Machine Learning Prediction of Coordination Number from EXAFS Spectra

Authors: Zeng Haitao, Hu Longfei, Yao Tao

Date: 2025-04-03T00:00:00+00:00

Abstract

X-ray absorption fine structure (XAFS) is an important structural analysis technique widely applied to investigate the oxidation state, coordination environment, and neighboring atomic characteristics of amorphous materials and disordered systems. However, due to the complexity of XAFS spectra, their interpretation relies on experienced researchers and is prone to inaccuracies. This study utilizes machine learning methods, particularly neural networks, bagged decision tree models, and random forest models, to analyze XAFS data for predicting the coordination number of absorbing elements. The research collected EXAFS data for fourth-period transition metal elements from the Materials Project database, encompassing various coordination environments, with a total of 13,374 valid data entries. The results demonstrate that both neural network and random forest models achieve high accuracy in predicting coordination numbers. By enhancing the generalization capability and interpretability of the models, this study provides a more efficient and reliable method for XAFS data analysis.

Full Text

Machine Learning Approach for Predicting Coordination Numbers from EXAFS Spectra

Haitao Zeng, Longfei Hu, Tao Yao
National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei 230029, China

Abstract

Background: X-ray Absorption Fine Structure (XAFS) is a vital technique for structural analysis, widely employed to investigate the oxidation state, coordination environment, and neighboring atom properties of amorphous materials and

disordered systems. However, the complexity of XAFS spectra often requires interpretation by experienced researchers, which can still lead to inaccuracies. **Purpose:** This study aims to use machine learning approaches to analyze XAFS data and predict the coordination number of absorbing atoms. **Methods:** First, a dataset of 13,374 valid EXAFS spectra of fourth-period transition metal elements was sourced from the Materials Project database. Second, this data was utilized to train three machine learning models: neural networks, bagging models, and random forest models. Finally, these models were applied to predict the coordination numbers of the absorbing atoms in the spectra. **Results:** The study achieved an average prediction accuracy of approximately 70%. Feature importance analysis revealed that data points within R $<$ 3.0 Å were critical for predictions, consistent with the prominence of short-range atomic interactions in EXAFS theory. **Conclusions:** This research enhances the efficiency and reliability of XAFS data analysis by improving model generalizability and interpretability.

Keywords: EXAFS, Coordination Number, Bagging, Random Forest

X-ray Absorption Fine Structure (XAFS) refers to the characteristic fine spectral features exhibited when X-ray energies approach or exceed the binding energy of atomic core levels. XAFS spectra are highly sensitive to the oxidation state of absorbing atoms, their coordination environment, and the types and distances of neighboring atoms. Moreover, since X-ray interactions with atoms impose no specific requirements on sample morphology, XAFS has become an essential analytical tool for investigating the structure of amorphous materials and disordered systems [**?**].

XAFS can be divided into two main regions: X-ray Absorption Near Edge Structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS). XANES primarily analyzes the oxidation state and local coordination geometry of absorbing atoms, while EXAFS focuses on determining coordination numbers, neighboring atom types, and distances between absorbing and neighboring atoms.

Due to its unique capabilities, XAFS finds widespread application across physics, chemistry, materials science, and biology. However, the complexity of XAFS spectra typically necessitates interpretation by experienced researchers. Even so, professionals may still reach inaccurate conclusions during analysis [**?**]. Consequently, developing more efficient and reliable analytical methods to improve the accuracy of XAFS data interpretation has become an important research direction.

Machine learning, as a representative data-driven research method in recent years, has emerged as a significant focus in scientific research. By inputting sufficient datasets and setting appropriate hyperparameters, researchers can enable computers to automatically learn underlying patterns in data and generate corresponding model parameters. Neural networks have attracted particular attention for their high prediction accuracy. However, due to their complexity

and "black box" nature, researchers often struggle to interpret the specific basis for model parameters and cannot intuitively understand the decision-making process. This lack of interpretability limits the application of neural networks in research scenarios requiring explicit feature importance analysis [**?**].

To improve the interpretability of machine learning models, researchers have developed various techniques, including Ridge Regression [**?**], LASSO (Least Absolute Shrinkage and Selection Operator) [**?**], SISSO (Sure Independence Screening and Sparsifying Operator) [**?**], and Decision Trees [**?**]. Decision trees have gained widespread attention due to their simple structure and high interpretability. Furthermore, based on Ensemble Learning [**?**], researchers have developed additional methods such as Bagging [**?**], Random Forest [**?**], and Boosting [**?**].

Neural network technology has demonstrated breakthrough research progress in materials science. As early as 1999, when artificial neural networks were still in their early stages, researchers successfully applied this technology to predict multi-scale material properties, including alloy mechanical strength, metal fatigue behavior, and phase transition temperatures [**?**].

In XAFS research, deep neural network technology has been widely applied to extended-edge spectral analysis and near-edge structure feature extraction. Timoshenko et al. developed an innovative neural network-based method to directly extract radial distribution functions (RDF) from EXAFS spectra without relying on prior structural assumptions. This method generated training data through molecular dynamics, optimized input features using wavelet transforms, and was successfully applied to analyze high-temperature bcc-fcc phase transitions in iron, with subsequent extension to nickel and cobalt systems, demonstrating its efficiency and versatility [**?**].

Beyond neural network models, other recent advances in machine learning have enhanced spectral data analysis capabilities. For example, Tetef et al. (2021) demonstrated that unsupervised methods such as t-SNE and VAE could effectively classify sulfur organic compounds based on XANES and VtC-XES spectra, revealing detailed chemical properties including oxidation states and aromaticity [**?**]. Guda et al. (2021) used a combination of linear regression and support vector machines to model relationships between XANES descriptors and structural parameters, achieving high prediction accuracy [**?**]. This XANES spectral analysis method is versatile and can be applied to various material systems, providing a powerful tool for structural characterization.

Decision tree models have also shown potential in XAFS spectral research. Torrisi et al. (2020) utilized random forest machine learning models to analyze X-ray Absorption Near Edge Structure (XANES) spectra, predicting coordination numbers, nearest-neighbor distances, and Bader charges of absorbing atoms in transition metal oxides [**?**]. Their study improved model interpretability through multi-scale polynomial feature engineering, revealing key spectral regions associated with different properties. Liu et al. (2024) introduced random forest, gradient boosting decision trees, support vector machines, and multi-grained cascade

forest (gcF) models to predict binding configurations of heavy metals on iron (oxyhydr)oxides by analyzing EXAFS data. Their research demonstrated that the gcF model achieved optimal performance in predicting coordination types and Me-Fe shell types [**?**].

In previous studies using machine learning to analyze XAFS spectra, researchers typically modeled XAFS data for single materials under different conditions (such as temperature variations). Although such studies achieved certain results, the resulting machine learning models often lacked generalizability and were difficult to extend to other material systems. To address this issue, our study collected EXAFS data for fourth-period transition metal elements from the Materials Project database. This data encompasses spectral features of the same element under various coordination environments, aiming to extract common patterns from diverse coordination geometries. Our research focuses on the coordination number of absorbing elements as the key objective. Results demonstrate that both neural network and random forest models exhibit high accuracy in predicting coordination numbers. Furthermore, random forest analysis reveals that coordination number predictions primarily rely on intensity information from early data points in the Fourier-transformed EXAFS spectra. The workflow of this study is illustrated in Figure 1 [Figure 1: see original paper]. We collected a total of 15,498 XAFS data entries for fourth-period transition metal elements from the Materials Project, retaining 13,374 entries with integer coordination numbers after screening. Subsequently, we transformed these XAFS data from E-space to R-space and imported them into fully connected neural networks and decision tree-based bagging and random forest ensemble learning models. For each element, we trained and evaluated models separately.

1.1 X-ray Absorption Fine Structure (XAFS) and Extended X-ray Absorption Fine Structure (EXAFS)

The XAFS phenomenon was first discovered and scientifically documented by Friche and Hertz in 1920. A breakthrough in XAFS research emerged from the innovative method proposed by Sayers, Stern, and Lytle—using Fourier transform techniques to effectively analyze EXAFS oscillation signals in wavevector space [**?**]. This milestone work marked the entry of XAFS technology into quantitative analysis and initiated its systematic development for materials characterization [**?**].

XAFS technology provides detailed structural information about the local chemical environment around absorbing atoms. In recent years, atomic-level local property analysis based on XAFS has become an important approach in materials science and catalysis research. Specifically, XAFS technology has been applied to investigate coordination environments in metal clusters, explore local structural distortions in ferroelectric materials, and analyze catalytic mechanisms in single-atom catalysts.

XAFS spectra are obtained by measuring a material' s X-ray absorption co-

efficient as a function of incident photon energy. XAFS spectra are typically divided into two characteristic regions: X-ray Absorption Near Edge Structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS). The theoretical expression for EXAFS can be represented as: $(k) = S_0^2 \Sigma (N f(k)/kR^2) e^{\wedge}(-2R/\lambda(k)) e^{\wedge}(-2k^2\sigma^2) \sin(2kR + \delta(k))$, where k represents the photoelectron wavevector, N denotes the coordination number of the j-th shell, R is the average distance from the absorbing atom to neighbors in the j-th shell, σ^2 characterizes the disorder of the j-th shell, f (k) represents the scattering amplitude of neighboring atoms, δ (k) is the phase shift, λ(k) denotes the photoelectron mean free path, and S_0^2 is the amplitude reduction factor.

The above theoretical expression indicates that Fourier transform can convert the EXAFS oscillation function from k-space to R-space, thereby obtaining radial distribution information about the local structure around absorbing atoms. However, since the EXAFS formula contains multiple coupled variable parameters and may involve superposition of multiple scattering paths, precise structural analysis based solely on EXAFS spectra faces severe ill-posed problems. Therefore, in practical EXAFS data analysis, researchers typically employ a "fingerprint comparison" approach—comparing experimental data with reference spectra of known structures or theoretical simulations to obtain relative models of local structure.

Although direct inversion of all structural parameters from XAFS spectra presents inherent difficulties, calculating theoretical XAFS spectra based on known structural models has become a reliable research method. In XAFS research, FEFF [**?**] and FDMNES [**?**] are two widely adopted computational packages. Both software tools employ first-principles (ab initio) calculations by solving the Schrödinger equation to accurately simulate XAFS spectral features. FEFF uses Real-Space Multiple Scattering (RSMS) theory, while FDMNES is based on the Finite Difference Method (FDM), providing reliable theoretical references for experimental data interpretation.

1.2 Larch [?]

Larch is a comprehensive toolkit specifically designed for XAFS and related spectral data analysis. Compared to the IFEFFIT package developed in FORTRAN, Larch' s reconstruction in Python significantly enhances large-scale data processing capabilities and data visualization functions. In this study, we utilized core functions from the Larch toolkit for data processing: the autobk function performs background subtraction and normalization to convert EXAFS spectra from E-space to k-space, while the xftf function executes Fourier transform to convert EXAFS spectra from k-space to R-space.

1.3 Coordination Number

Coordination number is a key structural parameter describing the local coordination environment of atoms in materials and is crucial for studying mate-

rial properties. In traditional studies, coordination numbers were typically determined through intuitive judgment of coordination conditions for individual atoms in materials. However, with the development of high-throughput materials computing technology and the generation of large-scale datasets, developing reliable and automated coordination number calculation algorithms has become an urgent research need.

Currently, various coordination number calculation methods have been proposed in materials science [**?**]. Simpler algorithms are based on empirical parameters for interatomic distances and tolerance values, such as BrunnerNN [**?**] and MinimumOKeeffeNN [**?**]. These methods determine coordination numbers by comparing actual interatomic distances in models with their tolerance values. Although these algorithms are relatively simple and intuitive, they are sensitive to small atomic perturbations and changes in empirical tolerance parameters, where minor atomic variations may lead to different coordination number assignments.

O' Keeffe' s VoronoiNN algorithm [**?**] provides an alternative solution. Based on geometric principles, this algorithm employs Voronoi decomposition [**?**] to treat the environment around a central atom as a polyhedron and performs weighting based on spatial angles to determine the number of neighboring atoms. Similarly, this study uses the CrystalNN algorithm [**?**]. The CrystalNN method is also based on Voronoi decomposition and is widely recognized for its high computational accuracy. This method calculates probabilities of possible coordination environments through Voronoi decomposition and ultimately selects the result with the highest probability as the atom' s coordination environment. In this paper, we systematically calculated coordination numbers of absorbing atoms in various material systems using the CrystalNN algorithm from the pymatgen materials computing package [**?**].

1.4 Neural Networks

As an important branch of machine learning, neural networks (deep learning) have become a core area of scientific research and technological application due to their exceptional performance in regression and classification tasks. Neural network models achieve feature extraction and pattern recognition through numerous adjustable parameters that are automatically optimized during training via backpropagation algorithms. Researchers primarily optimize model performance by tuning hyperparameters including batch size, learning rate, and network architecture [**?**].

A single hidden layer feedforward neural network can be formally defined by the mathematical expression: $f(X) = \beta_0 + \Sigma\ \beta\ g(\alpha_0 + \Sigma\ \alpha\ X)$, where $X = (X_1, X_2, \cdots, X)$ represents the p-dimensional input feature vector. Model parameters include: α representing connection weights from the j-th input feature to the k-th hidden neuron, β denoting weight coefficients from the k-th hidden neuron to the output layer, and β_0, α_0 corresponding to bias terms for

linear transformations.

Although this architecture appears as a linear superposition in form, nonlinear activation functions g($\cdot$) must be introduced to effectively approximate nonlinear functional relationships in real-world scenarios. In machine learning, the Sigmoid function and ReLU (Rectified Linear Unit) function [**?**] are two classic activation functions. The Sigmoid function exhibits S-shaped saturation characteristics with an output range of (0,1), which made it widely used in early neural network research. However, this function suffers from vanishing gradients when input values approach positive or negative infinity, where derivatives approach zero and significantly reduce parameter update efficiency during backpropagation. In contrast, the ReLU function, defined as g(x) = max(0,x), ensures constant derivatives in the positive region due to its piecewise linear nature, thereby effectively mitigating the vanishing gradient problem. Additionally, ReLU' s computational complexity is significantly lower than that of the Sigmoid function, enabling superior computational efficiency and convergence properties in deep neural network training.

1.5 Decision Trees [?]

Despite neural networks' significant advantages in prediction performance, their inherent "black box" characteristics result in insufficient model interpretability. Specifically, the mechanism for setting numerous neuron parameters in neural networks is difficult to interpret clearly, prompting researchers to seek alternative analysis tools with stronger interpretability. In this study, we employ decision trees as an analysis tool due to their intuitive structure and ease of interpretation. Through decision trees, researchers can clearly identify relationships between data features and results. However, traditional decision tree methods exhibit obvious deficiencies in prediction accuracy. To overcome this limitation, researchers have developed various ensemble learning methods, including Bagging [**?**], Random Forest [**?**], and Boosting [**?**]. While these ensemble methods significantly improve prediction performance, they sacrifice the intuitive interpretability of single decision trees. Nevertheless, by analyzing the Gini Index, we can still effectively evaluate the importance of each feature in classification decisions.

Decision tree modeling follows a recursive feature space partitioning strategy. The algorithm employs preset tree-structure hyperparameters (such as maximum depth) and uses supervised learning to extract optimal splitting criteria from training datasets, thereby decomposing the p-dimensional feature space (assuming input samples are p-dimensional feature vectors) into several mutually exclusive geometric subspaces. Theoretical studies show that this model exhibits excellent classification performance for datasets with explicit hyperplane partitioning characteristics, but its classification effectiveness decreases significantly when handling nonlinearly separable data or classification problems with complex decision boundaries. However, empirical studies demonstrate that when processing tabular data, ensemble learning algorithms derived from deci-

sion tree architectures show superior prediction accuracy compared to neural network models. For example, the XGBoost algorithm, an optimized gradient boosting framework, achieves significant performance breakthroughs in multiple benchmark tests through iterative decision tree integration and regularization strategies [**?**].

1.6 Materials Databases

Constructing robust predictive models in materials science machine learning research heavily depends on support from large-scale, high-quality datasets [**?**]. Current mainstream materials informatics platforms include Open Catalyst [**?**], Materials Project [**?**], and the Open Quantum Materials Database (OQMD) [**?**]. This study selected Materials Project [**?**] as the data source based on the following considerations: the database integrates structural information for over 160,000 materials and provides XAFS spectral data calculated based on FEFF theory, which is valuable for establishing structure-property relationships. Additionally, its open Python Application Programming Interface (API) significantly improves data acquisition efficiency and enables seamless integration with high-throughput computational workflows. Data acquisition was performed through the pymatgen [**?**] materials analysis toolkit, with all processed structured data stored in MongoDB document databases for efficient retrieval.

2 Dataset and Preprocessing

The data for this study was sourced from the Materials Project database [**?**]. We systematically retrieved materials containing fourth-period transition metals and extracted EXAFS calculation data. All EXAFS data were generated through the FEFF computational package. To ensure data quality, we excluded samples where different absorbing atoms within the same material exhibited different coordination numbers, as such data could interfere with coordination number classification.

For periodic EXAFS data, we employed Fourier transform for preprocessing. For large-scale data processing, we used the Larch toolkit [**?**] combined with Python programming to achieve automated processing. Specifically, the autobk function from Larch was utilized to convert raw E-space data to k-space data, followed by the xftf function to Fourier transform k^2-weighted k-space data into R-space data.

Regarding R-space data processing, we extracted intensity values at intervals of 0.030 Å as features. Additionally, we used Python' s Peak package to calculate R-coordinates and intensity values corresponding to each characteristic peak, incorporating these parameters as additional features into the dataset.

3 Model Construction and Training

To accomplish coordination number prediction, this study constructed a prediction model based on fully connected neural networks. As shown in Figure 2 [Figure 2: see original paper], the neural network' s input layer contains 326 features corresponding to intensity values from EXAFS R-space spectra at intervals of 0.030-0.031 Å. The network employs a dual hidden-layer architecture with 512 neurons in each layer. Each hidden layer is followed sequentially by ReLU activation layers and Dropout regularization layers: ReLU activation functions enhance the model' s nonlinear expressive capability, while Dropout regularization mitigates overfitting. The output layer parameter k varies according to the coordination number range for different elements. During model optimization, the model with optimal validation set performance was selected as the final model through five-fold cross-validation, effectively improving model generalizability.

To enhance model interpretability and validate neural network predictions, this study simultaneously constructed ensemble learning models based on decision trees. We first implemented two ensemble models using the Scikit-learn [**?**] machine learning framework: decision tree ensembles based on Bagging and Random Forest. Bagging generates sub-training sets equal to the data volume through Bootstrap Sampling, trains independent decision tree models on each subset, and integrates predictions through voting mechanisms. Random Forest introduces feature randomness on top of Bagging—during each node split, only sqrt(d) features (where d is the total feature count) are randomly selected for optimal partitioning, thereby enhancing model diversity. Both models use the Gini Index as the splitting criterion. As shown in Figure 3 [Figure 3: see original paper], each box in the diagram represents a split node. "Freq_r" in the boxes indicates the intensity corresponding to coordinate r in EXAFS R-space, while "samples" represents the data volume contained in that node before splitting.

This study systematically analyzed EXAFS spectra of fourth-period transition metal materials using neural networks and decision tree algorithms. As shown in Figure 4 [Figure 4: see original paper], for each metal, the models from left to right are: neural network, bagging model, random forest model, bagging model with peak height and position features, and random forest model with peak height and position features. Results indicate that both approaches achieved approximately 70% prediction accuracy, with neural networks showing slightly superior performance compared to decision tree models. Notably, the accuracies of both methods exhibit significant correlation, showing synchronous increase and decrease trends with element type. In ensemble learning applications, random forest models improved prediction accuracy by approximately 1-3 percentage points compared to bagging models. Additionally, we examined feature datasets incorporating peak position and height information, finding no significant difference in decision tree model prediction accuracy compared to using the original dataset.

Research results demonstrate that model prediction accuracy exhibits significant element dependence in coordination number classification. Vanadium (V) showed the best predictive performance in the neural network framework, achieving 81.74% accuracy, while cobalt (Co) exhibited relatively lower accuracy at only 67.39%. Further analysis revealed that this difference may relate to coordination number distribution characteristics in the database. Specifically, V maintains a relatively stable six-coordinate structure in most materials, whereas Co shows more diverse coordination number variations, which may be the primary reason for its lower prediction accuracy.

This study further constructed feature importance distribution maps for each element' s EXAFS spectra based on Gini indices from decision tree ensemble models. Figure 5 [Figure 5: see original paper] illustrates feature importance and its correspondence to R-space for Co and Mn elements in bagging and random forest models. Subfigure (a) shows bagging Gini index for Co; (b) shows random forest Gini index for Co; (c) shows bagging Gini index for Mn; (d) shows random forest Gini index for Mn. Since the R-space dataset contains 326 equally spaced coordinate points, maximum single feature importance values are generally below 0.1. Based on EXAFS spectral processing experience, we only discuss regions with $R < 6.0$ Å. Comparative analysis of relative importance distributions reveals that feature importance weights concentrate primarily in the short-range interaction region with $R < 3.0$ Å, consistent with empirical understanding in EXAFS theory that short-range atomic interactions dominate spectral features.

Additionally, this study observed that Bagging significantly amplifies importance differences among features in feature importance distribution maps, while random forest models show more balanced feature importance distributions. This phenomenon may relate to bagging models' tendency to overfit. Specifically, bagging removes only one feature per iteration for optimization, causing features with higher importance to be further amplified through multiple iterations. In contrast, the random forest model used in this study simultaneously screened 14 features per iteration for optimization, thereby avoiding excessive enhancement of single feature importance and effectively improving model generalizability. This finding indicates that random forest models can better balance feature importance and reduce overfitting risks when handling high-dimensional data, providing more robust solutions for EXAFS spectral analysis.

This study successfully developed and validated a machine learning approach for predicting coordination numbers of fourth-period transition metal elements from EXAFS spectra. By collecting large-scale EXAFS data from the Materials Project database and combining neural networks with decision tree algorithms, our models achieved approximately 70% average prediction accuracy in coordination number classification. Notably, neural network models performed exceptionally well for certain elements (such as vanadium) with accuracy up to 81.74%, demonstrating their powerful potential for processing complex spectral data.

Results show that random forest models not only match neural networks in prediction performance but also reveal critical information in EXAFS spectra through feature importance analysis. Specifically, intensity information from data points at R < 3.0 Å in Fourier-transformed R-space is crucial for coordination number prediction. This finding aligns with the EXAFS theory perspective that short-range atomic interactions dominate spectral features, providing theoretical support and interpretability for model decision-making processes.

However, the study also found significant element dependence in model prediction accuracy. For example, cobalt exhibited lower accuracy (67.39%), possibly due to diverse coordination number distributions in the database. Additionally, feature importance analysis revealed that peak position and intensity parameters for certain elements (such as cobalt) were less important than expected, suggesting potentially underutilized spectral features. This phenomenon indicates room for improvement in current feature extraction methods.

Future research directions should include optimizing feature engineering, exploring more sophisticated spectral feature extraction techniques, and experimenting with other advanced machine learning algorithms to improve prediction accuracy and generalizability across different elements. Additionally, hybrid approaches combining physical models with machine learning warrant further investigation to enhance the efficiency and reliability of EXAFS spectral analysis. The results of this study provide new tools and insights for rapid analysis of local material structures, holding significant scientific importance and application value.

References

[1] M. Newville, "Fundamentals of XAFS," *Reviews in Mineralogy and Geochemistry*, vol. 78, no. 1, pp. 33–74, Jan. 2014, doi: 10.2138/rmg.2014.78.2.

[2] J. Terry, "Analysis of extended X-ray absorption fine structure (EXAFS) data using artificial intelligence techniques," *Applied Surface Science*, 2021.

[3] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: with Applications to Python*. in Springer Texts in Statistics. Cham: Springer International Publishing, 2023. doi: 10.1007/978-3-031-38747-0.

[4] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Feb. 1970, Accessed: Mar. 03, 2025. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634

[5] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] R. Ouyang, "SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates," *Phys. Rev. Mater.*, vol. 2, no. 8, 2018, doi: 10.1103/PhysRevMaterials.2.083802.

[7] J. R. Quinlan, "Induction of decision trees," *Mach Learn*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[8] R. Maclin and D. Opitz, "Popular Ensemble Methods: An Empirical Study," *jair*, vol. 11, pp. 169–198, Aug. 1999, doi: 10.1613/jair.614.

[9] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[10] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[11] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[12] H. K. D. H. Bhadeshia, "Neural Networks in Materials Science.," *ISIJ International*, vol. 39, no. 10, pp. 966–979, 1999, doi: 10.2355/isijinternational.39.966.

[13] J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans, and A. I. Frenkel, "Neural Network Approach for Characterizing Structural Transformations by X-Ray Absorption Fine Structure Spectroscopy," *Phys. Rev. Lett.*, vol. 120, no. 22, p. 225502, May 2018, doi: 10.1103/PhysRevLett.120.225502.

[14] S. Tetef, N. Govind, and G. T. Seidler, "Unsupervised machine learning for unbiased chemical classification in X-ray absorption spectroscopy and X-ray emission spectroscopy," *Phys. Chem. Chem. Phys.*, pp. 23586–23601, 2021, doi: 10.1039/D1CP02903G.

[15] A. A. Guda et al., "Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms," *npj Comput Mater*, vol. 7, no. 1, pp. 1–13, Dec. 2021, doi: 10.1038/s41524-021-00664-9.

[16] S. B. Torrisi et al., "Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships," *npj Comput Mater*, vol. 6, no. 1, p. 109, Jul. 2020, doi: 10.1038/s41524-020-00376-6.

[17] J. Liu, "Predicting the binding configuration and release potential of heavy metals on iron (oxyhydr)oxides: A machine learning study on EXAFS," *Journal of Hazardous Materials*, 2024.

[18] D. E. Sayers, E. A. Stern, and F. W. Lytle, "New Technique for Investigating Noncrystalline Structures: Fourier Analysis of the Extended X-Ray—Absorption Fine Structure," *Phys. Rev. Lett.*, vol. 27, no. 18, pp. 1204–1207, Nov. 1971, doi: 10.1103/PhysRevLett.27.1204.

[19] Z. Sun, Q. Liu, T. Yao, W. Yan, and S. Wei, "X-ray absorption fine structure spectroscopy in nanomaterials," *Sci. China Mater.*, vol. 58, no. 4, pp. 313–341, Apr. 2015, doi: 10.1007/s40843-015-0043-4.

[20] J. J. Rehr, J. J. Kas, F. D. Vila, M. P. Prange, and K. Jorissen, "Parameter-free calculations of X-ray spectra with FEFF9," *Phys. Chem. Chem. Phys.*, vol. 12, no. 21, p. 5503, 2010, doi: 10.1039/b926434e.

[21] Y. Joly, "X-ray absorption near-edge structure calculations beyond the muffin-tin approximation," *Phys. Rev. B*, vol. 63, no. 12, p. 125120, Mar. 2001, doi: 10.1103/PhysRevB.63.125120.

[22] M. Newville, "Larch: An Analysis Package for XAFS and Related Spectroscopies," *J. Phys.: Conf. Ser.*, vol. 430, p. 012007, Apr. 2013, doi: 10.1088/1742-6596/430/1/012007.

[23] H. Pan et al., "Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures," *Inorg. Chem.*, vol. 60, no. 3, pp. 1590–1603, Feb. 2021, doi: 10.1021/acs.inorgchem.0c02996.

[24] G. O. Brunner, "A definition of coordination and its relevance in the structure types AlB2 and NiAs," *Acta Cryst A*, pp. 226–227, 1977, doi: 10.1107/S0567739477000461.

[25] M. O' Keefe and N. E. Brese, "Atom sizes and bond lengths in molecules and crystals," *J. Am. Chem. Soc.*, vol. 113, no. 9, pp. 3226–3229, Apr. 1991, doi: 10.1021/ja00009a002.

[26] M. O' Keeffe, "A proposed rigorous definition of coordination number," *Acta Crystallogr A Cryst Phys Diffr Theor Gen Crystallogr*, vol. 35, no. 5, pp. 772–775, Sep. 1979, doi: 10.1107/S0567739479001765.

[27] G. Voronoi, "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites.," *Journal für die reine und angewandte Mathematik*, vol. 133, pp. 97–178, 1908.

[28] N. E. R. Zimmermann and A. Jain, "Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity," *RSC Adv.*, vol. 10, no. 10, pp. 6063–6081, 2020, doi: 10.1039/C9RA07755C.

[29] S. P. Ong et al., "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis," *Computational Materials Science*, vol. 68, pp. 314–319, Feb. 2013, doi: 10.1016/j.commatsci.2012.10.028.

[30] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU),"Feb. 07, 2019, arXiv: arXiv:1803.08375. doi: 10.48550/arXiv.1803.08375.

[31] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[32] K. L. K. Lee, C. Gonzales, M. Nassar, M. Spellings, M. Galkin, and S. Miret,

"MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling," Sep. 12, 2023, arXiv: arXiv:2309.05934. doi: 10.48550/arXiv.2309.05934.

[33] L. Chanussot et al., "Open Catalyst 2020 (OC20) Dataset and Community Challenges," *ACS Catal.*, vol. 11, no. 10, pp. 6059–6072, May 2021, doi: 10.1021/acscatal.0c04525.

[34] A. Jain et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, Jul. 2013, doi: 10.1063/1.4812323.

[35] S. Kirklin et al., "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies," *npj Comput Mater*, vol. 1, no. 1, p. 15010, Dec. 2015, doi: 10.1038/npjcompumats.2015.10.

[36] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Jun. 05, 2018, arXiv: arXiv:1201.0490. doi: 10.48550/arXiv.1201.0490.

Note: Figure translations are in progress. See original paper for figures.