7917	302	3873	3635	7810	4031
Iron	6744	1052	3087	2534	6673	3908
Cobalt	3453	538	1047	1851	3436	2075
Nickel	3396	206	972	2183	3361	2224
Copper	3444	468	2257	651	3376	2167

The same spectra were used for the coordination classification as for the mean nearest-neighbor distance regression.

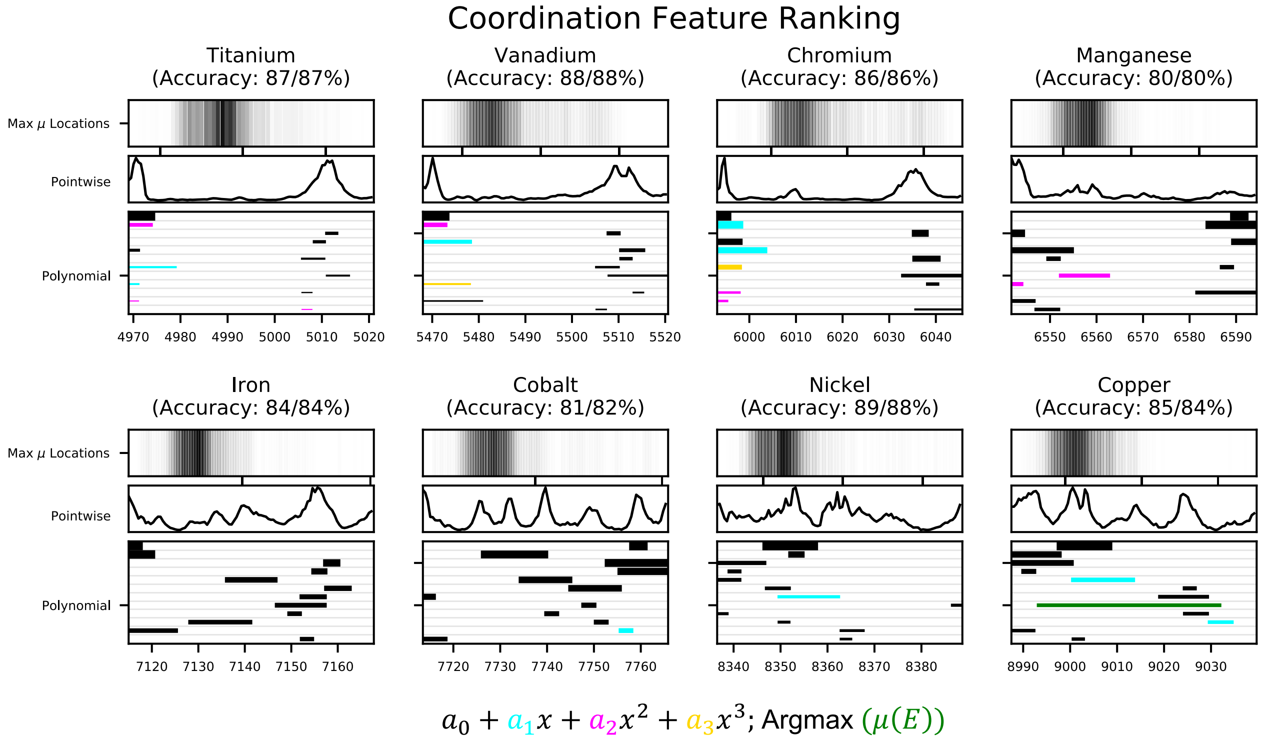

Fig. 5 By absorbing element, the pointwise/polynomial featurized model accuracy, the white line energy (energy of maximum absorption) distribution and feature rankings for coordination number models. Top: at each energy, the proportion of spectra with the maximum absorption occurring at that energy (darker means more spectra). Middle: feature importance of absorption at each energy using pointwise featurization. Bottom: the highest ranked polynomial features, with the most important at the top and 12th most important at the bottom, and relative importance indicated by bar thickness. Bar width illustrates the energy range of the partition where the polynomial feature was fit, and coefficient type (constant, linear, quadratic, cubic) is indicated by color. For the white line energy feature, the bar represents the spread of different energy values at which the maximum absorption occurred.

Mean nearest-neighbor distance

For the mean nearest-neighbor distance, we see strong model performance; the mean absolute error (MAE) for both featurizations is at or below 0.02 Å for all but Cu, which is around 0.05 Å (See Supplementary Tables III and IV). The pointwise and polynomial featurizations, respectively result in an R^2 value of 0.93/0.93 for V and 0.92/0.91 for Mn, 0.71/0.69 for Cu and 0.84–0.91 for all other elements. The test set performance of pointwise models are presented in finer detail in Fig. 6, a parity plot with a different random forest model trained for each transition metal element. We note that the performance is strongest for average-case distances, and that most of the errors arise in underestimation of nearest-neighbor distance for outlier atoms with further-spaced or nearer-spaced neighbors.

Pointwise features mostly exhibit a sharp increase in importance as the energy values begin to enter the EXAFS region, but also attributes some importance closer to the edge energy. The polynomial fits make it easier to see how to interpret regions of interest. In contrast to coordination, which was dominated by the constant absorption (a_0 terms) for important post-edge features, in Fig. 7, the linear terms are among the most important features for all mean nearest-neighbor distance predictions, except for Ti and Cu. This lines up with intuition that the direction of increase or decrease in the post-edge region corresponds to ‘shells’ of increasing radius about the absorbing atom that produce the first interference pattern of the pre-EXAFS region, as the location of the first post-edge oscillation will correlate with the nearest neighbors. Higher energies correspond to a larger ‘shell’ about the absorbing atom: sinusoidal oscillation patterns in EXAFS originate in the spacing of nearby neighbors (see Eq. 2 of ref. [25]) and the wave number k associated with a photoexcitation.

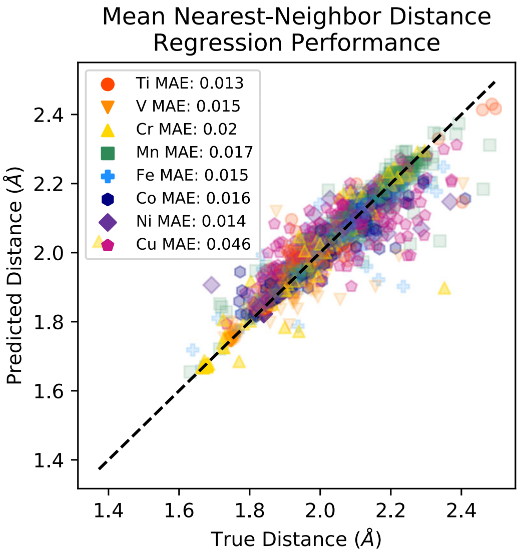

Fig. 6 Parity plot of mean nearest-neighbor distances. Each point represents a spectrum-property pair in the test set, and compares the predicted distance from pointwise featurized random forest models to the “true” distance computed from OQMD and MP structures.

Because the models are trained on thousands of spectra, heuristics which correlate peak location and nearest-neighbor distance may be found that are applicable for diverse transition metal oxide (TMO) structures. Models can hone in on physics-based spectral trends discernible to the human eye while also

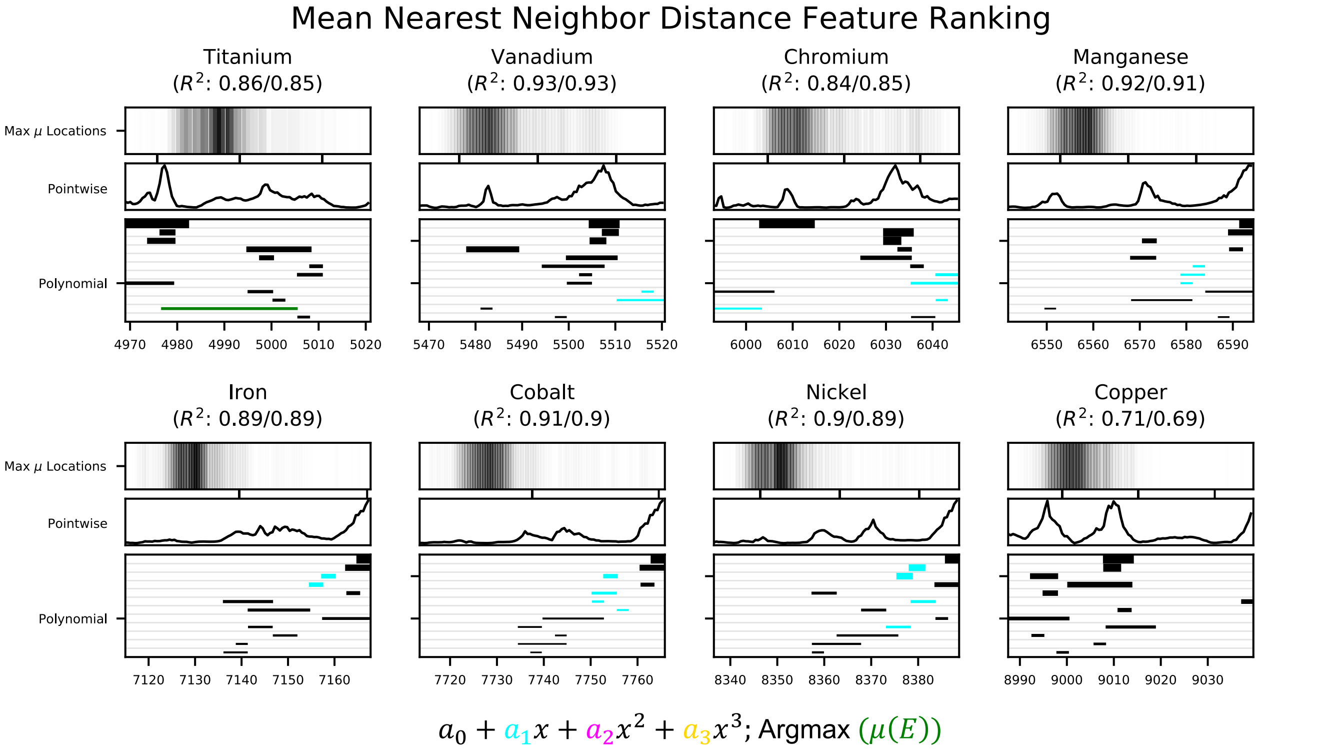

Fig. 7 By absorbing element, the pointwise/polynomial featurized model accuracy, the white line energy distribution and feature rankings for mean nearest neighbor distance models. See the caption of Fig. 5 for a detailed description of the plots.

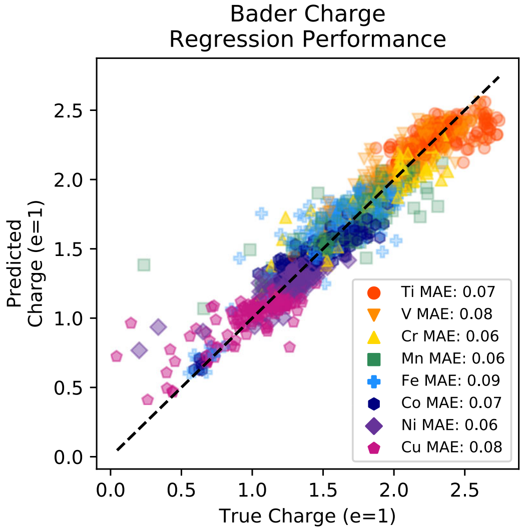

Fig. 8 Parity plot of Bader charges. Each point represents a spectrum-property pair in the test set, and compares the predicted charge from pointwise featurized random forest models to the "true" Bader charge computed from OQMD and MP electron densities.

using information from the rest of the spectrum, such as pre-edge linear and constant trends for Cr being important alongside post-edge trends. This contrasts with the common human bias of choosing easily discernible spectral features localized to one part of the spectrum and applicable to a smaller set of model systems.

Bader charges

Previous work has shown that oxidation number can be predicted from XANES spectra via ML classification models[56]. Here, we show that this principle is extensible to the continuous regression of Bader charges. We obtain varying degrees of success based on the element of the absorbing atom, as shown in Fig. 8. For the polynomial featurization, V and Cr showcase the best performance, with R^2 near 0.83 and 0.85, respectively, with Fe, Co, and Mn presenting values around 0.80. The charges of Ti, Ni, and Cu are not well reproduced by our models, with R^2 scores in the range of 0.49–0.69. This means that the model interpretability for these metals is possibly less trustworthy; however, the MAE is on average 0.07, which is a natural bound for the resolving power of our model due to how the data was collected (see "Methods" section for more details).

For Bader charges, Fig. 9 shows how including information about the spectral derivatives alongside magnitude is helpful for interpretation: coefficients which describe linear or quadratic curvature share importance with the constant polynomial term. Previous work has shown that the first derivative of the spectrum often coincides with the oxidation state of the absorber[61]. Since the Bader charge correlates with the oxidation state in solids, our suggested importance of linear trends lines up with existing scientific intuition.

Because the number of valence electrons affects the binding energy of core electrons in an atom[62], the rising edge energy is often used to experimentally detect oxidation state of metal atoms. For a limited set of compounds, this heuristic can work well: for example, the shift of the edge energy was successfully used as a fingerprint for oxidation state of Mn in silicate glasses[63]. But other studies, such as those for chromium chloride compounds[64], demonstrate the difficulty of determining oxidation state of Cr from edge energy alone. The white line energy usually occurs shortly after the rising edge energy, and, surprisingly, the white line energy does not occur as one of the top 12 features for Bader charge in this study. This suggests there is sufficient information in the rest of the spectrum to make accurate predictions, and also shows the limitation of relying on peak energies for determining metal oxidation states.

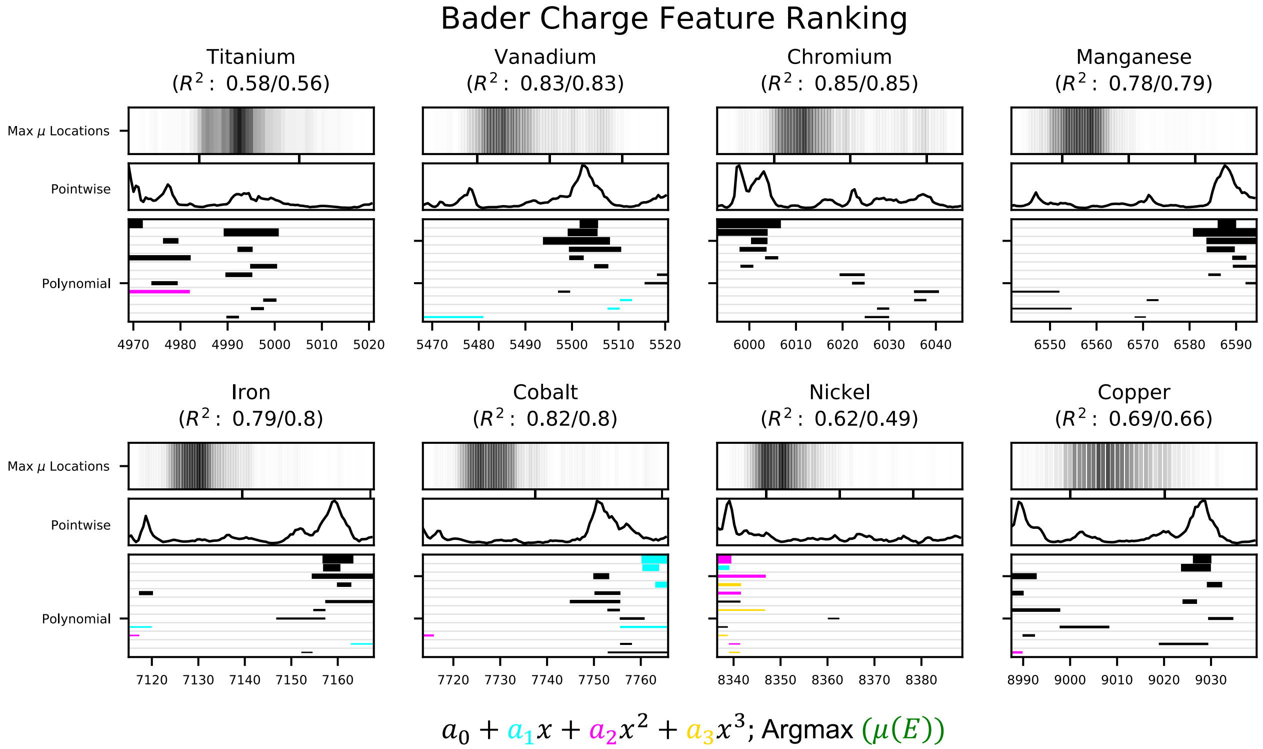

Fig. 9 For each absorbing element, the pointwise/polynomial featurized model accuracy, the white line energy distribution and feature rankings for Bader charge models. See the caption of Fig. 5 for a detailed description of the plots.

There are factors for certain elements which could explain the comparatively weaker Bader charge model performance. The majority of the Ti found in materials are in the oxidation state Ti (IV), and the rising edge energies are sensitive to coordination number[31]. The confluence of these factors could obscure subtle pre-edge trends which correlate with the oxidation state, and help explain the poor Bader charge prediction. It has also been shown that the spin state of Ni and Co atoms can have a great impact on their host geometries[65]: Ni may undergo ligand-based redox in some systems[66], which leads to little change on the rising edge upon redox state changes[67] while still affecting the underlying charge distributions. This could therefore lead to inaccurate predictions on the Bader charge for Ni. For Cu, the lower performance is possibly due to varied hybridization between Cu and O that affect its oxidation state.

For compounds of V, Cr, and Mn, where Bader charges were most accurately predicted, correlation of the peak position with oxidation state may assist characterization of materials, as charge densities (and therefore, charge transfer processes) can influence e.g., catalytic or battery efficiency and leave detectable signatures within XAS spectra[11,29,68].

DISCUSSION

This work represents an advance in the scope of ML applications for XANES and the use of feature ranking for generating XANES insights. Our extension of ML-XANES predictions to mean nearest-neighbor distance and Bader charges help to expand the space of inverse problems solvable via data-driven methods. Our study of the importance of featurization also helps to demonstrate how it can improve performance and assist human intepretability.

Feature ranking of pointwise descriptor vectors occurs in the finest possible features for the data set (on the level of individual 'pixels' of absorption). However, spectral trends of interest to us, such as the presence or absence of a small peak, or the local curvature at a given energy range, are not discernible at that level of feature ranking. While these trends are possibly captured in the structure of decision trees, it is not obvious how to extract them from feature rankings which only highlight the importance of individual values of an input $\vec{\mu}(E)$ vector. In contrast, multiscale polynomials concisely describe local trends across varying energy ranges of the spectrum. Many chemical features such as hybridization and forbidden transitions have information embedded in the shape of the spectrum. Furthermore, work in the saliency map literature found that identically performing models with more easily understandable function were rated by untrained users as more trustworthy[69]; this may justify the appeal of polynomial featurization even when performance is comparable to pointwise featurization.

We also anticipate that our work on featurization could come with other benefits. Experimental XAS data often suffers from noise and systematic variation in absorption which can vary even on the same material on the same beamline. Ways to 'coarsen' spectra into features are thus desirable to preserve transferability of model function, and model interpretability, as trends associated with individual regions can be more precisely described compared to pointwise featurization. This interpretability could also be of critical importance in assessing the transferability of a model. If a model with good performance in predicting a given property was found to be depending largely on one feature (say, the presence or lack thereof of a given peak in the spectrum) and that feature was known to be irrelevant (perhaps due to systematic error in a calculation or being associated with the substrate of a material), then it would be unlikely that this model could be reliably transferred to arbitrary experiments that exhibit no such feature.

Our polynomial featurization framework captures the importance of coarse, as well as fine region splits simultaneously, building on previous work by Zheng and Chen et al. that studied the relative information content of different spectra, but only on a per-region basis (pre-edge, edge, and post-edge regions), and for predicting coordination number[47]. Our work is further made distinct by our focus on exploring how alternate featurization can

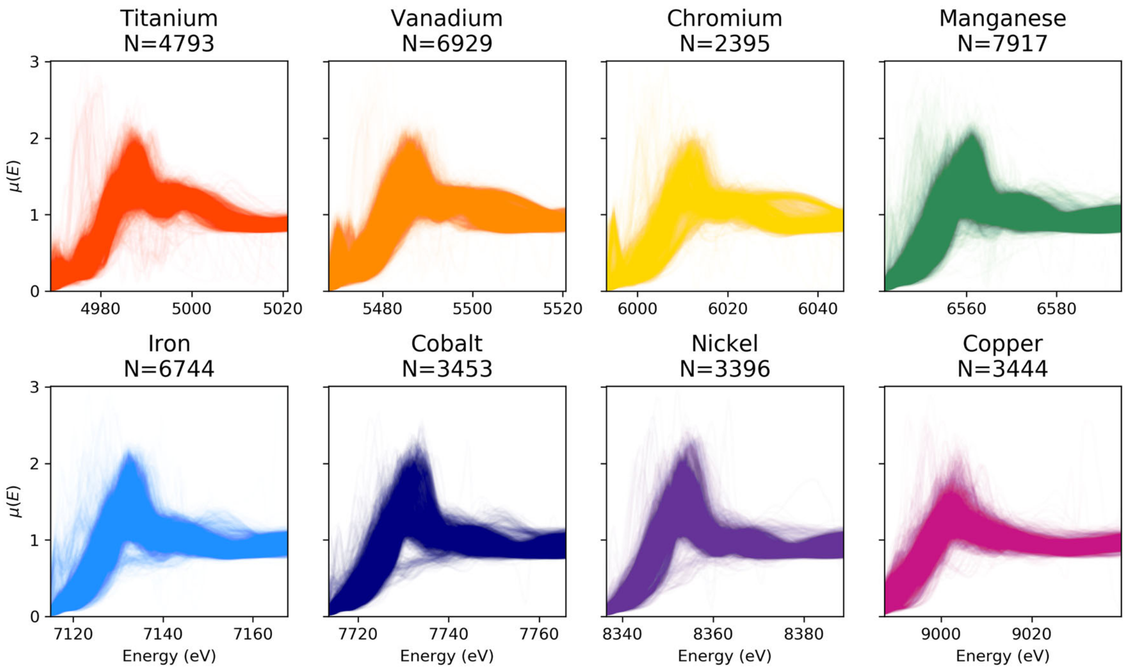

Fig. 10 All spectra in the XANES dataset. Spectra are shown separately for each absorbing element, with N indicating the total number of spectra per element and greater opacity indicating a greater number of spectra overlapped in a given region.

aid model interpretability. Because ML as applied to the analysis of XANES spectra is a budding field, the agreements between our work and theirs is heartening for the reproducibility and transferability of both of our findings.

We also found, critically, that the normalization can not only affect model performance but can heavily affect the feature ranking, introducing an important caveat for future interpretability studies: Carbone et al.[46] and Zheng and Chen et al.[47] both normalized each spectrum to the maximum value of the spectrum, and we compared their max-normalization to post-edge normalization, with spectra normalized to the value of absorption 50 eV after the rising edge value (which FEFF 9 performs by default). While normalization to the peak constrains the range of the spectrum, there is no clear physical basis for normalization to that value. For mean nearest-neighbor distance and Bader charge, the relative importances of a_0 and a_1, a_2 swung based on the normalization decision, highlighting the importance of physically relevant normalization.

Future work could focus on improving accuracy by generating a more class-balanced dataset and by better understanding of the XAS data itself, perhaps via an active learning loop. Improved target labels may also help; for example, with a coordination number of 4, the metal center could possess T_d, D_{4h}, D_{2d}, C_{3v}, and other possible symmetries. Also, as mentioned earlier, continuous descriptors of coordination[70] as a target property appear to produce more robust structure-property prediction[47]. Most importantly, future models could leverage known correlations between multiple materials properties. For XAS, the intensity, shape and the energy of both pre-edge and rising edge are usually codeterminant with many more factors than coordination number, oxidation state and average bond distance alone. For example, in silicate glasses, Mn^{3+} ions are only found in 6-fold coordinated environments while Mn^{2+} can be found in 4-fold, 5-fold, or 6-fold structures[63]; in this way, understanding the coordination environment could help conclusively identify the oxidation number. Other correlations could be exploited: higher oxidation state atoms tend to have shorter bonds, and certain spin states favor certain geometry.

By using random forests, we trade a rigorous reconstruction of the chemical environment for bypassing the iterative process of computing additional XANES spectra. Evaluation of a random forest model will in general be orders of magnitude faster than a full multiple scattering calculation. While coordination, Bader charge, and nearest-neighbor distances do not always fully characterize the environment around the absorbing atom, they provide useful insights into structure-property relationships[11,29,68], and reduce the number of candidates necessary for complete local structure determination. At their current accuracies and efficiencies, models developed within this work could be used as a pre-processing step into these more rigorous workflows—those involving FEFF, MXAN, or Pyfitit—to narrow down initial guesses for a chemical structure when no a priori structure knowledge is available[41–45].

FEFF-computed spectra are efficient to generate and are understood to present qualitative agreement with experiment. Transfer learning from theoretical to experimental spectra would require careful consideration of any systematic errors associated with the calculations, as well as corrections which can be applied to improve these errors; For instance, the contributions of the real-space Green's function approximation and muffin-tin potential calculation[25] corrections could be applied that include a Hubbard-like U parameter[71], vibrational contributions[72] or others (see Supplementary Information for more details). For a dedicated use-case, a more costly but quantitatively accurate method of generating reference theoretical XANES spectra based on solving the Bethe-Salpeter equation, could be used to supplement experimental data[73–77].

For integration with experiments, one could imagine a 'meta-decision-tree' or a Bayesian algorithm in conjunction with an automated characterization probe in real time, scanning energy ranges with high expected information value until sufficient signal is achieved to determine structural properties of interest with reasonably high accuracy, analogously to an active learning loop as used in other contexts for ML and materials science[1,12,78,79]. The important energy ranges in Figs 5, 7, and 9 could be used to target measurements of different energy regions when certain properties are desired for structural characterization. Appropriate system-specific and apparatus-specific calibration of relevant spectra would be necessary for a careful comparison to experiment. We anticipate that easily implementable and interpretable models in the vein of the random forests discussed in this paper, once adapted to experimental data input will help provide real-time feedback for experimentalists to maximize the information extracted per experiment.

METHODS

Dataset: computed structures, properties

Our study focuses on transition metal oxide materials, of which there are thousands of structures available in the Open Quantum Materials Database (OQMD)[8] and the Materials Project (MP)[7], with geometries optimized via density functional theory (DFT). Our datasets consist of materials' unit cell structure (from which we extract coordination number and nearest neighbor distances), their electron densities (from which we extract Bader charges), and their XANES spectra (which we either download from the Materials Project[56,80] or compute ourselves).

The chosen structures are comprised of unit cells with at least one transition metal in the set {Co, Fe, V, Cu, Ni, Cr, Mn, Ti}, and at least one oxygen atom. For any duplicate structures between MP and OQMD—identified using the pymatgen structure matcher[81], as well as ICSD numbers associated with each structure—we keep only the MP structure. Coordination numbers are computed from these unit cells for all materials, and mean nearest-neighbor distances are computed for all absorbing atoms that are four, five, and six-fold coordinated.

Spectra for these unit cells are obtained either through querying the MP API or by computing XANES spectra using FEFF 9[40] wrapped by Atomate workflows[82] with the same parameters as the MP to ensure transferability (cataloged in pymatgen at time of writing as MPXANESSet[55,56,81]). We ultimately computed over 23,000 new XANES spectra to supplement the spectra downloaded from the Materials Project database. Duplicate spectra and those that appear to be from an unconverged calculation (i.e., the computed spectra appear unphysical, such as those with anomalously high peaks in the pre-edge region) are left out from our dataset. The unphysical spectra filtering and de-duplication are similar to the data preparation in earlier work[46], except here the similarity cutoffs for duplicate data are more restrictive. See SI for the exact removal criteria. In addition, due our focus on interpreting spectral features, we only keep spectra with maximum summed errors less than 0.1 (in absorption units) between individual piecewise polynomial fits and their respective domains in the max-normalized spectrum (See SI). Special SciPy functions[83] and the NumPy library[84], were instrumental to all parts of analysis. All spectra are visualized in Fig. 10.

We performed our entire analysis twice over, using both the spectrum normalized to the post-edge absorption value (to 50 eV after the edge location, as determined by FEFF) and the spectrum normalized to the maximum value of absorption (as performed by Carbone[46] and Zheng and Chen et al.[47]). All figures from analysis using the max-normalized spectra are in Supplementary Figs. 76–139 and Supplementary Tables VI and VII.

The Bader charge of each atom is computed by partitioning a three-dimensional charge density distribution of a material at its zero-flux surfaces, and computing the total charge within each partition[85]. Electron densities data from the MP and the OQMD, both circa 2017, are used for our Bader charge dataset. This data is only available for a subset of the structures used for the XANES, coordination number, and mean nearest neighbor datasets. For just the Bader charge prediction model, we add another screening criteria for the dataset, only allowing structures which had similar Bader charge values (less than 0.07 difference) between atoms of the same species within the unit cell. See the SI for more preprocessing details. We also note that the MP and OQMD DFT calculations use slightly different parameters, such as plane-wave basis set cutoff fidelity, pseudopotential choice, k-point grid, or convergence criteria. There is a resulting mean absolute difference of 0.07 charge units for Bader charges between MP and OQMD structures identified as identical.

For each absorbing element, the total number of XANES spectra and each target property are presented in Table 1. The same datasets are used for pointwise and polynomial featurized model training and testing.

Model training and performance

We used random forest classifiers to predict coordination number, and random forest regressors to predict mean nearest-neighbor distance and Bader charge, both in the scikit-learn implementation (For this work, version 0.21.3)[86]. We quantify classifier performance using the *F*1 score for each class. For regressors, performance is quantified using the coefficient of determination R^2 (both defined in SI).

Coordination classification is complicated by the class imbalances inherent to our data set, since 4-fold coordination is, in general, under-represented compared to 5-fold and 6-fold coordination in TMOs. In training our models, we use random over-sampling (over-sampling the minority class with replacement until parity is reached with the number of the majority class) to ameliorate the effect of class imbalance, and found that this improved model performance on the validation set.

In order to prevent overestimation of model accuracy, we performed an 80-10-10 training-validation-test set split for each element, with all sets randomly chosen. We then studied model performance on a 10% validation subset of the overall data, and gauged the importance of class imbalances, pre-processing choices, and other hyperparameters on this validation set. All performance reported from this paper is from performance on the test set. All models were trained 10 times on the same data and different random seeds; all error bars seen in the manuscript represent ±1 standard deviation.

Featurization and Interpretability

We featurize XANES spectra in two ways in this work. The "pointwise" featurization is the straightforward use of a vector of all the values of $\vec{\mu}(E)$, interpolated on 100 equally spaced energy values. For the polynomial featurization, polynomials are fit (see Fig. 2) to four partitions of the energy range: four-fold, and five-fold partitions capture coarse trends in the spectrum, and 10-fold and 20-fold splits capture the finer features. We add a physically meaningful value, the energy value E of the peak $\vec{\mu}(E)$ i.e., Argmax($\vec{\mu}(E)$), which is commonly known as the white line energy, as an additional feature to assist with fitting and interpretability. In summary, our feature map transforms a single spectrum of 100 values of $\vec{\mu}(E)$ into 157 multiscale features.

A relative importance score can be computed for each feature vector component to a random forest model by considering the features which are associated with the greatest reductions in Gini impurity (for classification) and MSE (for regression). Because the ranking occurs by comparing individual values of the input feature vector, careful choice of feature vectors is necessary to capture physically relevant spectrum-property features. During our 10 re-train cycles of each RF, as expected, there was some variability in the feature ranking values due to the randomness of the fitting procedure. However, trends were generally consistent across the fits. The variability associated with each feature vector for each property, featurization, and normalization is plotted in Supplementary Figs 20–67 for each transition metal we considered, respectively.

Anion coincidence in structures

Every structure that we considered in this study had at least one oxygen atom present within the cell. Because the co-existence of other anions can change the charge distribution behavior, we indicate in Supplementary Table V the number of spectra which (i.) Have associated Bader charge values and were used in the study and (ii.) contain a given anion, due to the fact that different anions can induce different charge transfer behaviors.

Bader charge vs. oxidation state

In order to demonstrate the correspondence between Bader charges and oxidation state, we present figures in the Supplementary Information juxtaposing the Bader charges associated with given structures and the oxidation states approximately guessed by pymatgen. A correlation between the Bader charge and predicted oxidation state can be observed, though the agreement is not strictly quantitative.

DATA AVAILABILITY

The feature vectors (XANES spectra projected onto 100 grid points and polynomial coefficient vectors), and their associated label values, are shared publicly on TRI's data-sharing website, https://data.matr.io.

CODE AVAILABILITY

The code used to train the models and generate the figures in this publication is publicly available at the TRIXS (Toyota Research Institute X-ray Spectroscopy) repository at https://github.com/TRI-AMDD/trixs.

Received: 19 February 2020; Accepted: 22 June 2020;
Published online: 29 July 2020

REFERENCES

1. Häse, F., Roch, L. M. & Aspuru-Guzik, A. Next-generation experimentation with self-driving laboratories. *Trends Chem.* **1**, 282–291 (2019).
2. Roch, M. L. et al. ChemOS: an orchestration software to democratize autonomous discovery. *Chemrxiv* 10.26434/chemrxiv.5952655 (2018).
3. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
4. Haber, J. A. et al. Discovering Ce-rich oxygen evolution catalysts, from high throughput screening to water electrolysis. *Energy Environ. Sci.* **7**, 682–688 (2014).
5. Stein, H. S. et al. Functional mapping reveals mechanistic clusters for OER catalysis across (Cu-Mn-Ta-Co-Sn-Fe)O: X composition and pH space. *Mater. Horiz.* **6**, 1251–1258 (2019).
6. Kluender, E. J. et al. Catalyst discovery through megalibraries of nanomaterials. *Proc. Natl Acad. Sci. USA* **116**, 40–45 (2019).
7. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
8. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
9. Suram, S. K. et al. Automated phase mapping with AgileFD and its application to light absorber discovery in the V-Mn-Nb oxide system. *ACS Comb. Sci.* **19**, 37–46 (2017).
10. Gomes, C. P. et al. CRYSTAL: a multi-agent AI system for automated mapping of materials' crystal structures. *MRS Commun.* **9**, 600–608 (2019).
11. Yu, Y. et al. Revealing electronic signatures of lattice oxygen redox in lithium ruthenates and implications for high-energy Li-ion battery material designs. *Chem. Mater.* **31**, 7864–7876 (2019).
12. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 1–7 (2014).
13. Pendleton, I. M. et al. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Commun.* **9**, 846–859 (2019).
14. Sayers, D. E., Stern, E. A. & Lytle, F. W. New technique for investigating noncrystalline structures: fourier analysis of the extended x-ray-absorption fine structure. *Phys. Rev. Lett.* **27**, 1204–1207 (1971).
15. Eisenberger, P. & Kincaid, B. M. EXAFS: new horizons in structure determinations. *Science* **200**, 1441–1447 (1978).
16. Huang, W. C. et al. A facile method for sodium-modified Fe2O3/Al2O3 oxygen carrier by an air atmospheric pressure plasma jet for chemical looping combustion process. *Chem. Eng. J.* **316**, 15–23 (2017).
17. Alalwan, H. A., Mason, S. E., Grassian, V. H. & Cwiertny, D. M. *α*-Fe2O3 nanoparticles as oxygen carriers for chemical looping combustion: an integrated materials characterization approach to understanding oxygen carrier performance, reduction mechanism, and particle size effects. *Energy Fuels* **32**, 7959–7970 (2018).
18. Peng, H., Ndione, P. F., Ginley, D. S., Zakutayev, A. & Lany, S. Design of semiconducting tetrahedral Mn1-xZnxO alloys and their application to solar water splitting. *Phys. Rev. X* **5**, 021016 (2015).
19. Weckhuysen, B. M. Chemical imaging of spatial heterogeneities in catalytic solids at different length and time scales. *Angew. Chem. Int. Ed.* **48**, 4910–4943 (2009).
20. Beale, A. M., Jacques, S. D. & Weckhuysen, B. M. The role of synchrotron radiation in examining the self-assembly of crystalline nanoporous framework materials: From zeolites and aluminophosphates to metal organic hybrids. *Chem. Soc. Rev.* **39**, 4656–4672 (2010).
21. Meirer, F. & Weckhuysen, B. M. Spatial and temporal exploration of heterogeneous catalysts with synchrotron radiation. *Nat. Rev. Mater.* **3**, 324–340 (2018).
22. Ankudinov, A. L., Rehr, J. J., Low, J. J. & Bare, S. R. Sensitivity of Pt x-ray absorption near edge structure to the morphology of small Pt clusters. *J. Chem. Phys.* **116**, 1911–1919 (2002).
23. Kuzmin, A. & Chaboy, J. EXAFS and XANES analysis of oxides at the nanoscale. *IUCrJ* **1**, 571–589 (2014).
24. Lee, P. A., Citrin, P. H., Eisenberger, P. & Kincaid, B. M. Extended x-ray absorption fine structure its strengths and limitations as a structural tool. *Rev. Mod. Phys.* **53**, 769–806 (1981).
25. Rehr, J. J. & Albers, R. C. Theoretical approaches to x-ray absorption fine structure. *Rev. Mod. Phys.* **72**, 621–654 (2000).
26. Rehr, J. J. et al. Ab initio theory and calculations of X-ray spectra. *Comptes Rendus Phys.* **10**, 548–559 (2009).
27. Yano, J. & Yachandra, V. K. X-ray absorption spectroscopy. *Photosynth. Res.* **102**, 241–254 (2009).
28. Wong, J., Lytle, F. W., Messmer, R. P. & Maylotte, D. H. K-edge absorption spectra of selected vanadium compounds. *Phys. Rev. B* **30**, 5596–5610 (1984).
29. Mueller, D. N., MacHala, M. L., Bluhm, H. & Chueh, W. C. Redox activity of surface oxygen anions in oxygen-deficient perovskite oxides during electrochemical reactions. *Nat. Commun*. **6**, 1–8 (2015).
30. Farges, F., Brown, G. E., Navrotsky, A., Gan, H. & Rehr, J. J. Coordination chemistry of Ti(IV) in silicate glasses and melts: II. Glasses at ambient temperature and pressure. *Geochim. Cosmochim. Acta* **60**, 3039–3053 (1996).
31. Farges, F. & Brown, G. E. Ti-edge XANES studies of Ti coordination and disorder in oxide compounds: comparison between theory and experiment. *Phys. Rev. B Condens. Matter Mater. Phys.* **56**, 1809–1819 (1997).
32. Farges, F., Brown, G. E., Petit, P. E. & Munoz, M. Transition elements in water-bearing silicate glasses/melts. Part I. A high-resolution and ancharmonic analysis of Ni coordination environments in crystals, glasses,and melts. *Geochim. Cosmochim. Acta* **65**, 1665–1678 (2001).
33. Jackson, W. E. et al. Multi-spectroscopic study of Fe(II) in silicate glasses: implications for the coordination environment of Fe(II) in silicate melts. *Geochim. Cosmochim. Acta* **69**, 4315–4332 (2005).
34. Hanson, H. & Beeman, W. W. The Mn K absorption edge in manganese metal and manganese compounds. *Phys. Rev.* **76**, 118–121 (1949).
35. Cotton, F. A. & Hanson, H. P. Soft X-Ray absorption edges of metal ions in complexes. III. Zinc (II) complexes. *J. Chem. Phys.* **28**, 83–87 (1958).
36. Cotton, F. A. & Hanson, H. P. Soft X-ray absorption edges of metal ions in complexes. III. Zinc (II) complexes. *J. Chem. Phys.* **28**, 83–87 (1958).
37. Farges, F. Ab initio and experimental pre-edge investigations of the Mn K -edge XANES in oxide-type materials. *Phys. Rev. B Condens. Matter Mater. Phys.* **71**, 155109 (2005).
38. Wilke, M., Farges, F., Petit, P. E., Brown, G. E. & Martin, F. Oxidation state and coordination of Fe in minerals: an Fe K-XANES spectroscopic study. *Am. Mineral.* **86**, 714–730 (2001).
39. Wilke, M., Hahn, O., Woodland, A. B. & Rickers, K. The oxidation state of iron determined by Fe K-edge XANES -application to iron gall ink in historical manuscripts. *J. Anal. Spectrom.* **24**, 1364–1372 (2009).
40. Rehr, J. J., Kas, J. J., Vila, F. D., Prange, M. P. & Jorissen, K. Parameter-free calculations of X-ray spectra with FEFF9. *Phys. Chem. Chem. Phys.* **12**, 5503–5513 (2010).
41. Benfatto, M. & Della Longa, S. Geometrical fitting of experimental XANES spectra by a full multiple-scattering procedure. *J. Synchrotron Radiat.* **8**, 1087–1094 (2001).
42. Benfatto, M., Congiu-Castellano, A., Daniele, A. & Della Longa, S. MXAN: a new software procedure to perform geometrical fitting of experimental XANES spectra. *J. Synchrotron Radiat.* **8**, 267–269 (2001).
43. Benfatto, M., Longa, S. D. & Natoli, C. R. The MXAN procedure: a new method for analysing the XANES spectra of metalloproteins to obtain structural quantitative information. *J. Synchrotron Radiat.* **10**, 51–57 (2003).
44. Martini, A. et al. PyFitit: The software for quantitative analysis of XANES spectra using machine-learning algorithms. *Comput. Phys. Commun.* **250**, 107064 (2020).
45. Guda, A. A. et al. Quantitative structural determination of active sites from in situ and operando XANES spectra: from standard ab initio simulations to chemometric and machine learning approaches. *Catal. Today* **336**, 3–21 (2019).
46. 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.* **3**, 33604 (2019).
47. 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* **1**, 100013 (2020).
48. Trejo, O. et al. Elucidating the evolving atomic structure in atomic layer deposition reactions with in situ XANES and machine learning. *Chem. Mater.* **31**, 8937–8947 (2019).
49. Timoshenko, J. et al. Probing atomic distributions in mono- and bimetallic nanoparticles by supervised machine learning. *Nano Lett.* **19**, 520–529 (2019).
50. Timoshenko, J. et al. Probing atomic distributions in mono- and bimetallic nanoparticles by supervised machine learning. *Nano Lett.* **19**, 520–529 (2019).
51. Liu, Y. et al. Mapping XANES spectra on structural descriptors of copper oxide clusters using supervised machine learning. *J. Chem. Phys.* **151**, 160901 (2019).
52. Timoshenko, J. & Frenkel, A. I. "Inverting" X-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *ACS Catal.* **9**, 10192–10211 (2019).

53. Timoshenko, J., Lu, D., Lin, Y. & Frenkel, A. I. Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles. *J. Phys. Chem. Lett.* **8**, 5091–5098 (2017).
54. Timoshenko, J. et al. Neural network approach for characterizing structural transformations by X-ray absorption fine structure spectroscopy. *Phys. Rev. Lett.* **120**, 225502 (2018).
55. Mathew, K. et al. Data descriptor: high-throughput computational X-ray absorption spectroscopy. *Sci. Data* **5**, 180151 (2018).
56. Zheng, C. et al. Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Comput. Mater.* **4**, 12 (2018).
57. Umehara, M. et al. Analyzing machine learning models to accelerate generation of fundamental materials insights. *npj Comput. Mater.* **5**, 34 (2019).
58. Frey, N. C. et al. Prediction of synthesis of 2D metal carbides and nitrides (MXenes) and their precursors with positive and unlabeled machine learning. *ACS Nano* **13**, 3031–3041 (2019).
59. Hoyt, R. A. et al. Machine learning prediction of H adsorption energies on Ag alloys. *J. Chem. Inf. Model.* **59**, 1357–1365 (2019).
60. Yamamoto, T. Assignment of pre-edge peaks in K-edge X-ray absorption spectra of 3d transition metal compounds: electric dipole or quadrupole? *X-Ray Spectrom.* **37**, 572–584 (2008).
61. Yildirim, B. & Riesen, H. Coordination and oxidation state analysis of cobalt in nanocrystalline LiGa5O8 by X-ray absorption spectroscopy. *J. Phys. Conf. Ser.* **430**, 012011 (2013).
62. Kunzl, V. A linear dependence of energy levels on the valency of elements. *Collect. Czechoslov. Chem. Commun.* **4**, 213–224 (1932).
63. Abuín, M., Serrano, A., Chaboy, J., García, M. A. & Carmona, N. XAS study of Mn, Fe and Cu as indicators of historical glass decay. *J. Anal. Spectrom.* **28**, 1118–1124 (2013).
64. Tromp, M., Moulin, J., Reid, G. & Evans, J. Cr K-edge XANES spectroscopy: ligand and oxidation state dependence-what is oxidation state? *AIP Conf. Proc.* **882**, 699–701 (2007).
65. Ha, Y. et al. The electronic structure of the metal active site determines the geometric structure and function of the metalloregulator NikR. *Biochemistry* **58**, 3585–3591 (2019).
66. Sarangi, R. et al. Sulfur K-edge X-ray absorption spectroscopy as a probe of ligand-metal bond covalency: metal vs ligand oxidation in copper and nickel dithiolene complexes. *J. Am. Chem. Soc.* **129**, 2316–2326 (2007).
67. Németh, Z., Szlachetko, J., Bajnóczi, É. G. & Vankó, G. Laboratory von Hámos X-ray spectroscopy for routine sample characterization. *Rev. Sci. Instrum*. **87**, 103105 (2016).
68. Grimaud, A. et al. Double perovskites as a family of highly active catalysts for oxygen evolution in alkaline solution. *Nat. Commun*. **4**, 1–7 (2013).
69. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
70. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
71. Ahmed, T., Kas, J. & Rehr, J. Hubbard model corrections in real-space x-ray spectroscopy theory. *Phys. Rev. B Condens. Matter Mater. Phys.* **85**, 165123 (2012).
72. Vila, F. D., Rehr, J. J., Rossner, H. H. & Krappe, H. J. Theoretical x-ray absorption Debye-Waller factors. *Phys. Rev. B Condens. Matter Mater. Phys.* **76**, 014301 (2007).
73. Gulans, A. et al. Exciting: A full-potential all-electron package implementing density-functional theory and many-body perturbation theory. *J. Phys. Condens. Matter* **26**, 363202 (2014).
74. Vorwerk, C., Cocchi, C. & Draxl, C. Addressing electron-hole correlation in core excitations of solids: An all-electron many-body approach from first principles. *Phys. Rev. B* **95**, 155121 (2017).
75. Vorwerk, C., Aurich, B., Cocchi, C. & Draxl, C. Bethe-Salpeter equation for absorption and scattering spectroscopy: implementation in the exciting code. *Electron. Struct.* **1**, 037001 (2019).
76. Gilmore, K. et al. Efficient implementation of core-excitation Bethe-Salpeter equation calculations. *Comput. Phys. Commun.* **197**, 109–117 (2015).
77. Vinson, J., Rehr, J. J., Kas, J. J. & Shirley, E. L. Bethe-Salpeter equation calculations of core excitation spectra. *Phys. Rev. B Condens. Matter Mater. Phys.* **83**, 115106 (2011).
78. Vandermause, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 1–11 (2020).
79. Mrdjenovich, D. et al. propnet: a knowledge graph for materials science. *Matter* **2**, 464–480 (2020).
80. Mathew, K. et al. High-throughput computational X-ray absorption spectroscopy. *Sci. Data* **5**, 180151 (2018).
81. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
82. Mathew, K. et al. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. *Comput. Mater. Sci.* **139**, 140–152 (2017).
83. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
84. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
85. Henkelman, G., Arnaldsson, A. & Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Comput. Mater. Sci.* **36**, 354–360 (2006).
86. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

ACKNOWLEDGEMENTS

This work was supported by Toyota Research Institute through the Accelerated Materials Design and Discovery program. S.B.T. and M.R.C. received support from the U.S. Department of Energy through the Computational Science Graduate Fellowship (DOE CSGF) under grant number DE-FG02-97ER25308. S.B.T. and B.A.R. were supported during part of this work by an internship at Toyota Research Institute's Accelerated Materials Design and Discovery program. Y.H. is funded by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Vehicle Technologies of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. J.Y. acknowledges Strategic Partnership Projects Agreement No. FP00007929 between Lawrence Berkeley National Laboratory and Toyota Research Institute. The authors gratefully acknowledge helpful discussions with John Rehr, Li Cheng Kao, Arjun Bhargava, Chirranjeevi Gopal, Daniel Schweigert, Muratahan Aykol, Emily Jia, Yang Shao-Horn, Livia Giordano, and Jiayu Peng.

AUTHOR CONTRIBUTIONS

S.B.T. was the principal author of the manuscript, calculated the additional XAS spectra, and performed the random forest data processing and analysis. M.R.C. provided detailed advising on the machine learning and data processing. S.B.T. conceived the work with S.K.S. and L.H., who supervised the project and provided guidance. J.H.M. and B.A.R. performed the calculations of the Bader charges for the Materials Project and Open Quantum Materials Databases, respectively. Y.H. and J.Y. provided discussions about interpretation and relevant literature. All authors provided contributions to the manuscript and discussed the results.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00376-6.

Correspondence and requests for materials should be addressed to S.B.T., S.K.S. or L.H.

Reprints and permission information is available at http://www.nature.com/reprints