

Robust Machine Learning Inference from X-ray Absorption Near Edge Spectra through Featurization

Yiming Chen,^{†,‡} Chi Chen,[†] Inhui Hwang,[¶] Michael J. Davis,[§] Wanli Yang,^{||}
Chengjun Sun,[¶] Gihyeok Lee,^{||} Dylan McReynolds,^{||} Daniel Allan,[⊥] Juan
Marulanda Arias,[⊥] Shyue Ping Ong,^{*,†} and Maria K.Y. Chan^{*,‡}

[†]*Department of NanoEngineering, University of California San Diego, 9500 Gilman Dr, La
Jolla, CA 92093, United States*

[‡]*Center for Nanoscale Materials, Argonne National Laboratory, 9700 South Cass Avenue,
Lemont, IL 60439, United States*

[¶]*Advanced Photon Source, Argonne National Laboratory, 9700 South Cass Avenue,
Lemont, IL 60439, United States*

[§]*Division of Chemical Sciences and Engineering, Argonne National Laboratory, 9700 South
Cass Avenue, Lemont, IL 60439, United States*

^{||}*Advanced Light Source, Lawrence Berkeley National Laboratory, 6 Cyclotron Rd, Berkeley,
CA 94720, United States*

[⊥]*National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY
11973, United States*

E-mail: ongsp@ucsd.edu; mchan@anl.gov

Abstract

X-ray absorption spectroscopy (XAS) is a commonly-employed technique for characterizing functional materials. In particular, x-ray absorption near edge spectra (XANES) encodes local coordination and electronic information and machine learning approaches to extract this information is of significant interest. To date, most ML approaches for XANES have primarily focused on using the raw spectral intensities as input, overlooking the potential benefits of incorporating spectral transformations and dimensionality reduction techniques into ML predictions. In this work, we focused on systematically comparing the impact of different featurization methods on the performance of ML models for XAS analysis. We evaluated the classification and regression capabilities of these models on computed datasets and validated their performance on previously unseen experimental datasets. Our analysis revealed an intriguing discovery: the cumulative distribution function (CDF) feature achieves both high prediction accuracy and exceptional transferability. This remarkably robust performance can be attributed to its tolerance to horizontal shifts in spectra, which is crucial when validating models using experimental data. While this work exclusively focuses on XANES analysis, we anticipate that the methodology presented here will hold promise as a versatile asset to the broader spectroscopy community.

Introduction

X-ray absorption spectroscopy (XAS) is a versatile characterization technique to probe the oxidation states,¹ spin states,² and coordination environment³ in materials. A typical XAS spectrum can be divided into two regions depending on the energy range. The X-ray absorption near-edge structure (XANES), the region within 50 eV of absorption onset, is more sensitive to the oxidation states and coordination environments. The extended X-ray absorption fine structure (EXAFS), on the other hand, encodes information about the neighboring atoms and excited states.⁴ While the quantitative analysis of EXAFS is relatively well-established with explicit equations to fit computed spectra, the analysis for the XANES is

still constrained by the limited number of reference spectra.^{5,6}

Nevertheless, advances in operando measurements, computational approaches⁷⁻⁹ and computing power in recent years have greatly alleviated such data scarcity. For example, in previous studies,^{10,11} some of the present authors have developed the XASDB, the world’s largest database of computed XANES hosting approximately 500,000 K-edge and 140,000 L-edge spectra. Such large, computational XAS databases provide a highly useful complement to experimental XAS data. While experimental XAS data collection typically focuses on limited chemistries (usually a single phase or phase mixture) under a wide variety of conditions, computational databases such as the XASDB provide spectra on a broad diversity of structures and chemistries, albeit under limited conditions (typically 0K structures from density functional theory (DFT) calculations).

With the increasing availability of experimental and computed XANES data, there have been significant research efforts into the application of machine learning (ML) techniques to extract insights and make predictions from XAS. For instance, Guda et al. applied multivariate curve resolution methods on operando XANES spectra to isolate individual species/phases from the multicomponent data mixture in a catalyst system.¹² Supervised machine learning techniques have been widely applied to establish a correspondence between spectra and target properties, either in a forward or inverse direction.¹³ A large body of work focused on the accurate inference of target properties from XANES. For example, electronic structures such as oxidation states and electronic configuration of *d* states can be determined from K and L-edge XANES spectra through ML-based approaches.^{14,15} The determination of coordination environment has also been investigated in several studies.¹⁶⁻¹⁸ Despite using different coordination environment descriptors, remarkably high accuracy of above 80% has been demonstrated in all cases. Another study also reported a >80% accuracy in determining whether an inorganic material is topological from the XANES data. In addition to inverse property determination from spectra, researchers also explored how to perform forward modeling of XAS spectra based on structural information using neural

networks to achieve quantitative accuracy and derive uncertainty matrix.^{19,20} Beyond XAS, other techniques such as X-ray photoelectron spectroscopy (XPS) and phonon density of states (DoS) prediction were made possible through machine learning approaches.^{21,22} These studies exemplify what ML models are capable of in terms of both accuracy and generalizability.

The majority of the ML models in XAS analysis have thus far utilized the raw spectra, i.e., the paired values of energies and intensities, as the input features. Only a few studies have investigated how the ML models could benefit from additional transformations and dimensionality reduction of the raw spectra. One such example is that of Torrisi et al., who found that polynomial-fitted features from XAS could aid model interpretability by incorporating local trends and focusing on spectral shape that was buried under individual intensity points.⁴ Similarly, Tetef et al. showed that a t-distributed stochastic neighbor embedding (t-SNE) of XANES not only achieves superior prediction accuracy in classifying aromaticity, but is also able to distinguish the finer sub-classes for sulforganics.²³ Latent representation was also explored to generate a low-dimensional representation of Pd K-edge XANES spectra that could maintain the spectrum-structure relationship and provide an innovative pathway to identify the key factors for spectral changes.²⁴

In this work, we investigated different featurization approaches and benchmarked their effect on the classification and regression performance of various ML models for XAS analysis. We have selected the Ni K-edge XANES of $\text{Li}_z\text{Ni}_x\text{Mn}_y\text{Co}_{1-x-y}\text{O}_2$ (NMC) as our system of interest. NMC is a family of cathodes of major importance in rechargeable Li-ion batteries because of their high energy density and long-term cyclability.²⁵ During the cycling of a battery where Li is extracted or inserted, the transition metals, in particular, Ni, undergo oxidation state changes and with it, corresponding bond length changes.²⁶ The K-edge XANES from operando or ex-situ XAS experiments is often used to track such changes in NMC cathodes during, or at different stages of, battery cycling. Evaluating different ML models and featurization approaches based on computed data, we find that ensemble tree-

based methods such as gradient boosting or random forest models tend to outperform other ML models in terms of predicting bond length regression and oxidation state classification, in line with the findings of previous studies.^{4,12,18} Several featurization approaches result in similarly accurate inference on computed XANES, such as cumulative distribution function (CDF), peak feature and continuous wavelet transform (CWT). However, when performing inference on experimental data, we find that the performance of tree-based models can be enhanced by performing a cumulative distribution function (CDF) transformation on the XAS.

Methods

Overall workflow

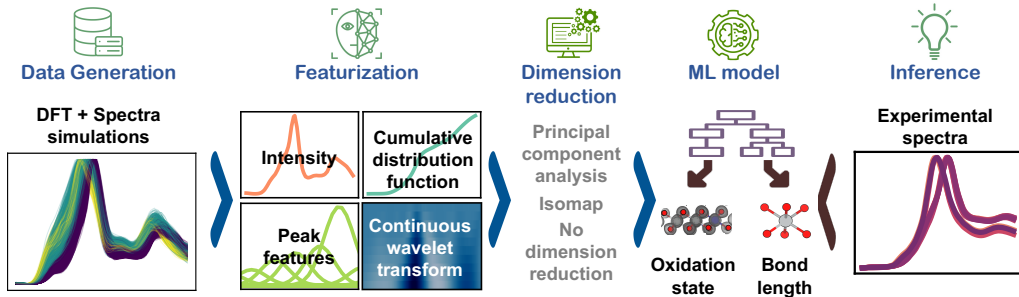


Figure 1: Schematic diagram of benchmarking feature space for supervised machine learning.

Figure 1 shows the overall workflow for this work. The raw XANES were first interpolated and projected onto the same energy scale with 100 individual intensities, which were then normalized to the maximum intensity. The initial feature vectors were then derived by performing four different transformations (see Figure 2):

1. Original intensity: The original vector of 100 intensities for each spectrum was used.

2. Cumulative distribution function (CDF): The CDF of each spectrum was computed and normalized to the maximum value.
3. Peak feature: Each spectrum was decomposed into a sum of 20 Gaussian peaks using non-linear least square fitting in scikit-learn.²⁷ The information for each Gaussian peak was simplified to three parameters: peak center, peak amplitude and peak width. During fitting, each peak center was limited to vary within a 5 eV range and each peak width, FWHM, was bounded between 0 and 6 eV. For each decomposed spectrum, a total of 20 Gaussian peaks are ordered by their peak energies in an increasing order and converted into a vector containing 60 values in the form of $[\text{center}_1, \text{amplitude}_1, \text{width}_1 \dots \text{center}_{20}, \text{amplitude}_{20}, \text{width}_{20}]$.
4. Continuous wavelet transformation (CWT): The CWT is a common transformation used in EXAFS analysis.^{28–30} However, relatively few studies have explored its application in XANES analysis. A Ricker wavelet function and widths of integers from 1 to 10 were applied for CWT transformation. The resultant CWT transformation is a 2d array with dimension (10, 100).

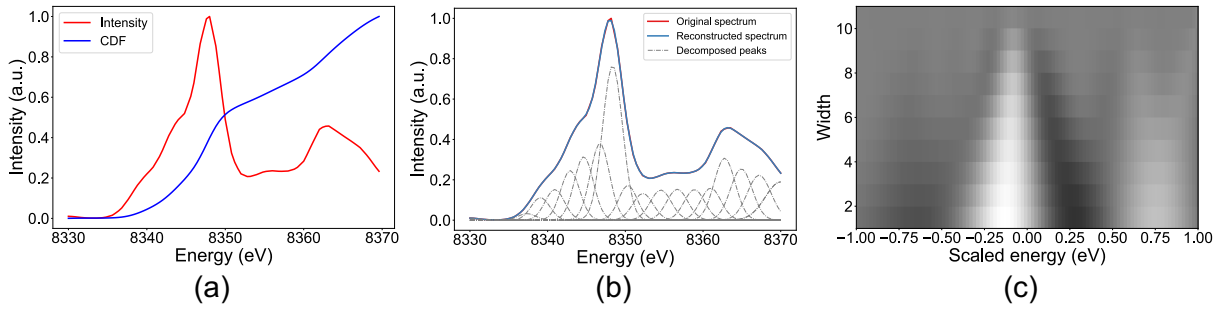


Figure 2: Visualization of spectral transformation for (a) intensity and CDF, (b) peak extraction and (c) CWT.

For each featurization approach, we then optionally applied three dimensionality reduction techniques - principal component analysis (PCA), Isomap, or Autoencoder (AE).

PCA^{31,32} is a linear decomposition technique that has been applied widely in the spectroscopy field for data analysis.^{23,33} Similarly, Isomap³⁴ is a non-linear isometric mapping dimensionality reduction technique to capture the underlying geometry of a data manifold. Autoencoder aims to learn how to efficiently compress and encode data so that it can reconstruct the data from the latent space representation in a way as close to the original input as possible. Tensorflow³⁵ was used to build AE models. To allow a fair comparison between different dimensionality reduction techniques, all features were reduced to 3-dimensional ones. The python code to featurize the spectra, example inputs, trained models and computational dataset are open to the public on Github: github.com/MaterialEyes/FeatureXAS. The dataset is also hosted on AIMMDB: aimm.lbl.gov/ui/browse/nmc_sim_vasp and Zenodo: zenodo.org/records/10476278.

Machine learning models

All the ML models used in this work were implemented using the scikit-learn package.²⁷ Four supervised ML models - ridge regression, gradient boosting (GB), random forest (RF), and multi-layer perceptron (MLP) - were assessed in terms of the performance in inferring oxidation states and bond length from XANES. In addition, we define a “baseline” (dummy - DUM) performance as the ratio of most abundant category for classification or the error based on always predicting the mean of dataset for regression.

Machine learning targets

We selected two targets for machine learning in this work. The first regression target is the average Ni-O bond length $\bar{L}$ of each NiO₆ octahedra, i.e.,

$$\bar{L} = \frac{\sum_{i=1}^6 L_i}{6} \tag{1}$$

where L_i is the length of the i th bond in an NiO₆ octahedra.

The second classification target is the oxidation state of Ni, which is determined from the integrated DFT spin density. Integrated spin density measures the magnetic net moment up to a certain radius (e.g., $2\text{\AA}$ in this study) around the atom and this method has been widely used in computational materials science to determine the oxidation states of transition metal elements.^{36–38} In this specific case, Ni^{2+} , Ni^{3+} and Ni^{4+} have electronic configurations of $t_{2g}^6e_g^2$, $t_{2g}^6e_g^1$ and $t_{2g}^6e_g^0$, respectively and those configurations represent 2, 1 and 0 net moment.

Dataset

The computed XAS dataset was obtained by performing density functional theory (DFT) calculations on NMC structures. An NMC material is typically abbreviated based on the relative ratios of transition metals in its formula. For example, NMC811 and NMC111, two common cathode compositions, refer to $\text{LiNi}_{0.8}\text{Mn}_{0.1}\text{Co}_{0.1}\text{O}_2$ and $\text{LiNi}_{1/3}\text{Mn}_{1/3}\text{Co}_{1/3}\text{O}_2$, respectively. The supercell used in this work is a $5\sqrt{3} \times 2\sqrt{3} \times 1/3$ cell of the LiCoO_2 conventional cell (structure prototype: $\alpha\text{-NaFeO}_2$, space group: $R\bar{3}m$), as shown in Figure 3. The Co is then replaced with an appropriate mixed occupancy of Ni:Mn:Co to obtain the NMC622, NMC811 and NMC721 compositions and the Li site is replaced with a partial occupancy to obtain lithiation levels in intervals of 0.1. An enumeration is then carried out to obtain all symmetrically-distinct orderings of transition metals and Li/vacancy. The detailed structure distribution can be found in Figure S1. All structures were fully relaxed before XAS computations (see below). A total of 2831 site-specific Ni K-edge XANES spectra were obtained for around 700 NMC structures. To account for the offset in energies between VASP-computed and experimental NMC spectra, all computed spectra were shifted higher in energy by a constant 120 eV. The spectra were also normalized such that the maximum peak intensity has a value of 1. One-dimensional interpolation was applied to ensure the same energy grid (i.e., 0.4 eV increment) and energy range (i.e., 8330–8370 eV) for all spectra in the dataset. Gaussian broadening with a full width half maximum (FWHM) of 3 eV

was applied on the raw spectra to mimic the instrumental broadening. The whole site-wise spectra dataset was split into a train and a test dataset with an 80:20 ratio.

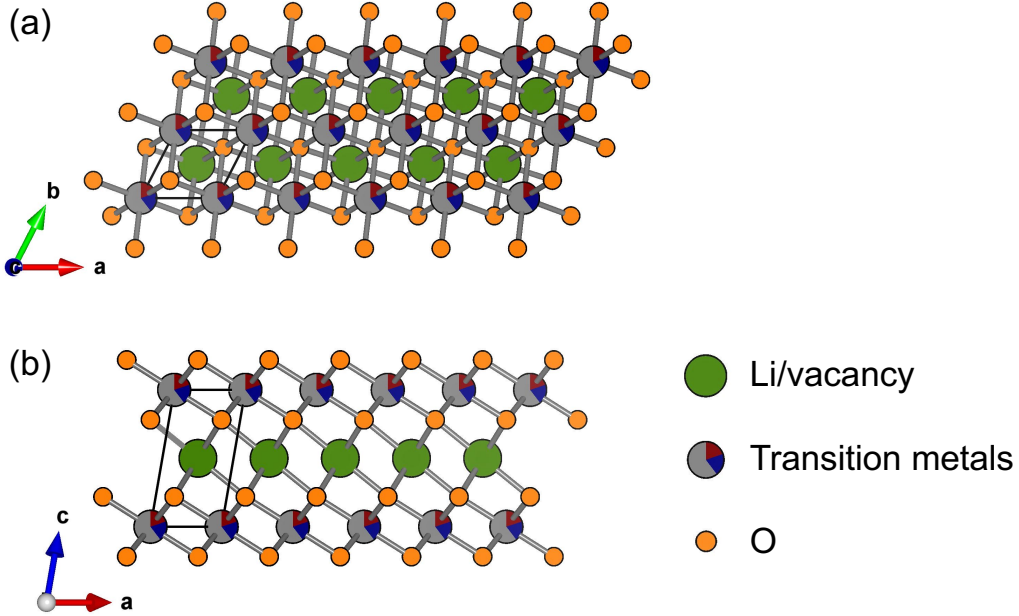


Figure 3: Crystal structure for the NMC supercell along (a)c-axis and (b)b-axis. Each supercell contains 10 formula units of $\text{Li}_z\text{Ni}_x\text{Mn}_y\text{Co}_{1-x-y}\text{O}_2$.

The experimental NMC datasets comprise ex-situ Ni K-edge XANES measurements taken for NMC622 and NMC721 batteries. The incident X-ray energy was selected using a three-quarters-tuned Si(111) double crystal monochromator and Harmonic rejection mirror at the 20-BM beamline of the Advanced Photon Source (APS). The XAFS data processing was performed using the Demeter package, a software package for the analysis of X-ray absorption spectroscopy, following standard analysis procedures.³⁹ We also collected two independent datasets for two Li-rich systems, $0.5\text{Li}_2\text{MnO}_3 \cdot 0.5\text{LiMn}_{0.5}\text{Ni}_{0.35}\text{Co}_{0.15}\text{O}_3$ ⁴⁰ and $\text{Li}_{1.2}\text{Mn}_{0.6}\text{Ni}_{0.2}\text{O}_2$,⁴¹ using WebPlotDigitizer.⁴² To compare with the experimental spectra, we constructed three site-averaged datasets from site-wise spectra. The first dataset contained only site-averaged spectra for NMC622 structures. The second dataset was for the NMC721 system, while the third one contained all site-averaged spectra for NMC622, NMC721, and NMC811. These datasets comprised 181, 115, and 409 spectra, respectively.

Density functional theory calculations

All spin-polarized DFT calculations were performed using version 6.1 of the Vienna Ab initio Simulation package (VASP)⁴³ within the projector-augmented wave approach.⁴⁴ The exchange-correlation functional used for structural relaxation and spectroscopy calculation was strongly-constrained and appropriately normed (SCAN) functional.⁴⁵ Hubbard U values of 2.43 eV, 2.93 eV, and 2.86 eV were applied for the *d* orbitals of Ni, Mn and Co, respectively, based on previous work by Wang et al.⁴⁶ who developed Hubbard U correction for SCAN functional using regular PAW potentials. All calculations were initialized in a ferromagnetic configuration with Co in a low spin state and Ni and Mn in a high spin state, consistent with previous studies.^{47–49} The plane wave energy cutoff was set to 450 eV and k-points density was 1500/ (# of atoms), which is similar to parameters used in Materials Project.⁵⁰ The energy and force convergence criteria were 1×10^{-4} eV and -0.05 eV $\text{\AA}^{-1}$, respectively. All input generation, and output analysis were performed using the open-source Python Materials Genomics (pymatgen) package.⁵¹

VASP6 was also used to compute the Ni K-edge XAS using the super-cell core-hole method⁵² that creates a 1s hole and places that electron into the conduction bands.. For spectroscopy calculations, the GW PAW potential, which is a harder pseudopotential that includes more electronic states, was used instead of the typical PAW ones used for structural relaxations. In addition to VASP, we also performed XAS calculations using three well-established computational codes - FEFF,⁵³ FDMNES,⁵⁴ and OCEAN⁵⁵ - for benchmarking purposes. A detailed discussion of computational XAS theory can be found in several excellent reviews.^{7–9} Briefly, FEFF and FDMNES employ real-space full multiple-scattering theory within the muffin-tin approximation, which simplifies calculations for complex systems.⁵³ FDMNES also supports the full-potential finite difference method, which avoids potential limitations from muffin-tin approximation by constructing a totally free potential shape. Such a non-muffin-tin effect is key to nanocluster simulations where the contribution from surface atoms is essential.⁵⁶ OCEAN employs DFT calculations with the Bethe-Salpeter

equation approach, which includes excitonic effects and better reproduces the L_3/L_2 ratios of light transition metals.⁵⁷ We adopted local density approximation (LDA) exchange correlation functional for OCEAN computation and SCAN for VASP computations.

Results

Benchmarking of computed XANES

Figure 4 compares the computed Ni K-edge XANES for NiO and LiNiO₂ from different codes. The computed spectra were horizontally shifted to align the position of the maximum peak in the experimental spectrum. For both NiO and LiNiO₂, we found that all four codes - VASP, FEFF, FDMNES, and OCEAN - produced XANES that are in good agreement with experiments. Using the Pearson correlation coefficient as a metric, FEFF and VASP produced XANES that have the highest similarity with the experimental spectra. Given that VASP was already used for structural relaxations, we adopted VASP for all subsequent XANES computations for ease of computational workflow.

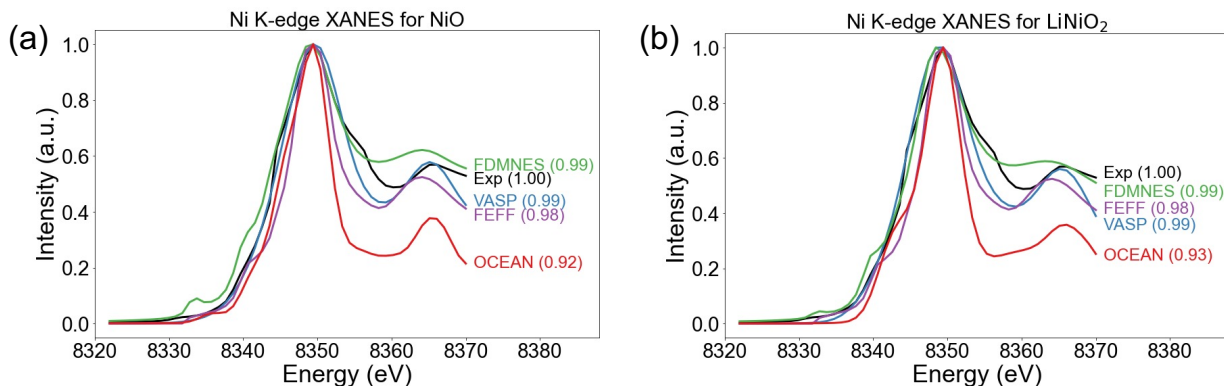


Figure 4: Ni K-edge XANES spectra for (a) NiO and (b) LiNiO₂ computed using VASP 6.1, FEFF 9.6, FDMNES, and OCEAN 2.5, as well as measured experimentally. The values in brackets indicate the Pearson correlation between the computed and experimental spectra. A higher Pearson correlation indicates a higher similarity.

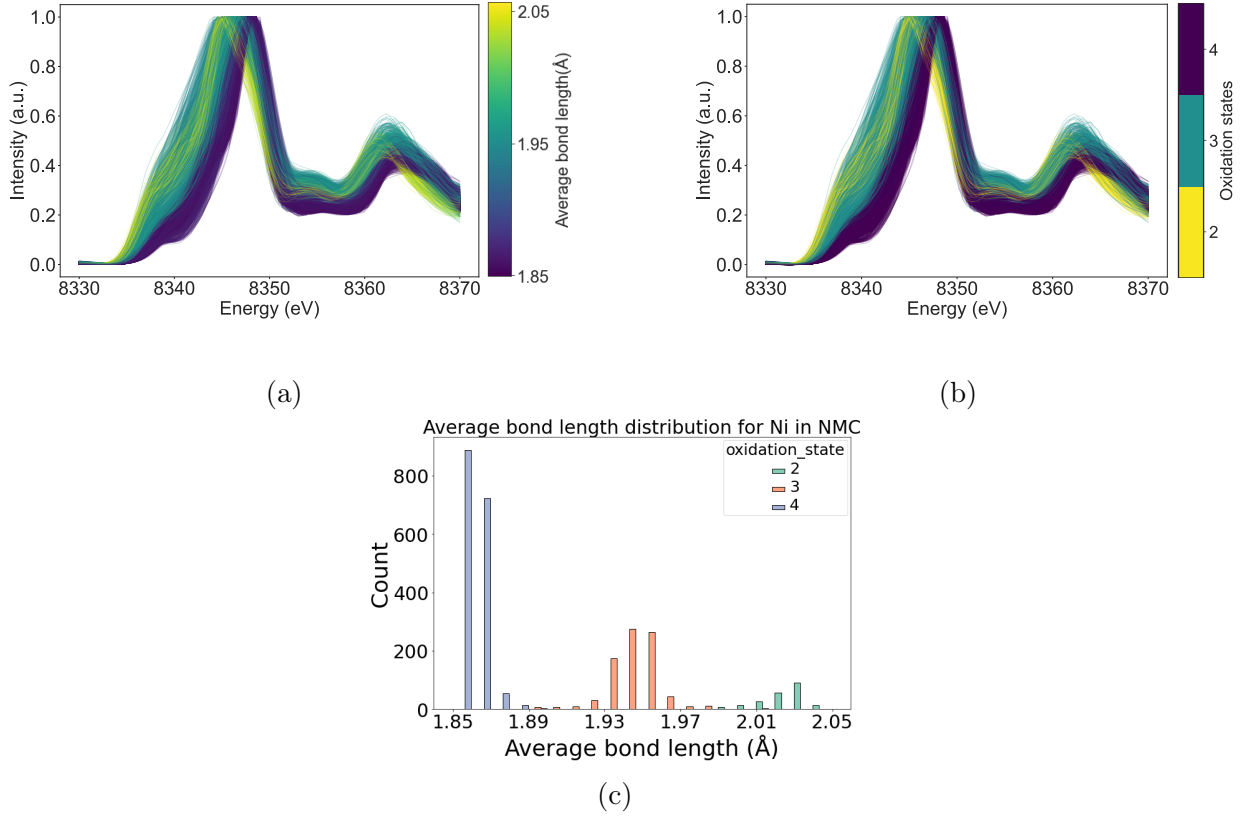


Figure 5: (a) Computed Ni K-edge XANES for NMC compounds, colored by average Ni-O bond length of absorbing atoms. (b) Computed Ni K-edge XANES for NMC compounds, colored by oxidation state of absorbing atoms. (c) Average bond length distribution for computed site-wise Ni K-edge XANES spectra.

Initial target analysis

Figure 5(a) and (b) plot the computed site-wise XANES in the train dataset, colored by the average Ni-O bond lengths and formal Ni oxidation states, respectively. It may be observed that the spectra corresponding to shorter average bond lengths and higher Ni oxidation states tend to shift towards higher absorption energies. This is consistent with the fact that Ni atoms in higher oxidation states tend to have shorter bond lengths due to stronger electrostatic attraction, which in turn results in higher energy necessary to excite the outermost electrons in the absorbing atom. Figure 5c plots the distribution of average bond lengths for different corresponding formal Ni oxidation states. It may be observed that the average Ni-O bond lengths for each formal Ni oxidation state are well separated.

Oxidation states classification

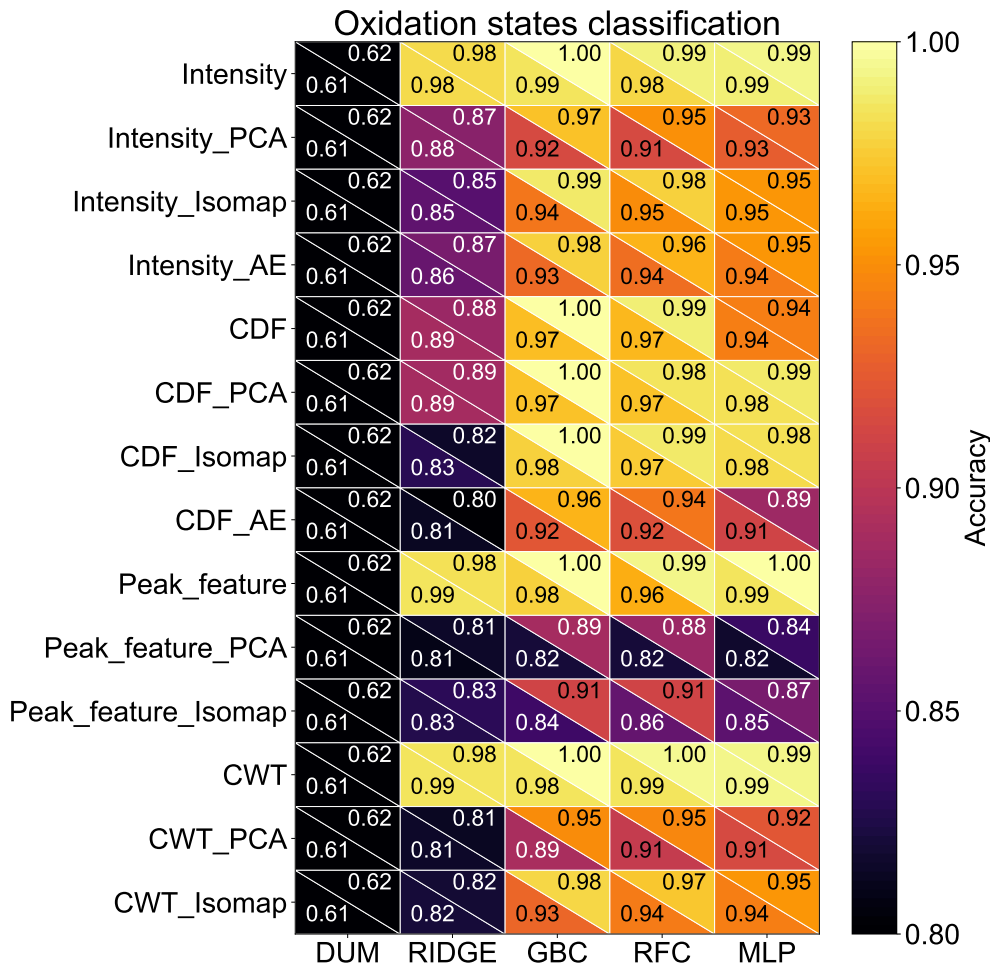


Figure 6: Accuracy for oxidation states classification. The labels on the bottom represent various ML models for the ML tasks (DUM = dummy or baseline, RIDGE = ridge regression, GBC = gradient boosting classifier, RFC = random forest classifier, and MLP = multi-layer perceptron). The side labels are for different features (CDF = cumulative distribution function, CWT = continuous wavelet transform, PCA = principle component analysis, AE = autoencoder). The upper right and lower left triangles within each cell represent train and test errors, respectively.

Figure 6 compares the classification accuracy of different ML models and feature transformations in predicting the formal oxidation state from the computed site-wise XANES. All raw features (i.e., CWT, CDF, and peak feature) exert a similar level of accuracy as compared to the baseline input, intensity. This indicates featurization of the original spectrum

preserves vital information to establish the relation between spectral shape and oxidation states. Generally, the train and test errors are similar, suggesting that there is little to no overfitting. The introduction of dimensionality reduction inevitably decreases the prediction accuracy.

However, the features reduced using Isomap experienced less decrease than PCA and AE. This suggests the possibility that the linear relationship between spectroscopy and properties of interest is not well-established. Such a non-linear relationship is also implied by the inferior performance when using linear machine learning models such as ridge regression. CDF experiences a marginal decrease after dimensionality reduction, suggesting that CDF plus dimensionality reduction techniques can be an optimal combination for large-scale applications.

Average bond length regression

The traditional way to obtain the average bond length around an absorbing atom is to perform fitting based on the EXAFS. However, in this study, we showcased that such information can be directly obtained from the XANES region of the spectra using ML without the EXAFS data. Although a similar schema was applied to this average bond length problem as to the oxidation states task, a regression problem that uses root mean squared error (RMSE) as the error metric can more appropriately describe the problem.

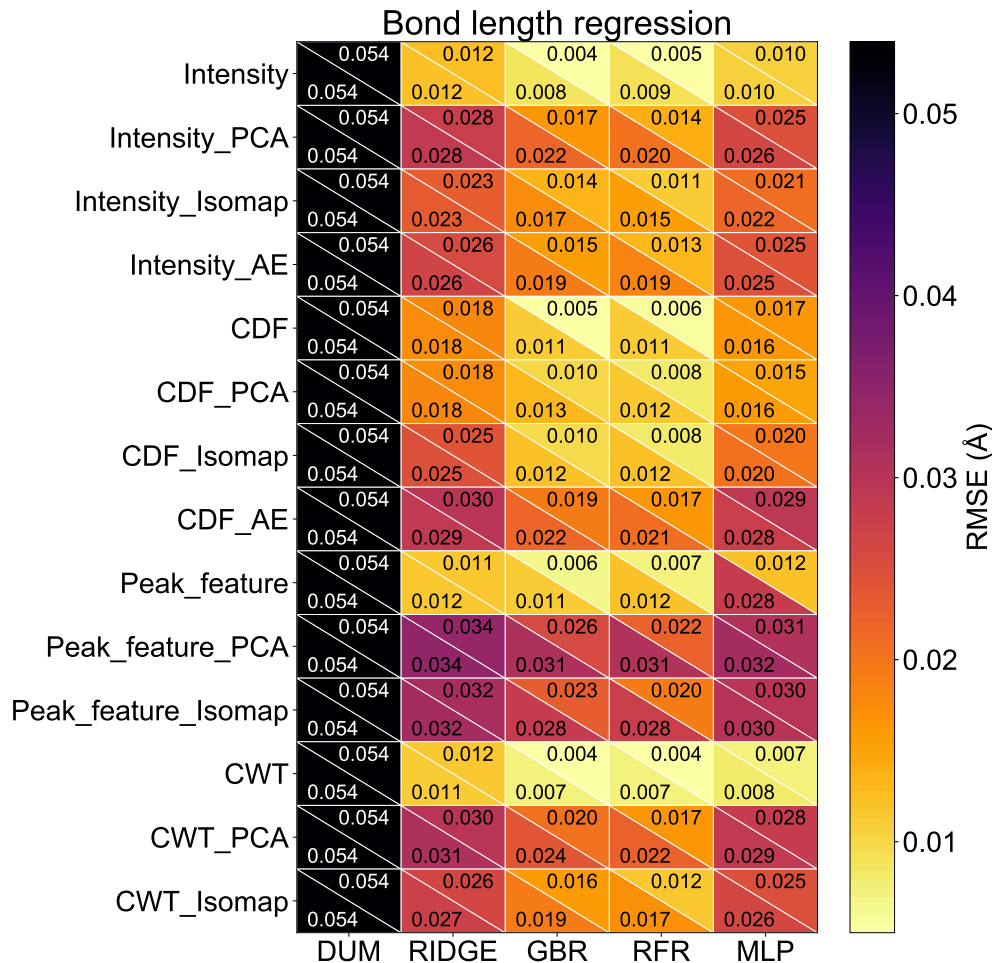


Figure 7: Root-mean-square errors for average bond length regression. The x label represents various ML models for the ML tasks and y labels are for different features, similar to Figure 6. The upper right and lower left triangles within each cell represent train and test errors, respectively.

The RMSEs for the train and test dataset are shown in Figure 7. Similar to the layout for the oxidation states classification, a lighter color in the heatmap represents a lower RMSE, hence a more accurate prediction. The major conclusions found in oxidation states classification align with those found in bond length regression, partially because the oxidation states and average bond length are highly correlated (as demonstrated in Figure 5c). Because ensemble models (i.e., RFR and RFC) achieved the highest prediction accuracy for both regression and classification tasks, we decided to adopt random forest models for later analysis.

Inference on Experimental Data

While the above results show that different featurization approaches (intensity, CDF, CWT) have similar performance when it comes to bond length and oxidation state inference on computed site-wise spectra, it is important to determine if these results hold when pre-trained ML models are applied towards data not used in training, especially experimentally-obtained spectra. In this section, we consider the performance of the featurization approaches on three experimental datasets. Different from the previous section where we trained models using only computational site-wise spectra, we used computational compound datasets for training purposes in this section.

Generally, all experimental spectra are site-averaged over all atoms of the same element. Therefore, compound or site-averaged spectra are needed for a direct comparison to experiment ones. The detailed procedures to collect experimental spectra and to construct compound datasets have been described in the Methods section.

As suggested in Figure 8(a), the experimental spectra for NMC622 samples form two distinct groups. Spectra 1, 4, and 8 form one group with lower peak energy, while the rest belong to the other group. Figure 8(c) indicates the average Ni-O bond lengths from EXAFS fitting and predicted values using various features and the pre-trained RF models. The CDF feature achieves a quantitative agreement with our reference, the bond length from EXAFS fitting, with an average 2% decrease compared to the values obtained from EXAFS fitting. This constant decrease may be attributed to the difference between DFT lattice parameters and experimental values, which is also around 2%. A detailed lattice parameter comparison is available in Figure S3. Moreover, Figure 8(c) displays both the corresponding voltage when measuring the XANES spectra and the inferred average oxidation states from compound spectra. While the ground truth average oxidation states are not known from these experimental data, we expect that the oxidation state will be higher for the higher voltage samples due to the removal of Li ions at high voltages. Moreover, as discussed before and shown in Figure 5c, shorter bond lengths should correspond to higher oxidation

states. Although raw features including intensity, CDF, and CWT performed satisfactorily on the computed dataset (see Figure 6), only CDF gives the expected variations in the inferred oxidation state. The predicted oxidation states range from 3+ to 4+, which is also consistent with expectation.

The results in Figure 8(d)-(f) for NMC721 lead to similar observations as NMC622 systems, supporting the feasibility of using CDF to infer average bond length and oxidation states in unseen experimental data. Additionally, Figure S5 shows the t-SNE distributions for compound NMC622 spectra. While t-SNE distributions for intensity, peak feature, and CWT form a continuous band in 2D projections, the distribution for CDF is more separated, implying a higher possibility of distinguishing different oxidation states.

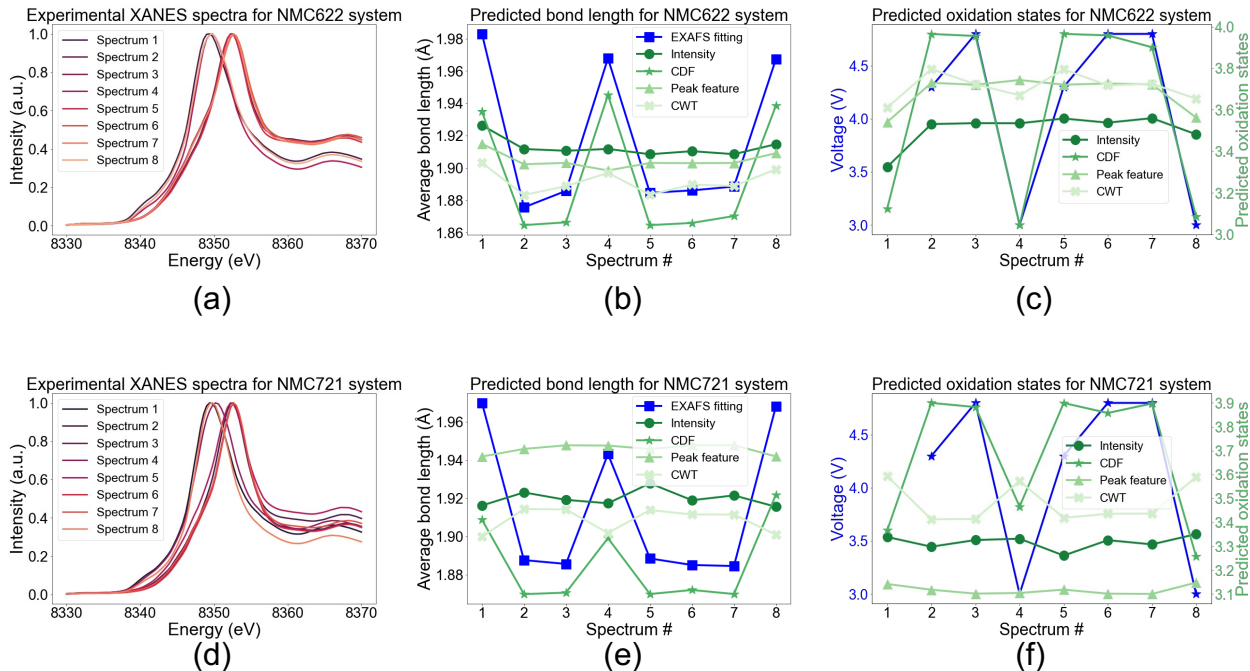


Figure 8: Performance of the trained random forest models on experimental NMC datasets. (a)-(c) show NMC622 results and (d)-(f) show NMC721 results. (a, d) Ex-situ measurements of Ni K-edge XANES. (b, e) Average bond length from EXAFS analysis and predicted average bond length from trained models. (c, f) Voltage during battery cycling where ex-situ samples were taken for XAS measurements and predicted oxidation states. The voltage for spectrum 1 was left blank because it corresponded to pristine sample.

An independent dataset of Ni K-edge XANES spectra for a Li-rich system was collected

from Reference⁴⁰ and replotted along with the computed site-wise spectra in Figure 9(a). This chemical system, which consisted of $0.5\text{Li}_2\text{MnO}_3 \cdot 0.5\text{LiMn}_{0.5}\text{Ni}_{0.35}\text{Co}_{0.15}\text{O}_3$, was similar to previous NMC systems but has a different transition metal ratio, as well as contains additional Li_2MnO_3 . Since the chemical system was different and there was no reference spectrum available in the computational dataset, no additional horizontal shift was applied, resulting in a several-eV shift between the computed and experimental spectra. However, CDF was still able to generate oxidation states that were expected for the measured voltage (Figure 9(b)), while raw intensity, CWT, and peak feature produced unusable results. Similar observations were also found from another independent dataset from Li et al. who investigated $\text{Li}_{1.2}\text{Mn}_{0.6}\text{Ni}_{0.2}\text{O}_2$ (LMNO) system.⁴¹ The corresponding results were shown in Figure S4. Despite the challenges posed by a different chemical system and lack of precise energy alignment, CDF demonstrated its robustness as a featurization approach for applying models trained from computed spectra to real-world data. Applying the model on chemistries outside of the training data gives qualitatively reasonable results, but for more quantitative inference, transfer learning with additional training data should be considered.

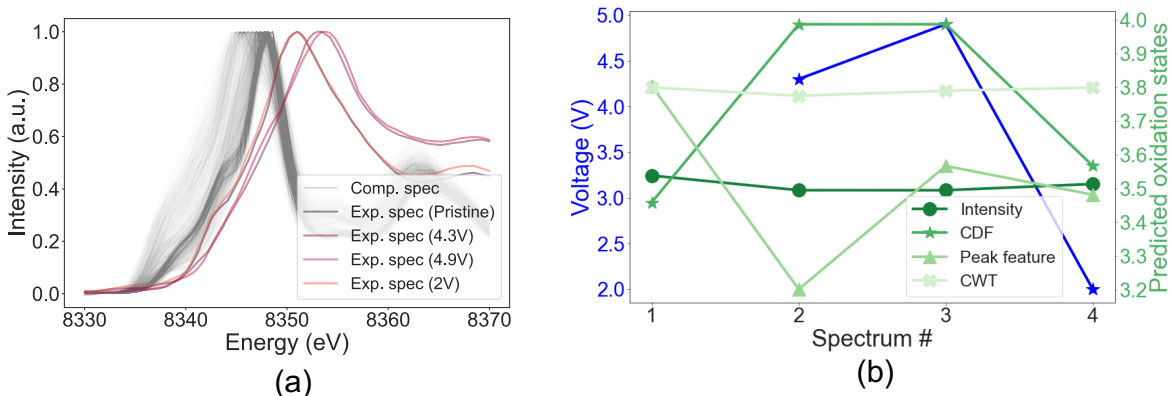


Figure 9: Performance of the trained random forest models on experimental Li-rich $0.5\text{Li}_2\text{MnO}_3 \cdot 0.5\text{LiMn}_{0.5}\text{Ni}_{0.35}\text{Co}_{0.15}\text{O}_3$ dataset.⁴⁰ (a) Visualization of experimental spectra and computed spectra. (b) Relation between corresponding voltage and predicted oxidation states. The voltage for spectrum 1 was left blank because it corresponded to pristine sample.

Model interpretability

Energy alignment is a challenge for XAS researchers, as its causes can be complex and difficult to pinpoint. Factors such as sample preparation, XAS measurements, and post-processing may all contribute to the horizontal shift if the data are not well-processed/treated. Consequently, an energy adjustment is typically necessary when comparing two datasets. In this section, we investigated the tolerance to energy shift for the intensity, CDF, peak feature, and CWT. For each featurization approach, we calculated the Pearson correlation between the shifted spectra, with the Ni K-edge XANES spectrum for pristine NMC622 serving as the reference. As shown in Figure 10, CDF demonstrated the least decrease in Pearson correlation when comparing shifted and non-shifted spectra, indicating its superior tolerance to energy shifts. Even in extreme cases such as 5 eV and -5 eV shift, the Pearson correlations for CDF were extremely high, at 0.97 and 0.98, respectively. In contrast, the baseline feature, intensity, experienced a much more drastic decrease (e.g., from 1.0 to 0.5 with a 5 eV shift). This finding explains why CDF worked best for the Li-rich dataset, which had several eV difference between the computed and experimental datasets. To the best of our knowledge, a 5 eV shift after manual alignment accounts for most deviations between experiments and computations. The success of CDF is a milestone in tackling specific alignment issues in ML spectral fingerprinting, but accurate computational prediction, as well as careful experimental calibration, of absolute spectral energy position will still be important for benchmarking and evaluation of future simulations and ML models.

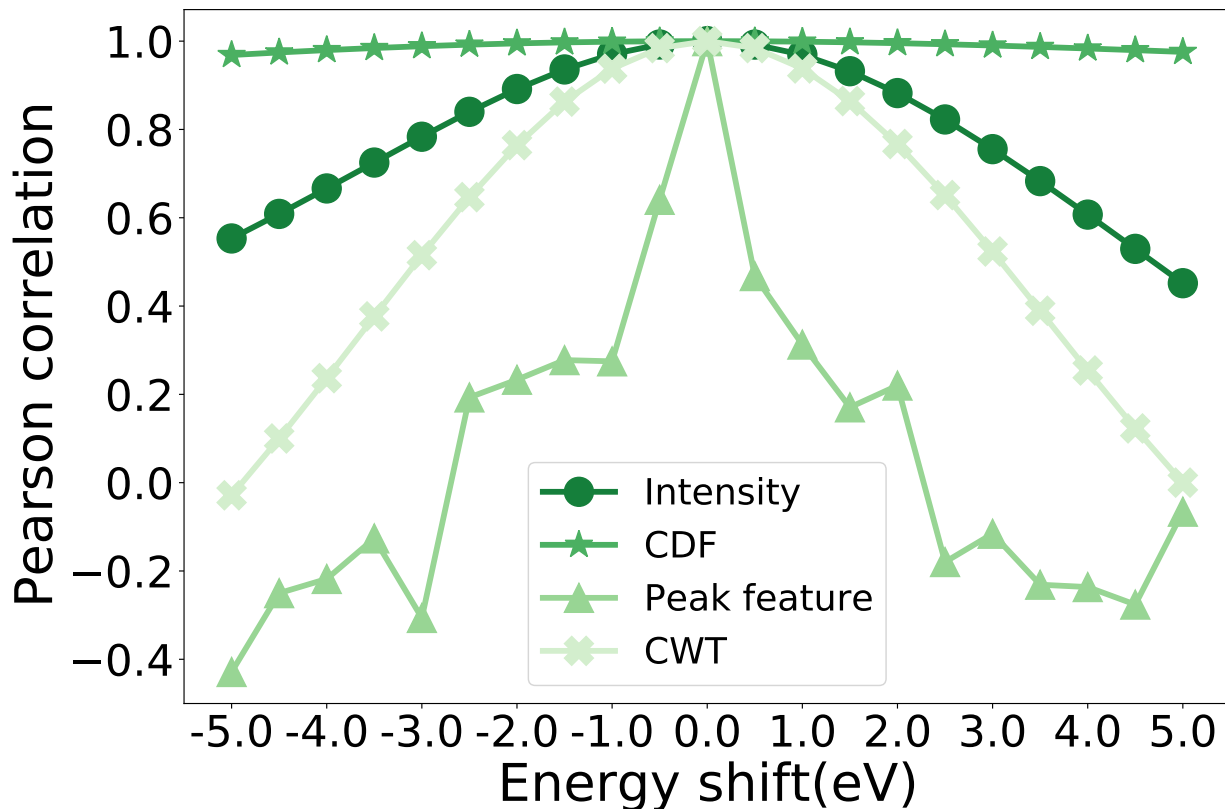


Figure 10: Pearson correlation between spectra with and without horizontal shift. 0 eV energy shift (no energy shift) was set as the reference to compute Pearson correlation for each feature.

Conclusion

In summary, this work addresses a crucial gap in data-driven approaches for XANES analysis through an in-depth benchmark for spectra featurization. In contrast to most previous studies that focused on spectral intensity, we investigated different ways to featurize the spectra and discovered that the CDF feature achieves a delicate balance between high prediction accuracy and excellent transferability. This excellent robustness is ascribed to its tolerance to energy shifts in spectra, which is critical when validating models trained on unseen experimental spectra. While CDF appears to alleviate the need for precise energy alignment, the success of model trained on simulated spectra towards inference from experimental measure-

ments still depends heavily on the accuracy of the simulated spectral shape. Although this study focused on a specific technique (XANES) and a specific edge for a family of battery materials, the use of CDF may be generally advantageous for spectroscopic studies, as shown in recent work on quantitative metrics for comparing molecular spectra.^{58–60}

Despite the continuous progress in both computational theory and computing power, the gap between experimental and computational spectroscopy remains. Such a gap has been impeding the broader application of ML models trained on computed datasets on experimental data, and the availability of experimental data with known ground truth is severely limited. The current study suggests that the use of CDF may be a useful strategy in bridging this gap and enabling ML models to harness the wealth of computed data while also making robust and accurate inference when applied to experimental data.

Acknowledgement

We thank Deyu Lu and Matthew Carbone for helpful discussion on spectral simulation and machine learning approaches. This work is supported by the U.S. Department of Energy (DOE) Office of Science Scientific User Facilities project titled “Integrated Platform for Multimodal Data Capture, Exploration and Discovery Driven by AI Tools”. We also acknowledge the support provided the Data Infrastructure Building Blocks (DIBBS) Local Spectroscopy Data Infrastructure (LSDI) project funded by National Science Foundation (NSF), under Award Number 1640899. M.K.Y.C. and Y.C. acknowledge the support from the BES SUFD Early Career award. M.J.D. was supported by the U. S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences operating under Contract Number DE-AC02-06CH11357. Work performed at the Center for Nanoscale Materials and the Advanced Photon Source, both U.S. Department of Energy Office of Science User Facilities, was supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. This research used resources of

the Advanced Light Source, which is a DOE Office of Science User Facility under contract No. DE-AC02-05CH11231. We gratefully acknowledge the computing resources provided on Bebop, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. We also acknowledge the computational resources provided by Triton Shared Computing Cluster (TSCC) and Expanse at University of California, San Diego. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions

M.K.Y.C., S.P.O., and Y.C proposed the concept. Y.C carried out the calculations and following analysis with the help of C.C., and M.J.D. C.J., W.Y., I.H., and G.L prepared the sample and performed XANES measurements. D.M., D.A., and J.M.A developed AIMMDB and enabled data dissemination. Y.C prepared the initial draft of the manuscript. All authors contributed to the discussion and revisions of the manuscript.

References

- (1) Ravel, B.; Scorzato, C.; Siddons, D. P.; Kelly, S. D.; Bare, S. R. Simultaneous XAFS measurements of multiple samples. *Journal of Synchrotron Radiation* **2010**, *17*, 380–385.
- (2) Boillot, M.-L.; Zarembowitch, J.; Itié, J.-P.; Polian, A.; Bourdet, E.; Haasnoot, J. G. Pressure-induced spin-state crossovers at room temperature in iron(II) complexes: comparative analysis; a XANES investigation of some new transitions Dedicated to the memory of Professor Olivier Kahn. *New Journal of Chemistry* **2002**, *26*, 313–322.
- (3) Hudson-Edwards, K. A.; Byrne, P.; Bird, G.; Brewer, P. A.; Burke, I. T.; Jamieson, H. E.; Macklin, M. G.; Williams, R. D. Origin and Fate of Vanadium in the Hazeltine Creek Catchment following the 2014 Mount Polley Mine Tailings Spill in British Columbia, Canada. *Environmental Science & Technology* **2019**, *53*, 4088–4098.
- (4) 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 Computational Materials* **2020**, *6*, 109.
- (5) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **2022**, *8*, 59.
- (6) Newville, M. Fundamentals of XAFS. *Reviews in Mineralogy and Geochemistry* **2014**, *78*, 33–74.
- (7) Rehr, J. J.; Albers, R. C. Theoretical approaches to x-ray absorption fine structure. *Reviews of Modern Physics* **2000**, *72*, 621–654.

- (8) Groot, F. d. Multiplet effects in X-ray spectroscopy. *Coordination Chemistry Reviews* **2005**, *249*, 31–63.
- (9) Laskowski, R.; Blaha, P. Understanding the L 2 , 3 x-ray absorption spectra of early 3 d transition elements. *Physical Review B* **2010**, *82*, 205104.
- (10) Mathew, K.; Zheng, C.; Winston, D.; Chen, C.; Dozier, A.; Rehr, J. J.; Ong, S. P.; Persson, K. A. High-throughput computational X-ray absorption spectroscopy. *Scientific Data* **2018**, *5*, 180151.
- (11) Chen, Y.; Chen, C.; Zheng, C.; Dwaraknath, S.; Horton, M. K.; Cabana, J.; Rehr, J.; Vinson, J.; Dozier, A.; Kas, J. J.; Persson, K. A.; Ong, S. P. Database of ab initio L-edge X-ray absorption near edge structure. *Scientific Data* **2021**, *8*, 153.
- (12) Guda, A. A.; Guda, S. A.; Lomachenko, K. A.; Soldatov, M. A.; Pankin, I. A.; Soldatov, A. V.; Braglia, L.; Bugaev, A. L.; Martini, A.; Signorile, M.; Groppo, E.; Piovano, A.; Borfecchia, E.; Lamberti, C. Quantitative structural determination of active sites from in situ and operando XANES spectra: From standard ab initio simulations to chemometric and machine learning approaches. *Catalysis Today* **2019**, *336*, 3–21.
- (13) Unruh, D.; Kolluru, V. S. C.; Baskaran, A.; Chen, Y.; Chan, M. K. Y. Theory+AI/ML for microscopy and spectroscopy: Challenges and opportunities. *MRS Bulletin* **2022**, *47*, 1024–1035.
- (14) Miyazato, I.; Takahashi, L.; Takahashi, K. Automatic oxidation threshold recognition of XAFS data using supervised machine learning. *Molecular Systems Design & Engineering* **2019**, *4*, 1014–1018.
- (15) Lüder, J. Determining electronic properties from L -edge x-ray absorption spectra of transition metal compounds with artificial neural networks. *Physical Review B* *103*, 045140.

- (16) Li, L.; Lu, M.; Chan, M. K. Y. A Deep Learning Model for Atomic Structures Prediction Using X-ray Absorption Spectroscopic Data. *arXiv* **2019**, arXiv:1905.03928, (accessed 2019-12-13).
- (17) Carbone, M. R.; Yoo, S.; Topsakal, M.; Lu, D. Classification of local chemical environments from x-ray absorption spectra using supervised machine learning. *Physical Review Materials* **2019**, *3*, 033604.
- (18) 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*, 100013.
- (19) Carbone, M. R.; Topsakal, M.; Lu, D.; Yoo, S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Physical Review Letters* **2020**, *124*, 156401.
- (20) David, T.; Nik Aznan, N. K.; Garside, K.; Penfold, T. Towards the automated extraction of structural information from X-ray absorption spectra. *Digital Discovery* **2023**, *2*, 1461–1470.
- (21) Golze, D.; Hirvensalo, M.; Hernández-León, P.; Aarva, A.; Etula, J.; Susi, T.; Rinke, P.; Laurila, T.; Caro, M. A. Accurate Computational Prediction of Core-Electron Binding Energies in Carbon-Based Materials: A Machine-Learning Model Combining Density-Functional Theory and *GW*. *Chemistry of Materials* **2022**, *34*, 6240–6254.
- (22) Chen, Z.; Andrejevic, N.; Smidt, T.; Ding, Z.; Xu, Q.; Chi, Y.; Nguyen, Q. T.; Alatas, A.; Kong, J.; Li, M. Direct Prediction of Phonon Density of States With Euclidean Neural Networks. *Advanced Science* **2021**, *8*, 2004214.
- (23) Tetef, S.; Govind, N.; Seidler, G. T. Unsupervised Machine Learning for Unbiased Chemical Classification in X-ray Absorption Spectroscopy and X-ray Emission Spectroscopy. *Physical Chemistry Chemical Physics* **2021**, *15*.

- (24) Routh, P. K.; Liu, Y.; Marcella, N.; Kozinsky, B.; Frenkel, A. I. Latent Representation Learning for Structural Characterization of Catalysts. *The Journal of Physical Chemistry Letters* **2021**, *12*, 2086–2094.
- (25) Manthiram, A. A reflection on lithium-ion battery cathode chemistry. *Nature Communications* **2020**, *11*, 1550.
- (26) Chakraborty, A.; Kunnikuruvan, S.; Kumar, S.; Markovsky, B.; Aurbach, D.; Dixit, M.; Major, D. T. Layered Cathode Materials for Lithium-Ion Batteries: Review of Computational Studies on $\text{LiNi}_{1-x-y}\text{Co}_x\text{Mn}_y\text{O}_2$ and $\text{LiNi}_{1-x-y}\text{Co}_x\text{Al}_y\text{O}_2$. *Chemistry of Materials* **2020**, *32*, 915–952.
- (27) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
- (28) Funke, H.; Chukalina, M.; Scheinost, A. C. A new *FEFF* -based wavelet for EXAFS data analysis. *Journal of Synchrotron Radiation* **2007**, *14*, 426–432.
- (29) Muñoz, M.; Argoul, P.; Farges, F. Continuous Cauchy wavelet transform analyses of EXAFS spectra: A qualitative approach. *American Mineralogist* **2003**, *88*, 694–700.
- (30) Timoshenko, J.; Kuzmin, A. Wavelet data analysis of EXAFS spectra. *Computer Physics Communications* **2009**, *180*, 920–925.
- (31) Hotelling, H. ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS. *Journal of Educational Psychology* **1933**, *24*, 417.
- (32) Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572.

- (33) Chen, Z.; Andrejevic, N.; Drucker, N. C.; Nguyen, T.; Xian, R. P.; Smidt, T.; Wang, Y.; Ernstorfer, R.; Tennant, D. A.; Chan, M.; Li, M. Machine learning on neutron and x-ray scattering and spectroscopies. *Chemical Physics Reviews* **2021**, *2*, 031301.
- (34) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323.
- (35) Abadi, M. et al. TensorFlow: A system for large-scale machine learning. *arXiv* **2016**, arXiv:1605.08695, (accessed 2022-05-23).
- (36) Liu, H. et al. Elucidating the Limit of Li Insertion into the Spinel $\text{Li}_4\text{Ti}_5\text{O}_{12}$. *ACS Materials Letters* **2019**, *1*, 96–102.
- (37) Walsh, A.; Sokol, A. A.; Buckeridge, J.; Scanlon, D. O.; Catlow, C. R. A. Electron Counting in Solids: Oxidation States, Partial Charges, and Ionicity. *The Journal of Physical Chemistry Letters* **2017**, *8*, 2074–2075.
- (38) Feng, X.; Chien, P.-H.; Zhu, Z.; Chu, I.-H.; Wang, P.; Immediato-Scuotto, M.; Arabzadeh, H.; Ong, S. P.; Hu, Y.-Y. Studies of Functional Defects for Fast Na-Ion Conduction in $\text{Na}_{3-y}\text{PS}_{4-x}\text{Cl}_x$ with a Combined Experimental and Computational Approach. *Advanced Functional Materials* **2019**, *29*, 1807951.
- (39) Ravel, B.; Newville, M. *ATHENA*, *ARTEMIS*, *HEPHAESTUS*: data analysis for X-ray absorption spectroscopy using *IFEFFIT*. *Journal of Synchrotron Radiation* **2005**, *12*, 537–541.
- (40) Ates, M. N.; Mukerjee, S.; Abraham, K. M. A high rate Li-rich layered MNC cathode material for lithium-ion batteries. *RSC Advances* **2015**, *5*, 27375–27386.
- (41) Li, Q.; Ning, D.; Wong, D.; An, K.; Tang, Y.; Zhou, D.; Schuck, G.; Chen, Z.; Zhang, N.; Liu, X. Improving the oxygen redox reversibility of Li-rich battery cathode materials via Coulombic repulsive interactions strategy. *Nature Communications* **2022**, *13*, 1123.

- (42) Rohatgi, A. Webplotdigitizer: Version 4.6. 2022; <https://automeris.io/WebPlotDigitizer>, (accessed 2022-11-02).
- (43) Kresse, G.; Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Physical Review B* **1996**, *54*, 11169–11186.
- (44) Blöchl, P. E. Projector augmented-wave method. *Physical Review B* **1994**, *50*, 17953–17979.
- (45) Sun, J.; Ruzsinszky, A.; Perdew, J. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Physical Review Letters* **2015**, *115*, 036402.
- (46) Wang, Z.; Guo, X.; Montoya, J.; Nørskov, J. K. Predicting aqueous stability of solid with computed Pourbaix diagram using SCAN functional. *npj Computational Materials* **2020**, *6*, 160.
- (47) Mueller, T.; Hautier, G.; Jain, A.; Ceder, G. Evaluation of Tavorite-Structured Cathode Materials for Lithium-Ion Batteries Using High-Throughput Computing. *Chemistry of Materials* **2011**, *23*, 3854–3862.
- (48) Tang, H.; Deng, Z.; Lin, Z.; Wang, Z.; Chu, I.-H.; Chen, C.; Zhu, Z.; Zheng, C.; Ong, S. P. Probing Solid–Solid Interfacial Reactions in All-Solid-State Sodium-Ion Batteries with First-Principles Calculations. *Chemistry of Materials* **2018**, *30*, 163–173.
- (49) Guo, X.; Wang, Z.; Deng, Z.; Wang, B.; Chen, X.; Ong, S. P. Design Principles for Aqueous Na-Ion Battery Cathodes. *Chemistry of Materials* **2020**, *32*, 6875–6885.
- (50) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

- (51) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.
- (52) Karsai, F.; Humer, M.; Flage-Larsen, E.; Blaha, P.; Kresse, G. Effects of electron-phonon coupling on absorption spectrum: K edge of hexagonal boron nitride. *Physical Review B* **2018**, *98*, 235205.
- (53) Rehr, J. J.; Kas, J. J.; Vila, F. D.; Prange, M. P.; Jorissen, K. Parameter-free calculations of X-ray spectra with FEFF9. *Physical Chemistry Chemical Physics* **2010**, *12*, 5503.
- (54) Bunău, O.; Joly, Y. Self-consistent aspects of x-ray absorption calculations. *Journal of Physics: Condensed Matter* **2009**, *21*, 345501.
- (55) Vinson, J.; Rehr, J. J.; Kas, J. J.; Shirley, E. L. Bethe-Salpeter equation calculations of core excitation spectra. *Physical Review B* **2011**, *83*, 115106.
- (56) Kravtsova, A. N.; Lomachenko, K. A.; Soldatov, A. V.; Meyer, J.; Niedner-Schatteburg, G.; Peredkov, S.; Eberhardt, W.; Neeb, M. Atomic and electronic structure of free niobium nanoclusters: Simulation of the M 4,5 -XANES spectrum of Nb 13 +. *Journal of Electron Spectroscopy and Related Phenomena* **2014**, *195*, 189–194.
- (57) Poths, P.; Alexandrova, A. N. Theoretical Perspective on *Operando* Spectroscopy of Fluxional Nanocatalysts. *The Journal of Physical Chemistry Letters* **2022**, *13*, 4321–4334.
- (58) Seifert, N. A.; Prozument, K.; Davis, M. J. Computational optimal transport for molecular spectra: The fully discrete case. *The Journal of Chemical Physics* **2021**, *155*, 184101.

- (59) Seifert, N. A.; Prozument, K.; Davis, M. J. Computational optimal transport for molecular spectra: The semi-discrete case. *The Journal of Chemical Physics* **2022**, *156*, 134117.
- (60) Seifert, N. A.; Prozument, K.; Davis, M. J. Computational optimal transport for molecular spectra: The fully continuous case. *The Journal of Chemical Physics* **2023**, *159*, 164110.

Supporting information

Detailed NMC structure distribution; spectra distribution; DFT functional comparison on lattice parameter; inference on LMNO dataset; and T-SNE distribution

