Title

Machine Learning of Big Materials Data

Author

Zheng, Chen

Publication Date

2019

UNIVERSITY OF CALIFORNIA SAN DIEGO

Machine Learning of Big Materials Data

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

NanoEngineering

by

Chen Zheng

Committee in charge:

Professor Shyue Ping Ong, Chair
Professor Zheng Chen
Professor Gary Cottrell
Professor Oleg Shpyrko
Professor Kesong Yang

2019

The Dissertation of Chen Zheng is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To

my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to my advisor, Dr. Shyue Ping Ong. The last five year is an unforgettable journey. He is my role model because of his professionalism in everything. His enormous support and guidance throughout my graduate studies go far beyond my research and have a substantial impact on my future career.

I would also like to take this opportunity and thank my committee Dr. Zheng Chen, Dr. Gary Cottrell, Dr. Oleg Shpyrko, and Dr. Kesong Yang for their assistance and support in the completion of this dissertation.

My special appreciation also goes to my colleagues in MAVRL group. My special thanks to Dr. Chi Chen. It is such a pleasant and fruitful collaboration with him and the other collaborators. Without Dr. Chi Chen's help, my graduate life will definitely be more challenge.

The last part is reserved for my parents' endless love and support. And the one who was there when I was in the dark.

Chapter 2, in full, is a reprint of the material "Effects of Transition-Metal Mixing on Na Ordering and Kinetics in Layered *P*2 Oxides" as it appears in Physical Review Applied, Chen Zheng, Balachandran Radhakrishnan, Iek-Heng Chu, Zhenbin Wang and Shyue Ping Ong, 2017, 7 (6). pp 064003. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material "Deep learning driven study of high entropy cathode $\mathrm{Na}_x\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$", Chi Chen, Chen Zheng, and Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material "High-throughput computational X-ray absorption spectroscopy" as it appears in Scientific Data, Kiran Mathew, Chen Zheng, Donald Winston, Chi Chen, Alan Dozier, John J. Rehr, Shyue Ping Ong, and Kristin A. Persson, 2018, 5, pp 180151. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material "Automated generation and ensemble-learned matching of X-ray absorption spectra" as it appears in *npj* Computational Materials, Chen Zheng, Kiran Mathew, Chi Chen, Yiming Chen, Hanmei Tang, Alan Dozier, Joshua J. Kas, Fernando D. Vila, John J. Rehr, Louis F.J. Piper, Kristin A. Persson, and Shyue Ping Ong, 2018,
4 (12). The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is currently being prepared for submission for publication of the material "Accurate Chemical Environment Classification from X-ray Absorption Near-Edge Structure using a Random Forest Model", Chen Zheng, Chi Chen and Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

VITA

2010	B.E. in Materials Science and Engineering, Beijing University of Technology, China
2012	M.S.E in Materials Science and Engineering, University of Pennsylvania, USA
2018	M.S in Electrical Engineering, University of California San Diego, USA
2019	Ph.D. in NanoEngineering, University of California San Diego, USA

PUBLICATIONS

C. Zheng, C. Chen and S. P. Ong, “Accurate Chemical Environment Classification from X-ray Absorption Near-Edge Structure using a Random Forest Model”, *In preparation.*

C. Zheng, B. Radhakrishnan, I.-H. Chu, Z. Wang and S. P. Ong, “Effects of Transition-Metal Mixing on Na Ordering and Kinetics in Layered *P*2 Oxides”, *Phys. Rev. Appl.* **2017**, *7* (6), 064003.

C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson and S. P. Ong, “Automated Generation and Ensemble-Learned Matching of X-Ray Absorption Spectra”, *npj Comput. Mater.* **2018**, *4* (1), 12.

K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong and K. A. Persson, “High-Throughput Computational X-Ray Absorption Spectroscopy”, *Sci. Data* **2018**, *5*, 180151.

ABSTRACT OF THE DISSERTATION

Machine Learning of Big Materials Data

by

Chen Zheng

Doctor of Philosophy in NanoEngineering

University of California San Diego, 2019

Professor Shyue Ping Ong, Chair

In the past decades, theoretical calculations of materials properties have become more accurate and accessible due to the successful development of *ab initio* codes, as well as advances in computational power. With the booming development of high-throughput computational materials repositories, opportunities have emerged in the area of data-driven discovery of new materials guided by machine learning. However, the interpretation of large materials data sets needs to be performed from an integrated perspective of statistics and materials science intuition. In this thesis, we will address this challenge by demonstrating how the integration of high-throughput software workflows, automated data generation, and machine learning can yield powerful new approaches to materials analysis and optimization. This thesis is broadly divided into two topics.

In the first topic (Chapters 2 and 3), we present comprehensive first-principle investigations of the effect of transition metal mixing on layered *P*2 oxides, using *P*2 $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ and *P*2 $\mathrm{Na}_x\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$ as model systems. Our results show that transition metal mixing significantly suppresses the formation of strongly ordered intermediates. Using *ab initio* molecular dynamics simulations and the climbing image nudged elastic band method, we reveal that transition metal substitution has a pronounced effect on the Na site occupancy energy and Na diffusion energy barriers. By employing a site percolation model, we derive theoretical upper and lower bounds on the concentration of transition metal species in the layered *P*2 oxides based on their effects on Na diffusion energy barriers. Another key innovation is the use of the MatErials Graph Network (MEGNet) model, a graph-based deep learning approach recently developed in our group, on layered *P*2 oxides for accurate energy prediction, which we will apply to study mixing energies in a "high-entropy" *P*2 $\mathrm{Na}_x\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$.

In the second topic (Chapters 4, 5, and 6), we present the development of a first-of-its-kind computational reference XAS database (XASdb). More importantly, we have developed a novel Ensemble-Learned Spectra IdEntification (ELSIE) algorithm that leverages on ensemble learning techniques to match an unknown target K-edge XANES spectra with computed spectra in XASdb. We will also discuss the development of general machine learning approaches to rapidly and efficiently identify the coordination environment of absorbing atoms from K-edge XANES.

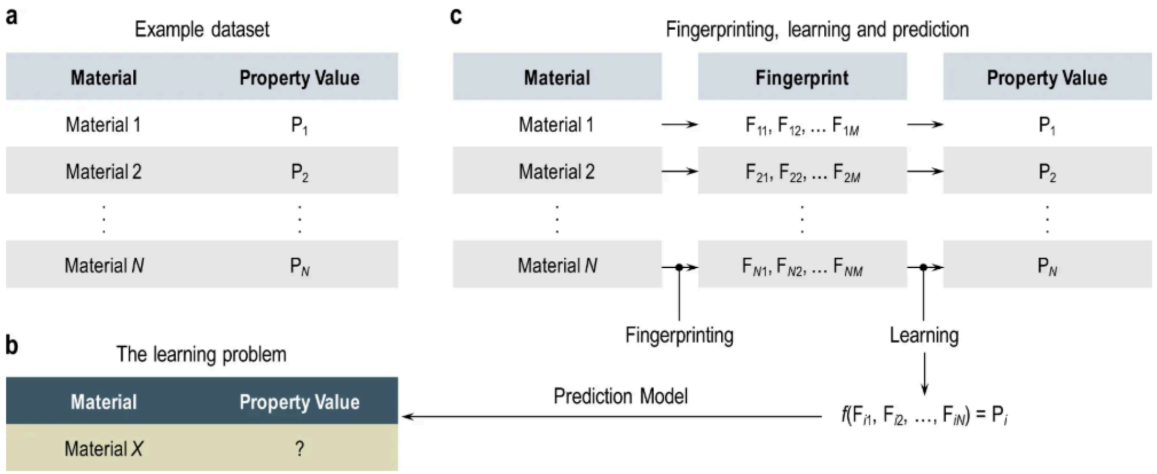
Chapter 1 Introduction

1.1 Background

Materials innovations are vital to the advancement of new technologies. Admittedly, materials development is complex, making it difficult to achieve breakthroughs and commercialize over a short period of time.[1] Many prospective materials technologies take approximately 20 years or even longer before being commercialized.[2] This tardiness is partly due to the experimental input-centric procedure followed by researchers in the development of new materials. Materials synthesis and characterization are laborious tasks and cannot be routinely conducted in a high-throughput manner. At present, there is no unified protocol applicable for experimentally screening prospective materials across different application domains. The as-developed high-throughput synthesis and characterization tools[3–6] are generally chemical system-specific. For example, in the case of hydrothermal and solvothermal synthesis, it can often be exceptionally difficult to achieve accurate control of reaction conditions at different scales.[7]

In the recent decade, high-throughput computations have emerged as a complementary approach to experimental approaches for materials discovery. Maturing theoretical tools and the advent of inexpensive computational resources offer researchers a more cost-effective solution to the materials design problem. Electronic structure codes, especially those based on Kohn-Sham density functional theory (DFT),[8,9] can be reliably applied to predict and assess a set of materials properties.[10] Further, the advancement of computational workflow management software[11], materials analysis packages[12], just-in-time job management tools, and *ab initio* computational code has enabled a growing trend in data-driven materials design and discovery.[13,14] By leveraging the growing and scalable computational power, large databases[20] containing electronic and thermodynamic properties of existing and hypothetical materials are now being constructed at unprecedented rates. For example, the Materials Project[17] database, a core part of the materials genome initiative, contains computed properties of more than 130,000 inorganic compounds at the time of writing, including elasticity[18], surface energy[19], and band structures[17]. There are also many other publicly available materials databases spanning a wide range of applications.[15,16]

The explosion in the quantity of computed materials data has shifted the paradigm in materials discovery and is playing a central role in the study of materials properties. However, it is equally true that large data sets bring their own set of unique challenges. Contemporary artificial-intelligence methods become promising approaches for studying the properties of the materials and have the potential to revolutionize the materials discovery paradigm.


a
Example dataset
c
Fingerprinting, learning and prediction
Fingerprinting
Learning
b
The learning problem
Prediction Model
$f(F_{i1}, F_{i2}, \ldots, F_{iN}) = P_i$

Material	Property Value
Material 1	P_1
Material 2	P_2
⋮	⋮
Material *N*	P_N

Material	Fingerprint	Property Value
Material 1	$F_{11}, F_{12}, \ldots F_{1M}$	P_1
Material 2	$F_{21}, F_{22}, \ldots F_{2M}$	P_2
⋮	⋮	⋮
Material *N*	$F_{N1}, F_{N2}, \ldots F_{NM}$	P_N

Material	Property Value
Material *X*	?

Figure 1.1: The key elements of machine learning in materials science. **a** Schematic view of an example data set, **b** statement of the learning problem, and **c** creation of a surrogate prediction model via the fingerprinting and learning steps. N and M are, respectively, the number of training examples and the number of fingerprint (or descriptor or feature) components.[20]

Machine learning algorithms can be divided into two broad categories: *supervised* and *unsupervised* learning. For supervised learning, the training datasets are well labeled and take the form of a collection of (x, y) pairs. The fundamental goal of supervised learning algorithm is to generalize beyond the training examples and be able to predict y^* in response to the new input sets x^* (unseen data). Unsupervised learning, on the other hand, typically involves the analysis of unlabeled training data and identification of hidden patterns in the input data.[21] In unsupervised learning, we generally have no access to the 'correct' output of the training dataset.

In the field of computational materials science, most research problems can be classified into supervised learning. Figure 1.1 shows a schematic view of the elements of machine learning within materials science. In the first part, a sufficiently large input dataset is first constructed, either from experiments or more commonly, using high-throughput first principles calculations. This dataset is essentially a mapping between a material description (e.g., composition, atomic arrangement of atoms, local environments) and properties such as formation energies, mechanical properties, X-ray absorption spectra, etc. The result of running the machine learning algorithms can be expressed as establishing a mapping between new yet-to-be-synthesized materials' attributes, such as compositional information or the electronic charge density distributions, and any or all of their properties. Subsequently, once the "hidden" rules that govern the materials properties have been discovered, the trained models can then predict the properties of a vast number of new materials at negligible additional computational cost, thereby by-passing the laborious and time-intensive computations.

The past few years have seen a rapid increase in the amount of research related to the application of machine learning approaches in the materials science field. For example, the Meredig *et al.*[22] have shown that the thermodynamic stability of ternary compounds can be predicted by leveraging a linear regression model trained on a database of thousands of first principles calculations. The Raccuglia *et al.*[7] have demonstrated that the support vector machine derived decision tree models can predict reaction outcomes with a success rate of 89%, and reveal the chemical principles governing reaction outcomes. Other materials properties such as bandgap energy[23], formation energy[24,25], elastic moduli[26] can also be determined with properly constructed machine learning models. Analyzing high-throughput computational datasets has now become a key component of materials property investigations, underpinning new waves of materials innovation.

In this thesis, we seek to address the challenges of high-throughput computational materials studies through machine learning approaches on two topics: (i) the stability of *P*2-type layered sodium transition-metal (TM) oxides and (ii) local environment determination from X-ray absorption spectroscopy (XAS). In each topic, its prominent material property assessment approaches and machine learning methods will be discussed.

1.2 *P*2-type layered sodium transition-metal oxides

1.2.1 Motivation and overview

Sodium ion secondary batteries (SIB) are considered to be promising candidates for large-scale applications due to the larger abundance (Clark number: 2.63) and lower cost of sodium compared to lithium (Clark number: 0.006). [27,28] In the last decade, new research lies in sodium ion battery has risen explosively.[29–31] Among a vast range of possible cathodic materials sodium ion batteries, layered sodium TM oxides $\mathrm{Na}_x\mathrm{MO}_2$ (*M* = Co, Ni, Mn, Fe, V, Cr, etc and their combinations) have been demonstrated to have among the most promising electrochemical performance.[27,32,33] Most common $\mathrm{Na}_x\mathrm{MO}_2$ compounds could be categorized into two polymorphs, *P*2 and *O*3, according to the classification by Delmas *et al*.[34] *P*(prismatic) and *O*(octahedral) denote the alkali atom occupation environments between $M\mathrm{O}_6$ octahedra form stacking sheets. 2 and 3 describe the repetition number of $M\mathrm{O}_6$ layers. *P*2 structure materials, in general, outperforms *O*3 structure materials in terms of higher reversible capacity and better cyclability.[31]

While the benefits of $P2$-$\mathrm{Na}_x\mathrm{MO}_2$ are well established, potential drops related to the formation of biphasic states in single TM layered sodium oxides $P2$-$\mathrm{Na}_x\mathrm{CoO}_2$[35] and $P2$-$\mathrm{Na}_x\mathrm{MnO}_2$[36] make it difficult to put them into practical use. Thus, to eliminate the potential drops caused by the formation of dominant Na orderings, various first and second row transition elements have been introduced as substitutes of Co. The TM mixing strategy has proved to be effective in suppressing the formation of biphasic states during the charging and discharging process, as evidenced by the solid-solution behavior sloping electrochemical curves and superior electrochemical performance of various mixed-TM *P*2 compounds.[30,32,37,38] However, the conventional way of discovering new *P*2 materials still involves a significant number of trail-and-error attempts, making it difficult to predict and identify prospective compounds. Taking the twenty first and second row transition elements into consideration, the composition space of mixed-TM *P*2 compounds is enormous as the ratios of TM elements are adjustable as well.

1.2.2 Challenges and opportunities of machine learning in layered *P*2 oxides research

At the same time, studies of mixed-TM layered *P*2 oxides from computational approaches are rarely found. Current computational studies of $P2$-$\mathrm{Na}_x\mathrm{MO}_2$ materials are restricted to single TM oxides systems.[39–41] The lack of computational studies prevents researchers from delving deeper into the mechanisms of how TM substitution can suppress the occurrence of phase transition and improve the electrochemical performance of single-TM *P*2 $\mathrm{Na}_x\mathrm{MO}_2$ materials. It is still unclear how TM substitution and mixing could affect Na diffusion kinetics of mixed-TM $P2$ $\mathrm{Na}_x\mathrm{MO}_2$ materials. Thus, *ab initio* study on mixed-TM $P2$ $\mathrm{Na}_x\mathrm{MO}_2$ with accurate control on stoichiometry is crucial.

One key limitation of previous computational studies is that a pure *ab initio* calculation strategy is limited in speed. Thus, the cluster expansion approach is usually adapted in combination with first-principles calculations to parametrize the energies of structures with different sodium concentrations and orderings. However, the number of candidate clusters to compute grows exponentially with the number of species, the supercell size, possible oxidation states, and spin states of elements.[42]

1.2.3 Approach to study of layered *P*2 oxides

Over the past few years, dramatic advances have been made in calculating materials properties quickly and accurately based on quantum-mechanical approaches. To accelerate the process of materials discovery, first principles computational workflows based on density functional theory (DFT) have been widely applied to the investigation and discovery of new materials. At the heart of DFT are the Hohenberg-Kohn (HK) theorem and Kohn-Sham (K-S) equation, which states that ground-state observables of any system of interacting particles are functionals of electron density $n(\mathbf{r})$.[43] In other words, the ground state energy of a system could be determined or approximated once the system's unique electron density in three spatial coordinates is obtained. The total energy from K-S equation is written as:

$$E_{KS}[n] = T_s[n] + \int V_{ext}(\mathbf{r})n(\mathbf{r})d^3\mathbf{r} + E_H[n] + E_{xc}[n], \tag{1.1}$$

where V_{ext} is the external potential acting on the electrons due to the nuclei, T_s is the independent-particle kinetic energy

$$T_s = \frac{1}{2}\sum_i \int |\nabla \emptyset_i|^2 d^3\mathbf{r}, \tag{1.2}$$

E_H is the Hartree (or Coulomb) energy of electrons

$$E_H = \int \frac{n(\mathbf{r})n(\boldsymbol{r}')}{|\mathbf{r} - \boldsymbol{r}'|} d^3\mathbf{r} d^3\boldsymbol{r}', \tag{1.3}$$

and E_{xc} is the exchange and correlation energy accounting for the interactions among electrons.

In the study of layered *P*2 oxides, the DFT-based energy calculation of different Na vacancy ordering structures is the key part. Through performing the energy calculation of different Na vacancy ordering structures, we could estimate the stability of Na ordering configurations with different Na concentrations. The pseudobinary 0-K stability diagram for each mixed-TM composition could then be constructed. The average Na intercalation potential (voltage profile) of cathode materials can be derived using the computed energies as well.

Though the Hohenberg-Kohn (HK) theorem and Kohn-Sham (K-S) equation have made the calculation of periodic systems such as crystalline solids' energy a practical and routine procedure, DFT simulations are generally restricted to the scale of hundreds of atoms. A single energy calculation of a structure takes hours to complete. For ternary or quaternary mixed-TM layered *P*2 oxides, the number of candidate structures to be computed is $> 10^9$ taking the possible arrangement of TMs and Na-vacancy orderings into consideration. Ternary or quaternary mixed-TM layered *P*2 oxides systems are thus cost-prohibitive and cannot be satisfactorily addressed with the traditional DFT approach due to their enormous configuration spaces.

It is therefore our view that this gap can be addressed via the integration of DFT calculations and the graph network machine learning (ML) approach (MatErials Graph Network). Graph neural networks are a new ML framework that operates on a graph and supports both relational reasoning and combinatorial generalization.[24] For crystals, graph-based representations are a natural choice. Essentially, the atoms form the nodes in the graph and the bonds form the edges. Given a graph G, a node $\mathbf{v}_i$ is characterized as an atomic attribute vector for atom i. Edge $E = \{(e_k, r_k, s_k)\}_{k=1:N}^{e}$ denotes the bond between atoms, where e_k is the bond attribute vector for bond k, r_k and s_k are the atom indices forming bond k, and N^e is the total number of bonds. To represent the crystal level or state attributes such as temperature of the system, the global state vector $\mathbf{u}$ is included and updated in the series of update operations. The architecture of the graph neural network models is provided in Figure 1.2.

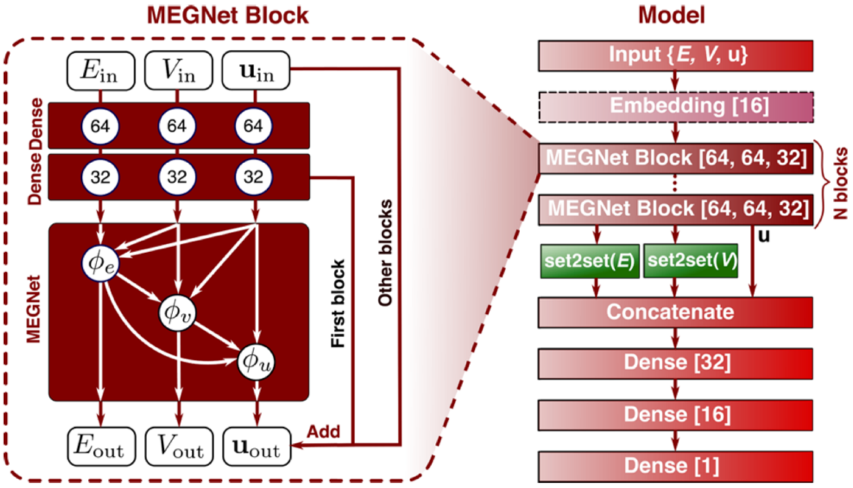


Figure 1.2: Architecture for the MatErials Graph Network (MEGNet) model. Each model is formed by stacking MEGNet blocks. The embedding layer is used when the atom attributes are only atomic numbers. In the readout stage, a set2set neural network is used to reduce sets of atomic and bond vectors into a single vector. The numbers in brackets are the number of hidden neural units for each layer. Each MEGNet block contains a MEGNet layer as well as two dense layers. The “add” arrows are skip connections to enable deep model training. Reprinted with Permission from Ref. [24]. Copyright 2019 American Chemical Society.

In this topic, we will use $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ and $\mathrm{Na}_{0.6}\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$ as model systems to demonstrate how investigation of Na ordering can be carried out using a combined first-principles computational and MEGNet ML approach. We will first conduct high-throughput DFT calculations of mixed-TM layered *P*2 oxides with different cell sizes and Na vacancy ordering patterns to generate high quality training and validation data. All the generated structures and their corresponding DFT results will then be applied to the construction of a predictive graph-based deep learning model. The optimized model will be extended to predict the energy from structure and search new ground state configurations. The two sub-projects under the layered *P*2 oxides topic are presented as follows.

Project 1: Detailed study of transition-metal mixing on Na ordering and kinetics in $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ using first principles calculations. In this project, we use DFT calculations to probe the fundamental relationships between transition-metal mixing, phase diagrams and Na diffusion kinetics in the *P*2 $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ model system. By employing *ab initio* molecular-dynamics simulation and nudged elastic-band calculations, we seek to identify the relative influence of TM composition on Na site energies and Na diffusion barrier. This fundamental investigation aims at providing a theoretical framework for optimization of mixed-TM compositions and filling the knowledge gap in layered *P*2 oxides.

Project 2: Deep learning driven study of $\mathrm{Na}_{0.6}\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$ cathode material In this work, we will present our research efforts to overcome the imperfections in the conventional first-principle computational approaches utilizing deep learning neural networks. We extend the universally generalizable, high-performance MatErials Graph network (MEGNet) model[24] to the high-entropy $\mathrm{Na}_{0.6}\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$ model system for fast and accurate calculation of different Na-vacancy ordering configurations' formation energies. This investigation is expected to provide a robust foundation for further integration of novel machine learning techniques into simulations of complex chemistry.

1.3 X-ray Absorption Spectroscopy

1.3.1 Motivation and overview

X-ray absorption corresponds to an intrinsically quantum mechanical phenomenon based on the X-ray photoelectric effect. When incident X-ray photons are absorbed by an atom, the core-level electron is removed from its quantum level. Due to the photoelectric effect, the absorption will not occur when the binding energy of the electron is less than the energy of the incident X-ray. In XAS, the absorption coefficient, $\mu(E)$ is measured as a function of X-ray energy E. Detailed descriptions of X-ray absorption theory and equation have been included in many excellent books and review papers.[44,45]

X-ray absorption spectroscopy (XAS) has been widely used in the investigation of the properties, physical states and the local environments of materials.[46–48] The X-ray absorption fine structure (XAFS) is typically divided into two regimes: X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS).[49] The XANES is a fingerprint of oxidation states and coordination chemistries of the absorbing atom. For example, K-edge XANES spectra have been widely used to probe the oxidation state changes of TMs during charging and discharging of mixed-TM layered $P2$ oxides.[38,50] Quantitative XANES interpretation is a challenging problem and is usually conducted in combination with principal component analysis or least-squares fitting.[51,52] The EXAFS, on the other hand, could be interpreted in a more quantitative and accurate way coupling with theoretically calculated XAFS spectra.[53]

1.3.2 Challenges and opportunities of machine learning in XAS

One of the main challenges of interpreting XANES and EXAFS lies in obtaining reference spectra to fit the unknown spectra; experimentally measuring XAFS spectroscopy is laborious and time-consuming,[54–56] requiring X-ray beams of finely tunable energy that are accessible only through synchrotron radiation facilities.[45] The measurement data are therefore sporadic.

Publicly available XAS databases have demonstrated to be valuable references for analysis. They are hosted across the world. However, existing XAS databases provide limited coverage of chemical space. To the authors' knowledge, the largest open reference database[54] for XAS is initiated in the 1990s, which contains a mere 271 experimental XAS spectra that cover 39 elements of the periodic table. For each element, only a few common compounds' spectra are available in the databases. Hence one main issue in XAS interpretation is the lack of available high-resolution reference spectra.

Another challenge of XAS spectra interpretation is that researchers rely on visual spectral comparisons for the interpretation of unknown spectra. Because of the steep learning curve, it takes years to acquire the spectral interpretation skills. Spectral interpretation knowledge sharing is limited across teams, organizations, and projects. As researchers are exposed to a small set of spectral data at a time, the dependence of spectral features on the absorbing species' coordination environments or chemical properties is usually evaluated with limited chemical compounds coverage.[57–60] Also, it explains why the search for generalizable relationships between spectral characteristics and the coordination environments of XAS absorbing species is so challenging.[61] Meanwhile, we observe a resurgence of interest in integrating machine learning with the quantitative and qualitative interpretation of XAS. For example, Timoshenko *et al.*[62] have recently shown that the coordination number of Pt atoms could be predicted by leveraging the neural network model trained on the K-edge XANES of Pt nanoparticles. The Lu group has demonstrated that the convolutional neural network can be used to extract the local coordination environment of $3d$ TM species from their K-edge XANES with high accuracy.[63] Nevertheless, these pioneering efforts are either validated on relatively small (~100) datasets or on datasets with limited chemical diversity.

1.2.3 Approach for calculations of XAS spectra

We have selected the latest version (v9) of FEFF for calculations of XAS spectra. The FEFF code is an implementation of the real-space Green's function (RSGF) approach. The FEFF-computed spectra have been shown to yield great agreement with experimentally measured spectra. The XAS spectra computations are relatively cost-effective and require minimum adjustable parameters. Schematically, the contribution to the X-ray absorption coefficient $\mu(\omega)$ at X-ray energy $\hbar\omega$ is proportional to the total absorption cross-section $\sigma(\omega)$. This cross-section could be computed based on Fermi's golden rule given by

$$\sigma(\omega) = 4\pi^2 \frac{\omega}{c} \sum_F |\langle 0|d|F\rangle|^2 \delta(\omega + E_0 - E_F), \tag{1.4}$$

where d represents the coupling to the X-ray field, E_0 is the ground-state energy, E_F is the excited-state energies.[64] The real-space Green's function (RSGF) approach is adopted to reduce the computational cost. In terms of the Green's function, $G(r, r'; E)$, the absorption coefficient, μ, from a given core level c is given by:

$$\mu = -\frac{1}{\pi} Im\langle c|\epsilon \cdot rG(r, r'; E)\epsilon \cdot r|c\rangle. \tag{1.5}$$

The spectral representation of the Green's function is an effective propagator in the presence of a core hole and multi-electron effects. The FEFF code computes the full propagator G incrementally using matrix factorization. This simplified and efficient *ab initio* approach makes FEFF possible to produce a wide variety of X-ray spectra.

1.2.4 Machine learning for XAS spectra interpretation

As XANES spectra provide precious information regarding the chemical environments of absorbing species, machine learning algorithms can be adapted to understand the quantitative relationship between XANES spectral features and the local chemical environment of absorbing species. In this topic, the computed spectral data is considered as a collection of the individual spectrum. Each spectrum is converted and presented as a vector of 200 intensity value, thus integrating seamlessly with off-the-shelf machine learning tools. The local environment features of absorbing species are preprocessed with the coordination environment assessment algorithm developed by Zimmermann *et al*.[65] This process converts a single absorption site's local chemical environment to a ranking label, which represents the mixed state of the site's local environment. These coordination environments ranking labels are then used as the target information for follow-up supervised classification.

We choose the five most commonly used classifiers includes random forest, k-Nearest Neighbor (kNN), multi-layer perceptron (MLP), support vector machine (SVM) and convolutional neural networks (CNNs) in the coordination environment classification problem. Brief summaries of each machine learning classifier are as follows.

The k-Nearest Neighbor classification is the most straightforward nonparametric decision rule. The classifier labels an unclassified observation by the majority label among its k-nearest neighbor in the training set. The distance between two points $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots y_n)$ in the training set is computed using distance metrics. Popular distance metrics for kNN classifiers are Euclidean distance ($D(X, Y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$), Manhattan distance ($D(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$), etc.

The random forest classifier is an ensemble classifier that uses the decision tree as base classifiers.[66] The decision trees are created by considering a random sample of m training predictors from the full set of p predictors through replacement. In other words, in building a random forest, at each split in the tree, the random forest algorithm is forced to consider only a subset of training samples. As each tree is independently produced based on a random subset of the predictors, decision trees in a random forest are weakly correlated. The random forest algorithm generates classification decisions by taking the average of the class assignment probabilities calculated by all tree.[67]

The SVM is a generalization of a simple and intuitive classifier called the maximal margin classifier.[68] SVM aims at solving the over-fitting problem as test observations are classified based on the adaptive margins. It can be seen as an extension of the support vector classifier by defining relevant kernel functions[69] in order to accommodate a non-linear boundary between the classes. The separating hyperplane of SVM is computed using a kernel function of the form $K(x_i, x_{i'})$. A kernel function is a function that quantifies the similarity of two observations. In SVM, non-linear kernels such as polynominal kernel or radial kernel are popular choices. The advantage of using a non-linear kernel rather than the standard linear kernel ($K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij}\, x_{i'j}$) is that the support vector classifier will have much more flexible decision boundaries. It can help to classify the non-linear data. In our topic, we adapt the radial kernel, which takes the form

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2\right). \tag{1.6}$$

MLP is one of the most common neural networks in use. Typically, an MLP network consists of multiple layers of connected neurons regarded as the processing units. By choosing an error function, the learning process of an MLP neural network is based on the minimization of the error function. In each iteration, the error at the output is fed backward through the network. The weights of neurons get updated adaptively. Training of the MLP neural network ends once the error reaches a convergence criterion.

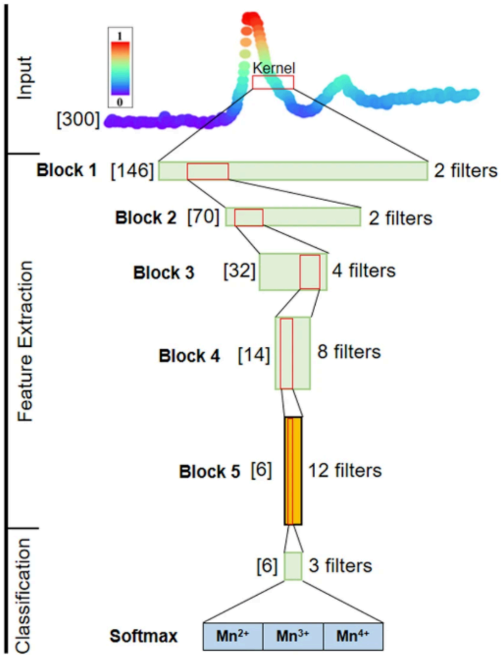


Figure 1.3: CNN spectrum classification pipeline. Reprinted with Permission from Ref. [70]. Copyright 2019 Nature Publishing Group

Deep CNNs were proposed by LeCun *et al.*[71] in the 1980s. The CNNs are feedforward networks consist of convolutional and pooling (or subsampling) layers. The CNNs require no manual feature engineering and can learn the representations from the image by only feeding the image itself into the model. Figure 1.3 illustrates the typical CNN architecture for a spectrum classification task. In the XANES spectra interpretation task, computed spectra could be considered as 1-D images. Using a 1-D convolutional layer as the first layer of CNNs, we can directly input a spectrum to the network, followed by several stages of convolution and pooling. The last fully connected layer outputs the spectrum's coordination environment class labels.

In this topic, we will capitalize on the recent advances in *ab initio* computations of X-ray spectra and machine learning techniques to address the needs of efficient XAS interpretation. The goal can be broken into three sub-projects.

Project 1: Developing a high-throughput framework to generate a reference XAS database (XASdb) spanning tens of thousands of materials. In this project, we will discuss the development of a high-throughput framework to generate a reference XAS database (XASdb) for all materials in the Materials Project[17] database. We select the latest version (9) of the popular FEFF program as our software of choice in this work. FEFF is a program for *ab initio* multiple scattering calculations of XAFS and various other spectra for clusters of atoms. The high-throughput framework combines the power of the Python Materials Genomics (pymatgen) materials analysis library[12] with the FireWorks workflow management software[11] to carry out hundreds of thousands of XAFS calculations using the FEFF9 code.[64]

Project 2: Implementing an automated XANES spectra matching algorithm capable of identifying similar XANES spectra from the computed reference XASdb. The second project involves the application of spectral matching and comparative algorithm on the computed reference XASdb. In this project, we will present the development of a novel automated XANES spectra matching algorithm that leverages on ensemble learning techniques to identify similar XANES spectra from our computed reference XASdb. We believe the combination of the XASdb with these machine-learned spectra matching tools will be an invaluable resource to the materials research community by significantly enhancing the efficiency at which experimental XAS spectra can be analyzed.

Project 3: Developing a novel spectral interpretation algorithm that allow for identification of local environments of absorbing atoms. The development of the largest computed reference XASdb also changes the approaches to the interpretation of XANES spectral data. In the classical analysis of XANES, researchers rely on comparisons between the experimentally measured sample spectrum and reference spectra of known compounds to estimate ratios of various standard compositions in a sample material. We believe the most useful way to view the theoretically computed XASdb is as an input dataset for machine learning based quantitative XANES characterization. In this project, we will discuss the state-of-the-art development of general machine learning approaches to identify the local structure motifs from XANES rapidly. Our work covers 33 elements in 259 distinct coordination environments. A broad repertoire of machine learning tools is systematically evaluated for XANES spectra characterization. We propose to use the random forest classifier on the local chemical environment characterization from the XANES. Finally, we demonstrate that the usage of feature importance measures opens the 'black box' implementation of machine learning techniques in XAS characterization and enhances the awareness of interpretability in materials informatics.

Chapter 2 Effects of Transition-Metal Mixing on Na Ordering and Kinetics in Layered *P*2 Oxides

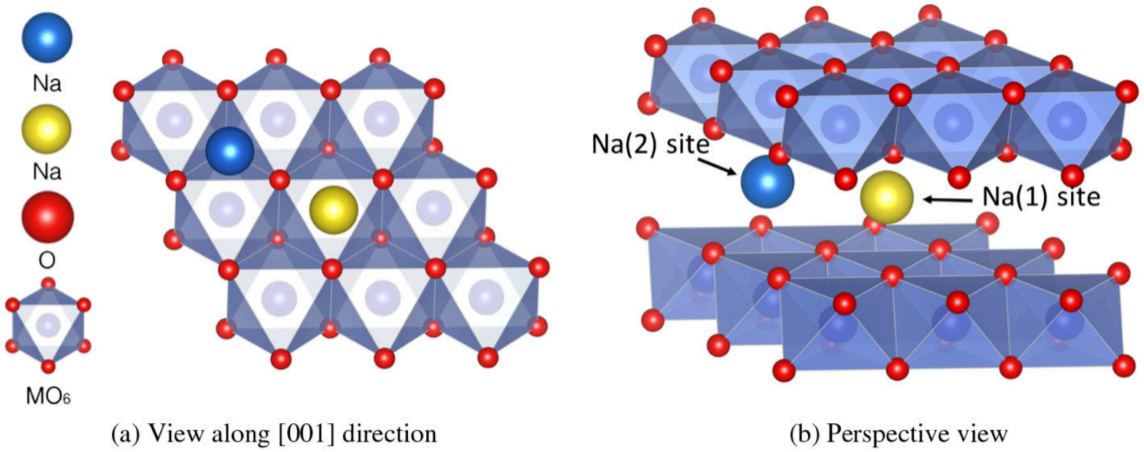
2.1 Introduction

Rechargeable sodium-ion batteries (SIBs) have recently emerged as promising candidates for large-scale energy-storage applications.[29–31] Sodium (2.3% of Earth's crust) is 3 orders of magnitude more abundant than lithium (0.0017%).[27,28] More importantly, sodium-ion battery chemistry enables new cell designs that can potentially yield significant advantages over lithium-ion chemistry. For instance, the possibility of using Al foil as an anode current collector instead of the more expensive Cu makes it probable that Na-ion batteries can be produced at less than half the cost of Li-ion ones.[72] Also, a great variety of sodium superionic conductors are known, [73–76] paving the way for the potential development of all-solid-state Na-ion batteries that may be safer with higher-energy density than traditional architectures based on organic liquid electrolytes.

One of the key challenges in SIBs is the development of cathodes with sufficiently high voltage and capacity. The most promising candidates are the layered sodium transition-metal (TM) oxides $\mathrm{Na}_x M\mathrm{O}_2$, which have been extensively investigated as cathodes in SIBs due to their excellent electrochemical performance.[32,33,77–86] Here, M can be either a single TM, e.g., Co, Ni, Mn, Fe, V, or Cr, or a mixture of these TMs, sometimes with other elements such as Li. Unlike the layered $\mathrm{Li}M\mathrm{O}_2$, which exists only in the *O*3 polymorph, $\mathrm{Na}_x M\mathrm{O}_2$ compounds exist in both the *P*2 and *O*3 stackings. In this stacking classification first proposed by Delmas, Fouassier, and Hagenmuller[34], *P* and *O* denote the environment occupied by the alkali atom (prismatic and octahedral, respectively) between the $M\mathrm{O}_6$ octahedral stacking sheets, while the numerals 2 and

3 refer to the number of $M\mathrm{O}_2$ layers per repeating unit. *P*2 $\mathrm{Na}_xM\mathrm{O}_2$ generally outperform their *O*3 analogs in terms of the reversible capacity and cyclability.[31]

The *P*2 $\mathrm{Na}_xM\mathrm{O}_2$ crystal structure is shown in Fig. 2.1. Na in *P*2 $\mathrm{Na}_xM\mathrm{O}_2$ can occupy two types of sites: The Na(1) site shares faces with two $M\mathrm{O}_6$ octahedra, and the Na(2) site shares only edges with $M\mathrm{O}_6$ octahedra. Adjacent Na(1) and Na(2) sites cannot be occupied simultaneously due to strong Coulombic repulsion.



(a) View along [001] direction

(b) Perspective view

Figure 2.1 Crystal structure of *P*2 $\mathrm{Na_xMO_2}$with ABBA-type layer stacking. Na occupies two distinct prismatic sites: The Na(1) site (yellow) shares faces with two $M\mathrm{O}_6$ octahedra, and the Na(2) site (blue) shares only edges.

For the single-TM *P*2 $\mathrm{Na}_x\mathrm{CoO}_2$ and *P*2 $\mathrm{Na}_x\mathrm{MnO}_2$,[35,36] potential drops related to the formation of biphasic states during the charge and discharge processes indicate the existence of dominant Na orderings across the entire Na insertion or deinsertion range. These orderings, especially for *P*2 $\mathrm{Na}_x\mathrm{CoO}_2$, have been extensively investigated using both experimental and computational approaches.[39,41,87–90] The presence of ordered intermediate phases plays a critical role in the electrochemical performance of *P*2 cathodes, as highly favored orderings may introduce kinetic limitations that may, in turn, limit the achievable capacity.

Various mixed-TM *P*2 $\mathrm{Na}_xM_yM'_{1-y}\mathrm{O}_2$ and $\mathrm{Na}_xM_yM'_{1-y}M''_{1-y-z}\mathrm{O}_2$ have been investigated[30,32,37,38,78,91,92] with the aim of eliminating phase transformations during the charge or discharge process and extending the stability of the *P*2 phase over a wider Na intercalation region. Studies of binary or ternary TM *P*2 compounds suggest that Mn mixing can suppress the occurrence of long-range sodium orderings.[31,32,92] Recent studies on a series of $P2$ $\mathrm{Na}_x M\mathrm{O}_2$ (*M* = Co, Mn) also suggest that a small amount of Co substitution with Mn in *P*2 $\mathrm{Na}_x\mathrm{CoO}_2$ reduces the formation of stable Na orderings at certain Na concentrations and results in the solid-solution-like behavior over a wide range of Na compositions.[48,93]

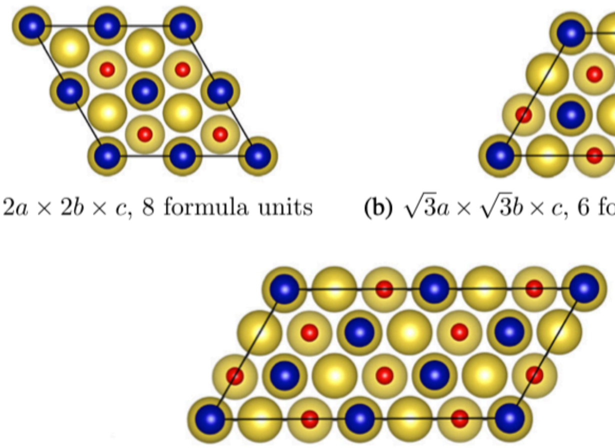
Previous computational studies have shown that the Na-ion diffusion in *O*3 layered $\mathrm{Na}_x M\mathrm{O}_2$ can be as facile as the Li analog,[29] even though the difference in the ionic radius between Na^+ (1.02 Å) and Li^+ (0.76 Å) is substantial.[94–96] Mo, Ong, and Ceder[96] have also demonstrated using *ab initio* molecular-dynamics (AIMD) simulations that *P*2 $\mathrm{Na}_x\mathrm{CoO}_2$ exhibit excellent Na conductivity over a wide range of Na concentrations. More recently, Guo and co-workers[97,98] showed that there is a strong correlation between the crystal structure and the Na diffusion in *P*-type $Na_{0.62}Ti_{0.37}Cr_{0.63}O_2$. In addition, the investigation of Na-ion conductivity in different com- pounds at the interphase layer of SIBs has been carried out using a combined experimental and theoretical approach.[99]

In this work, we present a density-functional-theory (DFT) study on the effects of transition-metal mixing on Na ordering and kinetics in layered *P*2 oxides using $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ as a model system. The choice of the *P*2 $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ system is motivated by the fact that this system has been well studied in experiments,[32,48,93] providing a wealth of data for comparison and validation. We demonstrate that Co-Mn mixing reduces the energetic differences between Na orderings and present a theoretical framework to tune mixed-TM compositions for optimal Na kinetics.

2.2 Methods

2.2.1 Structure enumeration

All symmetrically distinct Na orderings in $P2$ $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ for $y = 0, 1/3, 2/3, 1$ are enumerated using the algorithm of Hart and Forcade.[100] For the single-TM systems ($y = 0, 1$), enumerations are carried out at both $1/8$ and $1/6$ Na-concentration intervals, i.e., $x = 0, \frac{1}{8}, \frac{1}{6}, \frac{1}{4}, \frac{1}{3}, \frac{3}{8}, \frac{1}{2}, \frac{5}{8}, \frac{2}{3}, \frac{3}{4}, \frac{5}{6}, \frac{7}{8}, 1$. For the mixed-TM systems, enumerations are carried out only at $1/6$ Na-concentration intervals, i.e., $x = 0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}, 1$, due to the lower symmetry of these systems. For each y, we first determine the lowest-energy Na, Co, and Mn orderings at $x = 2/3$ within a $\sqrt{3}a \times \sqrt{3}b \times c$ supercell [Fig. 2.2(b)] to mimic initial synthesis Na concentrations of layered $P2$ cathodes. The supercell we use is comparable to those used in previous first-principles investigations of layered Na TM oxides.[88,101] The lowest-energy Co-Mn ordering at each y is then retained at all other Na concentrations x. We note that it has been well established experimentally that Co and Mn generally form a solid solution in this system,[93,102] and we discuss the implications of fixing the Co-Mn ordering in the results section.



(a) $2a \times 2b \times c$, 8 formula units

(b) $\sqrt{3}a \times \sqrt{3}b \times c$, 6 formula units

(c) $2\sqrt{3}a \times \sqrt{3}b \times c$, 12 formula units

Figure 2.2 Supercells of $P2$ $\mathrm{Na_xCo_{1-y}Mn_yO_2}$ used to enumerate Na and Co/Mn orderings, viewed along the [001] direction. Top view on the AB plane.

To keep the number of orderings manageable, we adopt the following constraints in performing the enumeration of Na orderings at different Na concentrations:

i. Three supercell sizes that comprise up to twelve formula units (f.u.), as shown in Fig. 2.2, are used.

ii. Each $P2$ cell comprises two Na layers. For the mixed-TM systems, the Na concentrations in both layers are constrained to be equal to limit the total number of orderings.

iii. Although Na can occupy both Na(1) and Na(2) sites, the simultaneous occupation of adjacent Na(1) and Na(2) sites is not allowed. Because of the large Coulombic repulsion between Na^+, these structures are likely to be of too high energy to be of any interest.

In total, structural optimization and total energy calculations of more than 5000 distinct structures at various Na concentrations and TM-mixing ratios are performed using an automated workflow implemented with the FIREWORKS scientific workflow software.[11] The lowest-energy structures of $P2$ $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ are then adopted in subsequent calculations.

2.2.2 Energy calculations

All DFT energy calculations are performed using the Vienna *ab initio* simulation package (VASP) within the projector-augmented-wave approach[101]. Spin-polarized calculations are performed using a k-point density of at least 1000/(number of atoms in the unit cell) and an energy cutoff of 520 eV. The exchange-correlation functional used is the Perdew-Burke-Ernzerhof (PBE)[103] generalized-gradient approximation and the Hubbard U extension to it (PBE + U)[104]. The spherically averaged scheme of the on-site Coulomb interactions is adopted[105,106], and the effective U values used for Mn and Co are 3.9 and 3.32 eV, respectively, similar to the values used in the Materials Project[17]. These effective U values are average values that have been well tested to reproduce the energies of redox energies involving Mn$^{3+/4+}$ and Co$^{3+/4+}$ in accordance with the approach proposed by Wang, Maxisch, and Ceder[105]. Co^{3+} and Co^{4+} are initialized in low spin, and Mn^{3+} and Mn^{4+} are initialized in high spin, which are found to yield the lowest energy in *P*2 $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$. All calculations are initialized in a ferromagnetic configuration[107], and the total magnetic moment of the unit cell is constrained to the expected value determined from the Na concentration and the consequent oxidation states of Co and Mn.

0-K stability diagram. —The pseudobinary 0-K stability diagram for each TM-mixing ratio y is constructed by plotting the formation energy of each ordering σ_i at Na concentration x with respect to the fully sodiated and desodiated end members, given by the following equation:

$$\Delta E_f^{\sigma_i} = E(\sigma_i) - xE\left(\mathrm{NaCo}_{1-y}\mathrm{Mn}_y\mathrm{O}_2\right) - (1-x)E(\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2) \qquad (2.1)$$

where $E(\sigma_i)$, $E\left(\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2\right)$, and $E(\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2)$ are the total DFT energies per f.u. of the ordering: σ_i , $\mathrm{NaCo}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$, and $\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$, respectively. The stable phases are then identified using the convex-hull construction[108]. Because no entropic effects (e.g., vibrational, configurational, etc.) are taken into account, these diagrams are by definition 0-K stability diagrams and not finite-temperature phase diagrams.

Intercalation potential.—The average intercalation potential V of the cathode between two stable Na ordered phases at Na concentrations x_1 and x_2 is calculated using the following expression[109]:

$$\mathrm{V} = -\frac{E\left(\mathrm{Na}_{x_2}\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2\right) - E\left(\mathrm{Na}_{x_1}\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2\right) - (x_2 - x_1)E(Na)}{(x_2 - x_1)e}, \qquad (2.2)$$

where E is the DFT total energy and e is the electronic charge.

2.2.3 Ab initio molecular-dynamics simulations

AIMD simulations are carried out in the constant volume (NVT) ensemble at 1000 K with a Nosé-Hoover thermostat[110,111]. The aim of these calculations is not to obtain converged statistics for an estimate of the diffusivity but rather to elucidate the Na site occupancies and diffusion mechanisms in $P2$ $\mathrm{Na_{1/2}Co_{1-y}Mn_yO_2}$, $0 \le y \le 1$, at nondilute Na concentrations. As such, these calculations are performed at a single, relatively high temperature (no melting is observed in our simulations), and simulations are carried out for a relatively short time of 80 ps. To reduce computational costs, all AIMD simulations are non-spin polarized, and a smaller plane-wave energy cutoff of 300 eV and Γ-centered $1 \times 1 \times 1$ k-point grid are employed. Supercells of $4 \times 4 \times 1$ (32 formula units) and $3\sqrt{3} \times 2\sqrt{3} \times 1$ (36 formula units) are used for the single- and mixed-TM systems, respectively. The time step of simulations is 2 fs. The initial models for the simulations are obtained by removing half of the Na atoms in each Na layer from fully sodiated structures to model compounds at Na concentration $x = 1/2$.

2.2.4 Climbing-image nudged elastic-band calculations

Na migration barriers are calculated using the climbing-image nudged elastic-band method (CINEB)[112,113]. The PBE generalized gradient approximation functional is adopted in the NEB calculations to exclude the impact of electron transfer on the diffusion barrier calculation[29]. Supercells of $4 \times 4 \times 1$ (32 formula units) and $2\sqrt{3} \times 2\sqrt{3} \times 1$ (24 formula units) are used for mixed- and single TM systems, respectively. A Γ-centered $2 \times 2 \times 2$ k-point grid is used, and each interpolated image is relaxed until the forces on each atom are less than 0.05 eV $\mathrm{Å}^{-1}$.

At a dilute Na concentration, to isolate the role of the transition metal on Na diffusion, different Na(1) site configurations are created in the lattice of $\mathrm{CoO_2}$ with Ni, Mn, and Fe as dopants. Na migration barriers from a Na(2) site to its nearest Na(2) site via different Na(1) sites are evaluated. We also adopt the $\mathrm{Co_{2/3}Mn_{1/3}O_2}$ framework to study the influence of the Ni dopant on the Na migration energy in the *P*2 ternary TM oxides.

All analyses are performed using the Python Materials Genomics (PYMATGEN) package[12].

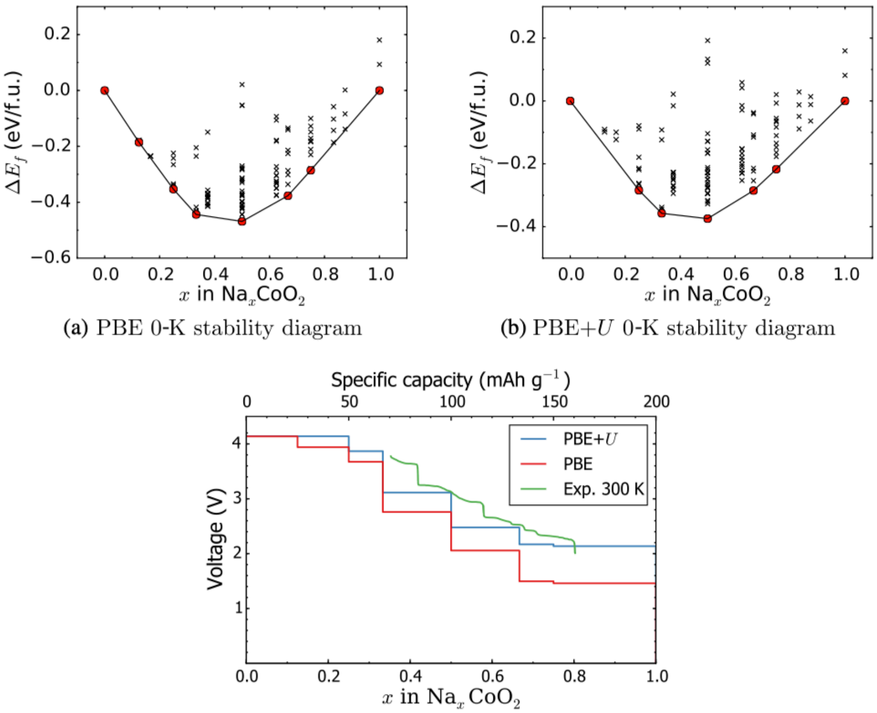
2.3 Results

2.3.1 0-K stability diagram and Na ordering

1. *NaCoO₂*

The stability diagram, Na ordering, and diffusion in *P*2 $\mathrm{Na_xCoO_2}$ have been extensively investigated through DFT calculations[39,41,88–90,96] and electrochemical characterization[35,114]. Figure 2.3 shows the computed 0-K stability diagrams of $\mathrm{Na_xCoO_2}$ using the PBE and PBE + U functionals. These 0-K stability diagrams are in good agreement with previous studies. Stable orderings of $\mathrm{Na_xCoO_2}$ are found at $x = \frac{1}{8}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$ with the PBE approximation, as shown in Fig. 2.3(a). Although our study does not identify any stable orderings at $x = 5/6$ and $x = 7/8$, the formation energy of the lowest-energy orderings at these concentrations are within 10 meV f.u.$^{-1}$ of the convex hull. The two stable Na orderings previously reported by Berthelot, Carlier, and Delmas[35] at these concentrations require larger supercells than those used in our high-throughput study. Nevertheless, these high-Na-concentration compositions are of less practical interest given that the layered *P*2 materials are typically synthesized at $x \leq 0.75$. For the PBE + U functional, stable Na orderings are found at the same Na concentrations as the PBE functional, with the exception of $x = 1/8$. Also, the lowest-energy orderings at $x = 5/6$ and $x = 7/8$ are significantly higher in energy above the convex hull (56 and 45 meV f.u.$^{-1}$, respectively), consistent with previous works[35,41].

Despite the better qualitative agreement of the stable Na orderings calculated without the Hubbard U parameter, especially at lower Na concentrations, we find that PBE + U reproduces absolute voltages much better than PBE [Fig. 2.3(c)]. This is due to the better self-interaction error cancellation for the redox reaction with the application of the Hubbard U.[115] The agreement between our DFT-predicted voltage profiles and experimental studies is on par with the previous first-principles investigation of the $P2$ systems.[41]



(a) PBE 0-K stability diagram

(b) PBE+U 0-K stability diagram

(c) Voltage profile

Figure 2.3 (a) PBE and (b) PBE + U 0-K stability diagrams of Na_xCoO_2. Black line, convex hull; red dots, stable orderings; black cross, unstable orderings. (c) Comparison of PBE and PBE + U voltage profiles with the experimental data from Ref.[35].

2. *NaMnO₂*

Figures A.1(a) and A.1(b) show the PBE and PBE + U 0-K stability diagrams of Na_xMnO_2. The stable Na orderings with both functionals are given in Figs. A.2 and A.3 in Appendix A. Although the Na concentrations at which stable orderings occur are the same in

PBE and PBE + U, we find considerable differences in the predicted stable Na orderings at $x = 1/2$. The PBE + U ground state for $\mathrm{Na_{1/2}MnO_2}$ has only Na(2) sites occupied, while the PBE ground state shows characteristic "row" motifs formed by alternating Na(1) and Na(2) lines, similar to $\mathrm{Na_{1/2}CoO_2}$. The difference in orderings is likely due to a greater energy penalty associated with the face-sharing Na(1) sites due to the Jahn-Teller distortion of $\mathrm{Mn^{3+}}$, which is better captured with the application of the Hubbard U.[116,117]

To our knowledge, the experimental stability diagram of *P*2 $\mathrm{Na_xMnO_2}$ has not been previously reported because *O*3 $\mathrm{Na_xMnO_2}$ is significantly more stable than the *P*2 structure at low temperatures[36,118,119]. Similar to $\mathrm{Na_xCoO_2}$, we find that the PBE + U voltages are in much better agreement with experimental voltages compared to PBE [see Fig. A.1(c)].

3. $\boldsymbol{Na_xCo_{1-y}Mn_yO_2, y = \frac{1}{3}, \frac{2}{3}}$

For $\mathrm{Na_xCo_{1-y}Mn_yO_2}$, we discuss mainly the PBE + U results. With the presence of a non-negligible amount of Mn, these systems are likely to exhibit strong 3d localization, for which the application of the Hubbard U is appropriate. The PBE (no U) results are given in Figs. A.4 and A.5 in Appendix A for interested readers. Figure A.6 shows the lowest-energy $\mathrm{Na_xCo_{1-y}Mn_yO_2}$ structures at $x = 2/3$ and $y = \frac{1}{3}, \frac{2}{3}$. We find that the lowest-energy structures in both instances exhibit a hexagonal ordering pattern similar to that observed in other mixed-TM layered *P*2 systems. It is well established in the experimental literature [14,38,47] that the mixed Co-Mn system tends to exhibit a solid-solution behavior; i.e., no superstructure ordering is observed for Co and Mn. Indeed, we find that the Co-Mn ordering has a small effect on relative energies, regardless of the specific Na ordering (see Tables A.1 and A.2 in Appendix A).

From Figs. 2.4(a) and 2.4(b), we make the observation that the mixed-TM $\mathrm{Na_xCo_{1-y}Mn_yO_2}$ are characterized by the presence of many metastable orderings whose energies are within 30 meV f.u.$^{-1}$ of the convex hull. This is unlike the single-TM $\mathrm{Na}_x M\mathrm{O}_2$ [Figs. 2.3(b) and A.1(b)], which exhibit distinct stable Na orderings that are substantially lower in energy compared to other orderings; i.e., there is a large energy gap between the ground-state ordering and the next lowest-energy ordering.

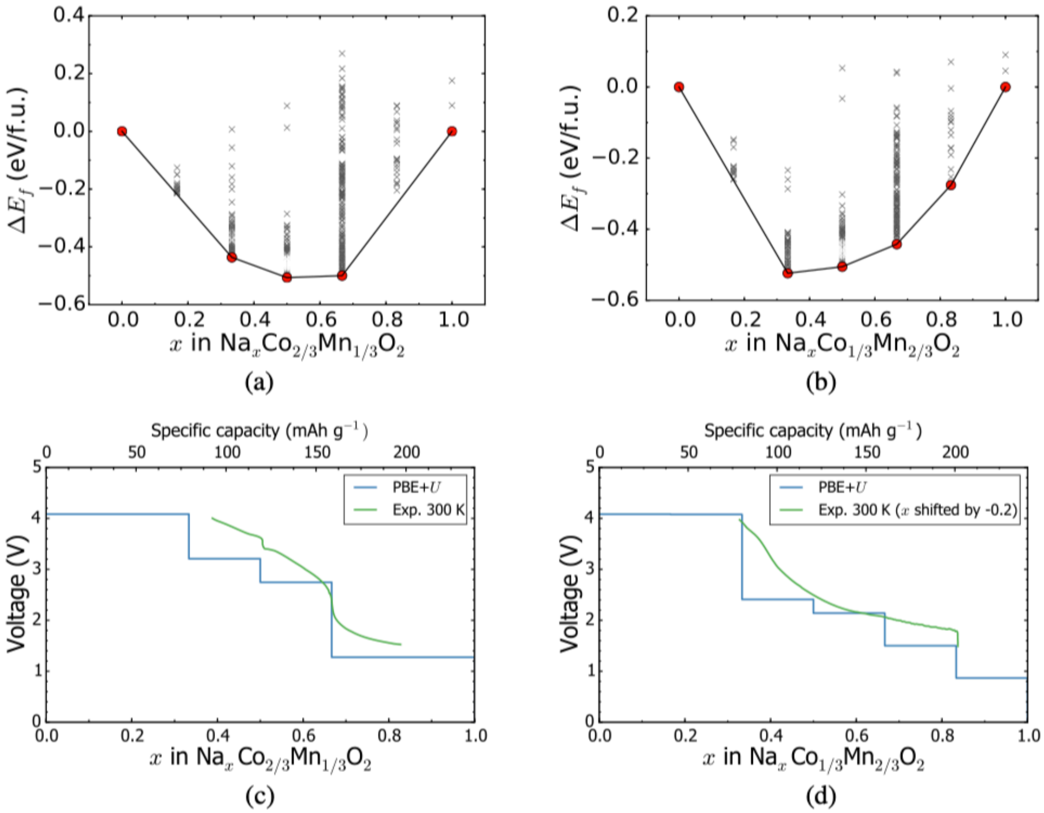


Figure 2.4 PBE + U 0-K stability diagrams of (a) $\mathrm{Na}_x\mathrm{Co}_{2/3}\mathrm{Mn}_{1/3}\mathrm{O}_2$ and (b) $\mathrm{Na}_x\mathrm{Co}_{1/3}\mathrm{Mn}_{2/3}\mathrm{O}_2$. The solid black line shows the convex hull with red dots representing stable orderings on the hull. The black cross dots show unstable orderings. (c) and (d) show the comparison between PBE + U and experimental voltage profiles of $\mathrm{Na}_x\mathrm{Co}_{1-\mathrm{y}}\mathrm{Mn}_y\mathrm{O}_2$ for $\mathrm{y} = 1/3$ and $2/3$. Experimental voltage profiles for $\mathrm{Na}_x\mathrm{Co}_{2/3}\mathrm{Mn}_{1/3}\mathrm{O}_2$ and $\mathrm{Na}_x\mathrm{Co}_{1/3}\mathrm{Mn}_{2/3}\mathrm{O}_2$ (x shifted by -0.2) are from Refs.[93]and [32], respectively.

Figures 2.4(c) and 2.4(d) compare the PBE + U voltage profiles for $\mathrm{Na}_x\mathrm{Co}_{2/3}\mathrm{Mn}_{1/3}\mathrm{O}_2$ and $\mathrm{Na}_x\mathrm{Co}_{1/3}\mathrm{Mn}_{2/3}\mathrm{O}_2$ with the electrochemically measured voltage profiles of Wang et al.[32]. Again, we find that the calculated voltages are in relatively good agreement with the experimental ones. For $\mathrm{Na}_x\mathrm{Co}_{2/3}\mathrm{Mn}_{1/3}\mathrm{O}_2$, the small potential drop at $x = 1/2$ in the PBE + U

voltage profile indicates the formation of a stable ordered phase during the deintercalation or intercalation process, in agreement with previous experiments[32,93]. It should be noted that, for $\mathrm{Na}_x\mathrm{Co}_{1/3}\mathrm{Mn}_{2/3}\mathrm{O}_2$, we shift the experimental voltage curve by -0.2 Na concentration to properly align the computed and experimental voltage profiles. This shift accounts for the artificial oversodiation (which results in Na stoichiometry > 1 in experiments) observed in the electrochemical measurements[32].

2.3.2 Na diffusion kinetics

To identify the effect of TM mixing on Na kinetics, we perform AIMD simulations on $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ at a single Na concentration of $x = 1/2$ and Mn concentrations of $y = 0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$, and 1 in our study. The choice of $x = 1/2$ is motivated by the fact that strong Na orderings are typically observed at that Na concentration in the single TM, and even in mixed-TM systems, which may have a significant effect on Na kinetics.

We classify the Na sites of $\mathrm{Na}_{1/2}\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ into four types according to our Co-Mn frameworks:

i. Na(2), which shares only edges with *M*O6 octahedra;

ii. $\mathrm{Na(1)}_{\mathrm{Mn\text{-}Mn}}$, which shares faces with two MnO_6 octahedra;

iii. $\mathrm{Na(1)}_{\mathrm{Co\text{-}Co}}$, which shares faces with two CoO_6 octahedra; and

iv. $\mathrm{Na(1)}_{\mathrm{Co\text{-}Mn}}$, which shares faces with one CoO_6 octahedra and one MnO_6 octahedra.

Figure 2.5 shows the isosurfaces of the Na probability density distribution extracted from the AIMD simulations of the ground-state TM orderings. The probability density distribution is calculated by averaging the Na occupation on a uniform grid over the trajectories during AIMD simulations of 25 ps. As can be observed from Fig. A.8, thermal equilibration is achieved for all systems after 10 ps of simulation time, and the relevant statistics are obtained after this equilibration period. In the ground-state orderings of $y = 1/3$ and $2/3$, all Na sites either share edges with MO$_6$ octahedra or share faces with only one type of TM; i.e., there are no Na(1)$_{\text{Co-Mn}}$ sites. We relax this constraint shortly. We may observe that the Na trajectories in the single TM [Figs. 2.5(a) and 2.5(b)] exhibit the typical honeycomb network observed in previous studies[96].

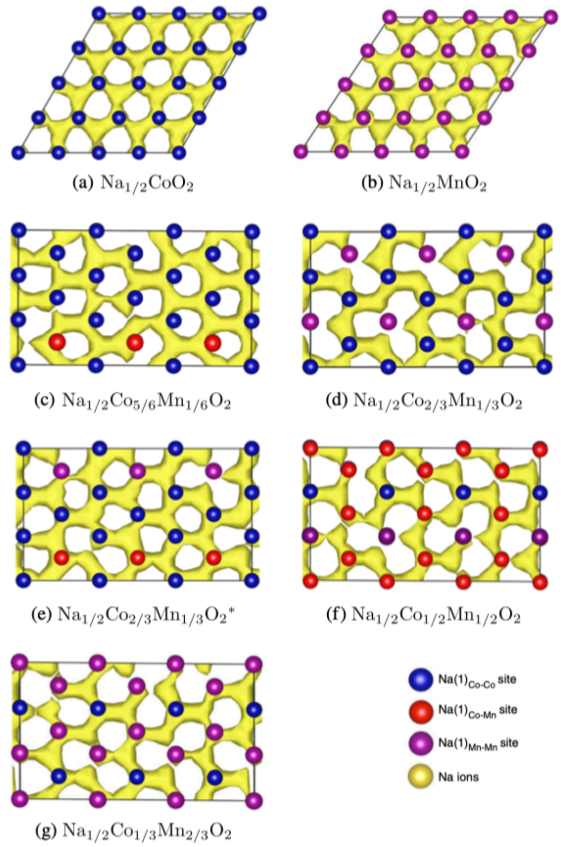

(a) $Na_{1/2}CoO_2$
(b) $Na_{1/2}MnO_2$
(c) $Na_{1/2}Co_{5/6}Mn_{1/6}O_2$
(d) $Na_{1/2}Co_{2/3}Mn_{1/3}O_2$
(e) $Na_{1/2}Co_{2/3}Mn_{1/3}O_2{}^*$
(f) $Na_{1/2}Co_{1/2}Mn_{1/2}O_2$
(g) $Na_{1/2}Co_{1/3}Mn_{2/3}O_2$
Na(1)$_{\text{Co-Co}}$ site
Na(1)$_{\text{Co-Mn}}$ site
Na(1)$_{\text{Mn-Mn}}$ site
Na ions

Figure 2.5 Isosurfaces of Na ion (yellow) probability density distribution P at $P = P_{\text{max}}/12$ for $P2$ $Na_{1/2}Co_{1-y}Mn_yO_2$ at 1000 K, top view from the AB plane. Metallic blue dots indicate the positions of Na(1)$_{\text{Co-Co}}$ sites, and red dots represent the positions of Na(1)$_{\text{Co-Mn}}$ sites. Magenta circles correspond to Na(1)$_{\text{Mn-Mn}}$ site positions. All simulations are carried out using ground-state Co-Mn frameworks of $P2$ $Na_{1/2}Co_{1-y}Mn_yO_2$ except $P2$ $Na_{1/2}Co_{2/3}Mn_{1/3}O_2^*$. The $Na_{1/2}Co_{2/3}Mn_{1/3}O_2^*$ framework in (e) is randomized such that it contains all three types of Na(1) site. This framework is different from the ground-state Co-Mn framework of $Na_{1/2}Co_{2/3}Mn_{1/3}O_2$ as shown in Fig. A.6(a).

At $y = 1/3$ [Fig. 2.5(d)], we observe that there is a clear preference for Na diffusion to occur via the Na(2) and Na(1)$_{\text{Co-Co}}$ sites, which form a percolating diffusion network, while the Na(1)$_{\text{Mn-Mn}}$ is clearly less preferred. At a higher concentration of Mn [$y = 2/3$, Fig. 2.5(g)], the Na trajectories once again follow the honeycomb topology given that a percolating diffusion network cannot be formed from Na(2) and Na(1)$_{\text{Co-Co}}$ sites.

To obtain a more quantitative assessment of site preferences, the Na site-occupancy fractions (SOFs) are estimated from the AIMD trajectories for the four kinds of sites in $\mathrm{Na_{1/2}Co_{1-y}Mn_yO_2}$. Each Na is assigned to the closest site within the Na honeycomb network. The average Na SOFs are then given by

$$SOF(i,t) = \frac{1}{tN_d}\int_0^t N_i(t')dt', \tag{2.3}$$

wherein $N_i(t')$ is the total number of Na^+ ions in sites of type i at time t' and N_d is the total number of Na^+ ions in the system.

Figure 2.6 summarizes the average Na SOFs. Here, we present the results for an additional AIMD simulation of the $y = 1/3$ structure where the Co-Mn ordering is randomized in such a way that it contains all four types of Na sites (denoted as $y = 1/3^*$). We notice that the randomized TM ordering and the ground-state one are close in energy for $\mathrm{Na_{1/2}Co_{2/3}Mn_{1/3}O_2}$ with identical Na ordering (< 10 meV atom^{-1}). We may make the following observations.

i. At all Mn concentrations, the Na(2) site is clearly the most preferred site, with an average SOF of approximately 0.6.

ii. At $y\ < 1/2$, we find that Na(1)$_{\text{Co-Co}}$ is the next most-preferred site, followed by Na(1)$_{\text{Co-Mn}}$. Na(1)$_{\text{Mn-Mn}}$ is the least preferred. The higher SOF of Na(1)$_{\text{Mn-Mn}}$ compared to Na(1)$_{\text{Co-Mn}}$ at $y = 1/3$ is an artifact, as the ground-state $y = 1/3$ ordering does not contain Na(1)$_{\text{Co-Mn}}$ sites. When the Co-Mn ordering is randomized ($y = 1/3^*$), the SOF of Na(1)$_{\text{Co-Mn}}$ is clearly higher than Na(1)$_{\text{Mn-Mn}}$.

iii. At $y = 1/2$, the SOF of the Na(1)$_{\text{Co-Co}}$ decreases substantially given that there are more Na(1)$_{\text{Co-Mn}}$ sites. Nevertheless, both Na(1)$_{\text{Co-Mn}}$ and Na(1)$_{\text{Co-Co}}$ exhibit significantly higher SOFs than Na(1)$_{\text{Mn-Mn}}$.

iv. Finally, at $y = 2/3$, the much higher concentration of Mn results in an almost equal Na(1)$_{\text{Co-Co}}$ and Na(1)$_{\text{Mn-Mn}}$ SOF.

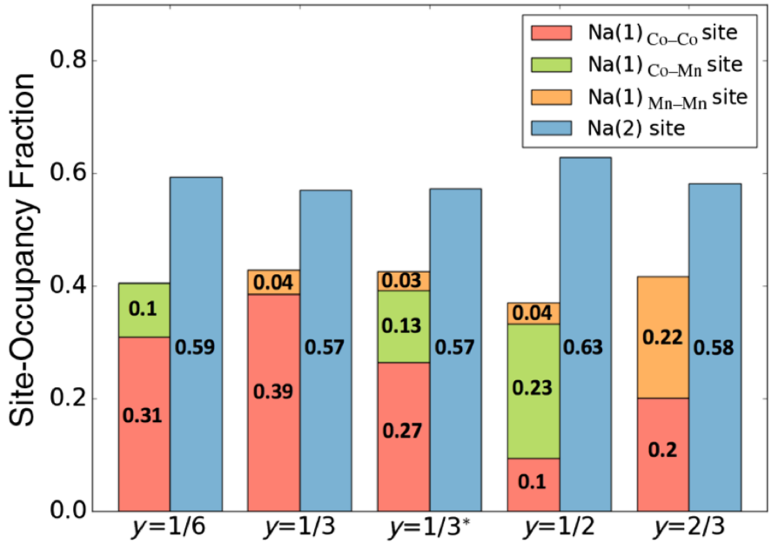


Figure 2.6 Average Na SOFs in *P*2 $\mathrm{Na_{1/2}Co_{1-y}Mn_yO_2}$ from AIMD simulations at 1000 K. The average SOFs are estimated from 25-ps AIMD simulation results. All AIMD simulations are carried out using ground-state Co-Mn frameworks of *P*2 $\mathrm{Na_{1/2}Co_{1-y}Mn_yO_2}$ except y = $1/3^*$, in which the Co-Mn ordering is randomized such that it contains all four types of Na sites.

2.4 Discussion

A. Na ordering in *P*2 $\mathrm{Na_xCo_{1-y}Mn_yO_2}$

We provide an overview of the *P*2 $\mathrm{Na_xCo_{1-y}Mn_yO_2}$ 0-K stability diagram as a function of x and y in Fig. 2.7, with the stable orderings indicated by blue circles. The background color indicates the formation energy relative to the fully sodiated and desodiated single-TM end points, calculated as

$$\begin{aligned}\Delta E_f\left(\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2\right) \\ &= E\left(\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2\right) \\ &- x[(1-y)E(\mathrm{NaCoO}_2) + yE(\mathrm{NaMnO}_2)] - (1 \\ &- x)[(1-y)E(\mathrm{CoO}_2) + yE(\mathrm{MnO}_2)\end{aligned} \tag{2.4}$$

where $E(X)$ refers to the energy of phase X. Values between data points are linearly interpolated.

In general, we find that both PBE and PBE + U give stable orderings at similar Na concentrations for each y, especially in the Na-concentration range of interest $\frac{1}{3} \leq x \leq \frac{2}{3}$. Though there are minor differences in the actual stable ordered structures, the qualitative features of the 0-K stability diagrams are generally consistent between PBE and PBE + U. The notable exception is the $\mathrm{Na}_x\mathrm{CoO}_2$ system, for which it has been well established that the application of the Hubbard U leads to 0-K stability diagrams that are in disagreement with experiments at lower x.[41]

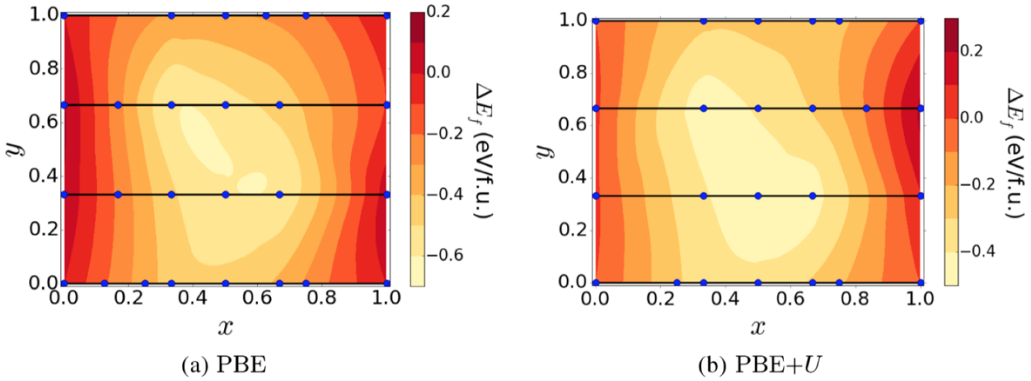


Figure 2.7 $P2$ $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ 0-K stability diagrams calculated using (a) PBE and (b) PBE + U. Stable orderings are indicated by blue circles. Background color indicates the formation energy per formula unit with respect to the fully sodiated and fully desodiated structures. See the text for details.

From Fig. 2.7, we observe that the formation of mixed Co-Mn phases is very unfavorable (positive formation energies relative to the single-TM end members) at close to full sodiation ($x \sim 1$) and desodiation ($x \sim 0$) at 0 K. Formation of mixed Co-Mn phases is most favorable at $0.25 \leq x \leq 0.75$. Most mixed-TM layered $P2$ materials are typically synthesized at $x \sim 0.67 - 0.75$,[32,78,85,86,92,93] which is within the range of x where the predicted formation energies relative to the end members are negative. We also note that the mixed Co-Mn phases are known to be disordered at finite temperatures[32], and configurational entropic effects are not taken into account in the 0-K stability diagram.

Previously, Wang et al.[32] have carried out an extensive experimental study of $P2$ $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$. Their findings were that, as Co is substituted by Mn, i.e., increasing y, the accessible capacity increases, and a number of ordered states, particularly that at $x = 2/3$, disappear. The computed 0-K stability diagrams also predict a large number of nearly degenerate Na orderings upon TM mixing, particularly at $x = 2/3$, which supports these experimental observations. We also find that Co-Mn ordering has a relatively small effect on total energies, regardless of Na ordering, which again supports experimental observations of Co-Mn disorder in this system.[102]

It should be noted that a key limitation of this work is that it is based on 0-K DFT calculations of Na orderings up to relatively small supercell sizes, and the effects of temperature are not considered. Na orderings at other compositions that require a larger supercell size are not included in our study. For the Na composition range of general interest (i.e., $0.33 < x < 0.75$), however, our predicted stability diagrams and intercalation voltage profiles are in reasonably good agreement with the experimental literature.[32,35] A possible extension to incorporate these effects to some degree is to fit a cluster expansion Hamiltonian[120] using the calculated energies and perform Monte Carlo simulations[121] on much larger supercell sizes to obtain finite-temperature voltage profiles and diffusivities. However, this effort would require accounting not just for Na or vacancy orderings but also electron or hole orderings for two transition metals. This significant undertaking is outside the scope of this work and would be the subject of future studies.

B. Na migration barriers

From the AIMD simulations, we have established that the Na(1) site that shares faces with two MnO_6 octahedra [$Na(1)_{Mn\text{-}Mn}$] is higher in energy relative to Na(1) sites that share faces with two CoO_6 [$Na(1)_{Co\text{-}Co}$] or one CoO_6 and one MnO_6 [$Na(1)_{Co\text{-}Mn}$]. We here generalize these results into a universal theoretical framework for the rational optimization of mixed-TM layered *P*2 oxides.

In the layered *P*2 oxides, Na diffusion occurs in a 2D honeycomb lattice[96], which can be decomposed into two intersecting triangular lattices comprising Na(2) and Na(1) prismatic sites.

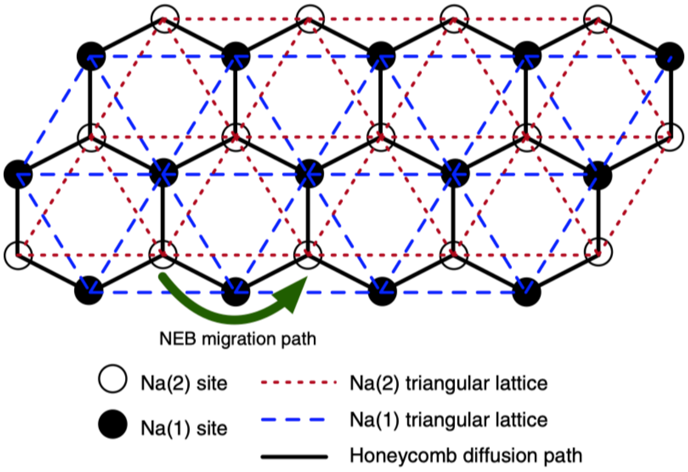


Figure 2.8 Schematic view of 2D Na diffusion pathways (top view on the AB plane) consisting of triangular lattices. Black dots are Na(1) sites. White circles represent Na(2) sites. Solid black lines highlight the two-dimensional honeycomblike diffusion pathways. Red dashed lines show triangular lattices formed by Na(2) sites. Blue dashed lines show triangular lattices consisting of Na(1) sites. Green arrows represent the Na migration pathway between two nearest-neighbor Na(2) sites via a Na(1) site from CINEB calculations.

The Na(2) sites are generally lower in energy, as they share only edges and not faces with MO_6 octahedra. However, diffusion between Na(2) sites must occur via the Na(1) sites. Hence, we may treat the triangular network of the higher-energy Na(1) sites as the effective diffusion topology for $P2$ layered oxides [Fig. 2.8]. From the site-percolation theory, a long-range percolating diffusion path exists at the macroscopic limit if the probability of site occupancy exceeds the percolation threshold p_c. For the triangular network, this threshold can be shown analytically to be 0.5[122].

Consider the introduction of a new TM species M' at some concentration z into a layered $P2$ oxide. M' can have either a beneficial or detrimental effect on Na diffusion by lowering or increasing the energy of Na(1) sites that share faces with it. Here, we assume that the new mixed-TM $P2$ layered oxide is disordered, which can occur because of either an intrinsically low enthalpy of mixing or through the synthesis or processing that usually takes place at elevated temperatures. There are four limiting cases.

i. M' has a substantial beneficial effect such that any Na(1) site that shares faces with at least one $M'O_6$ has a sufficiently low barrier for diffusion at 300 K. The probability of any Na(1) site having at least one $M'O_6$ is $2z(1-z)+z^2$. The condition for macroscopic fast diffusion is then $2z(1-z)+z^2 >$ 0.5, i.e., $z > 0.293$.

ii. M' has a moderate beneficial effect such that only Na(1) sites that share faces with two $M'O_6$ have a sufficiently low barrier for diffusion at 300 K. The probability of any Na(1) site having two $M'O_6$ is z^2. The condition for macroscopic fast diffusion is then $z^2 >$ 0.5, i.e., $z > 0.707$.

iii. M' has a substantial detrimental effect such that any Na(1) site that shares faces with at least one $M'O_6$ has a high barrier for diffusion at 300K.Theprobabilityof any Na(1) site having at least one $M'\text{O}_6$ is $2z(1-z)+z^2$. The condition for macroscopic fast diffusion is then $1-[\,2z(1-z)+z^2] > \ 0.5$, i.e., $z < 0.293$.

iv. M' has a moderate detrimental effect such that only Na(1) sites that share faces with two $M'\text{O}_6$ have a high barrier for diffusion at 300 K. The probability of any Na(1) site having two $M'\text{O}_6$ is z^2. The condition for macroscopic fast diffusion is then $1 - z^2 > 0.5$, i.e., z < 0.707.

Using CINEB calculations, we calculate the Na(2)-Na(1)-Na(2) (see Fig. 2.8 for the path) migration barriers for various Na(1) site compositions. These calculations are performed for a single Na hopping in an otherwise empty lattice of $P2$ Na_xCoO_2, i.e., at the fully charged limit, with different M' introduced at the Na(1) site. As can be seen from Table 2.1, Mn is indeed predicted to have a detrimental effect on Na diffusion, consistent with the results of the AIMD simulations and SOFs. We find that Fe and Ni are predicted to have a beneficial effect on Na diffusion, i.e., lowering of the Na(1) site energies relative to the pure NaCoO_2 at the end of the charge. We perform similar calculations using $\text{NaCo}_{2/3}\text{Mn}_{1/3}\text{O}_2$ [see Table A.4], and the same qualitative trends in the effect of M' are observed.

C. Na migration barriers

From the preceding analyses, we may surmise that TM mixing can have two effects: (i) the suppression of Na vacancy ordering, leading to a wider range of single-phase behavior, and (ii) the modification of the Na(1) site energies, and hence diffusion barriers, especially towards the end of the charge. The former effect is well established in the literature, with many experimental works conclusively demonstrating the suppression of Na-vacancy ordering in the Co-Mn[32,35], Co-Mn-Fe[92], Co-Mn-Ni[123], and other systems. However, the latter effect has not been explored in detail, and, indeed, most TM mixing in the search for compositions that offer a greater rate capability have been carried out mostly in an trial-and-error fashion thus far.

Table 2.1 Na migration barriers of different Na(1) site compositions.

Na(1) site TM	NEB barrier (meV)
Co-Co	76
Co-Mn	90
Mn-Mn	100
Co-Fe	63
Fe-Fe	61
Co-Ni	46
Ni-Ni	65

This work provides a rational basis for the selection of mixed-TM compositions by establishing the effect of different TMs on Na(1) site energies and the necessary minimum and maximum limits for beneficial and detrimental TM dopants, respectively. We have indirect evidence from the experimental literature supporting these conclusions. Most mixed-TM *P*2 layered materials reported with a reasonable rate capability thus far have a concentration of Mn, which a high Na(1) site energy, of around 0.67[33,37,78,124–127], close to the upper limit of 0.707 predicted in order for a percolating network of sites that do not contain two Mn to exist. In particular, two reported compositions with a high rate performance are $Na_{2/3}Mn_{1/2}Co_{1/4}Mn_{1/4}O_2$[92] and $Na_{2/3}Co_{2/3}Mn_{2/9}Ni_{1/9}O_2$[123], wherein the Mn concentration is well below 0.707. In the case of $Na_{2/3}Co_{2/3}Mn_{2/9}Ni_{1/9}O_2$, the Mn concentration is below 0.293, indicating that a percolating network of Na(1) without Mn exists in this material. Furthermore, the improved rate performance of both these materials compared to $Na_{2/3}CoO_2$[128], *P*2 $Na_{0.6}MnO_2$[36], and *P*2 $Na_{2/3}Co_{1-y}Mn_yO_2$[32]suggests that Ni and Fe do indeed have a beneficial effect as suggested by the CINEB results.

It should be noted that the above conclusions are based on CINEB calculations on model frameworks in the dilute Na (fully charged) limit and under the assumption of fully disordered TM mixing. No consideration is given to a possible effect of the TM mixing composition on interlayer spacing at various Na concentrations, a factor that is known to have a significant effect on diffusion barriers close to the end of the charge[129,130]. Nevertheless, we believe the results provide a useful rational framework to further explore TM composition tuning in the layered *P*2 oxides. We hope that future systematic experimental investigations would provide a quantitative verification (as opposed to the indirect evidence highlighted above) of the effects of the different TMs on diffusion barriers and the concentration limits suggested by our model.

2.5 Conclusion

In conclusion, we perform a first-principles investigation on the Na diffusion kinetics of mixed-TM *P*2 $Na_xCo_{1-y}Mn_yO_2$. The calculated 0-K stability diagrams suggest that Co-Mn mixing tends to decrease the energy difference between different Na orderings, which may suppress the formation of strongly ordered intermediates and promote single-phase behavior over a wider range of Na concentration. Using AIMD simulations and CINEB calculations, we show that the TM composition at a particular Na(1) site can have a profound effect on the Na site occupation energy. The presence of Mn is shown to lead to an increase in the Na(1) site energy, leading to higher diffusion barriers. Fe and Ni, on the other hand, are shown to lower Na(1) site energies and diffusion barriers relative to Co. By employing a site-percolation model based on

Na site occupancy and CINEB results, we also establish theoretical upper and lower bounds on the concentration of various TM species based on their beneficial or detrimental effect on Na diffusion barriers at the end of the charge. These results provide a useful rational framework for the further optimization of TM mixing composition in the *P*2 layered oxides.

Chapter 2, in full, is a reprint of the material “Effects of Transition-Metal Mixing on Na Ordering and Kinetics in Layered *P*2 Oxides” as it appears in Physical Review Applied, Chen Zheng, Balachandran Radhakrishnan, Iek-Heng Chu, Zhenbin Wang, and Shyue Ping Ong, 2017, 7 (6), pp 064003. The dissertation author was the primary investigator and author of this paper.

Chapter 3 Deep learning driven study of high entropy cathode $Na_xCo_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$

3.1 Introduction

Studying high entropy materials using density functional theory (DFT) is extremely challenging due to the relatively slow computing time of DFT (> hours for single energy calculation) and exponentially complex configuration space. A fast, general and accurate surrogate energy model replacing the time consuming DFT is mandatory to shed light on this type of problem. Recently we developed a MatErials Graph Networks (MEGNet) framework for general materials property predictions and the models show excellent agreement with the DFT results in terms of formation energy, band gap and elasticity on the entire Materials Project data base[131] containing around ~70,000 crystal structures[132]. In particular, the formation energy model error is as low as 28 meV/atom. This model is employed here for fast calculation of energies for different configurations of high entropy materials.

3.2 Methods

The structural formation energy for all configurations were first predicted using the MEGNet pre-trained model from our previous work[132]. Using the MEGNet surrogate model, a Monte Carlo simulated annealing method coupled with Metropolis sampling algorithm was used for finding the structural configurations with lowest energies. We started from the well-known Na vacancy configuration pattern for *P*2-$Na_{0.6}CoO_2$, and then we replaced Co with equal molar Co-Ti-Mn-Ni-Ru randomly. The chosen supercell contains 10 formula unit and 36 atoms. Note that the possible arrangement for transition metals (TMs) is 113400 if not considering symmetry. If the configuration of Na vacancies (38760) is also considered, the total configuration space is $> 4 \times 10^9$ even for this relatively small supercell. All distinct structures from simulated annealing trace were then calculated using DFT. Those represent the lowest energy structures for the entire configuration space. To make comparisons, two simulated annealing optimizations were performed with one only swapping the TMs and the other also flipping the Na occupancy.

The PBE generalized gradient approximation functional was adopted in density functional theory calculations. Na migration barriers were calculated using the climbing-image nudged elastic-band (CINEB).[112] In this study, supercells of $4 \times 4 \times 1$ (32 formula units) were used for mixed- and single TM systems. The k-point grid was generated following the automated k-mesh generation method. Each interpolated image was relaxed until the forces on each atom are less than 0.02 eV Å^{-1}. To investigate the role of the transition metal on Na diffusion, we performed CINEB calculations at the dilute Na concentration. In the $P2$ layered sodium TM oxides, there are two kinds of Na sites. The Na(f) site shares faces with two TM oxide octahedra, and the Na(e) site shares only edges with TM oxide octahedron. A total of 25 different Na(f) site configurations were created in the lattice of CoO_2 with Ni, Mn, Ru, Ti as dopants, and considering TM species combinations. We calculated the Na(e)-Na(f)-Na(e) migration barriers for various Na(f) compositions.

AIMD simulations were carried out in the constant volume (NVT) ensemble at 1000K with a Nosé-Hoover thermostat.[110] Instead of estimating the diffusivity of $P2$ $Na_{0.6}(CoMnNiTiRu)_{0.2}O_2$, we aimed to elucidate the Na site occupancies and diffusion mechanism of the material at non-dilute Na concentration. AIMD simulations were carried out for a relatively short time of 60 ps. We conducted non-spin-polarized simulations with a smaller plane-wave energy cutoff of 300 eV and Γ-centered $1 \times 1 \times 1$ k-point grid to reduce computational costs. Two simulated annealing optimized structures were adopted in the AIMD simulations. One lowest-energy structure of $P2$ $Na_{0.6}(CoMnNiTiRu)_{0.2}O_2$ was obtained by performing optimization with swapping the TMs only (TM ground-state structure), and the other one was obtained by flipping the Na occupancy as well (Na-TM ground-state structure).

3.3 Results and discussions

3.3.1 MEGNet results validation

As a first step, the MEGNet predictions were validated using DFT calculations. The pristine *P2*-$NaCoO_2$ structure from Materials Project was taken as the initial structure and then rescaled match experimental lattice parameters. Then different TMs were substituted into the Co site and the MEGNet predictions were performed. These results were compared to the DFT relaxation results, as shown in Table 3.1.

Table 3.1 MEGNet predicted results compared to DFT for $NaTMO_2$ (TM=Ti, Mn, Co, Ni or Ru).

TM	Ti	Mn	Co(Na$_e$)	Co(Na$_f$)	Ni	Ru
$\boldsymbol{E_f^{DFT}}$(eV/atom)	-2.79	-1.94	-1.59	-1.46	-1.18	-1.38
$\boldsymbol{E_f^{MEG}}$(eV/atom)	-2.78	-1.94	-1.51	-1.4	-1.21	-1.42

The MEGNet prediction results match well with the DFT values across all single-TM structures, with the largest error of only 80 meV/atom as found in Co case with edge sharing Na polyhedron Na(e). Most importantly, the energy trend and difference are in line with DFT predictions. In particular, the predicted Na face sharing energy is higher than the edge sharing energy by 110 meV/atom, close to the DFT values of 130 meV/atom. In *P2* systems, the Na/vacancy ordering patterns are largely determined by the Na interactions with the TM layer. Thus the matching differences between edge sharing and face sharing local environment predicted by MEGNet make it promising for studying the Na disordering behavior.

To further benchmark the MEGNet predictions in the high entropy $\mathrm{Na}_x\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$, we first obtained the energy range of all TM configurations using MEGNet and then found the structures with predicted energies lying on the energy quantiles (see Appendix B for details). Those structures were subsequently calculated using DFT.

Table 3.2 MEGNet predicted results compared to DFT for structures with predicted energy quantiles (0, 0.25, 0.5, 0.75, 1).

MEG-Order	**DFT-Order**	$\boldsymbol{E_f^{MEG}}$ (eV/atom)	$\boldsymbol{E_f^{DFT}}$ (eV/atom)
1	1	-1.921	-1.893
2	4	-1.909	-1.879
3	3	-1.918	-1.865
4	2	-1.919	-1.851
5	5	-1.864	-1.837

For the five structures, the MEGNet model predicts mostly the correct energy trend. However, the 2nd and 4th structures are swapped in positions. This is however understandable given their energy differences (28 meV/atom) within our model error. The energy range is predicted to be 57 meV/atom, close to the DFT range of 56 meV/atom.

3.3.2 The possibility of high entropy cathode formation

Two simulated annealing simulations were performed with one only swapping the TMs and the other also flipping the Na occupancy but keeping the overall Na concentration fixed. The sampled unique structures were subsequently computed using DFT. The distribution of DFT formation energies for both cases are shown in Figure 3.1.

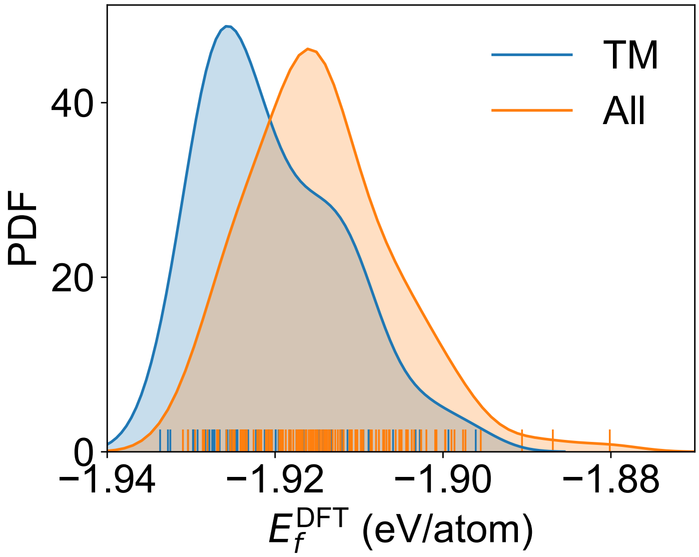
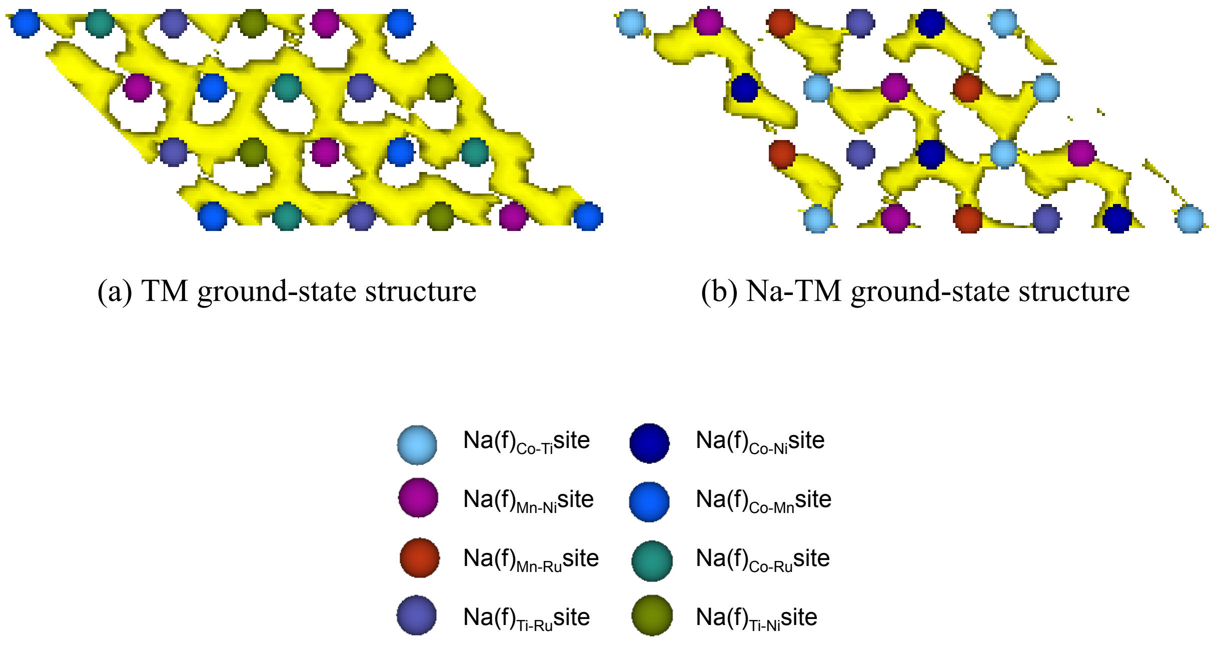


Figure 3.1 The probability distribution function (PDF) of energies for different orderings of TMs (TM) and for both Na ordering and TM ordering (All). The rugs indicate the DFT calculated values.

Here we show the distribution of structural DFT formation energies below -1.91 eV/atom (178 for All, and 65 for TM). The spreads of energies are within thermal energy at room temperature (~26 meV/atom). This small energy spread suggests that all those states can be reached via thermal excitation even at room temperature with considerate probability (>0.37). In addition, the entropy contribution to the overall free energy is -41.4 meV/atom for TM disordering and -30.2 meV/atom for Na disordering (Eq. B.1) in $\mathrm{Na_{0.6}Co_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2}$. Such large free energy driven force will smear the small enthalpy gap between disordered and ordered phase, favoring the formation of high entropy compounds.

3.3.3 Na diffusion kinetics

Figure 3.2 shows the isosurfaces of the Na probability density distribution extracted from the AIMD simulations of the ground-state TM orderings of $P2$ $Na_{0.6}Co_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$ resulted from two simulated annealing optimizations. We estimated the probability density distribution by averaging the Na occupation on a uniform grid over the trajectories during AIMD simulation of 20 ps following the thermal equilibration period.



(a) TM ground-state structure

(b) Na-TM ground-state structure

(c)

Figure 3.2 Isosurfaces of Na ion (yellow) probability density distribution P at $P = P_{\max}/32$ for $P2$ $Na_xCo_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$ at 1000K of ground state structures obtained by performing optimization with (a) swapping the TMs only, and (b) flipping the Na occupancy and TMs simultaneously. Top view on the AB plane. (c) The color scheme of circles correspond to different Na(f) site configurations.

As can be observed from Fig. 3.2(a), the Na trajectories in the TM ground-state structure exhibit the typical honeycomb topology consistent with previous studies.[96,133] For Na-TM ground-state structure, we notice that Na(f)$_{\text{Ti-Ru}}$ is the least preferred site and the formed percolating diffusion network bypasses the Na(f)$_{\text{Ti-Ru}}$ sites. It should be noted that due the limitations in computational resources, the ground-state structures were enumerated based on a relatively small supercell sizes (30 formula units). Generating structures including all 15 Na(f) sites' TM configurations requires a larger supercell size and is impractical for AIMD studies.

3.3.4 Na migration barriers results

Figure 3.3 summarizes the Na migration barriers of different Na(f) site compositions. Here, consider the migration barrier values. We could classify the TM species into 3 categories.

(1) Ni and Co: The Ni and Co have a substantial beneficial effect on Na diffusion as the diffusion barrier of Na(f) site that shares faces with at least one CoO_6 or NiO_6 has a low barrier for diffusion. For those Na(f) site that share faces with two NiO_6, the diffusion barrier drops to 89 meV.

(2) Ti and Mn: Comparing to the Ni and Co, these two TM species are predicted to have a moderate detrimental effect on Na diffusion. For Na(f) sites that share faces with one TiO_6 or MnO_6, the diffusion barriers are around 140 meV.

(3) Ru: Ru is predicted to have substantial detrimental effect such that Na(f) site that shares faces with at least one RuO_6 has a high barrier for diffusion. For those Na(f) site that shares two faces with RuO_6, the Na diffusion barrier spike to 306 meV, which is about 200 meV higher than the Na(f) site that shares faces with two beneficial TM species (Ni or Co).

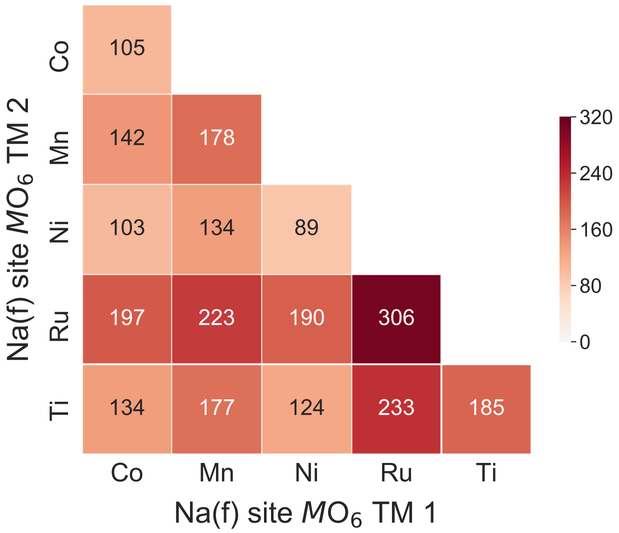


Figure 3.3 Na(e) - Na(f) - Na(e) single Na hopping migration barriers (meV) of different Na(f) site compositions at the fully charged limit.

In general, the TM concentrations of *P*2 $Na_xCo_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$'s are within the minimum and maximum limits estimated through the previously reported site-percolation model in the *P2* systems.[133] For macroscopic fast diffusion, the lower bound concentration *z* of TM with substantial beneficial effects is estimated to be 0.293[133]. In the *P*2 $Na_xCo_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$, z_{Ni+Co} is 0.4, which is above the percolation limit. For TM dopants with substantial or moderate beneficial effect on Na diffusion, their concentration should exceed 0.707[133]. In this case, the total concentration is 0.8 for the TM species Ni, Co, Mn, Ti, that have beneficial effect on Na diffusion. Once again, this concentration satisfies the necessary minimum limit estimated by the percolation model. From the preceding results, Ru is predicted to have a substantial detrimental effect on Na diffusion. Fortunately, z_{Ru} is below the upper limit of 0.293 predicted in the percolation model[133] in order for a percolating network of sites that do not contain two Ru to exist.

Chapter 3, in full, is currently being prepared for submission for publication of the material “Deep learning driven study of high entropy cathode $Na_xCo_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$”, Chi Chen, Chen Zheng, and Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

Chapter 4 High-throughput Computational X-ray Absorption Spectroscopy

4.1 Background

A crucial step in the process of novel materials discovery is the characterization of the synthesized material. There exists a wide array of tools and spectroscopic techniques that are used in the material identification process, e.g. X-ray diffraction (XRD), X-ray emission spectroscopy (XES), and X-ray absorption spectroscopy (XAS). XAS is widely-employed in the characterization of the local structural environment surrounding select elements within a material.

Great progress has been made over the past few years in the development of laboratory-based X-ray spectrometers for high-resolution x-ray absorption near edge structure (XANES) and X-ray emission spectroscopy (XES)[134]. The availability of relatively inexpensive laboratory-based XAFS system (http://easyxafs.com/) and third generation synchrotron facilities[135] have established the groundwork for the broad application of high-resolution XAS in characterization of materials. On the other hand, modern computational resources and methodologies have reached a level of maturity and efficiency to complement as well as to fast-track new discoveries. In the case of XAS, a variety of theoretical frameworks including time-dependent density-functional theory (TDDFT)[136,137], multiplescattering[64], and Bethe-Salpeter equation (BSE) based approaches[138] have been implemented, each exhibits its advantages and drawbacks. Leveraging spectroscopic simulation software with large crystal structure databases enables the computation of a large number of reliable theoretical spectra corresponding to well defined crystal structures[17], providing a broad reference dataset with clean unique structural fingerprints that can be used for identification purposes. With the help of carefully crafted software tools, these computations can be performed in a high-throughput fashion and can be used to scan the structural phase space for novel materials. In addition, through proper integration with modern database tools, these scans can be saved for future use and leveraged for training machine learning algorithms to assist the characterization process. Some examples of such publicly available spectroscopic database are the EELS Data Base[54], a compilation of valence and core-loss spectra from EELS and XAS experiments containing 271 spectra that covers 39 elements of the periodic table, and XCOM (https://www. nist.gov/pml/xcom-photon-cross-sections-database), which provides photon cross sections for scattering, photoelectric absorption and pair production, as well as total attenuation coefficients, for any element, compound or mixture. Other existing XAS databases[139,140], i.e. https:// www.cat.hokudai.ac.jp/catdb/ and http://cars.uchicago.edu/xaslib, covering a few hundred spectra, are hosted across the world and serve as valuable references for analysis.

The FEFF framework affords relatively inexpensive calculations compared to other approaches and requires minimum adjustable parameters. It provides an efficient means of generating high quality XAS spectra for a larger amount of chemical systems and structures. Hence, in our study, we selected the FEFF9 (ref. 5) program for the ab initio calculation of K-edge X-ray absorption near edge spectra (XANES). Using the parameter settings obtained from recent benchmarking work against experimental spectra[141] and the FEFF workflow infrastructure available in the open source materials science workflow package Atomate[142] , we generate spectra of all the materials available in the publicly accessible and widely used materials database, Materials Project (MP)[17].

A comprehensive database of computed XAS spectra enables comparison between different spectroscopic signatures across chemical systems and structures such that rapid determination of oxidation states, coordination environment, and other local atomic structure information can be obtained. Furthermore, using matching algorithms[141] or other machine learning methods[143], the data can be leveraged for on-the-fly characterization. Though the peak positions and amplitudes of the computational K-edge XANES spectral may exhibit differences compared to experimental spectra, theoretically computed XANES spectra provide sufficient information to identify oxidation state and coordination chemistry of the probe atom, and can be highly useful when experimental data are not available or scarce. For example, a previous study by Timoshenko et al.[62] showed that ab initio XANES spectra provide excellent input data for training supervised machine learning models aimed at reconstructing metal catalyst structures from their experimental XANES. The current authors have also shown in a previous study[144] that an ensemble-learned algorithm to match experimental K-edge XANES spectra in the EELS Data Base to computed spectra can achieve nearly 80% accuracy in identifying the correct oxidation state and coordination environment. In addition, the data is associated with download options and programmatic analyses tools for each structure in the Materials Project database, thereby making it accessible to the broader materials science community. Furthermore, the MP web application enables users to select spectra from the database, upload experimental spectra data and predict the material composition using the matching tool. To date, this is the largest computed XAS dataset available and it is still expanding.

The paper is organized as follows; first we briefly describe the XAS computation methodology as implemented in the FEFF code, and thereafter the high-throughput framework used in the generation of the spectra. We then describe the data storage and dissemination details, followed by the technical validation of the computational methodology and the high-throughput framework.

4.2 Method

4.2.1 Theory

The K-edge XANES spectra were computed using the FEFF[64] code which employs the Green's formulation of the multiple scattering theory to compute the spectra[64]. The X-ray absorption μ is computed in a manner similar to Fermi's golden rule when written in terms of the projected photoelectron density of final states or the imaginary part of the one-particle Green's function, $G(r,r';E)$. In terms of the Green's function, $G(r,r';E)$, the absorption coefficient, μ, from a given core level c is given by ref. 15.

$$\mu = -\frac{1}{\pi} Im < c|\varepsilon \cdot rG(r,r';E)\varepsilon \cdot r|c > \tag{4.1}$$

with the Green's function, $G(r,r';E)$ given by

$$G(r,r';E) = \sum_{f} \frac{\psi_f(r)\psi_f(r')^*}{E - E_f + i\Gamma} \tag{4.2}$$

where ψ_f are the final states, with associated energies E_f and net lifetime Γ, of a one-particle Hamiltonian that includes an optical potential with appropriate core hole screening.

The FEFF code computes the full propagator G incrementally using matrix factorization and uses the spherical muffin-tin approximation for the scattering potential[145]. For a more detailed description, we direct the readers to the review paper by Rehr *et al.*[145]

4.2.2 High-throughput Workflow

For the high-throughput XAS spectra generation, we use the FEFF workflow [Fig. 4.1] available in the open source computational materials science workflow package Atomate[142]. Atomate provides a high-level interface to compose workflows using open source materials science softwares such as pymatgen[12] , FireWorks[11] and Custodian (https://github.com/materialsproject/custodian). Each FEFF calculation involves the following 3 steps:

- Selection of the absorbing site and the cluster of atoms to be included in the scattering calculations.
- Generation of the FEFF input files for each site and its surrounding atomic cluster.
- Execution of the FEFF binary on the generated input files.

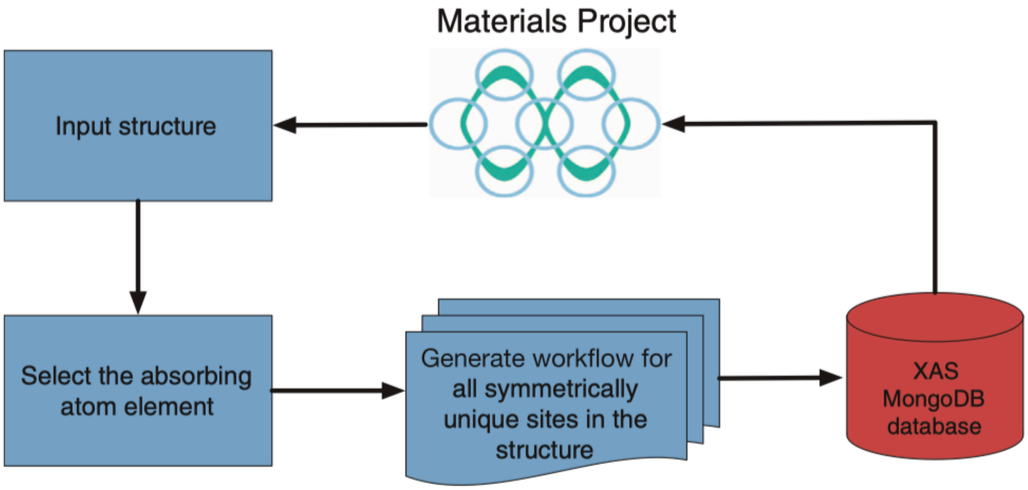

Materials Project
Input structure
Select the absorbing atom element
Generate workflow for all symmetrically unique sites in the structure
XAS MongoDB database

Figure 4.1: Schematic diagram of the high throughput framework employed in the generation of XAS spectra for the Materials Project.

As shown in Fig. 4.1, the workflow is initiated by importing a structurally optimized compound from MP. Each site in the downloaded structure is a possible absorbing center and

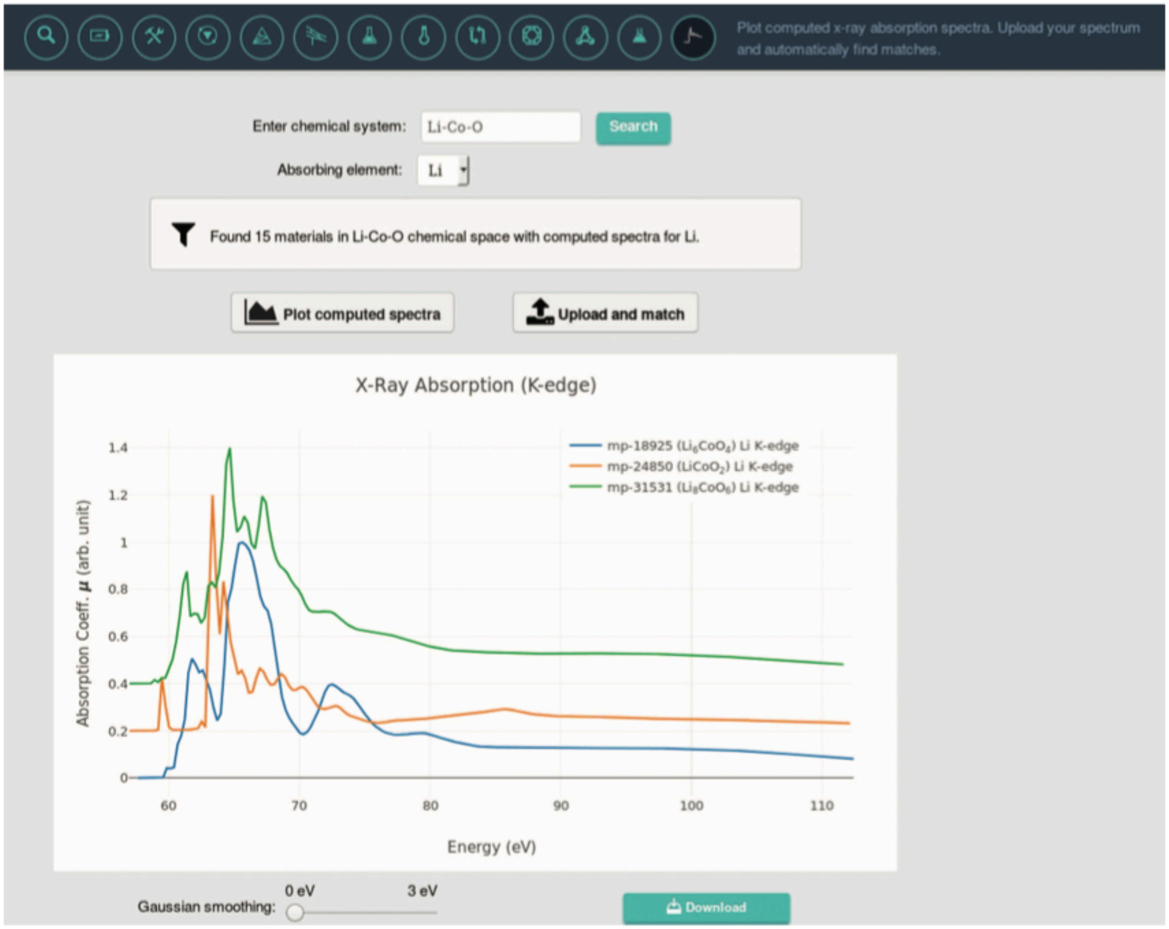
FEFF calculation sequence must be initiated for each site. However, the number of calculations can be reduced by considering only the symmetrically unique sites in the structure. The FEFF input files for each such symmetrically unique absorbing site are generated subsequently and the FEFF binary is invoked on each input set. In the final step, the computed spectra from each calculation is inserted into a MongoDB database and disseminated via the Material Project (https://materialsproject.org/) website.

4.2.3 Code availability

Except the for FEFF code, which is proprietary, all the other aforementioned packages used in the highthroughput XAS workflow are open source and can be found at https://github.com/ materialsproject and https://github.com/hackingmaterials/atomate.

4.3 Data Records

All the data described in this section can be found in the file, xas.json.tgz (Data Citation 1). The same data is also stored in the Materials Project database and is made freely available to the public. We also provide a user friendly web application called XAS Matcher (screenshot shown in Fig. 4.2) that enables user interaction with the computed data. The app can be reached at https://materialsproject.org/#apps/xas. Users can employ the app to search for computed XAS spectra, upload experimental spectra and find structures in the MP database whose computed spectra match that of the uploaded one. Details on spectra matching algorithm employed on Materials Project were published separately[141].



Plot computed x-ray absorption spectra. Upload your spectrum and automatically find matches.

Enter chemical system: Li-Co-O

Absorbing element: Li

Found 15 materials in Li-Co-O chemical space with computed spectra for Li.

Gaussian smoothing: 0 eV

Search
Plot computed spectra
Upload and match
X-Ray Absorption (K-edge)
mp-18925 (Li_6CoO_4) Li K-edge
mp-24850 ($LiCoO_2$) Li K-edge
mp-31531 (Li_8CoO_6) Li K-edge
Absorption Coeff. μ (arb. unit)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
60
70
80
90
100
110
Energy (eV)
0 eV
3 eV
Download

Figure 4.2: Screen shot of XAS Matcher web application. The web application is hosted at https://materialsproject.org/#apps/xas.

4.3.1 Data Representation

To date, spectra for more than half of the compounds(≈40000) in the Materials Project database are available, for all the symmetrically unique sites in each structure. Each structure dataset is stored in the database in the binary JavaScript Object Notation (BSON) format. The keys and respective descriptions are summarized in Table 4.1. Although the workflow yields separate spectra for each unique atomic site, the averaged absorption coefficient over all the sites in the structure with that element is presented on the MP website. This will facilitate comparison with experimental spectra, where the averaging over each element is unavoidable. However, the full data, e.g spectra for all unique sites, are available to the user for download and further analysis.

Table 4.1: Keys and their description for each spectra data JSON file.

Key	Data Type	Description
input_parameters	string	the FEFF input settings used in the computation of the spectrum.
xas_id	string	unique id for each spectrum, e.g. ‘mp-505011-28-XANES-K’.
spectrum_type	string	type of XAS e.g. ‘XANES’.
edge	string	absorption edge e.g. ‘K’.
mp_id	string	mp id of the structure
absorbing_atom	string	site index of the absorbing site in the structure
structure	string	the structure in dictionary format (can be loaded as a Structure object in pymatgen)
spectrum	float	array of shape (100, 6) where each codlumn means the following (in that order): Energy (eV), Energy with respect to the fermi level (eV), Wave number, μ (total absorption coefficient), μ_0 (the background absorption coefficient), x (normalized fine structure)

4.3.2 Data Download

The spectral data as well as the input parameters used for the calculations can be downloaded either directly from the Material Project website or using the REST Application Programming Interface (API) available in pymatgen[146]. Data can be downloaded for each element in the selected structure. The downloaded spectrum is provided in a tab separated file format and includes the spectral data for all the symmetrically unique sites of the selected element in the structure. The standard XAS data interchange (XDI) format[147] is also available for download, which can be directly imported into most existing XAFS data analysis programs[148] for further detailed analysis.

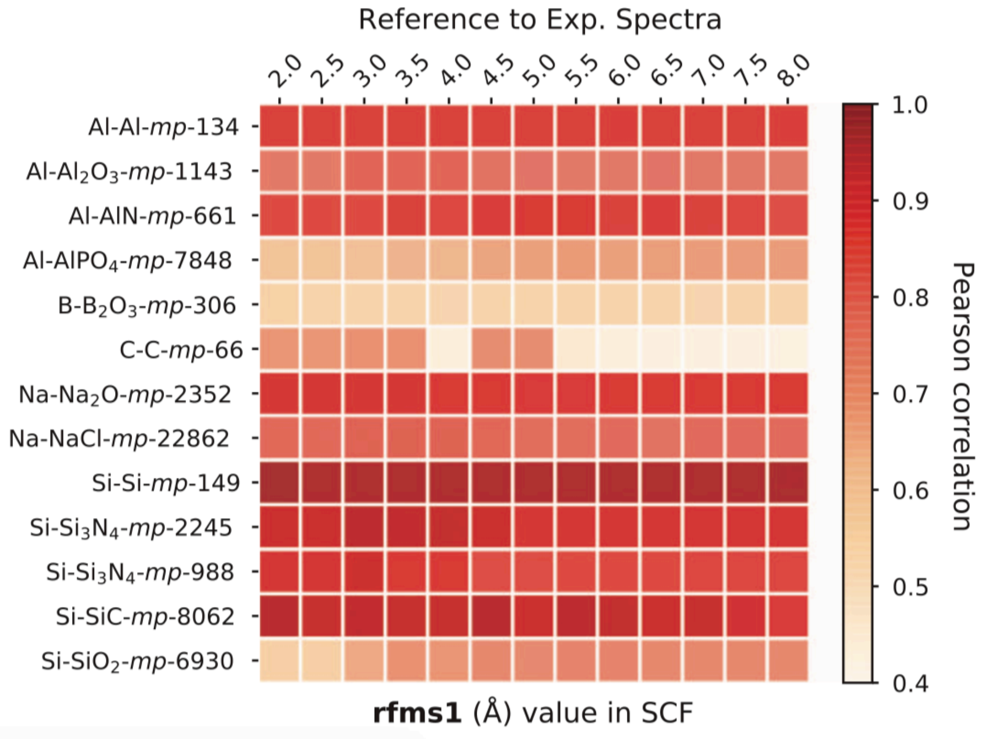


Figure 4.3: Benchmarking results of rfms1 parameter in the SCF card for K-edge XANES of various materials. Pearson correlation coefficients were calculated between spectra calculated at different rfms1 and the experimental reference provided by electron energy-loss spectroscopy (EELS) Data Base[54].

4.4 Technical Validation

4.4.1 Verification of the default parameter settings for the workflow

The workflow described above relies on the default FEFF input parameter settings to generate the K-edge XANES spectra in a high throughput fashion. In this section, we will briefly describe the major FEFF input parameters relevant to the calculation of the XANES spectra, the bench-marking procedure and sample validation cases against experimentally available XANES spectra.

FEFF9 is capable of achieving quantitative agreement with XAS experimental results with a minimal set of adjustable parameters. The development and implementation of parameter-free models within the FEFF9 code permit consistent calculations across different chemical systems and constitute the main advantage for high-throughput calculations. In the benchmarking process, we included 13 unique compounds and their corresponding high-quality K-edge XAS spectra available in the open EELS/XAS database[54], supplemented by 6 experimental XANES spectra of V_2O_5 , V_2O_3, VO_2, $LiNiO_2$, $LiCoO_2$, and NiO from previous studies[149,150]. Compounds included in the earliest commentary[64] of FEFF9 software were also evaluated. The benchmark compound dataset has a high structural diversity and covers a wide chemical space. Detailed benchmark information is provided in a previous publication[144]. For benchmark compounds that contain detailed structural information, we used structures from the Materials Project (https://materialsproject.org/) database that exhibit an optimized geometry with the same space group as the benchmark compound. For benchmark compounds without provided structural information, MP ground state structures with identical chemical compositions were used[144].

The following input fields in FEFF9 were subjected to convergence and optimization tests:

- *Self-consistent field (SCF)*: The SCF card controls FEFF automated self-consistent potential calculations. The self-consistent potential calculation is required in the XANES calculation for the Fermi level E_0 estimation. In the convergence test, we varied the number of atoms included in the self-consistent potential calculations through changing the rfms1 value from 2 Å to 8 Å at 1 Å interval.

- *Full multiple scattering (FMS)*: The FMS card is required in the XANES calculation as the multiple scattering (MS) expansion's convergence might not be stable in the XANES calculation[64]. To identify the effect of rfms field on XANES calculation results, we varied the rfms value from 3 Å to 11 Å at 1 Å interval.
- *EXCHANGE*: The EXCHANGE card specifies the exchange correlation potential model used for XANES calculation. The Hedin-Lundqvist self-energy is chosen as previously recommended for most applications[151].
- *COREHOLE*: The COREHOLE card is used for specifying how the core is treated during XANES calculation. The default choice in FEFF treats the core-hole interaction based on the Final State Rule (FSR), which could overestimate the strength of the core-hole and excludes the core-hole mixing effect[152]. To overcome these deficiencies and avoid possible break down of FSR for the L-shell metals[153], the random phase approximation (RPA) is used to approximate the core-hole interactions in our high-throughput K-edge XANES calculations.

Through the benchmarking study, a set of optimized FEFF parameters were determined to achieve the best balance between the computational cost and convergence performance. The Pearson correlation coefficient is used to compare spectra calculated using different parameters. The Pearson correlation coefficient between two same energy grid spectra, X_i and Y_i, is calculated using the following expression:

$$S_{Pearson}(X,Y) = \frac{\sum_{i=1}^{D}(X_i - \bar{X})(Y_i - \bar{X})}{\sqrt{(\sum_{i=1}^{D}(X_i - \bar{X})^2)(\sum_{i=1}^{D}(Y_i - \bar{Y})^2)}} \tag{4.3}$$

where X_i and Y_i are the corresponding absorption coefficients.

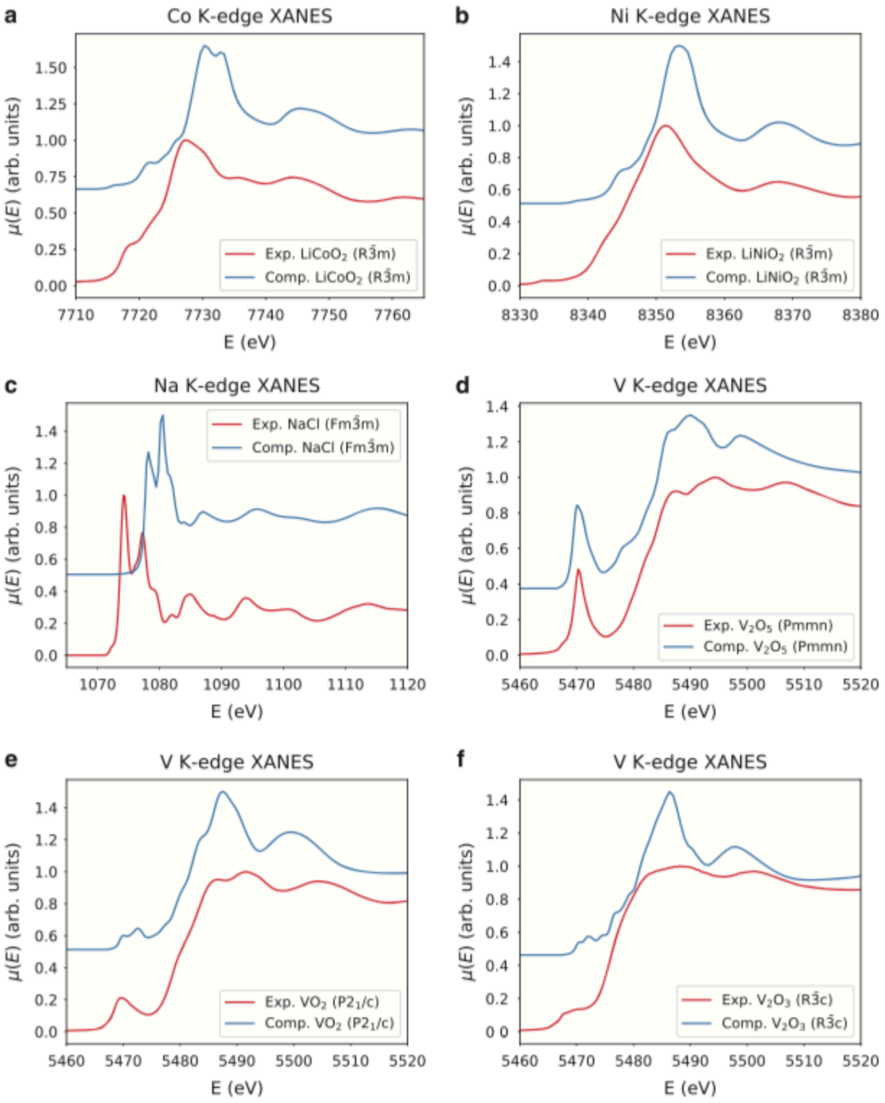


Figure 4.4: Sample comparisons of FEFF computed K-edge XANES spectra with the corresponding experimental ones for six different compounds. Computational spectra are shifted upwards by 0.5. (a) $LiCoO_2$, (b) $LiNiO_2$, (c) NaCl, (d) V_2O_5, (e) VO_2, (f) V_2O_3

We noticed that the Pearson correlation coefficients between FEFF computed spectra and experimental spectra obtained from EELS Data Base are above 0.85 in general (see Fig. 4.3). For C and B_2O_3, the FEFF-computed spectra are not in good agreement with experimental spectra. Possible solutions include the adoption of other higher-level real-space full-potential multiple scattering theory or first principles approaches[154], which are not ideal for high-throughput implementation due to their high computational cost. Figure 4.4 depicts some sample comparisons between the computed and the experimental K-edge XANES spectra. We note that the computed spectra match with that of the experimental ones only up to a constant shift in the energy. The computed K-edge XANES spectra of vanadium oxides given in Fig. 4.4d–f show a strong change in their first peak intensity. Reasonably good agreement between computational and experimental spectra was found.

4.5 Usage Notes

We present a database of K-edge XANES spectra computed using FEFF. The data is made freely available to all researchers via the Materials Project(www.materialsproject.org). Users can also download the data using the REST API that is part of pymatgen. All the codes used to create the high throughput are made freely available at Github (https://github.com/materialsproject and https://github.com/hackingmaterials/atomate). We hope that the users will find the data to be useful and will find novel ways to employ the data to accelerate their research. One such use case would be using machine learning techniques to predict structures from the experimentally measured spectra.

For users of FEFF and the spectra resulting in this study, it should be noted that K-edge XANES spectra computed by FEFF are more accurate for the investigation of elements in the periodic table up to the fifth-row. For excitations in heavier elements, e.g., the rare earth elements and 5*d* elements, L-edge XANES spectra are primarily used. FEFF is also applicable for the simulation of L-edge XANES spectra, though in certain cases[155–157] ground-state DFT methodologies need to be used for better agreement between computed spectra and experimental results. A detailed study of high-throughput FEFF calculation and implementation of L-edge XANES is currently being conducted by our research group. Furthermore, the analysis of

XANES is recommended for the identification of oxidation state and coordination chemistry of the absorbing atom[158]. We note that the quantitative accuracy of XANES calculations is not comparable to EXAFS in identification of the distances, coordination number, and species of the neighbors of the absorbing atom. The accurate and precise interpretation of EXAFS is routinely conducted coupled with well-established software packages[159] using the FEFF calculated EXAFS. The FEFF calculated K-edge EXAFS of all the materials available in the Materials Project database is underway, and a significant portion will be released in parallel with this publication.

Chapter 4, in full, is a reprint of the material “High-throughput computational X-ray absorption spectroscopy” as it appears in Scientific Data, Kiran Mathew, Chen Zheng, Donald Winston, Chi Chen, Alan Dozier, John J. Rehr, Shyue Ping Ong, and Kristin A. Persson, 2018, 5, pp 180151. The dissertation author was the primary investigator and author of this paper.

Chapter 5 Automated Generation and Ensemble-Learned Matching of X-ray Absorption Spectra

5.1 Introduction

X-ray absorption spectroscopy (XAS) is a widely used technique in the study of the properties, physical states and local environments of materials.[160–162] When incident X-ray photons with energy greater than the binding energy are absorbed by an atom, a core-level electron is removed from its quantum level. In XAS, the absorption coefficient, $\mu(E)$ is measured as a function of X-ray energy E. Detailed descriptions of X-ray absorption theory and equation have been included in many excellent books and review papers.[163,164]

The X-ray absorption fine structure (XAFS) is typically divided in to two regimes: X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS).[165] The XANES is a fingerprint of the oxidation states and coordination chemistries of the absorbing atom. Quantitative XANES analyses are typically difficult and are usually conducted in combination with principle component analysis or least-squares fitting. The EXAFS provides local atomic structure information, which can be extracted via coupling with theoretically calculated XAFS spectra using well-established software packages.[166] One of the main challenges of interpreting XANES and EXAFS lies in *obtaining reference spectra to fit the unknown spectra*; measuring XAFS spectroscopy experimentally is laborious and time-consuming, requiring X-ray beams of finely tunable energy that are accessible only through synchrotron radiation facilities.[164] To the authors' knowledge, open reference database usually contains at most hundreds of XAS spectra. For example, the Electron Energy Loss Spectroscopy (EELS) database[167] initiated in the 1990s contains 271 spectra, but only 21 of which are XAS spectra and 17 of which are K-edge spectra. EELS is theoretically equivalent to X-ray absorption[168] under common acquisition conditions, but is of lower quality in terms of signal to noise and energy resolution. Most XAS data are available only via publications in the literature, which cannot be extracted easily for comparison.

In recent years, theoretical calculations of XAFS have become more accurate and accessible due to the successful development of ab initio codes, such as the FEFF program[145,169], as well as advances in computing power. In this work, we will discuss the development of a high-throughput framework to generate a reference XAS database (XASdb) for all materials in the Materials Project[170] database. This framework combines the power of the Python Materials Genomics (pymatgen) materials analysis library[171] with the FireWorks workflow management software[172] to carry out hundreds of thousands of XAFS calculations using the FEFF9 code.[169] This framework has been implemented in the Atomate package.[142] More importantly, we have developed a novel automated XANES spectra matching algorithm that leverages ensemble learning techniques to identify similar XANES spectra from our computed reference XASdb. We believe the combination of the XASdb with these machine-learned spectra matching tools will be an invaluable resource to the materials research community by greatly enhancing the efficiency at which experimental XAS spectra can be analyzed. It should be noted that this work primarily focuses on common K-edge XANES spectra; higher edge XANES and EXAFS computations and analysis are currently ongoing and will be discussed in future publications.

5.2 Results and discussion

We have selected the latest version (v9) of the popular FEFF program as our software of choice in this work. FEFF is a program for *ab initio* multiple scattering calculations of XAFS and various other spectra for clusters of atoms. This choice is motivated by three factors: (i) FEFF-computed spectra has been shown to yield excellent agreement with experimentally measured spectra in a broad range of studies;[173–175] (ii) FEFF calculations are relatively inexpensive compared to other approaches for computing XAS spectra (e.g., a typical FEFF calculation takes < 1 hour on a single node, while multi-day, multi-core calculations are necessary for DFT-based spectra calculations); and (iii) FEFF requires minimal adjustable parameters. These three advantages make FEFF an ideal candidate for automation to generate XAS spectra across a broad range of chemistries. A key step in any automation framework is benchmarking of computational parameters for convergence and accuracy. The benchmarking dataset and criterion details are included in the methods section. The Pearson correlation coefficient, as given by the following expression, is used as the benchmarking criterion.

$$S_{Pearson}(X,Y) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{D}}(X_{\mathrm{i}} - \bar{X})(Y_{\mathrm{i}} - \bar{Y})}{\sqrt{(\sum_{\mathrm{i}=1}^{\mathrm{D}}(X_{\mathrm{i}} - \bar{X})^2)\,(\sum_{\mathrm{i}=1}^{\mathrm{D}}(Y_{\mathrm{i}} - \bar{Y})^2)}}, \tag{5.1}$$

where X_{i} and Y_{i} represent the absorption coefficients of two spectra on the same energy grid. The value of $S_{Pearson}$ can range from -1 to 1, with a value of 1 being a perfect match. Used in this context, the Pearson correlation coefficient is a similarity metric, *i.e.,* it measures the degree of similarity between two spectra.

We have tested the convergence of the FEFF calculated spectra with respect to four parameters: the radius of the cluster considered in the full multiple scattering calculation (**SCF rfms1**), the total number of multiple-scattering paths considered (**FMS rfms**), the exchange-correlation potential (**EXCHANGE**) and the treatment of the core (**COREHOLE**) (see Methods for a detailed description of the FEFF input file).

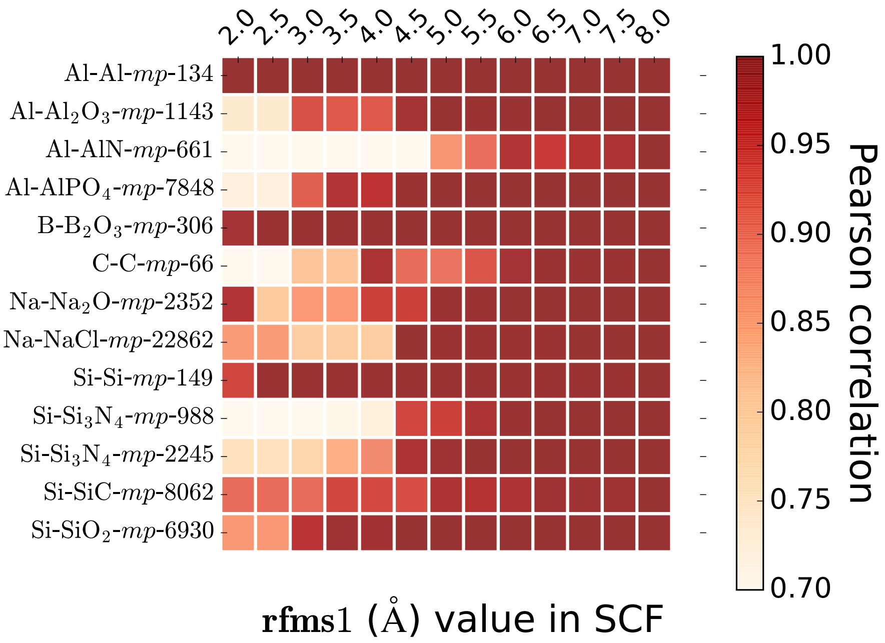


Figure 5.1: Benchmarking results of **rfms1** parameter in the SCF card for K-edge XANES of various materials. The **rfms1** parameter specifies the radius of the cluster considered for the full multiple scattering during self-consistent potential calculations. Pearson correlation coefficients were calculated between spectra calculated at different **rfms1** and the reference calculated at **rfms1** = 8.0 Å.

The **SCF rfms1** was varied from 2 Å to 8 Å, and the spectrum at the highest value (8 Å) was set as the reference for each material. Figure 5.1 shows the computed Pearson correlation coefficients between spectra computed at lower **rfms1** and the reference. We find that the computed spectra are converged ($S_{Pearson} > 0.95$) at around rfms1 = 6 Å for all material, though the Al K-edge for AlN is converged only for rfms1 = 6.5 Å. Given that the computational cost increases substantially for rfms1 > 7 Å (see Figure C.1), we have chosen **rfms1 = 7** Å **as the default setting for SCF in the high-throughput XANES computations**.

The **rfms** field in the FMS card was varied from 3.0 Å to 11.0 Å at 1.0 Å intervals, and the spectrum at the highest value (11 Å) is set as the reference for each material. We find that the computed spectra are converged ($S_{Pearson} > 0.95$) at around rfms = 9 Å for all materials (see Figure C.2). Since the computational cost increases substantially for rfms > 9 Å (see Figure C.3), we have chosen **rfms = 9** Å **as the default setting for FMS in the high-throughput XANES computations**.

In FEFF9, two approximations of the core-hole potentials have been implemented, i.e., a fully screened potential based on the final-state rule (FSR) and a linear random-phase-approximation (RPA) screening. Systematic reviews of these two approaches have been done by Rehr *et al*.[176] We evaluated the performance of all three core-hole options in FEFF9 on the computed K-edge XANES. As shown in Figure C.4(a), spectra obtained using both the FSR and RPA are in much better agreement with experimental results than ones without core-hole treatment. The spectra computed without a core-hole treatment lack the edge enhancement observed in the experiments. In general, spectra obtained using FSR and RPA are similar (Figure C.5). We have chosen **RPA screening as the default setting for the high-throughput XANES computations as the FSR might breakdown for the L-shell metals**.[177]

Similar evaluations of the EXCHANGE card options reveal that the default Hedin-Lundquist model is the best option (see Figure C.6).

5.2.1 Sensitivity of computed XAS spectra to lattice parameters

The FEFF code uses a self-consistent DFT calculation of the Fermi-energy based on the real-space Green's function (RSGF) approach with muffin-tin potentials for a given lattice structure. Comparing to the full-potential calculations, we find that the FEFF calculation of the densities of states is typically in fairly good agreement with DFT for many materials. In the Materials Project, the Perdew-Berke-Ernzerhof (PBE)[103] generalized gradient approximation functional was used as the default for all relaxation calculations. As it is well known that PBE leads to systematic errors of up to 5% in the lattice parameters (with a tendency to overestimate),[178–181] we tested the sensitivity of computed XANES spectra to ±5% changes in the lattice parameters. The results are shown in Figure 5.2.

We find that the Fermi energy level of the spectrum is sensitive to the lattice parameter variation [Figure 5.2(a)]. The Fermi energy level shifts towards lower energy as the lattice parameter increases, while the spacing of the spectral features contracts at the same time. An example for Na K-edge of Na_2O is shown in Figure 5.2(b), and additional examples are available in Figure C.7.

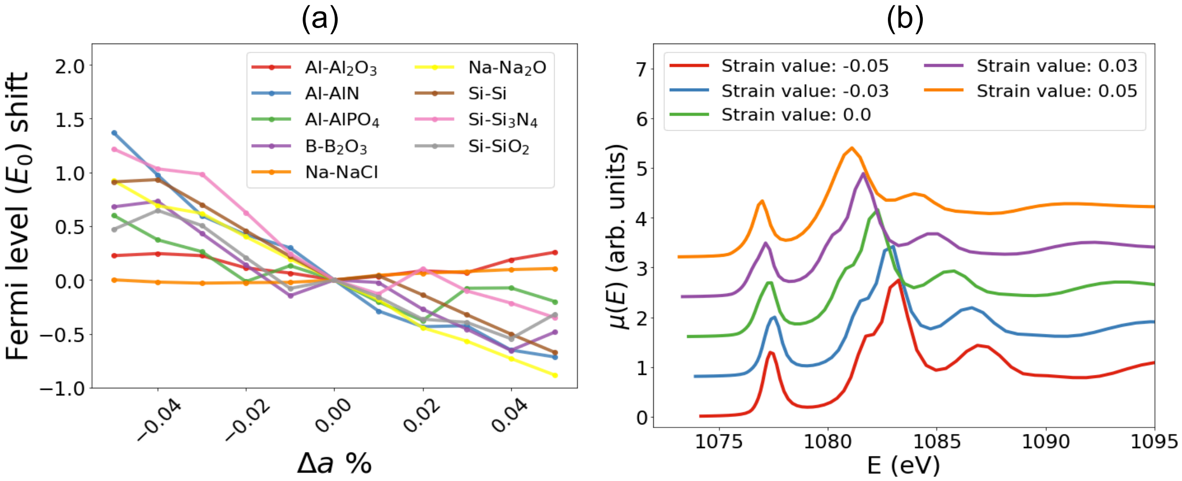


Figure 5.2: (a) Relationship between the Fermi energy level of K-edge XANES and **a** lattice parameter changes. Fermi energy levels of the unstrained structures are used as references. (b) Visualization of Na K-edge XANES spectra in Na_2O (mp-2352) calculated with different applied strain values.

A portion of the Fermi energy shift can be attributed to the artifacts of the FEFF's potential approximation model (see Figure C.9). Nevertheless, the shape of the spectra remains unchanged. While different corrections to eliminate the artificial component of the dependence have been reported, these approaches are not amenable to a high-throughput approach. Here, we note that due to the approximations used in FEFF, we need to calibrate the Fermi level with experimental spectra. Therefore, a pure energy shift only translates to an energy calibration value in the post processing.

In summary, the PBE-relaxed structures from the Materials Project can be used as the input for high-throughput XANES calculations, even though there are other functionals[183,184] that may provide better lattice parameters estimates.[185–188]

5.2.2 Workflow & Database

Using the high throughput parameters outlined above, we developed a high-throughput workflow for FEFF XAS calculations within the open source computational materials science workflow package Atomate[142]. Atomate provides a high level interface to compose workflows using the widely used open source materials science software such as Pymatgen[171], FireWorks[172] and Custodian. The proposed default FEFF9 parameters have been implemented as "input sets" in Pymatgen[171], which ensures reproducible and automated generation of standardized input files for any material. The compounds used in the high-throughput spectra generation were obtained from the Materials Project database[170]. For each compound, the K-edge XANES spectrum was computed with each symmetrically unique site in the structure as the absorbing atom.

All computed spectra, as well as accompanying meta-data (e.g., input structure, absorbing atom, materials project id, etc.), are stored in a MongoDB database for on-demand querying and retrieval of data. So far, K-edge XANES spectra have been computed for more than 40,000 unique materials in the Materials Project database, which amounts to over 800,000 K-edge spectra. This is by far the largest repository of XANES spectra in the world, and is growing rapidly. Future plans include the calculation of XANES for L, M, and N shells as well as EXAFS spectra.

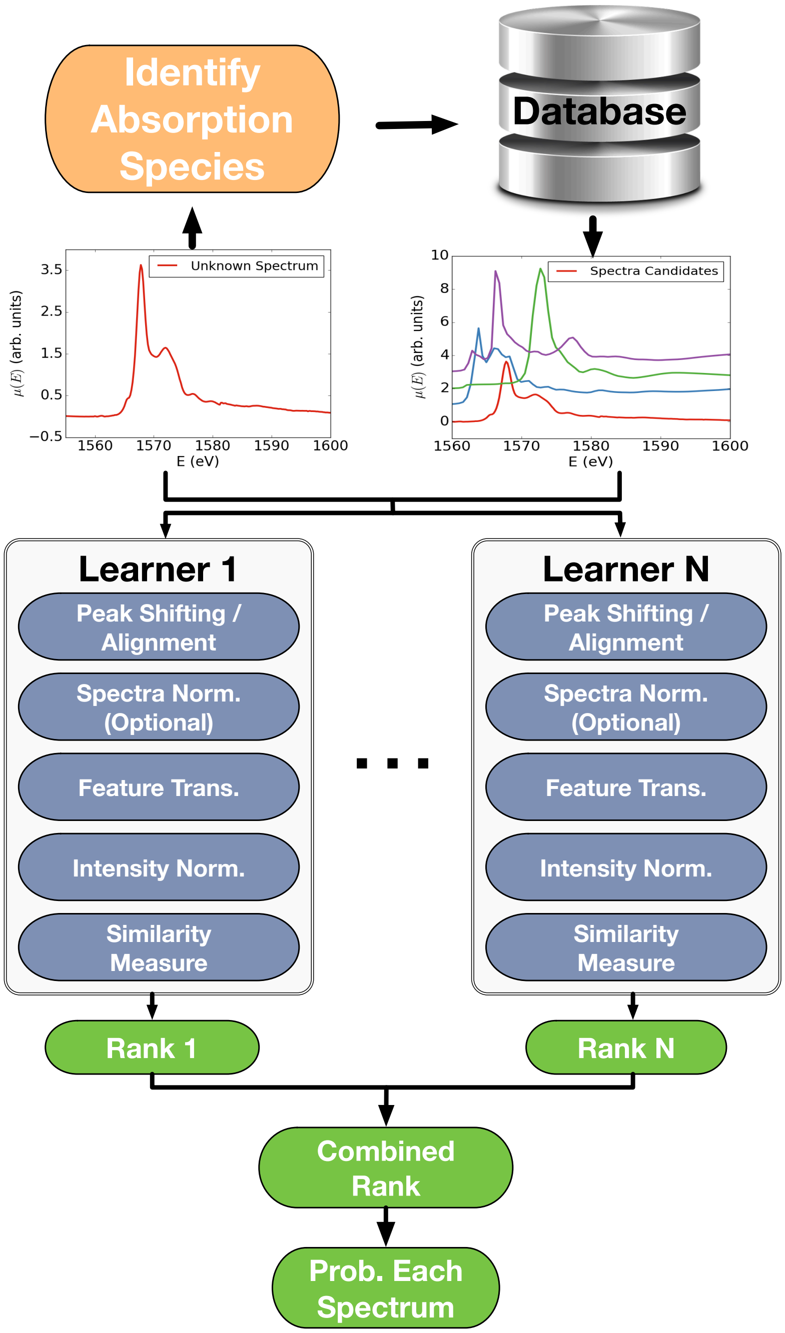


Figure 5.3: Workflow schema of the Ensemble-Learned Spectra IdEntification (ELSIE) algorithm. The algorithm consists of two steps. In the first step, the absorption species is identified and used to narrow down the candidate computed reference spectra. In the second step, the spectral matching ensemble yields a rank-ordered list of computational spectra according to similarity with respect to the target spectrum.

5.2.3 Spectra Matching using Ensemble Learning

To extract the most utility and power from the XASdb, we have developed a novel Ensemble-Learned Spectra IdEntification (ELSIE) algorithm that allows for rapidly identification of matching spectra for any experimental XAS spectra. The main goal of spectral matching is to obtain a list of compounds (the "hit list") whose spectra are most similar to that of the target spectrum. The success and failure of matching is defined by the characteristics of the spectrum. In the case of XANES spectra, the relevant information to be extracted is the coordination environment and oxidation state of the absorbing atom. As multiple materials can have atoms in the same oxidation state and coordination environment, we define the matching to be successful if the correct coordination environment and oxidation state are within the top entry.

The ELSIE algorithm uses the ensemble method to improve the robustness of XAS identification. In ensemble learning, the core concept is the combination of multiple weak learners to achieve superior performance. It relies on the assumption that each weak learner is better than a random guess and each weak learner captures different aspects of the problem. At the core of the algorithm is the process of building individual weak learners. Taking inspiration from the spectra matching algorithms for Raman spectroscopy[189] and other spectra[190,191], we broke down the problem of matching XAS spectra into two main steps, namely preprocessing and similarity computations. We define each weak learner to be a combination of a preprocessor (a specific series of preprocessing steps) with a similarity metric. Figure 5.3 provides an overview of the ELSIE algorithm (see Methods section for the details on the construction of the ELSIE algorithm).

We evaluated the ELSIE algorithm using 13 XANES spectra from EELSDb [Table C.1], supplemented by 6 high quality experimental XANES spectra of V_2O_5, V_2O_3, VO_2, $LiNiO_2$,

$LiCoO_2$, and NiO from previous studies.[149,150] The inclusion of this latter dataset is motivated by our desire to improve the diversity of the test data, especially with regards to transition metal species.

The first step is to narrow down the candidate computed reference spectra by the absorption element (A). Though this information is usually known *a priori*, the characteristic XAS absorption edge energy follows a power law with the atomic number,[164,165] which leads to clearly separated energy ranges. Hence, we can identify the absorption element with 100% accuracy by comparing the energy range of the target spectrum to tabulated X-ray absorption edge data.[192]

Once the absorbing element A is identified, the computed spectra of all materials within the same chemical system are queried from the XASdb. For example, for the Al *K*-edge of Al_2O_3, we include the Al *K*-edge spectra of all Al and Al_xO_y materials as reference spectra. We excluded compounds with energy above hull (E_{hull}) larger than 100 meV/atom since they are not likely to be stable.[193] For C K-edge XANES of the diamond structure ($Fd\bar{3}m$), we relaxed the constraint to 200 meV/atom as the corresponding entry (mp-66, diamond) has an E_{hull} of 136 meV/atom. It should be noted that though the individual absorption spectrum for each symmetrically distinct site was computed for all crystal structures in the Materials Project database, the reference spectra used for comparison with the target spectra are constructed by summing these individual spectra taking into account the site multiplicities.

To evaluate the overall performance of ELSIE, we looked at three key metrics: (i) whether the correct structure is within the top 5 ranked computed spectra, (ii) whether the top ranked entry has the absorbing species in the correct oxidation state, and (iii) whether the top ranked entry has the absorbing species in the correct coordination environment, i.e., coordination number and geometry. Where the exact structural information is not available (e.g., in the experimental spectra from EELSdb), it is assumed that those spectra correspond to the ground state structures in the Materials Project database with the same chemical composition. It should also be noted that some reference materials may have the same element in multiple oxidation states and coordination environments. Therefore, the application of metrics (ii) and (iii) merely indicates whether at least one of the distinct sites in the top entry have the correct oxidation state and coordination environment. The results are summarized in Table 5.1.

Of the 19 test spectra, we find that the correct structure is within the top 5 ranked structures for 11 systems, i.e., only 57.9% accuracy. However, the correct oxidation state and coordination environment are in the top entry for 16 and 15 systems, i.e., accuracies of 84.2% and 78.9%, respectively. The best coefficient α is found to be 0.01. Given that XANES is a technique primarily used to extract oxidation state and coordination environment information, these results are a major validation of the effectiveness of the ELSIE matching algorithm.

To emphasize the effectiveness of the ensemble approach, we also performed the same benchmark using a single learner utilizing just the sigmoid squashing function and cosine similarity measure on spectra that have been pre-normalized with respect to summed intensity. The ELSIE algorithm outperforms the single learner approach by **15.8%** in identifying both the correct oxidation state and coordination environment.

We will now illustrate the performance of our spectral matching algorithm with a few case studies on diverse chemistries. For all spectra, we have confined our comparison to the energy range from -10 eV to 45 eV from the absorption edge, which is the region typically referred to as XANES.

Table 5.1: Performance of ELSIE algorithm on 19 test spectra

Formula	**Space Group**	**Absorbing Species**	**Correct Structure within Top 5 Rank?**	**Correct Oxidation State in Top Entries?**	**Correct Coordination Environment in Top Entries?**
SiO_2	$P3_221$	Si	No	Yes	Yes
Si	$Fd\bar{3}m$	Si	Yes	Yes	Yes
$AlPO_4$	$I\bar{4}$	Al	No	Yes	Yes
SiC	$F\bar{4}3m$	Si	No	Yes	Yes
Al_2O_3	$R\bar{3}c$	Al	Yes	Yes	Yes
Al	$Fm\bar{3}m$	Al	Yes	Yes	Yes
Na_2O	$Fm\bar{3}m$	Na	Yes	No	No
C	$Fd\bar{3}m$	C	No	Yes	No
B_2O_3	$P3_221$	B	Yes	No	No
Si_3N_4	$P31c$	Si	Yes	Yes	Yes
Si_3N_4	$P6_3/m$	Si	Yes	Yes	Yes
AlN	$P6_3mc$	Al	Yes	Yes	Yes
NaCl	$Fm\bar{3}m$	Na	Yes	Yes	Yes
V_2O_5	$Pmmn$	V	No	Yes	No
VO_2	$P2_1/c$	V	No	Yes	Yes
V_2O_3	$R\bar{3}c$	V	No	Yes	Yes
$LiNiO_2$	$R\bar{3}m$	Ni	No	No	Yes
NiO	$Fm\bar{3}m$	Ni	Yes	Yes	Yes
$LiCoO_2$	$R\bar{3}m$	Co	Yes	Yes	Yes

Case study 1: Main group metals

Figure 5.4(a) and (b) shows the ELSIE spectral matching results of the Al K-edge XANES of α-Al_2O_3 and Na K-edge XANES of NaCl, respectively. For both target spectra, the correct

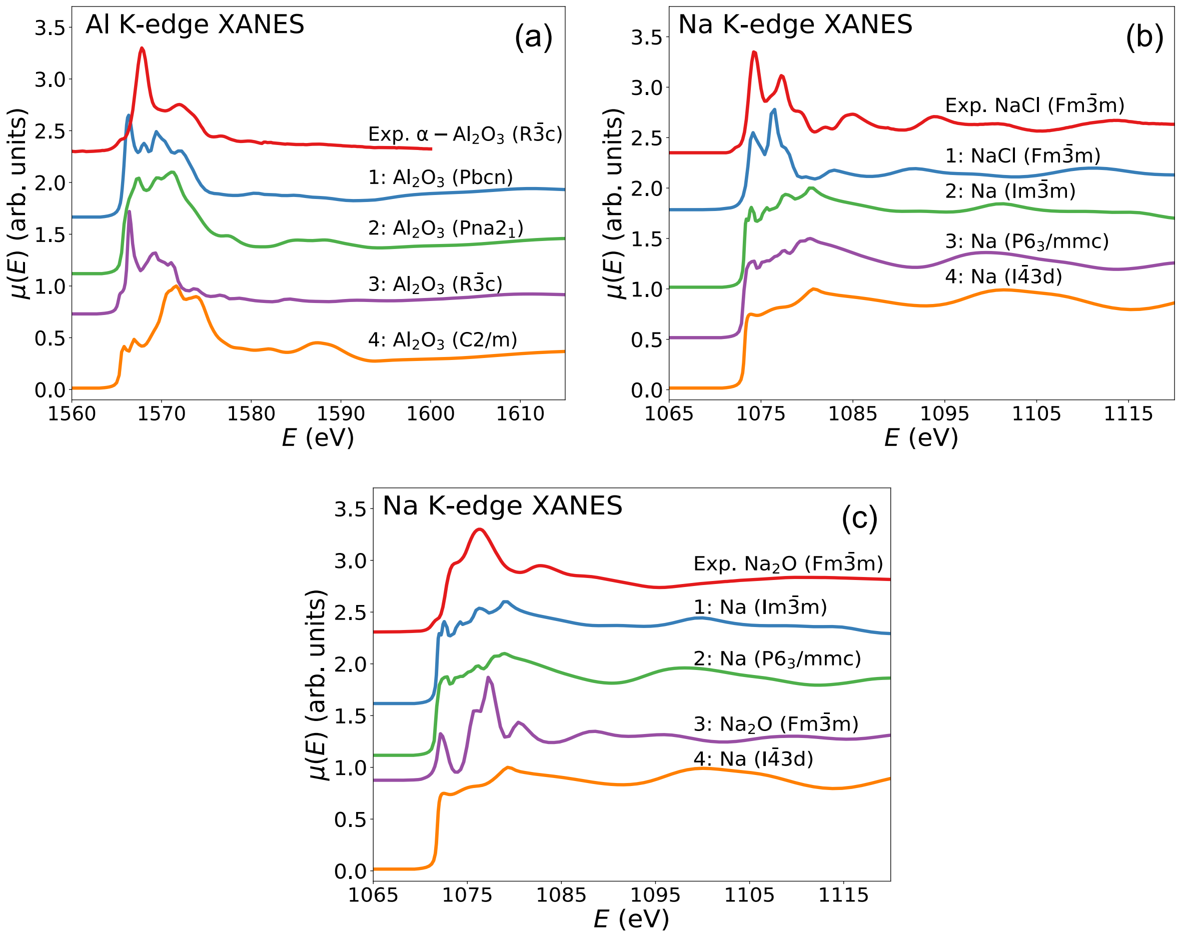


Figure 5.4: Results of the similarity ranking returned by the ELSIE matching algorithm on (a) Al K-edge XANES of α-Al_2O_3 entry; (b) Na K-edge XANES of NaCl; and (c) Na K-edge of Na_2O. Detailed information about the retrieved compounds can be found in the Materials Project website, (a) Al_2O_3 ($Pbcn$, mp-1938), Al_2O_3 ($Pna2_1$, mp-2254), Al_2O_3 ($R\bar{3}c$, mp-1143) and Al_2O_3 ($C2/m$, mp-7048), (b) NaCl ($Fm\bar{3}m$, mp-22862), Na ($Im\bar{3}m$, mp-127), Na ($P6_3/mmc$, mp-10172) and Na ($I\bar{4}3d$, mp-567772) and (c) Na ($Im\bar{3}m$, mp-127), Na ($P6_3/mmc$, mp-10172), Na_2O ($Fm\bar{3}m$, mp-2352) and Na ($I\bar{4}3d$, mp-567772), in decreasing similarity order.

oxidation states and coordination environments are found in the top candidates. Furthermore, we may observe that our proposed peak shifting approach is effective in aligning the target and reference spectra.

Figure 5.4(c) shows a notable case – the Na K-edge of Na_2O – where the ELSIE algorithm fails. Here, the ELSIE algorithm returns elemental Na as the top ranked result, as opposed to Na_2O. The main reason for this failure is that the FEFF computed spectra is not in good agreement with experimental spectra (see Figure C.7 for this and a few other examples). Possible solutions include the use of real-space full potential multiple scattering theory or other first principle approaches.[194] For Na_2O in particular, we find that the experimental Na K-edge XANES of Na_2O is more similar to the computed Na K-edge XANES of Na_2CO_3 [Figure C.7(c)], which may indicate possible contamination by the atmosphere in experiments.

Case study 2: Transition metal oxides

Figure 5.5 shows the ELSIE spectra matching results of the Ni K-edge XANES in NiO, Co K-edge XANES in $LiCoO_2$. From Figure 5.5(a), we note that although the computed peak positions and amplitude are not in great quantitative agreement with the experimental measured spectra, the ground state NiO entry is nevertheless returned as the top ranked candidate. In particular, the small Ni 1s-3d peak at 8332 eV in the experimental Ni K-edge XANES of NiO is not present in the FEFF calculated spectra. There is, however, a small peak at 8337 eV in the FEFF calculated spectra, which we believe is the Ni 1s-3d peak. The inaccuracy in the position of the peak may be due to the muffin tin approximation used in FEFF.

For $LiCoO_2$ [Figure 5.5(b)], the ground state structure of $LiCoO_2$ ($R\overline{3}m$) is among the top five entries. All Co^{3+} ions in the top entry ($Li(CoO_2)_2$) are in octahedral coordination, i.e., the same coordination environment of Co^{3+} ions in $LiCoO_2$ ($R\bar{3}m$). We may therefore conclude that the ELSIE algorithm performs satisfactorily in both instances.

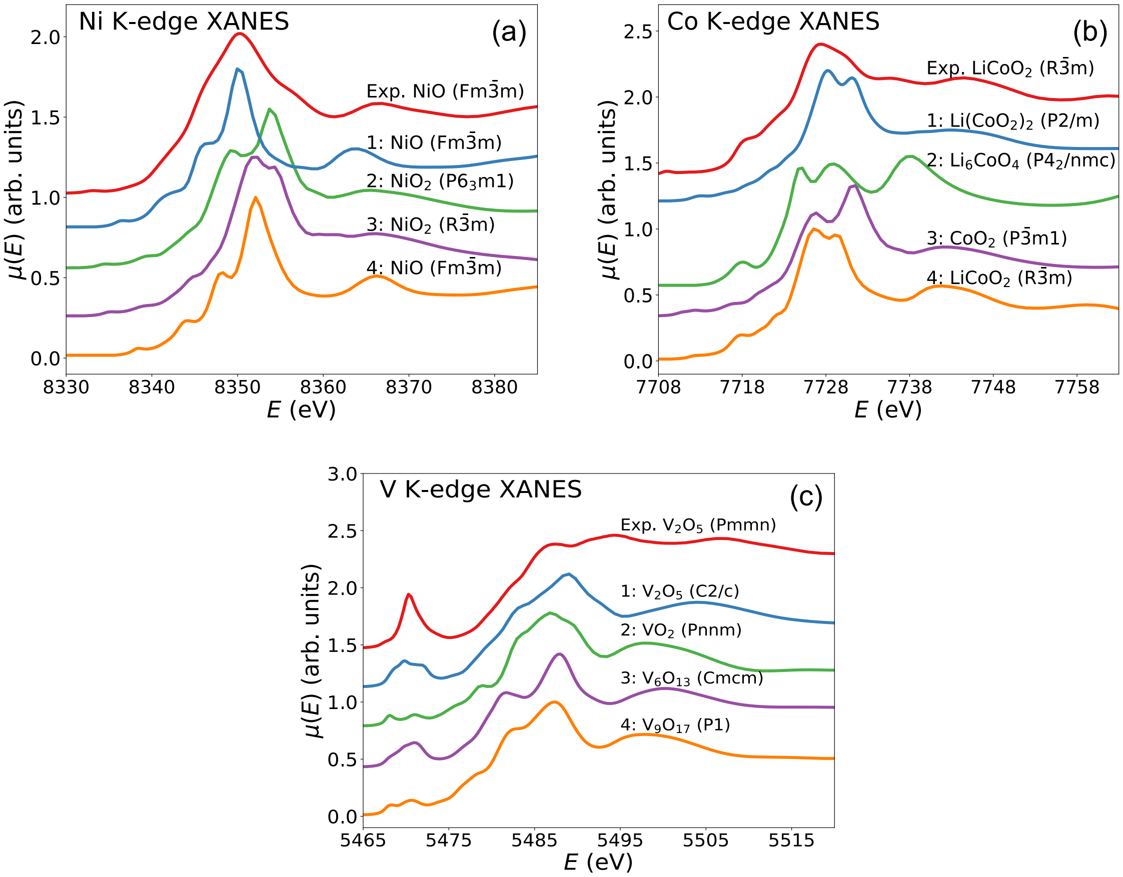


Figure 5.5: Results of the similarity ranking returned by the ELSIE matching algorithm on (a) Ni K-edge XANES of NiO; (b) Co K-edge XANES of $LiCoO_2$; and (c) V K-edge of V_2O_5. Detailed information about the retrieved compounds can be found in the Materials Project website, (a) NiO ($Fm\bar{3}m$, mp-19009), NiO_2 ($P6_3m1$, mp-543096), NiO_2 ($R\bar{3}m$, mp-25593) and NiO ($Fm\bar{3}m$, mp-715434), (b) $Li(CoO_2)_2$ ($P2/m$, mp-553952), Li_6CoO_4 ($P4_2/nmc$, mp-18925), CoO_2 ($P\bar{3}m1$, mp-714976) and $LiCoO_2$ ($R\bar{3}m$, mp-24850), and (c) V_2O_5 ($C2/c$, mp-542844), VO_2 ($Pnnm$, mp-714880), V_6O_{13} ($Cmcm$, mp-715617) and V_9O_{17} ($P1$, mp-716723), in decreasing similarity order.

Figure 5.5(c) shows the ELSIE spectra matching results for the V K-edge of V_2O_5 ($Pmmn$). The ELSIE algorithm fails to retrieve the correct square-pyramidal coordination environment of V^{5+} in V_2O_5 ($Pmmn$). Indeed, vanadium ions in the top five matches returned by the ELSIE algorithm are in octahedral coordination. Here, the relative similarity of the V K-edge spectra for the different V oxidation states and coordination environments seems to be the key issue. Further structural refinement based on EXAFS simulations therefore becomes critical, which will be available in the XASdb in the near future.

In conclusion, we have demonstrated the development of a large database for XAS using high-throughput FEFF calculations. Parameter benchmark results indicate that the overall quality of the FEFF9 calculations with default input parameters is in quantitative agreement with experiments, which is adequate for comparison purposes. We developed a novel spectra-matching algorithm – the Ensemble-Learned Spectra IdEntification (ELSIE) algorithm – that enables the rapid matching of computed reference spectra to any target spectra. The ensemble learning approach far outperforms any single approach based on a pre-defined set of preprocessing and similarity metric; outstanding ~84% and ~79% accuracies in identifying the correct oxidation state and coordination environment are demonstrated based on a diverse test set comprising 19 experimental XANES spectra. The XASdb with the ELSIE algorithm has been integrated into a web application in the Materials Project, providing an important new public resource for the analysis of XAS to all materials researchers, and the ELSIE algorithm itself has been made available as part of *veidt*, an open source machine learning library for materials science.

5.3 Methods

Benchmarking details

Robust, well-defined datasets are necessary for any benchmarking exercise. We have used the existing high quality K-edge XAS spectra available in the open EELS Data Base (EELSDb)[167] as reference data, and matched them with the corresponding materials in the Materials Project[170] using the Materials API[195] and pymatgen[171]. For materials in the EELSDb without structural information, ground state structures with identical chemical compositions in the Materials Project were used. For spectra in EELSDb taken using the same materials, we selected one and adopted it in our study. Table C.1 summarizes the 13 unique materials used in this work.

The FEFF software calculates X-ray absorption spectra using the real-space Green's function formulation of the multiple scattering theory.[145] The X-ray absorption μ is written in terms of the imaginary part of the one-particle Green's function $G(r, r'; E)$, which incorporate both the inelastic losses and other quasiparticle effects. In terms of $G(r, r'; E)$, μ is given by:

$$\mu = -\frac{1}{\pi} Im\langle c|\hat{\epsilon} \cdot rG(r, r'; E)\hat{\epsilon} \cdot r'|c\rangle \theta_{\Gamma}(E - E_F), \tag{5.2}$$

where θ_{Γ} is a broadened step function at the Fermi energy E_F. This yields a unified treatment of EXAFS and XANES. The treatment of X-ray absorption can then be separated into atomic and scattering parts, i.e., $G(r, r'; E) = G^c(r, r'; E) + G^{sc}(r, r'; E)$. The exact result of $G^{sc}(r, r'; E)$ is given by the full matrix inverse, or equivalently, a sum over all multiple-scattering paths. [196] For the XANES calculation, FEFF implements the full multiple scattering technique, which includes the contributions from all orders of scattering within a cluster containing the absorber and scatterers. The FEFF code also incorporates a GW-based self-energy based on the Hedin-Lundqvist plasmon-pole model which includes effects of electron-electron interactions such as mean-free paths and self-energy shifts. This method has been well tested and is usually a good approximation for EXAFS and reasonable for XANES. FEFF includes a screened corehole and gives results for excitonic enhancements comparable to GW/BSE calculations in many materials. FEFF can also incorporate Debye-Waller factors using correlated-Debye or more advanced models. Further details on the FEFF code and its theoretical foundations can be found in ref 11 for interested readers.

In the FEFF input file, parameters are specified in control “cards”. The following parameters in FEFF were tested for convergence.

i. **Self-consistent field (SCF)**: The **rfms1** field in the SCF card specifies the radius of the cluster considered in the full multiple scattering calculation. The higher the **rfms1** is, the greater the number of atoms is included in calculation.

ii. **Full multiple scattering (FMS)**: The **rfms** field in the FMS card determines the total number of multiple-scattering paths considered in the XANES calculation. Default values are used for the other five optional fields in the FMS card.

iii. **EXCHANGE**: The EXCHANGE card specifies the exchange correlation potential model used for XANES calculation. No shift was applied to the Fermi energy level in this work, i.e., the second and third fields of the EXCHANGE card were kept being 0.

iv. **COREHOLE**: The COREHOLE card is used to specify the treatment of the core during XAS calculations. ‘Core hole’ is the hole in the orbital formed by the excitation of a single electron from that orbital.[164] In FEFF9 code, a combination of Bethe-Salpeter equation (BSE) and time-dependent density functional theory (TDDFT) is used to improve the approximation of the core hole interactions.[169,177]

ELSIE algorithm construction

We adopted the concept of ensemble method to index the most similar spectra from the database with respect to a target spectrum. Each weak learner has a unique combination of a few spectral preprocessing techniques and one similarity metric, we will describe the preprocessing approaches and similarity metrics in turn.

Each preprocessor comprises a series of steps, designed to emphasize or weaken certain characteristics of the experimental and computed spectra. A preprocessor is generated as follows:

1) *Peak shifting and quantization*. This step is necessary to all preprocessors. Because of the differences in energy sampling intervals and energy ranges, linear interpolation was used to convert each spectrum to a vector of 200 intensity values with identical energy grid. The reference spectra are shifted such that the onset of absorption, which is well-defined by the photoelectric effect, is aligned with that of the target spectra. This onset is determined by ascertaining the lowest incident energy at which the computed absorption intensity reaches 6% of the peak intensity.

2) *Pre-normalization.* We included an optional pre-normalization step to rescale the intensity to a similar range. Given the spectrum X with X_i represents the ith intensity, four normalization approaches are adopted[197]:

$$X_i^{\mathrm{norm}} = \frac{X_i}{\sum X_i}. \tag{5.3}$$

$$X_i^{\mathrm{norm}} = \frac{X_i}{\sqrt{\sum X_i^2}} \tag{5.4}$$

$$X_i^{\text{norm}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}. \tag{5.5}$$

$$X_i^{\text{norm}} = (X_i - \mu)/\sigma \tag{5.6}$$

where $\mu = \sum X_i/n$ and $\sigma = \sqrt{\sum(X_i - \mu)^2/n}$.

3) *Feature transformation.* Several feature transformation functions were implemented in the third step, which include the square root and sigmoid squashing functions. The sigmoid squashed spectrum is calculated using $X' = \frac{1-\cos{(\pi X)}}{2}$. The squared root squashing uses $X' = \sqrt{X}$, where X' is the squashed new spectrum. This technique has shown to improve the response sensitivity with respect to different spectral features.[198] The feature transformation functions also include taking the first or second order derivative of spectrum, or weighted the spectra with the first and second order derivatives. This step is necessary to make distinct weak learners.

4) *Normalization.* This last step is for all preprocessors. The spectra are all normalized such that the sum of intensities is equal to 1, i.e. $\sum_{i=1}^{\mathrm{D}} X_i = 1$.

Both the computed and target spectra are processed using the same series of steps for each pre-processor.

The preprocessed target and computed spectra are then compared in a pairwise manner using a similarity metric. Only bin-to-bin similarity metrics are used in the ELSIE algorithm development as they are less computationally demanding for high-throughput datasets.[199] Four commonly used similarity metrics in the literatures are used in the ELSIE algorithm:

1) *Pearson correlation* as defined in the Benchmarking section.

2) *Euclidean similarity*. In the D-dimensional spectral feature space, the Euclidean distance between two spectra X and Y is given by the following equation:

$$d_{\mathrm{Euc}} = \sqrt{\sum_{i=1}^{\mathrm{D}} |X_i - Y_i|^2}\,. \tag{5.7}$$

The spectral similarity measure can be derived from the distance calculated using the following expression:

$$S_{\mathrm{Euc}}(X,Y) = 1 - \frac{d_{\mathrm{Euc}}(X,Y)}{d_{\mathrm{Euc}}^{\max}}, \tag{5.8}$$

where $d_{\mathrm{Euc}}^{\max}$ is the absolute maximum expected Euclidean distance between two probability mass functions.[199]

3) *Cosine similarity*. The cosine similarity measure is the normalized inner product and measures the angle between two spectral vectors.[200] The cosine similarity between two spectra can be calculated as:

$$S_{\mathrm{Cos}} = \frac{\sum_{i=1}^{\mathrm{D}} X_i Y_i}{\sqrt{\sum_{i=1}^{\mathrm{D}} X_i^2}\,\sqrt{\sum_{i=1}^{\mathrm{D}} Y_i^2}}. \tag{5.9}$$

4) *Ruzicka similarity*. The Ruzicka[199] similarity between two spectra is given by the following equation:

$$S_{\mathrm{Ruz}} = \frac{\sum_{i=1}^{\mathrm{D}} \min\,(X_i, Y_i)}{\sum_{i=1}^{\mathrm{D}} \max\,(X_i, Y_i)}. \tag{5.10}$$

The combination of preprocessors and similarity metrics results in a total of 168 learners that can potentially be used to construct the ELSIE algorithm. To make an ensemble that outperforms individual learners, one prerequisite is that each learner should have an error rate lower than random guessing. We therefore filtered the 168 leaners to 33 and adopted them in the ELSIE algorithm. The detailed filtering procedure can be found in the Appendix C.

For each target spectrum, each learner (one preprocessor + one similarity metric) outputs similarity scores for the reference spectra. However, the quantitative scores for different similarity metrics cannot be compared even for the same target spectrum. In the ELSIE algorithm, we instead **combine the reference spectra ranking from each learner to derive an ensemble result**. For a mixture of classifiers of various types, ranking-based combination methods have been shown to be more reliable.[201] Based on the rankings, we compute the Borda count, defined as the number of candidates that are ranked equal and below the specific candidate. For example, the top spectrum among 10 computed candidates would receive a Borda count of 10, while the second ranked spectrum has a Borda count of 9. For each target spectrum, the Borda counts of the reference spectra under all learners are then summed to arrive at a consensus ranking.[202]

Finally, the Borda ranks of all reference spectra are then combined with a penalty term for the peak shift and converted to a probabilistic estimate using the modified softmax function. The probability of a reference spectrum X^{k} is indicated by $P(X^{\mathrm{k}})$ where the superscript k indicates the k-th spectrum, and is calculated as follows:

1) The Borda count of each reference (R^{k}) is normalized with respect to the count sum: $R_{\mathrm{norm}}^{\mathrm{k}} = \frac{R^{\mathrm{k}}}{\sum R^{\mathrm{k}}}$. This step is required to avoid the exponential overflow.
2) $P(X^{\mathrm{k}})$ is then calculated by the following equation:

$$P(X^{\mathrm{k}}) = \frac{\exp(R_{\mathrm{norm}}^{\mathrm{k}}) \exp{(-\frac{\alpha|\Delta S^{\mathrm{k}}|}{\delta_S})}}{\sum \exp(R_{\mathrm{norm}}^{\mathrm{k}}) \exp{(-\frac{\alpha|\Delta S^{\mathrm{k}}|}{\delta_S})}}, \tag{5.11}$$

where ΔS^{k} could be calculated as $\Delta S^{\mathrm{k}} = S^{\mathrm{k}} - \bar{S}$. S^{k} is the peak shift amount between the reference spectrum X^{k} and the target spectrum. $\bar{S}$ is the mean peak shift of the reference spectra. δ_S is the standard deviation of S^{k}. Coefficient α is fitted to the test dataset. exp $(-\frac{\alpha|\Delta S^{\mathrm{k}}|}{\delta_S})$ is therefore a term that imposes a larger penalty on large peak shifts relative to smaller peak shifts.

The algorithm itself has been highly optimized by leveraging on well-established numerical packages such as numpy and scipy.[203,204] On a laptop computer with Intel i5 2.6GHz single CPU and 2 GB of RAM, the ELSIE algorithm can perform a comparison between a target and candidate spectrum in about 0.03 s. Typically, 20-30 spectra are selected for comparison according to the rules that the computational reference spectra should have identical absorption species, limited number of elements and $E_{\mathrm{hull}} < 100$ meV/atom. The overall time to perform a complete ranking is therefore around 1 s, which allows for on-the-fly matching of uploaded spectra.

Chapter 5, in full, is a reprint of the material "Automated generation and ensemble-learned matching of X-ray absorption spectra" as it appears in *npj* Computational Materials, Chen Zheng, Kiran Mathew, Chi Chen, Yiming Chen, Hanmei Tang, Alan Dozier, Joshua J. Kas, Fernando D. Vila, John J. Rehr, Louis F.J. Piper, Kristin A. Persson, and Shyue Ping Ong, 2018, 4 (12). The dissertation author was the primary investigator and author of this paper.

Chapter 6 Accurate Chemical Environment Classification from X-ray Absorption Near-Edge Structure using a Random Forest Model

6.1 Introduction

X-ray absorption spectroscopy (XAS) is an important technique for probing the local environments in a material, as it can provide information about atomic coordination symmetries, the number and chemical identities of neighboring atoms and oxidation states.[205–207] Depending on the energy range, XAS is divided into the X-ray absorption near-edge structure (XANES) at low energy and the extended X-ray absorption fine structure (EXAFS) at high energy. While the quantitative analysis of the EXAFS is relatively mature, the analysis of the XANES is challenging, partly due to its sensitivity to many factors including coordination number (CN)[208,209], orbital hybridization[210], spin state[211], oxidation state[212] and symmetry[213] of the central absorbing atoms. However, the XANES signal usually dominates the XAS spectrum and in principle, it provides richer information regarding the chemical environments compared to EXAFS.

In a typical analysis of XANES, researchers rely on comparisons between experimentally measured spectroscopy and spectra from well-known compounds.[58,214] There have been attempts for quantitative interpretations of XANES spectra using principal component analysis[215–217] (PCA) and linear deconvolution methods.[218] These approaches seek to deconvolute the XANES spectrum of a multi-component system into individual component spectra, which provide the statistical basis for estimating the presence and ratios of individual species. However, these techniques are difficult to apply on systems that do not have well-established reference spectra.

Theoretical calculations based on time-dependent density-functional theory (TDDFT),[219] multi-scattering[145,169] and Bethe-Salpeter equation (BSE) approaches[220] are an additional means of obtaining the XANES of any material. Recently, the current authors have developed a high-throughput workflow based on the FEFF multi-scattering code[64] to build a large, public database[144,221] of 580,000 K-edge XANES spectra of over 52,000 crystals in the Materials Project.[17] This database not only provides an important reference for experiments but also opens new paths for large-scale quantitative XANES analysis. For example, the authors have also shown that an ensemble-learning algorithm spectra matching algorithm can achieve a 84.2% accuracy in identifying oxidation state and local environment by matching unknown spectra with this large, open database.

The extraction of chemical environment information from the XANES is akin to that of image recognition, a field where machine learning (ML) techniques have made great strides and sometimes surpassing even human performance. Indeed, there have been attempts to apply ML to quantitative and qualitative XANES analysis. For example, Timoshenko *et al.*[62] have demonstrated that neural networks can be utilized to extract the coordination number of Pt atoms from L-edge XANES spectra of metallic nanoparticles. Carbone *et al.*[63] have shown that local coordination environment information of 3*d* transition metal species can be extracted from site-specific K-edge XANES spectra of 3*d* transition metal oxides using convolutional neural networks with high accuracy. It has also been reported that material information, such as chemical, elemental and geometric information, can be obtained from the interpretation of calculated oxygen K-edges ELNES/XANES spectra of metal oxides and SiO_2 based on decision tree methods.[222] Very recently, Suzuki *et al.*[223] use L-edge XAS or EELS spectra of MnO in conjunction with a regression model to capture the crystalfield parameters. However, all these studies are either restricted to relatively small (∼100) datasets or a few chemical species.

In this work, we present the state-of-the-art development of general ML approaches to accurately identify the local coordination environment of absorbing atoms from K-edge XANES using the largest computational XANES spectra database. We investigate and analyze a broad repertoire of ML tools applicable to spectral interpretation. Through careful analysis of the key performance drivers, our study provides a fundamental understanding of the use of machine learning techniques on local chemical environment characterization from x-ray absorption spectra. We show that random forest models trained on ∼ 190000 K-edge XANES of ∼ 22500 oxides compounds can achieve an environment prediction accuracy of ∼ 85.4%. The as-developed model is transferable to 33 elements, covering most technologically relevant elements including alkali, alkaline, metalloid, transitional metals, post-transition metals and carbon. This study marks by far the most thorough data-driven study of K-edge XANES. Lastly, the model's generalizability is further demonstrated on public available experimental data by showing consistent high accuracy. Incorporating variable importance measures into the random forest model performance interpretations, we are able to give a clear analysis of the correlations between spectral data and absorbing atoms' chemical properties, which re-establishes the link between spectra features and the coordination environments from a data-driven point of view. This work presents the synergy of model accuracy and interpretability as key focuses in the development XANES interpretation models, and provides valuable tools and data-driven insights for identifying and understanding the coordination environment from XANES for theorists and experimentalists.

6.2 Results

6.2.1 Training dataset and machine learning model construction

To demonstrate the generalization ability of the ML algorithm, we considered site-specific K-edge XANES of all cations in oxides available from the Materials Project Database.[17] Cations with atomic number larger than 52 were excluded due to the lack of distinguishable K-edge spectral features. We selected 0 to 45 eV energy window after the onset of spectra and converted it to a vector of 200 intensity values using linear interpolation. This is the strong scattering XANES region covering the pre-edge, main- and post-edge spectral features.[224] All three regions have shown to be critical for the identification of local chemical environments.[63] The intensity vector was then normalized so that the value of maximum magnitude equals 1. Since experimental XANES spectra are representations of site-averaged signals, for each compound, we also included the same absorbing species site-averaged ensemble spectra considering the site multiplicity.

In our previous investigation,[144] we found that the broadness of the computed XAS spectral feature is sensitive to the lattice parameter variation. We therefore randomly sampled 30% of spectra and stretched or compressed them to $\pm 5eV$ changes in energy range to mimic the variations in feature broadness. The augmented data were then added to the training set to improve the robustness of classification models. This spectral shape distortion corresponds to up to 7% variations in the lattice parameters, which exceeds the $\sim$5% systematic errors introduced by the Perdew-Berke-Ernzerhof (PBE)[103] generalized gradient approximation function used in the Materials Projects[17] for crystal structure optimization.

Our computed XAS spectra training dataset includes ~ 190000 spectra for 33 elements in more than 22500 oxides compounds. To the authors' best knowledge, our dataset represents the broadest element coverage to date in the study of XANES. To gain more insights into the relationship between chemical property and coordination environments, we divided the absorbing species into six groups according to their chemical properties:

i. Alkali group (lithium, sodium, potassium, and rubidium), 47789 spectra.

ii. Alkaline group (beryllium, magnesium, calcium, and strontium), 15246 spectra.

iii. Metalloid group (boron, silicon, and germanium), 19773 spectra.

iv. Carbon group, 7839 spectra.

v. Transition Metal (TM) (20 row 3 and row 4 transition metal elements), 86584 spectra.

vi. Post-transition metal (Post-TM) group (aluminum, indium, tin, and gallium), 9458 spectra.

In this study, we define the coordination environment as the union of coordination number (CN) and its coordination motif (CM). We adopted the coordination environment assessment algorithm developed by Zimmermann *et al.*[65] to represent the mixed state of local environments. In this algorithm, the first step identifies the number of bonded neighbors, i.e., CN, of an atom based on the Voronoi tessellations method implemented in pymatgen.[12] In the second step, the coordination pattern was evaluated based on pattern-matching to determine the CM. Twenty-five coordination prototype motifs work as candidates for assessing the first-shell atomic configuration. The resemblances, i.e., order parameters (OPs), are numerical values between 0 and 1, with 1 being a perfect match for CN or CM. We then transformed the CNs OPs and CMs OPs into ranking labels to represent the full picture of the coordination environment (see Methods section for the details on the construction of the ranking labels). In this way, the coordination environment recognition problem becomes a multi-label classification problem, where an absorption spectrum might reflect the existence of more than one coordination environment. This is an attractive problem transformation approach which provides both scalability and flexibility[225] and most off-the-shelf multi-label classification algorithms[226–228] could therefore be considered in our investigation.

We selected k nearest neighbor (kNN) classifier, random forest classifier, multi-layer perceptron (MLP) classifier,[71] convolutional neural network[229] (CNN) and support vector machines (SVM) to learn the mapping from spectral features to predefined coordination environment ranking labels. Model fitting and hyperparameter optimization were performed with five-fold cross validation using the high-throughput computational dataset,[230] excluding the experimental spectra. The adoption of this strategy is to minimize data leakage, which gives a more rigorous estimation of the model's generalizability.

As the characteristic XAS absorption edge energy follows a power law with atomic number and is well separated,[49] the absorbing species can be identified with 100% accuracy prior to the coordination environment classification. Hence, classifiers were optimized element-wisely. One benefit of the element-wise optimization approach is the high specificity of individual classifier. As the XAS database size continues to grow at a steady pace, the element-specific lightweight classifiers also provide the flexibility to be further optimized with evolving data streams.

Due to the limitation in computational resource, the dramatic hyper-parameter search space cannot be navigated using the grid search techniques. Thus, during the optimization process, we adopted the heuristic optimization approach and restricted certain ML parameters based on domain knowledge. The same hyper-parameter space was adopted in the optimization of ML models for each classification subtask (see Methods section for the details on the hyper-parameter optimization of ML algorithm). Figure 6.1 provides an overview of the spectrum-based coordination environment classification workflow. For each element, we separated the coordination environment classification model training process into two steps. In the first step, we trained models for identifying the CN ranking labels. We then trained models for determining the CN-type-specific CM. This training strategy was designed to improve the specificity of the classification models. Because the coordination environment of an absorbing atom needs to be represented by mixed states, data sets size under each label condition is intrinsically smaller. As depicted in Figure D.2, for the CN classification task, the median dataset size per CN ranking label per element is 99. For the coordination motif classification task, given a specific CN, the median dataset size per ranking label per element is 112. The dataset size distributions of both classification tasks showed a similar pattern. The two boxplots' interquartile ranges were ~ 300, which means that the sample size of half datasets is within the range from 50 to 350.

Similarly, the identification of an unknown spectrum's coordination environment was conducted following the two-step procedure. Element-specific CN classifiers were first applied to predict the number of bonded neighbors, i.e., CN ranking labels. Based on the predicted CN ranking label, the corresponding trained CM classifiers were then be utilized for CM ranking label assessment. We combined the results of both steps to generate the coordination environments label sets. Each coordination environment label is a representation of CN and CM.

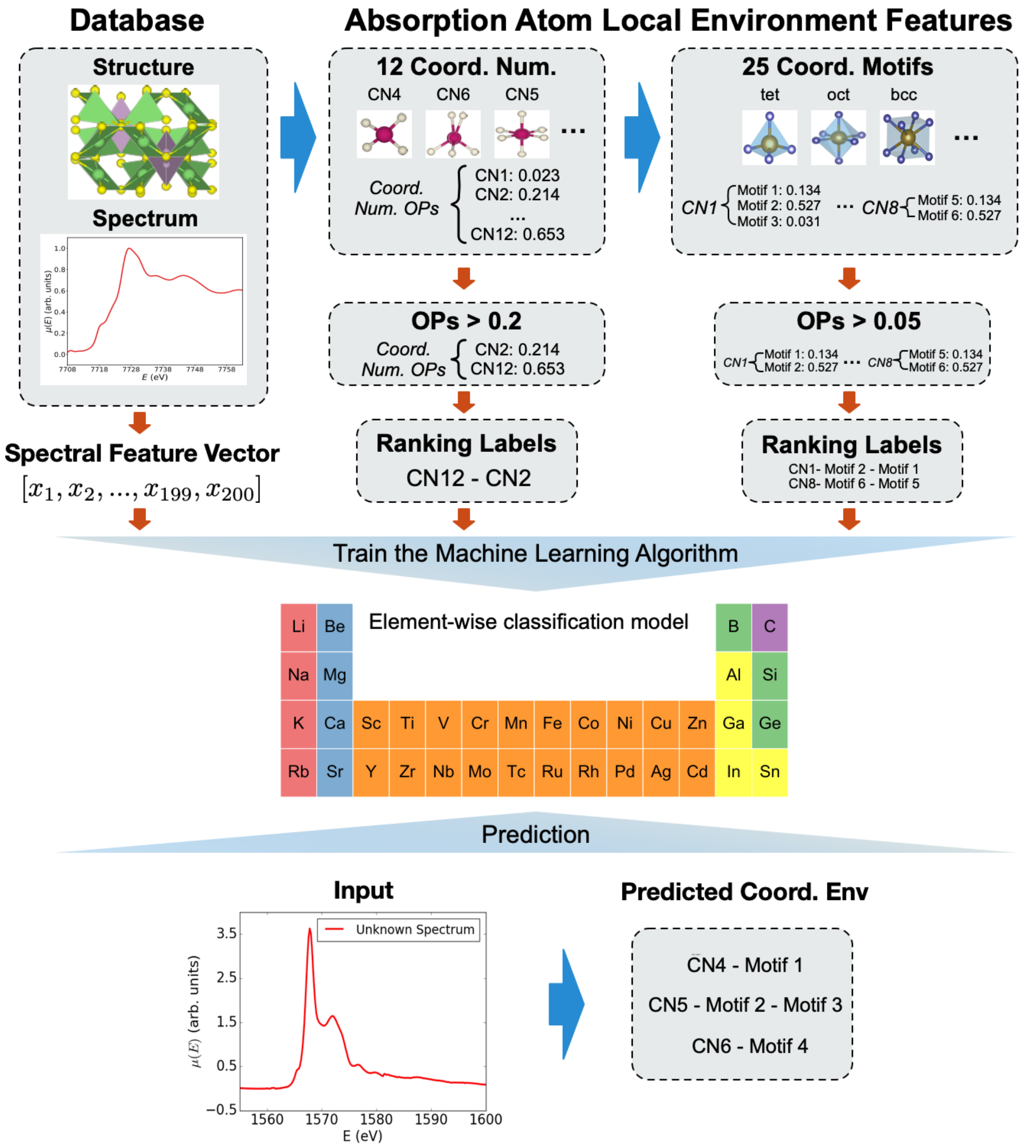

Database
Structure
Spectrum
Absorption Atom Local Environment Features
12 Coord. Num.
CN4
CN6
CN5
Coord. Num. OPs
CN1: 0.023
CN2: 0.214
...
CN12: 0.653
25 Coord. Motifs
tet
oct
bcc
CN1 Motif 1: 0.134 Motif 2: 0.527 Motif 3: 0.031
CN8 Motif 5: 0.134 Motif 6: 0.527
OPs > 0.2
Coord. Num. OPs CN2: 0.214 CN12: 0.653
OPs > 0.05
CN1 Motif 1: 0.134 Motif 2: 0.527
CN8 Motif 5: 0.134 Motif 6: 0.527
Spectral Feature Vector
$[x_1, x_2, ..., x_{199}, x_{200}]$
Ranking Labels
CN12 - CN2
Ranking Labels
CN1- Motif 2 - Motif 1
CN8- Motif 6 - Motif 5
Train the Machine Learning Algorithm
Element-wise classification model
Li Be B C
Na Mg Al Si
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn
Prediction
Input
Unknown Spectrum
$\mu(E)$ (arb. units)
E (eV)
Predicted Coord. Env
CN4 - Motif 1
CN5 - Motif 2 - Motif 3
CN6 - Motif 4

Figure 6.1: Workflow schema of the coordination environment identification algorithm.

The ranking of coordination environment labels was determined by the predicted CN ranking list as the CNs OPs are used as multiplying factors during the determination of CMs OPs.[65] Therefore, the top-ranked coordination environment label is supposed to represent the dominant coordination environment of the absorbing species.

6.2.2 Computational spectra classification performance

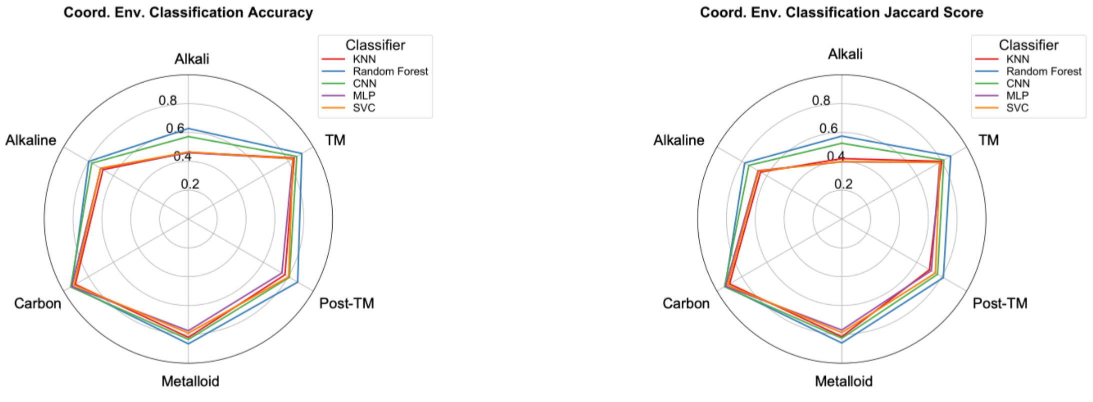
We systematically evaluated the performance of various classifiers using the Materials Project computational K-edge XANES spectra database.[230] We used the accuracy and Jaccard score as metrics to measure and compare the performances of different classifiers. The accuracy score is a way to measure how well an ML model performs in predicting the dominant coordination environment of the absorbing species. The sample-average Jaccard score is a more strict performance metric than accuracy and emphasizes the performance of ML models on identifying all coordination environments related to the absorbing species.

Figure 6.2 compares the accuracy and Jaccard score of various classifiers categorized by elemental groups. All five classifiers performed similarly on the carbon group, with classification accuracy of ~ 93.1%. The random forest classifier outperforms the other four classifiers significantly on the rest elemental groups and maintains a consistent level of performance. Overall, the random forest model achieves an accuracy of 85.4% on the top coordination environment prediction task and Jaccard score of 81.8% on the multi-label coordination environment classification problem.

As shown in Figure 6.2, we notice that the classification performance is highly correlated with elemental group. All five models suffer from performance drops on the alkali group. To elucidate the origin of the performance degradation, we borrow the concept of entropy to measure how disorganized the spectral coordination environment labels are in each group. The definition of entropy comes from information theory[231] and can be calculated using the following expression:

$$S = -\sum_{i} P_i log_2 P_i, \tag{6.1}$$

where P_i is the probability of a ranking label i out of all ranking labels. We computed the coordination environment ranking label entropy of each element. The entropy value will be high if the variability of the label values is high and vise versa.



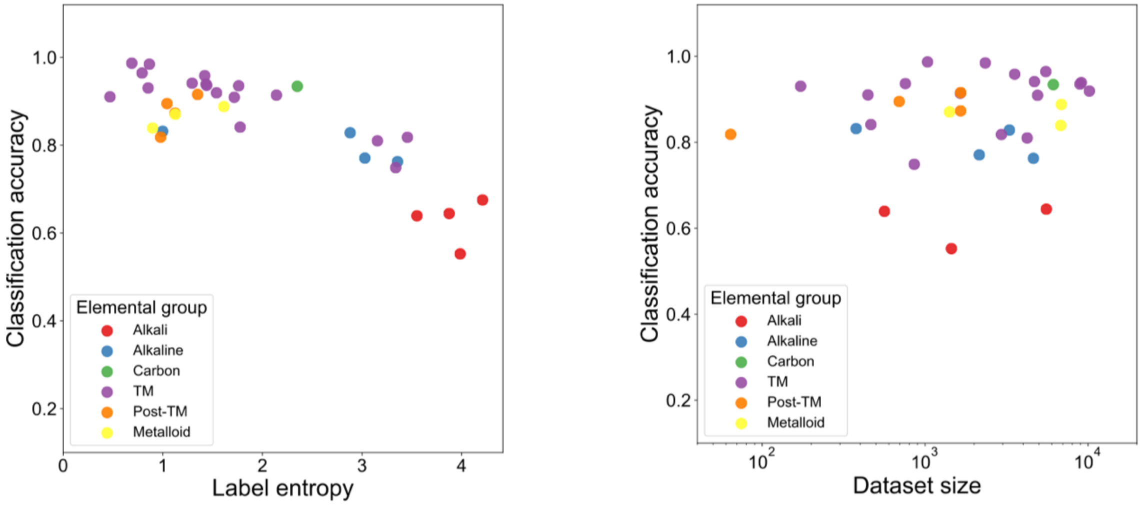
(a) Accuracy of top coordination environment classification

(b) Jaccard score of top coordination environment classification

Figure 6.2: Elemental group-wise coordination environment classification accuracy and Jaccard score derived from different ML classifiers.

Figure 6.3(a) shows a quasi-linear relationship between the random forest classifier's performance and label entropy. The label entropy values of alkali and alkaline groups are generally high, which indicates that these two sub-datasets are more diverse and make classification tasks more challenging. The results are expected, however, since the Materials Project database contains considerable amount of alkaline-conducting compounds, where the alkaline elements tend to form various local environments with low local symmetry.

The accuracy drop might also come from the inferior performance of FEFF in reproducing the spectra of light alkali elements.[232] In FEFF, the approximation of core hole potentials could result in too strong screening effects for the core hole, which causes a tendency to underestimate the white line intensity of light elements in compounds. In the previous study,[233] $Z + 1$ approach was adopted to enhance the intensities of the pre-edge features of K-edge XANES for lower Z elements. However, this strategy is not suitable for our large-scale high-throughput spectral simulations. As a comparison, Figure D.3(a) shows CNN's prediction accuracy as a function of label entropy values. We notice that the CNN classifier failed to deliver classification performances comparable to the random forest classifier. Besides, the CNN classifier showed a higher rate of prediction accuracy decreasing as label entropy increases.



(a)

(b)

Figure 6.3: Relationship between the random forest model's classification accuracy and (a) the label entropy; (b) training dataset size.

While the classification accuracy or Jaccard score may well be the main criterion that influence the selection of a ML model under most circumstances, care must be taken in model selection. In the context of selecting supervised learning algorithms for materials science applications, the importance of prior knowledge and problem analysis ought not to be overlooked.

To further validate this hypothesis, we produced Figure 6.3(b) and Figure D.3 depicting the relationship between dataset size and classifiers' performances. It is worth noting that the dataset size is not a deterministic factor affecting the classifier's performance. The correlation between dataset size and the model's performance is relatively weak. The random forest classifier was capable of achieving ≥ 80% accuracy in data-scarce regions. From Figure 6.3(b) and Figure D.3(b), we make the observation that CNN is more data hungry than the random forest model. Under the same training dataset conditions, CNN based models resulted in inferior performance than the random forest model.

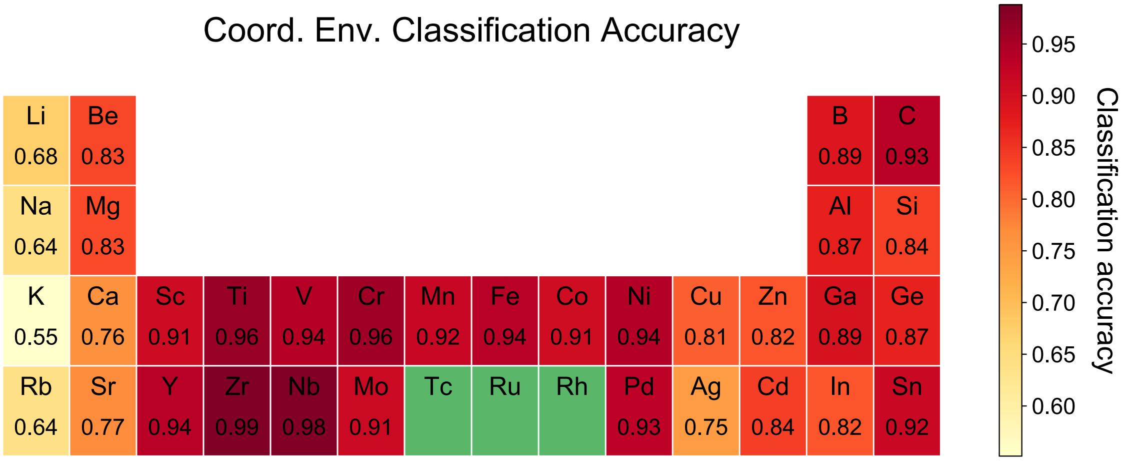


Figure 6.4: The random forest classifier's element-wise classification accuracy of top coordination environment. We do not have sufficient sample sizes of computed Technetium (Tc), Ruthenium(Ru) and Rhodium (Rh) K-edge XANES to form reliable training sets for classification tasks.

This once again suggests that random forest classifier is more applicable in the coordination environment identification problems. Thus, we adopt the random forest classifier for the identification of spectra's coordination environment. Element-wise random forest classifiers' accuracy and Jaccard scores of coordination environment classification tasks are provided in Figure 6.4 and Figure D.4. The models have also been made publicly available as a part of *veidt* (https://github.com/materialsvirtuallab/veidt), an open-source Python ML library.

6.2.3 Model insights

As ML models penetrate materials science community, the inability of humans to understand these models seems problematic. However, due to the underlying nonlinear structure of most ML models, they were usually applied in a black-box manner. We argue that ML implementations should not lack an explicit declarative knowledge and believe that the usage of high-throughput computational dataset ought not to be limited to the model-fitting stage. We need to understand which spectral attributes are the most important in determining the chances for successful identification since the interpretation of K-edge XANES tends to be an insight-driven activity.

In this section, we seek to clarify the relationship between K-edge XANES spectra and the geometrical structure around absorption atoms from a computational perspective. Here, we mainly consider the 4, 5 or 6 coordinated nine $3d$ transition metal elements (Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn) as extensive research has been carried out to reveal the link between their characteristic spectral features and coordination chemistry. To gain conceptual clarity of spectral characteristics related to coordination environment identification, we examined the variable importance measure of spectral coefficients.

In our study, we divide the K-edge XANES spectra into three regions of $0-15$ eV, $15-30$ eV and $30-45$ eV, which correspond approximately to the pre-edge region, main-edge region and post-edge region features, respectively. Instead of using the variable importance measures produced by the random forest model, we employed the brute force drop-variable importance mechanism. Its rationale is the following: by dropping a region of spectra, the prediction accuracy will change relative to the full-spectrum trained baseline model's performance. A reasonable measure for spectral feature importance is the absolute difference in prediction accuracy before and after dropping the spectral region. The advantage of the random forest drop-variable importance measure is that it provides the ground truth feature importance compared to alternative importance measures.[234]

Figure 6.5 shows the normalized feature importance of different spectral regions. We find that model performances decrease significantly when two spectral regions were eliminated. In most cases, models completely failed to identify the relevant coordination environment using only one of the three spectral regions. This finding demonstrates that the precise identification of coordination environment relies on full spectral characteristics, consistent with previous study[63] indicating the critical importance of features beyond the pre-edge region in accurately classifying local chemical environments for $3d$ transition metals.

The advanced feature importance figure also provides a pragmatic view of the informativeness of various spectral regions. We find that the importance of main-edge and post-edge regions' features was relatively high for the four coordinated early $3d$ transition metals (Ti, V, and Cr). As illustrated in Figure 6.5, six coordinated compounds exhibited higher main-edge and post-edge regions feature importance than four-coordinated compounds. It is well recognized in the experimental and computational literature that the pre-edge peak intensity of $3d$ elements' K-edge XANES decreases with an increase in the CN for $3d$ oxides. Thus, the decrease of pre-edge spectral features importance may be partially attributed to the drop in pre-edge peak intensity when species change from tetrahedral symmetry to octahedral symmetry.[212] It is important to note that features beyond pre-edge regions become more critical for coordination environment classifications in those six coordinated $3d$ transition metal oxides. These observations are in good agreement with previous studies[63,208,212] demonstrating that the pre-edge region feature intensities are a function of the number $3d$ electrons, which maximizes at d^0 and gradually decreases to zero at d^{10}. For five coordinated compounds, we observe that two spectral regions features are necessary for reliable classification, and the pre-edge region feature importance is weaker than in four, and six coordinated compounds.

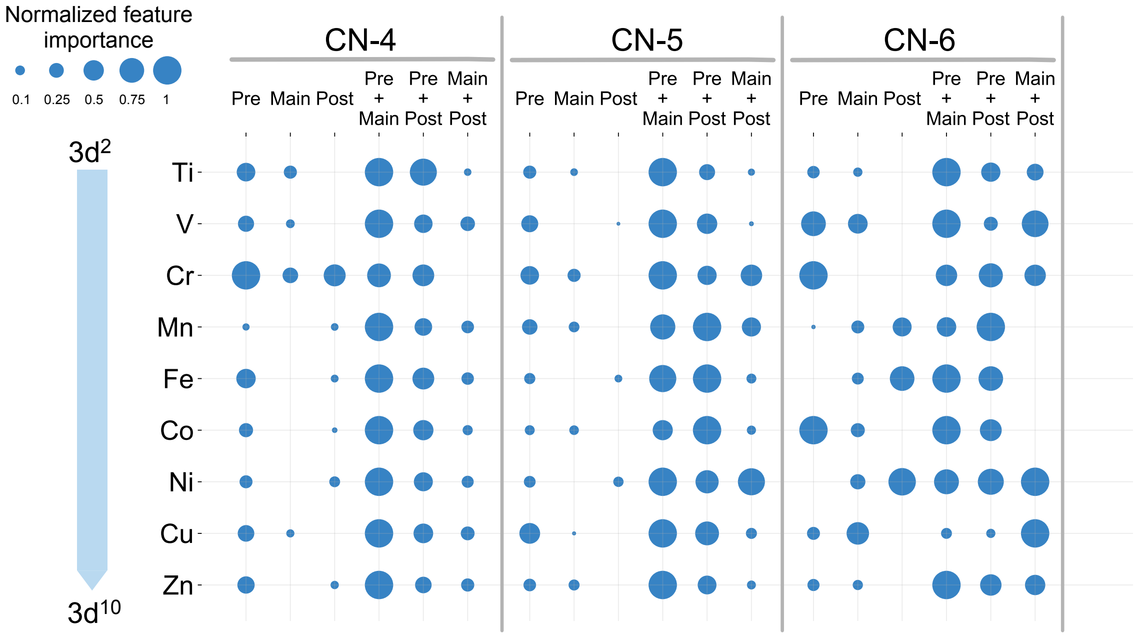


Figure 6.5: Normalized feature importance for different regions of spectra. The drop-variable feature importance is normalized with respect to the maximum importance of a spectral region element-wisely.

Furthermore, the feature importance measure allows us to study the dependence of Kedge XANES features on the CN for each $3d$ transition element. In the cases of Ti, Mn, Fe, Ni, Cu, and Zn, we observe that the dependence relationships are complicated. The low feature importance values of pre-edge and main-edge regions suggest that the use of these two spectral regions alone gives less robust information on the central atoms' coordination environment. Multiple spectral regions need to be considered together during classification learning. This provides the direct evidence that the spectral characteristics of 3d transition elements are influenced by valence,[211] symmetry,[235] electric quadrupole, and dipole transitions.[60,213] In V, Cr, and Co, we find that the pre-edge region is informative for six coordinated compounds. This might be due to the distortion of symmetry. For example, six-coordinated V compounds[236] have been confirmed to have intense pre-edge peaks caused by out-of-center displacements. Wong *et al.*[237] have also demonstrated that the pre-edge absorption features' intensity grows as the oxidation state of six coordinated V increases.

6.2.4 Coordination environment identification of experimental XANES spectra

The ultimate goal of ML classification algorithm development using computational XANES spectra is to identify the correlation between spectra and the local chemical environment of experimental XANES spectra. Generalizing ML model to the unseen and experimental dataset is not only the central concept of our work but also should be considered as the golden rule in evaluating the performance of computational data parameterized models. It is therefore critical to understand the generalization performance of the optimized models on open access experimental spectral datasets.

We evaluated the random forest classifiers using high-quality normalized XANES experimental spectra available from XAFS Spectra Library (http://cars.uchicago.edu/xaslib), EELS database,[54] and supplemented by six high-quality experimental XANES spectra of V_2O_5 , V_2O_3, VO_2 , $LiNiO_2$, $LiCoO_2$, and NiO from previous studies.[238,239] We selected the spectral region from -5 eV to 55 eV with reference to edge energy (E_0) determined by the MBACK algorithm.[240] As PBE usually leads to up to 5% lattice parameter overestimation error,[180,241] the expanded spectral region compromises this artificial spectral feature difference between computational and experimental XANES spectra.

We excluded the Al_2O_3, $SrCO_3$ and Mn_3O_4 spectra from the evaluation dataset. As for Al_2O_3, the Al is six coordinated by oxygen atoms, octahedral | pentagonal pyramidal | hexagonal planar is the only CM ranking label with sufficient computed spectral data. The rest CM ranking labels of six coordinated Al all had less than 30 spectral samples and were insufficient for training. Therefore, coordination motif classification for six coordinated Al could not be performed. For nine coordinated Sr^{2+} ion of $SrCO_3$, no CM type was provided by the coordination environment labeling tool.[65] In the experimental Mn_3O_4 entry, Mn^{4+} is four coordinated and there were only two computational XANES spectra with the same CM ranking label. Thus, the computational data is insufficient for proper training of the classifier. Nevertheless, the 27 spectra comprise a diversity dataset covering 13 chemical species for classifiers' performance assessment. For those experimental spectra from EELS database and XAFS Spectra Library without available structural information, we assumed those spectra correspond to the ground state structures in the Materials Project database with the same chemical composition.

On the out-of-sample test set, the random forest classifier successfully identified 22 of 27 top coordination environment ranking labels. The top coordination environment prediction accuracy is 81.5% and the coordination environment identification Jaccard score is 0.82. We find that the predictive capacity of the random forest classifier is on par with the performance obtained using the computational dataset. This indicates that models trained on high-throughput computational dataset exploit spectral features of the experimental dataset.

We observe that the classifier failed to retrieve the correct coordination environment of CuO, Na_2O and ZnO and two V_2O_5 spectra. It is worth noting that the classifiers were cable of predicting the dominating CN (CN with highest q_{CN}) with 100% of accuracy. Indeed, for V atoms, the classifier predicted that there exists a secondary CN condition. This may be primarily due to high similarity of the V K-edge spectra for different coordination environments. For V_2O_5, the classify successfully predicted the dominant CM, i.e., trigonal bipyramidal. We notice that the OPs difference between the second (i.e., $q_{\text{pentagonal planar}}$) and third (i.e., $q_{\text{square,pyramidal}}$) rank CMs is $\sim$ 0.029, and the OPs difference between the dominant and secondary CMs is $\sim$ 0.033. Thus, the coordination pattern of the five coordinated V can be identified as resembling three coordination motifs to the same extent. In ZnO, the coordination pattern of four coordinated Zn did not resemble any target CM to a great extent, i.e., all CM OPs < 0.22. Here, the relative low degree of resemblance between the absorbing atom's coordination pattern and target motifs seems to be the critical issue. For Na_2O spectrum, the failure of the model might be attributed to the possible contamination of the experimental sample.[144] For CuO, Cu^{2+} ion is four coordinated by oxygens and has a fairly complex coordination environment. The Cu^{2+} ion's coordination pattern is identified in a matching with 5 target motifs, where OPs of three CMs (i.e., rectangular see−saw−like, see−saw−like and square co−planar) exceed 0.5. In this case, the local environment investigation conducted using EXAFS measurements becomes critical.

6.3 Discussion

To summarize, we have shown that the random forest classifier can be a powerful tool for identifying the atomic coordination environments. The random forest classifier trained using high-throughput FEFF computed K-edge XANES spectra exhibits excellent generalizability and demonstrates the state-of-the-art performance on a diverse experimental spectra test set comprising 27 experimental XANES spectra of 13 chemical species. The model shows an outstanding ~ 81.5% top 1 accuracy in identifying the complex and mixed states of coordination environment from experimental spectra, respectively. This breaks the limitation of using the group theory derived character tables[242] in XANES spectra evaluation. Since in a real-world scenario, there are quite few compounds with completely regular symmetry. With the presence of unsymmetrical lattice vibration, short- and median-range disorders, the XANES spectral features might not perfectly reassemble those characteristics deduced from 1s-3d transition or d-p hybridization.[208,212]

In addition, we performed a comparative study on the performance of various modern modeling techniques and clarified how the sample sizes, classification problem complexity are related to the models' discriminatory ability. A key advantage of our random forest classifiers is that they are capable of delivering a high level of accuracy across 33 elements by training data size varying from 100 to 10000. Most importantly, we performed a systematic investigation on the feature importance of K-edge XANES 3*d* TMs and identified spectral attributes that are relevant for coordination environment classification. Using the feature importance measure, we show that the pre-edge spectral region is a relatively important factor in distinguishing four, or six coordinated absorbing species coordination environments. We reveal that the main-edge and post-edge spectral regions become more distinguished as CN changes from four to six. Accordingly, our interpretation process provides a useful rational framework for the element-wisely investigation of the relationship between spectral features and coordination environments by directly probing the models' variable importance measure. Understanding opacity in the ML algorithm will significantly advance the implementation of artificial intelligence technologies in the materials science regime. The knowledge mined from the high-throughput computational spectral database will guide researchers' decisions about where to dedicate their efforts to improve spectra interpretation. Moreover, understanding why ML models behave the way they do empowers materials scientists to utilize the classification model more effectively and efficiently.

6.4 Methods

6.4.1 Coordination environment ranking labels construction

Determining and feature engineering of the absorbing element's local environment is surprisingly a nontrivial task. This is because in the real material geometry, the structural distortion will introduce ambiguity to the identification of local environments. Using a deterministic label to represent an absorbing element's coordination environment from its site-averaged experimental XANES spectrum is ill-defined. A single absorption site's local chemical environment could resemble various coordination motifs (CMs) simultaneously. For instance, one site might have both tetrahedral and bcc-like coordination as the bcc motif can be viewed as two point-symmetric tetrahedra.[65] Therefore, it is insufficient to describe an absorbing specie's or site's coordination environment by utilizing one geometrical label.[63] More complex feature engineering techniques of the local chemical environment must be developed for accurate and reliable assessment.

To embed the nature of coordination environment conditions into the ML model, we quantified the coordination environment using order parameters[243] (OPs) and then transformed them into ranking labels. Here OPs are the similarity assessment scores used to quantify the degree of matching between local structural motifs and certain CNs or CMs, which range from 0 to 1. A numerical value of 1 represents perfectly resemblance of a CN or CM prototype, and vise versa. In our investigation, q_{CN} (e.g. q_{CN-1}, q_{CN-2}, etc.) is the CN specific order parameter that describes how consistent a site is with a particular CN. Symbols such as q_{tet}, q_{oct}, $q_{bcc}$and etc., are the motif-specific order parameters used to discern the similarity between coordination prototype motifs and local environments. We restricted our investigations to coordination environments with CN equals or less than 12 in our study. This cutoff was chosen as no coordination prototype motifs are given for geometries with CN greater than 12. A total of 25 perfect coordination motif prototypes were provided to assess the coordination patterns.

Given a vector $\widehat{\mathrm{OPs}} \in \mathbb{R}^L$ of real-valued order parameters (OPs) output of an absorption sites local chemical environment, each ith OPs represents how closely the sites local chemical environment resembles a CN condition or a certain CM. A threshold t is applied to $\widehat{\mathrm{OPs}}$ to create a bipartition of relevant and irrelevant CN and CM labels. The multi-label prediction $\hat{\mathbf{y}}$ can be obtained as:

$$\widehat{y_j} = \begin{cases} 1 \text{ if } \widehat{\mathrm{OPs}}_j \geq t \\ 0 \text{ if } \widehat{\mathrm{OPs}}_j < t. \end{cases} \tag{6.2}$$

Instead of using an arbitrary threshold like 0.5, we adopted the concept of label cardinality (LCard) and rigorously calibrated the threshold t to minimize the possibility of a spectrum being assigned to the no-label set. The LCard[244] is a standard measure of "multi-labeled-ness", which is simply the average number of labels associated with each example. For N examples, the LCard measure can be calculated as:

$$LCard = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_j^i\,. \tag{6.3}$$

The threshold t_1 for CN and threshold t_2 for CM are calibrated using the same procedure as follows:

$$t = argmin\left\|LCard\left(D_{site-specific}\right) - LCard\left(D_{site-averaged}\right)\right\|, \tag{6.3}$$

where $D_{site-specific}$ refers to the dataset consist of ~ 110000 site-specific computational Kedge XANES spectra and $D_{site-averaged}$ is the dataset of ~ 36000 site-averaged absorption spectra. The site-averaged spectral dataset is populated as the experimental XANES spectra of elements are the averaged absorption coefficients. The OPs of site-averaged spectra are obtained by summing those individual spectral OPs and normalizing using the site multiplicities. The calibration procedure results in the average observed label cardinality of site-specific spectra becomes close to the average label cardinality of site-averaged spectra. This calibration approach is found to be more effective and efficient in reducing the probability of empty-set prediction issues than simply using an arbitrary threshold value.[225]

To shine a light on the nature of mixed coordination environments for an absorption site, we evaluated the threshold value t_1 and t_2 from 0 to 0.4 at 0.01 intervals. The average number of labels associated with each spectrum will naturally decrease as the threshold value increases. Since the average number of CN labels associated with each spectrum drops below 1 when t_1 exceeds 0.37, the highest threshold value was set to 0.4. For the CN label set, we find that the LCard difference between the site-specific dataset and site-averaged dataset is minimized at t_1 = 0.2. The average number of CN labels associated with each spectral example is $\sim$ 1.2. For the CM label set, the two datasets' LCard difference reaches a minimum at t_2 = 0.05. The average number of coordination environment labels associated with each spectrum is $\sim$ 3.2.

After applying the calibrated thresholds, we then encoded the CN and CM label sets into the form of ranking labels in the descending order of computed OPs. By using 0.2 as cutoff for

CN OPs, the average number of CN ranking labels per element is 10. Note that the labels contain joint labels such as CN4-CN6. In the CM classification task, the average number of CM ranking labels is 5 per element per CN.

As expected, we observe that the distribution of relevant CN labels, i.e., CN with $q_{CN} \geq 0.2$, is inhomogeneous [Figure D.5]. In each elemental group, there are a few dominating CNs. The numbers of data points with prevailing CN were an order of magnitude more than the rest. In the CM classification problem, we could therefore restrict our consideration to those most abundant CN cases of each elemental group. We excluded CN nine to twelve from the CM classification task as no target CM was provided for those CNs.

As ML algorithms are highly ”data hungry”, for each absorbing specie, we excluded CN and CM ranking labels with less than 30 spectral samples. After we applied the minimum number of data points rule, all Tc and Rh ions are six coordinated. Therefore, we removed the Tc and Rh K-edge XANES from the first step CN classification task’s training dataset. For the second step CM classification task, we repeated this operation and excluded those sub-datasets (see Table D.1) associated with only one CM label from the training dataset as well. Finally, the following CN schemes in each elemental group were subjected to the coordination environment classification task.

i. Alkali group: The range of CN from 3 to 8

ii. Alkaline group: The range of CN from 4 to 8.

iii. Metalloid group: Values of CN at 3 and 4.

iv. Carbon group: The range of CN from 2 to 4.

v. Transition Metal (TM): The range of CN from 4 to 6.

vi. Post-transition metal (Post-TM) group: The value of CN at 4 and 6.

To validate the necessity of using ranking labels to represent the absorption elements' coordination environments, we visualized the joint distributions of the CN and CM OPs of the alkali and the transition metal elemental group [Figure D.6]. From Figure D.6, we observe that there are correlations across different CN OPs or CM OPs. The locations of those dark blue areas on the OPs distribution map are direct indicators of the coexistence of multiple coordination environments. We also note that the correlation between CM OPs is quite substantial. Most six coordinated transition metal ions' coordination patterns resemble two or more CMs to a great extent, i.e., various CM OPs exceed ≥ 0.4. These findings emphasize that labeling the absorbing sites' coordination environments with one label cannot adequately represent the full coordination environment schemes.

6.4.2 Hyper-parameter optimization of machine learning algorithm

As most binary class classification algorithm can be naturally extended to the multi-label classification problem, five most commonly used classifiers includes random forest, k-Nearest Neighbor (kNN), multi-layer perceptron (MLP), support vector machine and convolutional neural network (CNN) were chosen and modified to make multi-ranking label predictions. For each classification model, we first performed parameter optimization using the computational spectra through a heuristic approach. The parameters configurations with the highest accuracy were selected as the optimal parameter sets. We then compared the performance of the five parameter-optimized classification models across elemental groups in order to determine the optimal solution for the coordination environment classification task. The details on hyper-parameter space are as follows:

i. kNN: The k nearest neighbors classifier is optimized with respect to the number of neighbors (N) and the distance metric (p). The value of N is examined for 10, 20, 30, and 50. We restricted the minimum value of N to 10 during the parameter search to avoid overfitting and increase the generalizability of models. Manhattan distance and Euclidean distance were used to assess the distances metric effects. We did not vary the tree structure, leaf size, and the algorithm used to compute the nearest neighbors of kNN classifier as these three parameters are invented in a general sense to address the computational inefficiencies of kNN.[228]

ii. Random forest classifier: We varied the number of trees (n_estimators) in the forest, n_estimators$= 10, 20, 30, 50, 100, 200$. The rest parameters were kept at the empirical good default settings.

iii. MLP classifier: For the MLP classifier, we varied the number of hidden layers (L) from 1, 2 and 3. The number of hidden layer neurons was selected from 10 to 100.

iv. SVC: The penalty parameter C was drawn exponentially from 0.001 to 100.0. We restricted the maximum value of C to 100.0, as high C is prone to overfitting. We experimented two kernal coefficient (γ) values. The possible γ were (a) 1 divided by the number of features ($\gamma = 0.005$), and (b) 1 divided by the number of features multiplied by spectral absorption coefficients variance ($\gamma \simeq 0.013$). We chose the radial basis function (RBF) kernel as the number of observations is one or two orders of magnitude higher than the number of features in the training data. In addition, previous study 55 has shown that it is unnecessary to consider the linear kernel if the model selection is conducted using the RBF kernel.

v. CNN: In our study, we used the 2-layer convolutional neural network classifier. The two layers were fully connected, feed-forward hidden layers with 50 and 100 neurons, ending with a softmax output layer. The number of neurons in the output layer equals to the number of target ranking labels. It has been shown that the performance of the CNN-based model in the classification of XAS spectra is invariant across different neural network structures.[63]

For CN ranking labels classification, we find that the random forest classifier's performance converged at n_estimators=30 for all elemental groups. As the time it takes to train a random forest classifier increase with the number of trees in the forest, we chose n_estimators=30 for random forest classifier. In the case of kNN classifier, we find that the model using 10 nearest neighbors and manhattan distance performs the best. For the MLP classifier, the two-layer neural network architecture with ReLU activation function outperforms the rest models with tanh or logistic sigmoid neurons. The first layer has 50 neurons. The second layer consists 100 neurons. We find that increase in the number of hidden layers has a detrimental effect on classification performance. For the RBF SVC classifier, the model with $C = 100$ and $\gamma \simeq 0.013$ performs the best.

Chapter 6, in full, is currently being prepared for submission for publication of the material "Accurate Chemical Environment Classification from X-ray Absorption Near-Edge Structure using a Random Forest Model", Chen Zheng, Chi Chen and Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

Chapter 7 Summary and Outlook

Machine learning methods and artificial intelligence are beginning to make major inroads within modern materials science and engineering. During the last decades, significant advances in high-throughput computational techniques and architectures have already met phenomenal successes in the field of materials design and discovery, which turns the big materials data into the backbone of the next generation materials discovery. In this thesis, we report the development of machine learning frameworks for the investigations of $P2$ layered sodium TM oxides cathode materials and X-ray absorption spectroscopy.

For the layered sodium TM $P2$ oxides, we present a detailed first-principle DFT study of the phase diagram and Na-vacancy arrangement of $P2$ $\mathrm{Na}_x\mathrm{Co}_{1-y}\mathrm{Mn}_y\mathrm{O}_2$ system, with an emphasis on searching for stable ground state structures corresponding to different intermediate phases. Calculated voltage discharge curves were in excellent agreement with experimental results. DFT calculations were successfully applied to reveal the effect of TM substituents on the phase diagram, the tendency of forming ordered Na patterns at various Na concentrations. Our calculations demonstrate that mixing distinct TM elements results in significantly weaker and more consistent Na/vacancy ordering tendency in the entire sodium concentration region, compared with single TM Na layered oxides. Using *ab initio* molecular dynamics simulations and nudged elastic band (NEB) calculations, we elucidate that TM substitution has a pronounced effect on Na diffusion energy barriers and Na site occupation energy. Furthermore, by employing a site percolation model, we derive theoretical upper and lower bounds on the concentration of TM species in the $P2$ layered oxides based on their effects on Na diffusion energy barriers. To our knowledge, this is the first time that a universal framework to rationally tune mixed TM layered $P2$ compositions for optimal Na diffusion has been proposed.

We show that graph neural network technology, developed for applications such as computer vision, social network prediction, and recommender systems, can be used to capture the energy-density maps for quinary $P2$ layered TM oxides systems with negligible loss in accuracy. We attempt to generalize the MatErials Graph Network (MEGNet) model to the formation energy prediction task of high entropy cathode $P2$ $Na_xCo_{0.2}Mn_{0.2}Ti_{0.2}Ni_{0.2}Ru_{0.2}O_2$ materials, bypassing the need to solve the Kohn-Sham equations. MEGNet is able to correctly predict the energy trend of different Na/vacancy ordering configurations with various TM frameworks. These results indicate that graph neural network has the great potential to represent ground state wavefunctions and holds the promise of allowing larger systems to be tackled. This should yield orders of magnitude savings in computer time and now allows the first-principles investigation being performed in quaternary or quinary compound systems. We believe these advances would no doubt accelerate novel materials discovery.

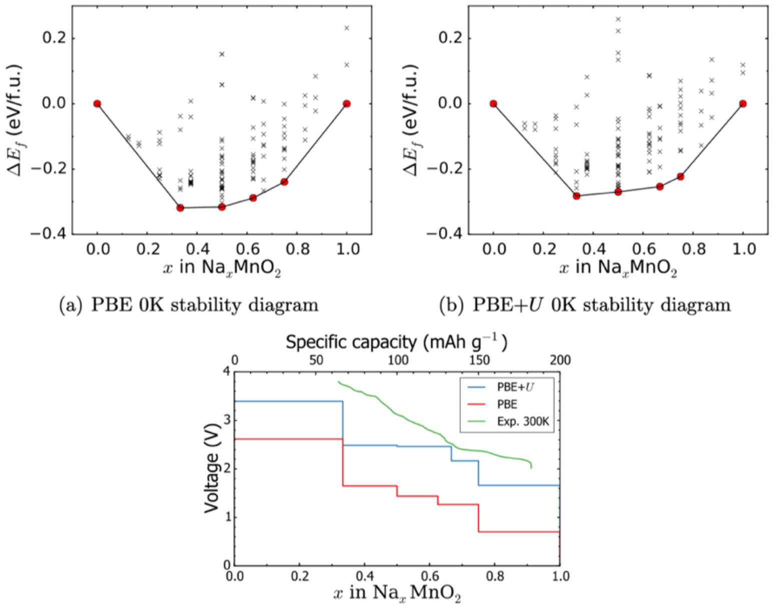
From chapter 4 to chapter 6, we attempt to solve X-ray absorption spectroscopy interpretation problems using machine-learning-based models, as the application of XAS for the screening of large quantities of materials remains to be relatively labor intensive and expensive. In the case of XAS, efficient computational codes like FEFF provide a means to generate high quality spectra computationally. In contrast to the experimental spectra, the computed spectra are unhampered by sample impurities, background noise, and equipment specific aberrations. We present the development of a high-throughput framework to generate a reference XAS database for all materials in the Materials Project database. This reference so far includes more than half millions of K-edge X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS) spectroscopy of close to 60,000 compounds. Input parameters used for the calculations can be downloaded directly from the Materials Project website or using the REST Application Programming Interface.

For the searching and interpretation of computational X-ray absorption spectroscopy, we separate our work into two sub-projects. We first demonstrate the development of a novel automated XANES spectra matching algorithm that leverages on ensemble learning techniques to find the similar XANES spectra from out computed reference XAS database. The as-developed spectral matching algorithm allows users to efficiently acquire similar spectra with respect to the query spectra through the Materials Project web page. The hit-list provided by the spectral matching algorithm will facilitate the drill-down analysis of materials in different aspects. In the second sub-project, we present the development of general machine learning approaches to rapidly and efficiently identify the local coordination environment of absorbing atoms from K-edge XANES. We evaluate and select the best machine learning algorithm from a broad repertoire of machine learning tools. A random forest model is built using the largest high-throughput computational K-edge XANES spectra dataset. The random forest model used to predict experimental spectra has an accuracy of 82% in identifying the absorbing atoms' coordination environments in mixed states. Finally, we demonstrate an approach that helps better understand what the random forest model has learned. Our study enhances the awareness of explainability in machine learning and artificial intelligence.

To conclude, the work outlined in this thesis represents a key step forward toward combining machine learning models with high-throughput materials data. We have demonstrated that machine learning techniques could be customized and integrated with massive materials datasets to make high-fidelity predictions on materials properties (including formation energy, structural properties, and coordination environment of X-ray absorbing species). The machine learning paradigm presented in this work can be considered as a stepping stone for future applications of machine learning methods within materials science.

To date, machine learning techniques have been widely applied to different areas of materials science research. Data-driven "materials informatics" strategies are now core parts of many materials studies. However, we notice that there remains significant scope for improvement for successful applications of machine learning methods within materials science. For instance, most predictive models only cover a small fraction of the properties in materials design. Machine learning algorithms were usually trained on a case-by-case manner. One well-trained model might perform poorly when the model is generalized to "unseen" datasets. Another challenging lies in the generation of large scale and high-fidelity computational datasets. Taking X-ray absorption spectroscopy computation as an example, even with modern efficient and well tested computational codes, approximations of the electronic structure using DFT-based approach still require substantial computational resources and thus are impractical in high throughput studies. For practical high-throughput materials modeling, a common strategy to reduce the computational cost is applying universal parameter sets and approximations across different systems. This procedure inevitably results in information loss and questionable accuracy associated with human intervention. Therefore, we hope that this thesis can be a first step in spurring further research on the adoptions of machine learnings techniques in materials science.

Appendix A Supplementary Information Effect of Transition Metal Mixing on Na Ordering and Kinetics in Layered P2 Oxides

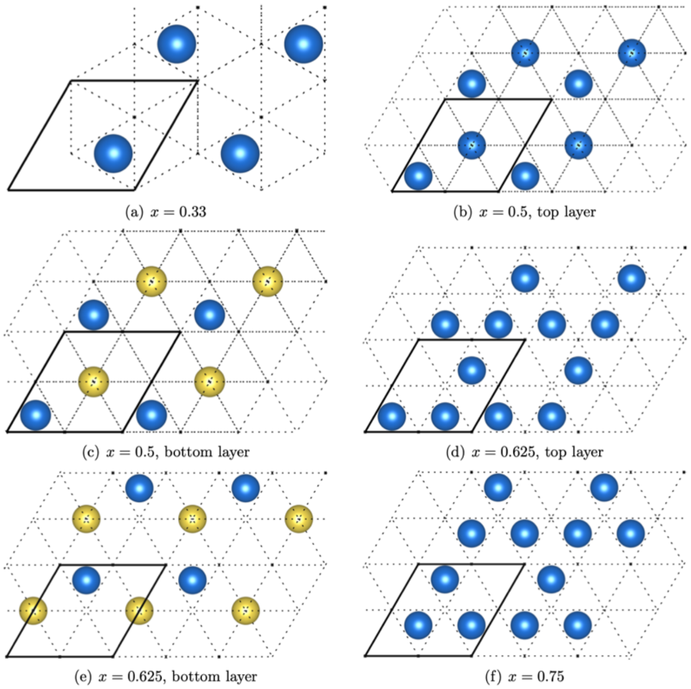


(a) PBE 0K stability diagram

(b) PBE+U 0K stability diagram

(c) Voltage profile

Figure A.1: (a) PBE and (b) PBE + U 0K stability diagrams of Na_xMnO_2. Black line: convex hull; red dots: stable orderings; black cross: unstable orderings. (c) PBE and PBE + U voltage profiles of Na_xMnO_2. Experimental voltage profile is obtained from Ref. 1.



(a) $x = 0.33$

(b) $x = 0.5$, top layer

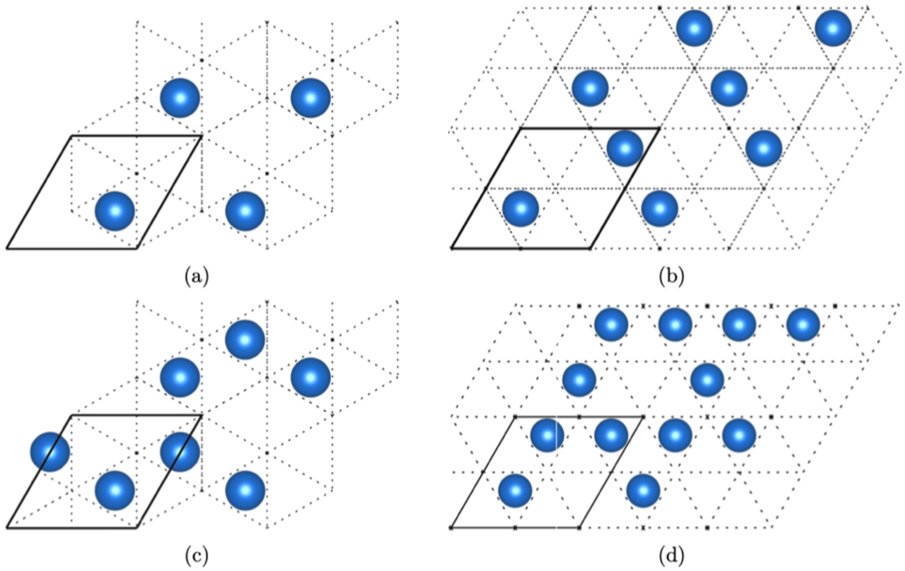
(c) $x = 0.5$, bottom layer

(d) $x = 0.625$, top layer

(e) $x = 0.625$, bottom layer

(f) $x = 0.75$

Figure A.2: Stable ground-state Na orderings of Na_xMnO_2 with PBE. Na(1) site (yellow circles); Na(2) site (blue circles).



(a)

(b)

(c)

(d)

Figure A.3: Stable ground-state Na orderings of Na_xMnO_2 in PBE + U at $x =$ (a) 0.33, (b) 0.5, (c) 0.67, (d) 0.75. Legend: Na(2) site (blue dot). Bold lines indicate unit cell.

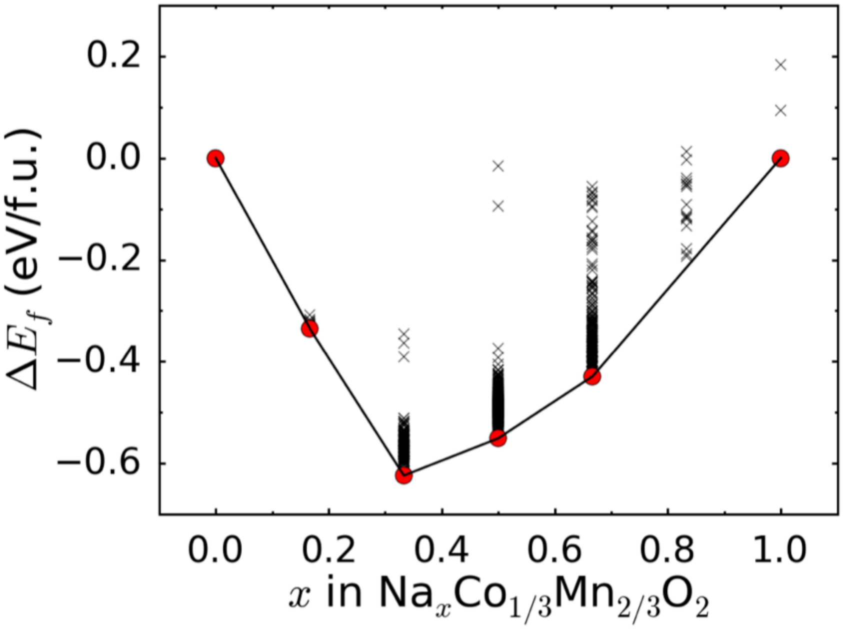


Figure A.4: PBE 0K stability diagram of $Na_xCo_{1/3}Mn_{2/3}O_2$. Black line: convex hull; red dots: stable orderings; black cross: unstable orderings.

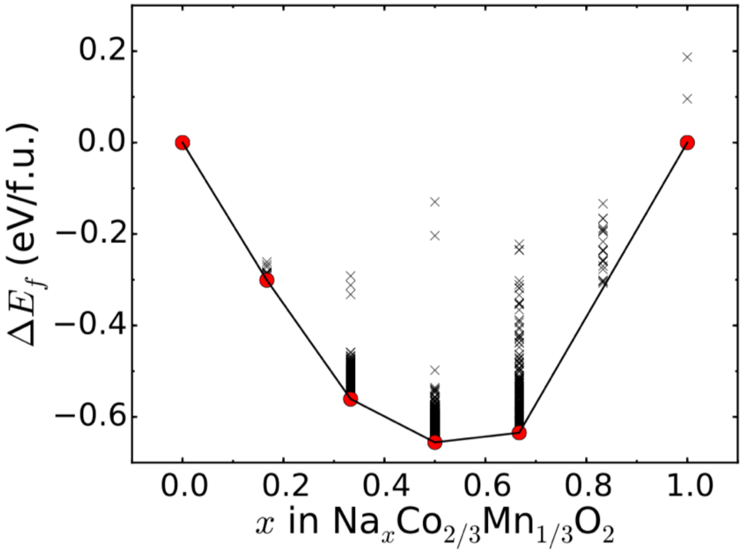
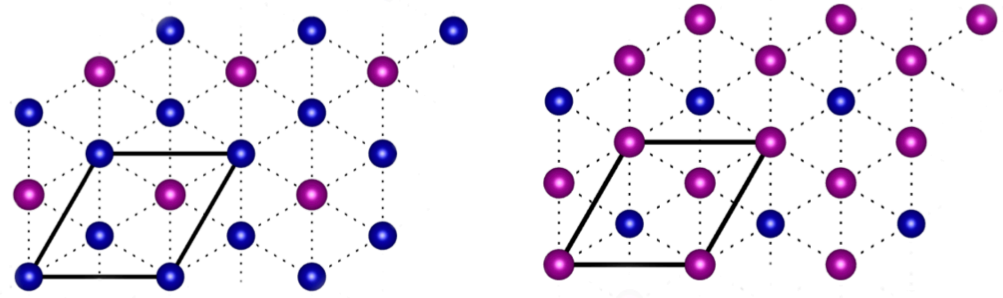


Figure A.5: PBE 0K stability diagram of $Na_xCo_{2/3}Mn_{1/3}O_2$. Black line: convex hull; red dots: stable orderings; black cross: unstable orderings.



(a) $Na_xCo_{2/3}Mn_{1/3}O_2$

(b) $Na_xCo_{1/3}Mn_{2/3}O_2$

Figure A.6: Lowest energy hexagonal Co-Mn orderings. Bold lines indicate the hexagonal ordering of Co-Mn. Co site (blue circles); Mn site (magenta circles).

Table A.1: Comparison of energies (meV atom^{-1}) of different Na orderings among different Co-Mn frameworks of $Na_{2/3}Co_{2/3}Mn_{1/3}O_2$. Each row has an identical Na ordering but different Co-Mn frameworks across columns. Each column has the same Co-Mn framework but different Na orderings. The value in each cell refers to the energy difference with respect to its corresponding Na ordering with the ground state Co-Mn framework. As can be observed, different Co-Mn orderings have a relatively small effect on the energy ($<$ 10 meV atom^{-1}).

Na orderings ID	Co-Mn Framework 1	Co-Mn Framework 2	Co-Mn Framework 3
1	0.0	0.0	1.4
2	0.3	0.3	7.6
3	0.0	5.7	6.1
4	0.7	0.8	11
5	0.4	1.7	7.9
6	0.9	0.2	0.8
7	0.1	0.3	0.5
8	0.1	0.2	0.1
9	0.2	0.0	0.3
10	0.5	0.0	0.2

Table A.2: Comparison of energies (meV/atom) of different Na orderings among different Co-Mn frameworks of $Na_{2/3}Co_{2/3}Mn_{1/3}O_2$. Each row has an identical Na ordering but different Co-Mn frameworks across columns. Each column has the same Co-Mn framework but different Na orderings. The value in each cell refers to the energy difference with respect to its corresponding Na ordering with the ground state Co-Mn framework. As can be observed, different Co-Mn orderings have a relatively small effect on the energy ($<$ 10 meV atom^{-1}).

Na orderings ID	Co-Mn	Co-Mn	Co-Mn
	Framework 1	Framework 2	Framework 3
1	-0.2	0.1	0.0
2	0.7	2.4	2.6
3	2.7	5.1	0.1
4	0.4	0.3	0.1
5	0.6	6.3	0.7
6	2.6	4.3	0.5
7	0.2	1.5	0.2
8	0.0	0.4	2.1
9	0.4	0.4	2.1
10	2.1	0.4	0.0

Table A.3: Average Na site occupancy fractions of $P2$ $\mathrm{Na_{1/2}Co_{1-y}Mn_yO_2}$ extracted from 25 ps of NVT AIMD simulations at 1000 K in Figure A.7. Site availability refers to the total proportion of such sites within the framework, while site occupancy refers to the actual occupancy during the AIMD simulations. The main observation is that sites containing Mn has a significantly lower site occupancy relative to site availability, suggesting that diffusing Na avoids such sites during the simulation.

Materials	Na site type	Site occupancy (simulation)	Site availability
$\mathrm{Na_{1/2}CoO_2}$	Na(1)$_{\text{Co-Co}}$	0.474	0.5
	Na(2)	0.526	0.5
$\mathrm{Na_{1/2}Co_{5/6}Mn_{1/6}O_2}$	Na(1)$_{\text{Co-Co}}$	0.31	0.333
	Na(1)$_{\text{Co-Mn}}$	0.096	0.167
	Na(2)	0.594	0.5
$\mathrm{Na_{1/2}Co_{2/3}Mn_{1/3}O_2}$	Na(1)$_{\text{Mn-Mn}}$	0.042	0.167
	Na(1)$_{\text{Co-Co}}$	0.387	0.333
	Na(2)	0.571	0.5
$\mathrm{Na_{1/2}Co_{2/3}Mn_{1/3}O_2^*}$	Na(1)$_{\text{Mn-Mn}}$	0.033	0.083
	Na(1)$_{\text{Co-Co}}$	0.265	0.25
	Na(1)$_{\text{Co-Mn}}$	0.128	0.167
	Na(2)	0.574	0.5
$\mathrm{Na_{1/2}Co_{1/2}Mn_{1/2}O_2}$	Na(1)$_{\text{Mn-Mn}}$	0.038	0.083
	Na(1)$_{\text{Co-Co}}$	0.095	0.083
	Na(1)$_{\text{Co-Mn}}$	0.238	0.333
	Na(2)	0.629	0.5
$\mathrm{Na_{1/2}Co_{1/3}Mn_{2/3}O_2}$	Na(1)$_{\text{Mn-Mn}}$	0.215	0.333
	Na(1)$_{\text{Co-Co}}$	0.202	0.167
	Na(2)	0.583	0.5
$\mathrm{Na_{1/2}MnO_2}$	Na(1)$_{\text{Mn-Mn}}$	0.484	0.5
	Na(2)	0.516	0.5

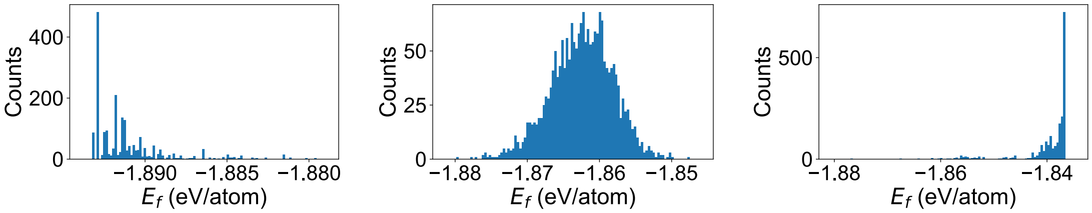
Table A.4: Na migration barriers of different Na(1) site configurations in selected $Co_{2/3}Mn_{1/3}O_2^*$ framework with and without Ni dopant.

Framework Materials	Dopant	Doping site	Na(1) site TM	NEB barrier (meV)
$Co_{2/3}Mn_{1/3}O_2^*$	N/A	N/A	Co-Co	102
	N/A	N/A	Co-Mn	153
	N/A	N/A	Mn-Mn	170
$Co_{2/3}Mn_{1/3}O_2^*$	Ni	Co	Ni-Co	114
	Ni	Co	Ni-Ni	81
$Co_{2/3}Mn_{1/3}O_2^*$	Ni	Mn	Ni-Mn	171
	Ni	Mn	Ni-Ni	151

Appendix B Supplementary Information Deep Learning Driven Study of High Entropy Cathode $\mathrm{Na}_x\mathrm{Co}_{0.2}\mathrm{Mn}_{0.2}\mathrm{Ti}_{0.2}\mathrm{Ni}_{0.2}\mathrm{Ru}_{0.2}\mathrm{O}_2$

Finding structures with predicted energy quantiles

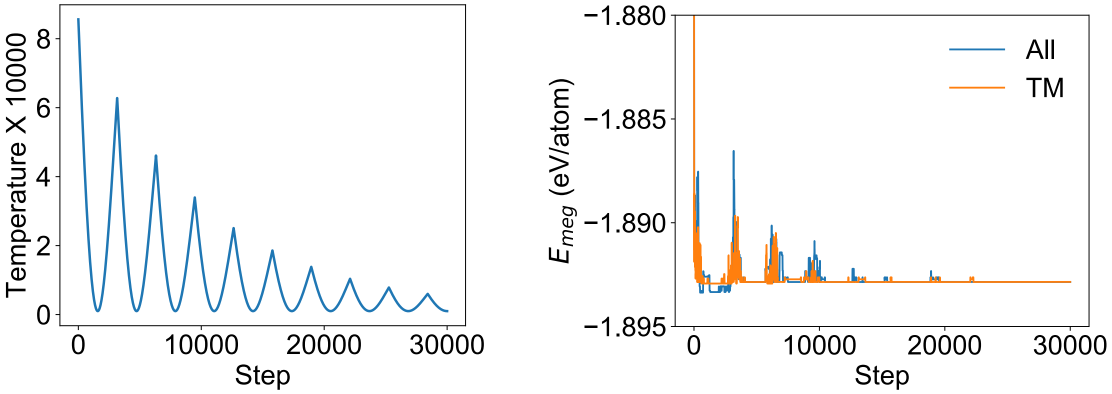
The structures with lowest energies [Figure B.1(a)] and highest energies [Figure B.1(c)] were find individually by simulated annealing and the structures with random energies were sampled randomly [Figure B.1(b)]. The structures with energy value quantiles were obtained by this method and DFT calculations were calculated.



(a) (b) (c)

Figure B.1: Simulated annealing and random sampling for determining the energy range for the disordered TMs. Simulated annealing is used to find the structures with lowest energies (a) and highest energies (c). Random sampling obtains structures with intermediate energies (b).

Finding lowest energy configurations



(a) (b)

Figure B.2: Temperature profile for simulated annealing (a) and MEGNet predicted energy evolution for only swapping TMs (TM) and also flipping the Na occupancy at the same time (All) (b)

Computation of entropy contributions

Here we assume an ideal mixing case, and the entropy contribution to the free energy can be computed as

$$-TS = -k_B T \sum_i x_i log x_i \tag{B.1}$$

Appendix C Supplementary Information Automated Generation and Ensemble-Learned Matching of X-ray Absorption Spectra

Table C.1: Materials used in benchmarking of FEFF parameters for K-edge XANES spectra calculations.[1]

Composition	Space Group	Materials Project Id	Absorbing Species	Reference
SiO_2	$P3_221$	mp-6930	Si	[167]
Si	$Fd\bar{3}m$	mp-149	Si	[167]
$AlPO_4$	$I\bar{4}$	mp-7848	Al	[167]
SiC	$F\bar{4}3m$	mp-8062	Si	[167]
Al_2O_3	$R\bar{3}c$	mp-1143	Al	[167]
Al	$Fm\bar{3}m$	mp-134	Al	[167]
Na_2O	$Fm\bar{3}m$	mp-2352	Na	[167]
C	$Fd\bar{3}m$	mp-66	C	[167]
B_2O_3	$P3_121$	mp-306	B	[167]
Si_3N_4	$P31c$	mp-2245	Si	[167]
Si_3N_4	$P6_3/m$	mp-988	Si	[167]
AlN	$P6_3mc$	mp-661	Al	[167]
NaCl	$Fm\bar{3}m$	mp-22862	Na	[167]

[1] There are two B K-edge XANES of B_2O_3, two C K-edge XANES of the diamond structure ($Fd\bar{3}m$) and three Si K-edge XANES of quartz alpha SiO_2 ($P3_221$) in the EELSDb. For spectra taken using the same structure, only one of them is adopted in our study, which reduces the number of K-edge XAS spectra included in the benchmarking dataset from 17 to 13.

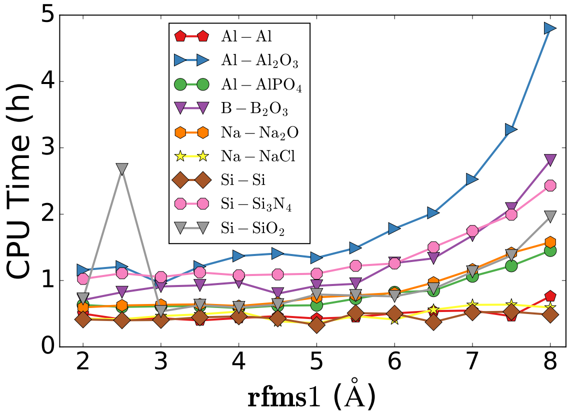


Figure C.1: Total computational time vs **rfms1** value in **SCF** card for K-edge XANES calculations. The first term in the label represents the absorption species. The second term represents the chemical compositions.

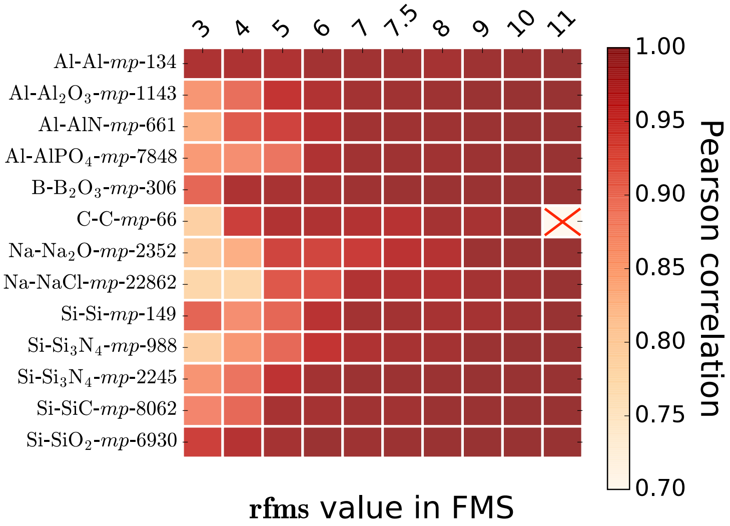


Figure C.2: Benchmark results of the K-edge XANES using different **rfms** value in the FMS card. Pair-wise spectra Pearson correlation coefficients are calculated considering **rfms** set to 11.0 as references. Calculation of C K edge XANES in diamond cubic crystal structure (mp-66) at **rfms** equals 11 fails due the insufficiency of node memory. For mp-66, the spectrum computed at **rfms** equals 10 is adopted as the converged referential spectrum. The first term in the label represents the absorption species. The second term represents the chemical formula. The third term corresponds to the mp-id.

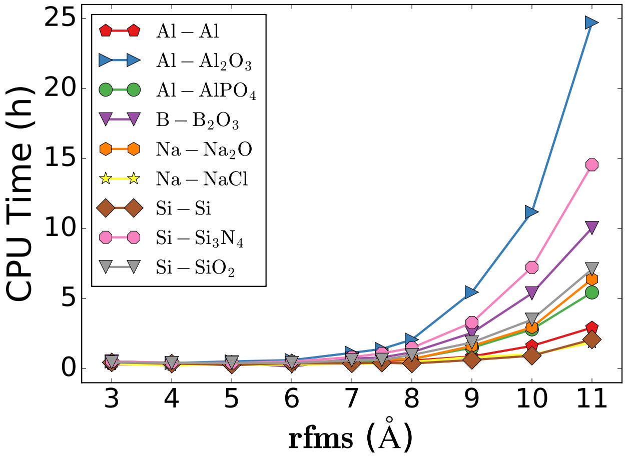
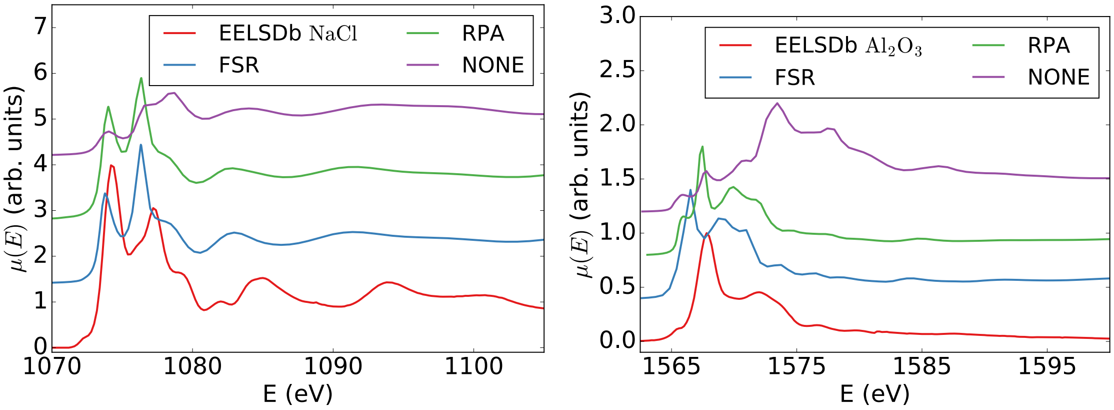


Figure C.3: Total computational time vs **rfms** value in FMS card. The first term in the label represents the absorption species. The second term represents the chemical formula.



(a)

(b)

Figure C.4: (a): Comparison of K-edge XANES spectra computed using different core-hole treatment approaches with experimental spectra from EELSDb[167] for (a) Na K-edge of NaCl and (b) Al K-edge of Al_2O_3. The theoretical curves are shifted vertically for clarity. The energy of the computed spectra of Na K-edge and Al K-edge has been shifted to align with experimental spectra (this shift will be discussed in the spectra matching section).

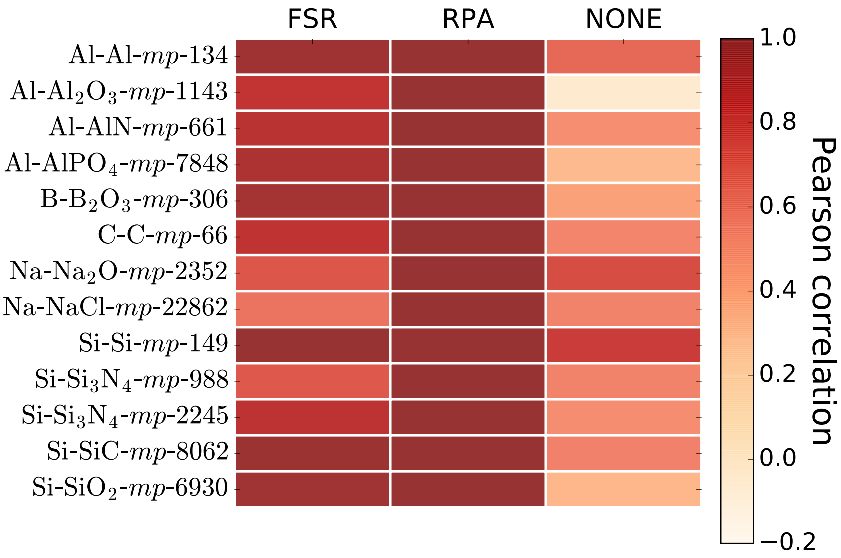


Figure C.5: Benchmark results of the K-edge XANES obtained with different core-hole treatment approaches. We use RPA as the default core-hole treatment setting.

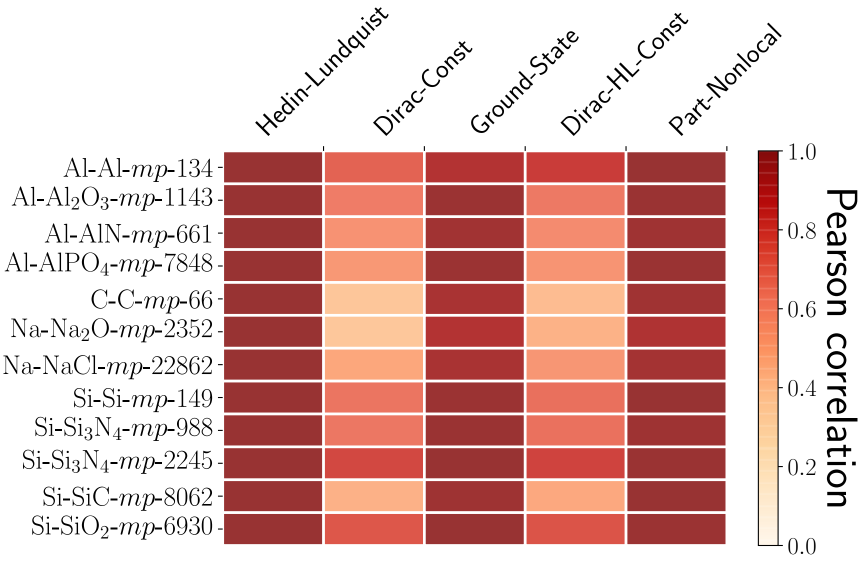


Figure C.6: Benchmark results of the K-edge XANES calculations obtained with different exchange models. We use the Hedin-Lundqvist functional as the default setting for energy dependent exchange correlation potential calculation. Using the Dirac-Hara exchange correlation potential causes computational instabilities of B K-edge XANES of B_2O_3, we excluded the B K-edge XANES results in this figure.

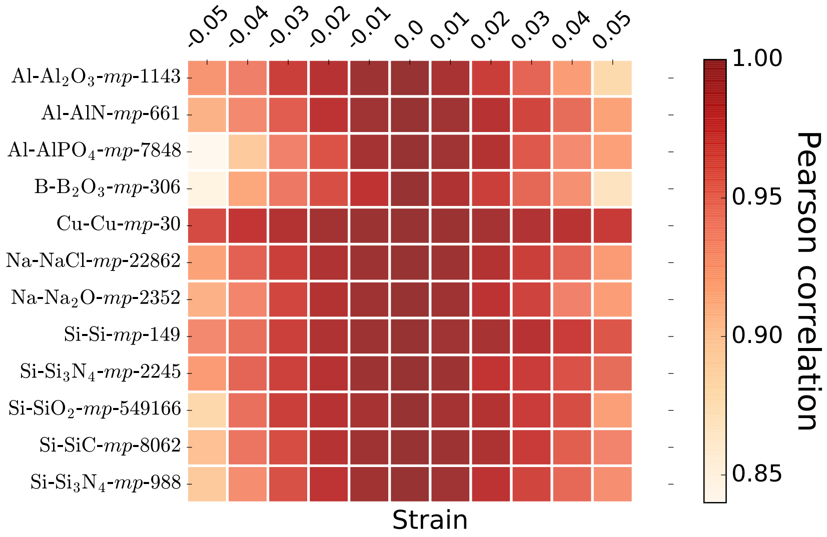


Figure C.7: Benchmark results of the K-edge XANES obtained using structures with applied strains. Pearson correlation coefficients between spectra are calculated using strain value set to 0.0 as reference. The relative Fermi energy shifts between the reference spectrum and the spectra calculated with applied strains are determined using the raw output data. Indeed, the Pearson correlation coefficients remain above 0.85 when the applied strain value ranges from -0.05 to 0.05 once the shift in Fermi energy is accounted for.

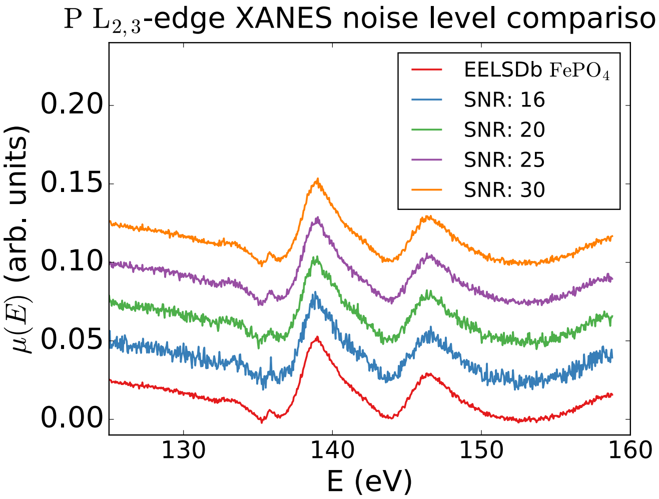
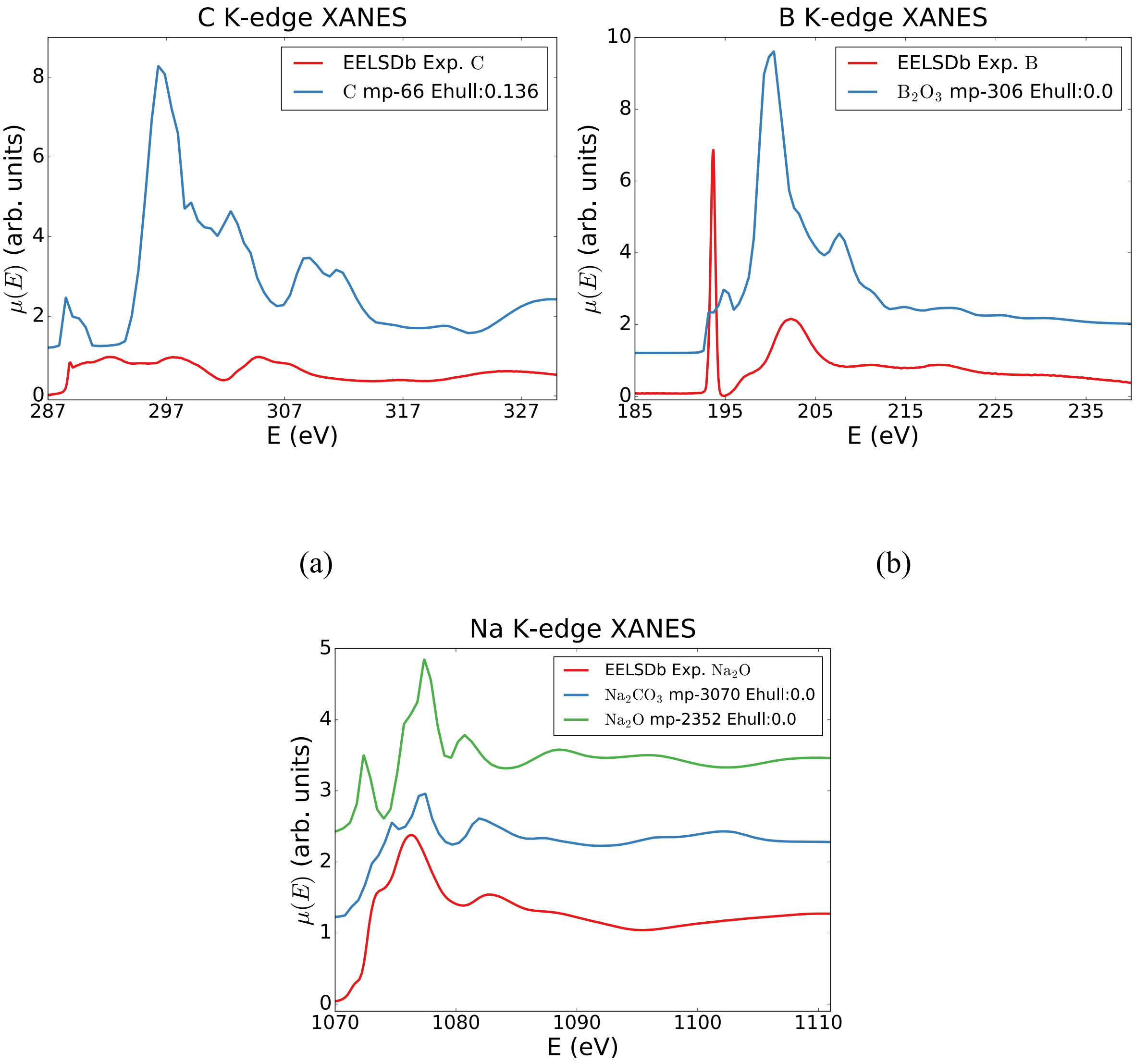


Figure C.8: Comparison of the P $L_{2,3}$-edge XANES for $FePO_4$ where Gaussian noise with different signal to noise ratio were manually added. The experiment reference P $L_{2,3}$-edge XANES $FePO_4$ is obtained from EELSDb.[167]



(a)

(b)

(c)

Figure C.9: Visualization of the divergence between the experimental target spectra and the FEFF computational spectra. Comparison between (a) the EELSDb experimental and the computed C K-edge XANES of the diamond structure ($Fd\bar{3}m$) and (b) the experimental and calculated B K-edge XANES spectra of B_2O_3. (c) Comparison between the Na K-edge XANES calculated using ground state Na_2CO_3, Na_2O structures in the Materials Project database and the EELSDb Na K-edge XANES of Na_2O. The energy of the computed spectra has been shifted to align with experimental spectra.

Selection of valid learners using distorted computational FEFF spectra

To examine and filter out learners with error rate below 0.5, distorted computational FEFF spectra were used for the valid learner selection. It is motivated by the fact that the spectra extracted from our computed spectra database provide a wealth of labeled data for training and validation.

We note that the XAS spectra in the EELSDb contains 7 distinguish absorbing species. We therefore constructed 7 computational spectra subgroups. Each subgroup is composed of all currently available K-edge XANES spectra from our XAS database with one absorbing specie included in EELSDb. We include carbon, oxygen and nitrogen in our chemical system construction to mimic the XAS measurement environment. There are total 453 spectra that meet the requirement [Table C.2]. In the following section, we will use these 453 spectra for the valid learner selection.

Gaussian noise and spectra shift were added to the 453 spectra. The signal to noise (SNR) ratio ranges from 16 to 30 at 2 intervals. The spectra shift values are 1, 3, 5 eV at two directions. A total number of 48 synthetic 'distorted' spectra of each spectrum were generated.

As each synthetic spectrum includes Gaussian noise, a Savitzky-Golay filter of width of 9 bins and polyorder of 4 is first applied to reduce the spectral noise. We note that this window width is much smaller than in the IFEFFIT software,[166] it is because most of our computed XANE spectra are featured with narrow overall energy range (< 100 eV), low SNR ratio (< 30) and shape peaks with narrow FWHM. Wider window width would therefore smear out the peak features and degrade the classification performance of the learner unexpectedly. The denoised spectra were then adopted in the spectral matching learner selection.

Considering each one of the 48 distorted computational spectrum as the target spectrum, the top 1 spectrum returned by a desired learner from our database is supposed to be the pristine spectrum we add distortions to. If the top one returned spectrum matches with the pristine spectrum, the correct count number of the learner increased by 1. For those learners included in spectral matching ensemble construction, the average correct count number over all 453 spectra needs to exceed 24. We note that this criterion is more strict than random guessing, as each valid learner is supposed to be capable of selecting the correct spectrum out from a spectra group with size range from 20 to 192.

It should be noted that, all original K-edge XAS of EELSDb show no noise. The noisiest XAS spectrum of EELSDb is the P $L_{2,3}$-edge XAS of $FePO_4$. As shown in Figure C.6, the SNR ratio of the P $L_{2,3}$-edge XAS is close to 30. According to our strain effect investigation, a 5 eV spectral shift is an indication of more than 10% difference in lattice constants. The spectrum matching ensemble constructed by the selected learners is expected to perform well under harsh conditions.

Table C.2: FEFF9 computed spectra used for learner validation and selection.

Chemical System	Absorbing Species	Chemical system spectra count
Al-P-O-N-C	Al	64
B-O-N-C	B	20
Na-O-C-N-Cl	Na	42
Si-O-C-N	Si	192
Li-Co-N-C-O	Co	31
Li-Ni-N-C-O	Ni	36
V-N-C-O	V	68

Appendix D Supplementary Information Accurate Chemical Environment Classification from X-ray Absorption Near-Edge Structure using a Random Forest Model

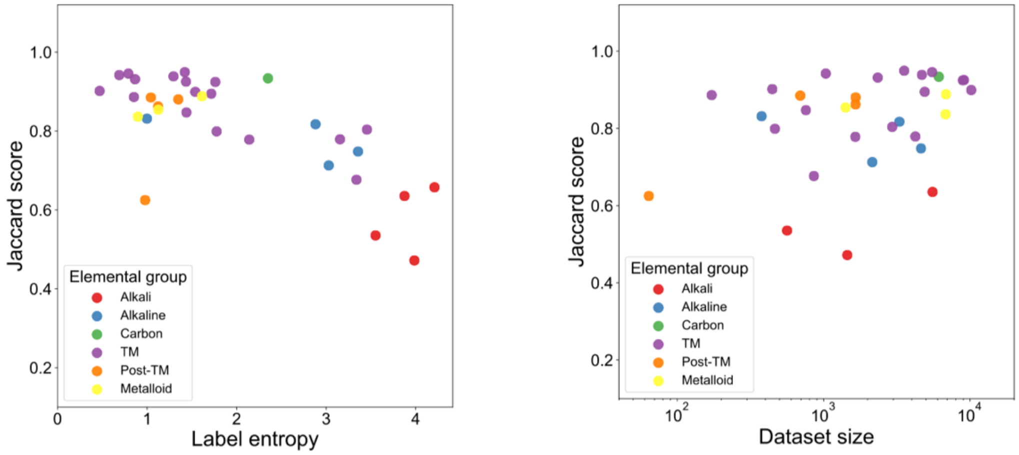


Figure D.1: Top coordination environment classification Jaccard scores of random forest classifier with respect to (a) datasets' label entropy categorized by elemental groups and (b) training dataset size.

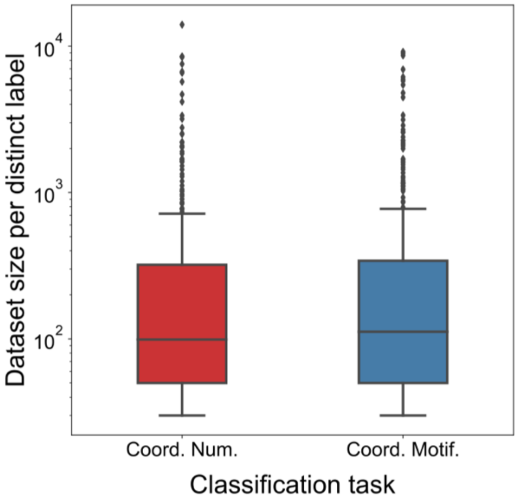


Figure D.2: Boxplots of dataset size distribution per distinct coordination number or coordination motif ranking label.

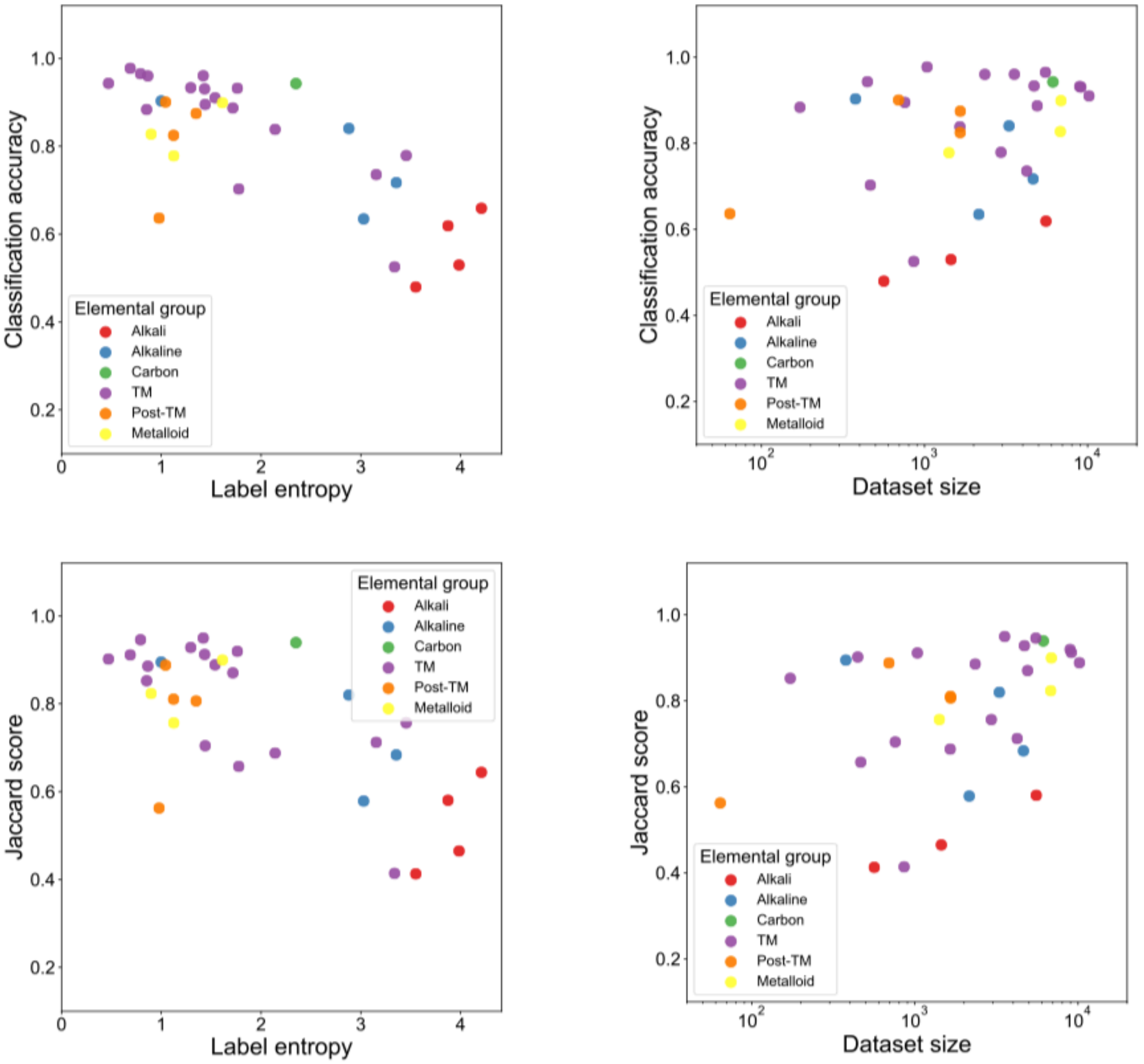


Figure D.3: Overview of convolutional neural network classifier's classification performance with respect to datasets' label entropy and training dataset size. Relationship between the top coordination environment classification accuracy and (a) datasets' label entropy and (b) training dataset size. Relationship between the Jaccard score and (c) datasets' label entropy and (d) training dataset size.

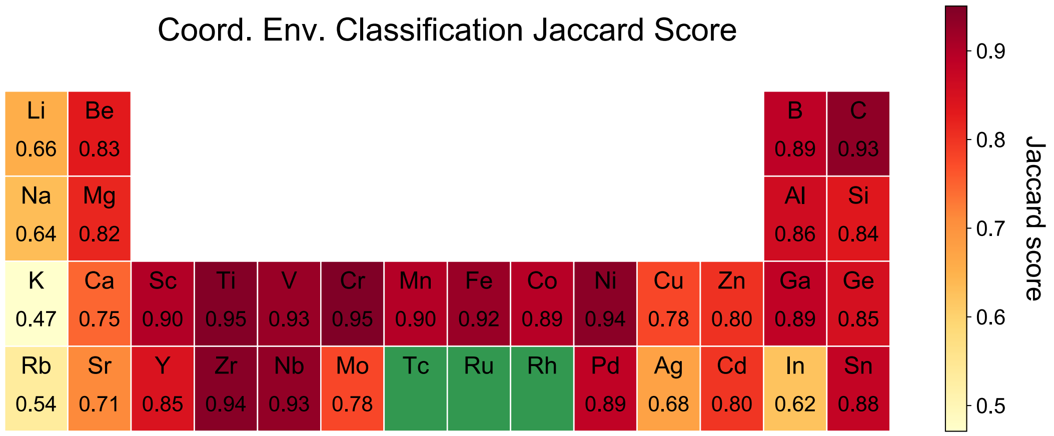
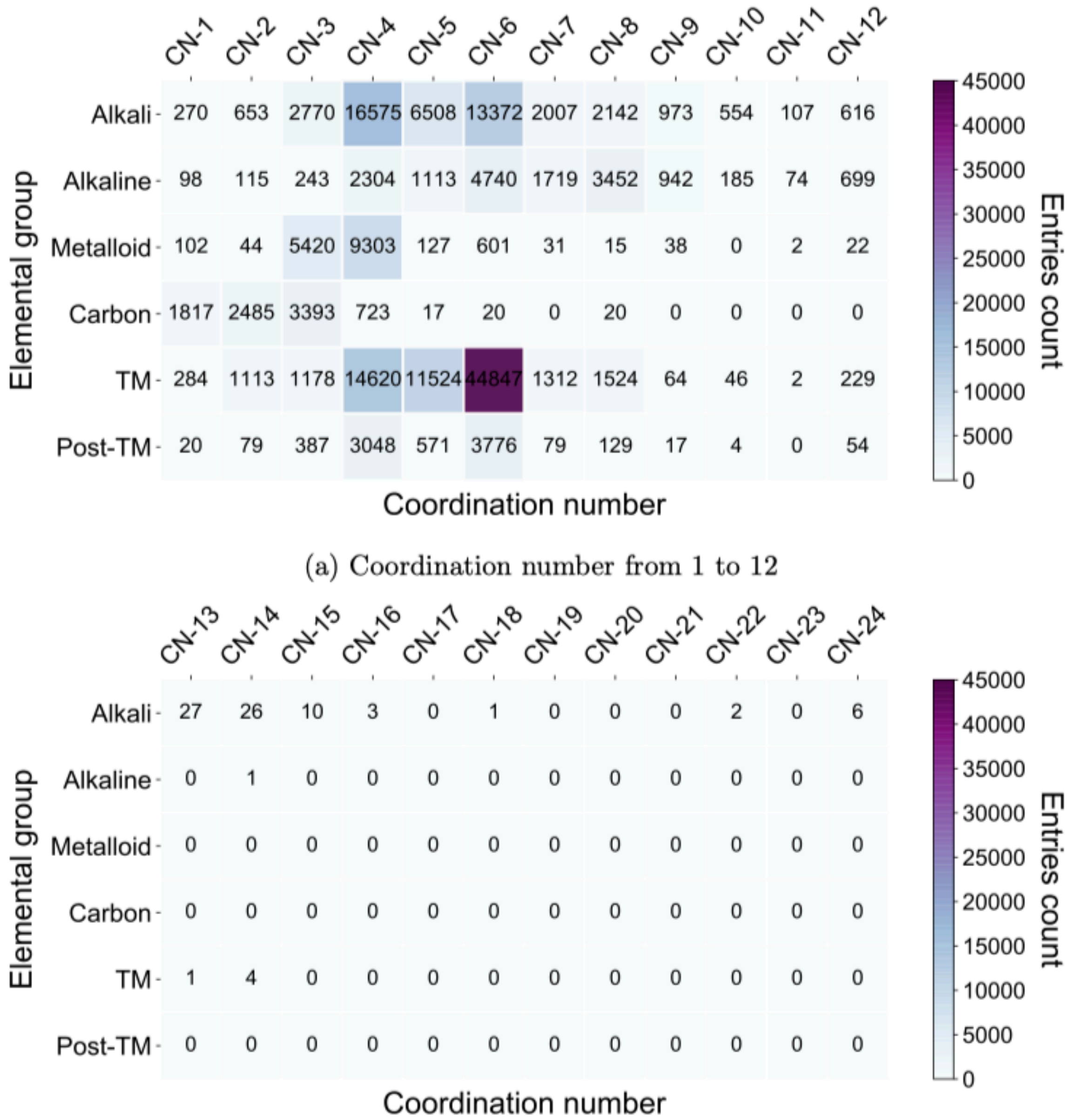


Figure D.4: The random forest classifier's element-wise classification Jaccard scores of coordination environment classification. We do not have sufficient sample sizes of computed Tc, Ru, Rh K-edge XANES to form a reliable training set for classification tasks.

Table D.1 Absorbing species and CNs with only one CM ranking label. Twelve coordinated ($q_{CN-12} \geq 0.2$) entries were excluded as their coordination environments all resemble the cuboctahedral coordination motif, i.e., $q_{\mathrm{cuboctahedral}} \geq 0.05$.

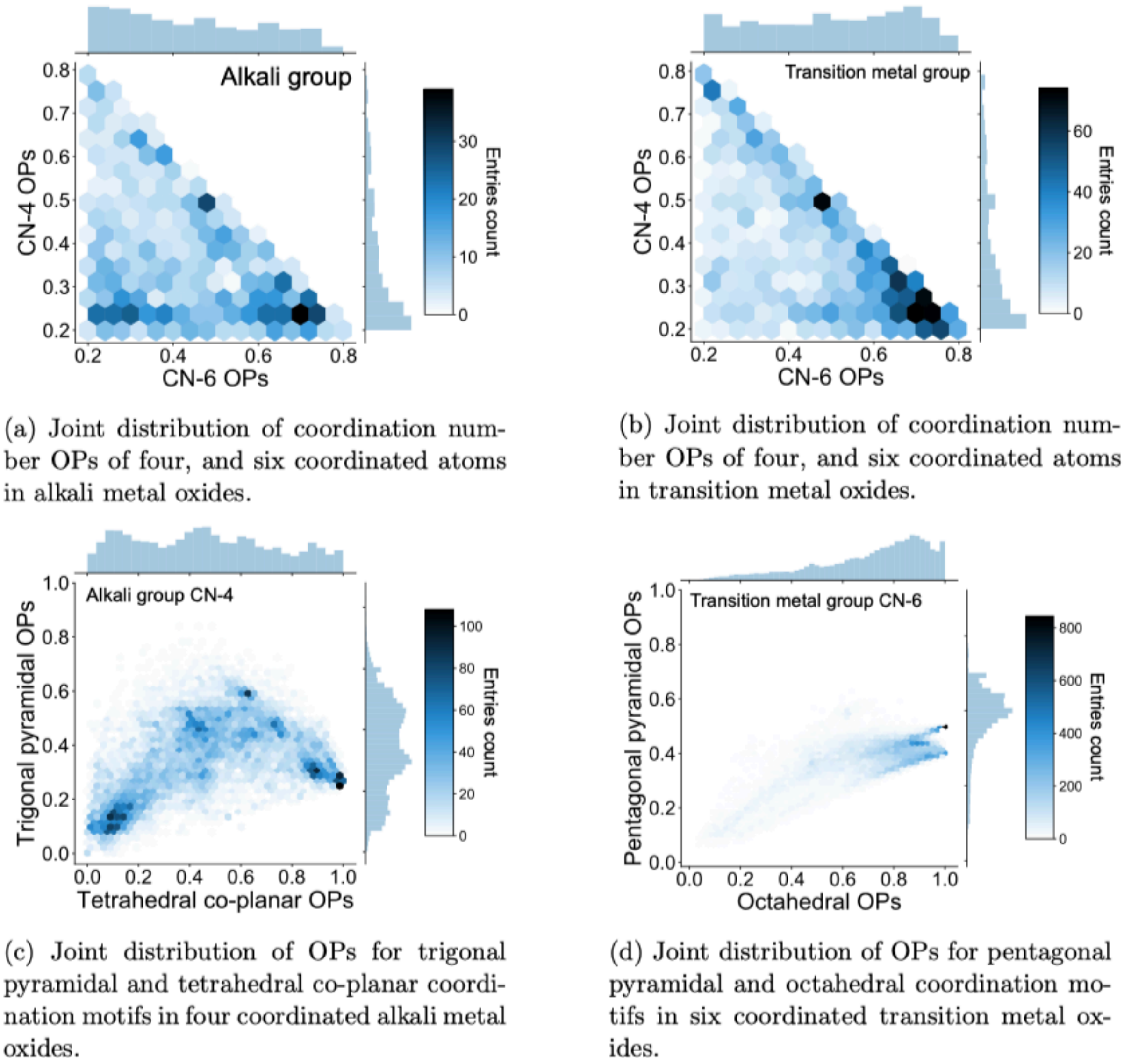
Absorbing specie	Coordination number ranking label	Coordination motif ranking label
Si	CN6	octahedral | pentagonal pyramidal | hexagonal planar
Al	CN6	octahedral | pentagonal pyramidal | hexagonal planar
Cd	CN5	trigonal bipyramidal | square pyramidal | pentagonal planar
In	CN6	octahedral | pentagonal pyramidal | hexagonal planar
Ge	CN6	octahedral | pentagonal pyramidal | hexagonal planar
Ru	CN6	octahedral | pentagonal pyramidal | hexagonal planar
Mg	CN7	pentagonal bipyramidal | hexagonal pyramidal
Sr	CN4	tetrahedral | trigonal pyramidal | seesaw like square co-planar
Mn	CN4	tetrahedral | trigonal pyramidal | seesaw like square co-planar
C	CN1	single bonds



(a) Coordination number from 1 to 12

(b) Coordination number from 13 to 24

Figure D.5: Counts of K-edge XANES entries with coordination number order parameters (OPs) larger than 0.2.



(a) Joint distribution of coordination number OPs of four, and six coordinated atoms in alkali metal oxides.

(b) Joint distribution of coordination number OPs of four, and six coordinated atoms in transition metal oxides.

(c) Joint distribution of OPs for trigonal pyramidal and tetrahedral co-planar coordination motifs in four coordinated alkali metal oxides.

(d) Joint distribution of OPs for pentagonal pyramidal and octahedral coordination motifs in six coordinated transition metal oxides.

Figure D.6: Joint distribution of CNs and CMs order parameters (OPs) of alkali group and transition metal group entries. Dark color represents high probability.

Table D.2: Coordination motif ranking labels prediction accuracy of optimized random forest classifiers on 27 experimental spectra

Formula	Space-group	Absorbing species	Correct CN labels	Correct CM labels	Correct CN-CM labels in top entry?	Data source
$LiCoO_2$	$R\bar{3}m$	Co	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	Ref.[238,239]
$LiNiO_2$	$R\bar{3}m$	Ni	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	Ref.[238,239]
NiO	$Fm\bar{3}m$	Ni	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	Ref.[238,239]
VO_2	$P2_1/c$	V	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	Ref.[238,239]
V_2O_5	$Pmmn$	V	CN-5	trigonal bipyramidal | pentagonal planar | square pyramidal	No	Ref.[238,239]
V_2O_3	$R\bar{3}c$	V	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	Ref.[238,239]
$AlPO_4$	$I\bar{4}$	Al	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	Yes	EELS Data Base[54]
B_2O_3	$P3_121$	B	CN-3	trigonal planar | trigonal non−coplanar | T−shaped	Yes	EELS Data Base[54]
SiO_2	$I\bar{4}2d$	Si	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	Yes	EELS Data Base[54]

Na_2O	$Fm\bar{3}m$	Na	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	No	EELS Data Base[54]
MnO	$Fm\bar{3}m$	Mn	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
MnO_2	$I4/m$	Mn	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
Mn_2O_3	$Pbca$	Mn	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
$K_2Cr_2O_7$	$P\bar{1}$	Cr	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	Yes	XAFS Library[140]
K_2CrO_4	$Pnma$	Cr	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	Yes	XAFS Library[140]
Cr_2O_3	$R\bar{3}c$	Cr	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
Na_2CrO_4	$Cmcm$	Cr	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	Yes	XAFS Library[140]
Fe_2O_3	$R\bar{3}c$	Fe	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
FeO	I4/mmm	Fe	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
ZnO	$P6_3mc$	Zn	CN-4	tetrahedral | trigonal pyramidal | see−saw−like | square co−planar	No	XAFS Library[140]

Ni_2O_3	$Cmcm$	Ni	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
CuO	$P4_2/mmc$	Cu	CN-4	square co−planar | rectangular see−saw−like | see−saw−like | trigonal pyramidal | tetrahedral	No	XAFS Library[140]
V_2O_5	$Pmmn$	V	CN-5	trigonal bipyramidal | pentagonal planar | square pyramidal	No	XAFS Library[140]
VO_2	$P4_2/mnm$	V	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
V_2O_3	$Ia3$	V	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
VO	$R\bar{3}m$	V	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]
CdO	$Fm\bar{3}m$	Cd	CN-6	octahedral | pentagonal pyramidal | hexagonal planar	Yes	XAFS Library[140]

Bibliography

(1) Jain, A.; Persson, K. A.; Ceder, G. Research Update: The Materials Genome Initiative: Data Sharing and the Impact of Collaborative Ab Initio Databases. *APL Mater.* **2016**, *4* (5), 053102.

(2) Eagar, T. W. Bringing New Materials to Market. *Technol Rev* **1995**, *98* (2), 42–49.

(3) Santanilla, A. B.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; et al. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015**, *347* (6217), 49–53.

(4) Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O'Keeffe, M.; Yaghi, O. M. High-Throughput Synthesis of Zeolitic Imidazolate Frameworks and Application to CO2 Capture. *Science* **2008**, *319* (5865), 939–943.

(5) Chung, Y. G.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Tool To Enable High-Throughput Screening of Nanoporous Crystals. **2014**.

(6) Stock, N.; Bein, T. High-Throughput Synthesis of Phosphonate-Based Inorganic–Organic Hybrid Compounds under Hydrothermal Conditions. *Angew. Chem. Int. Ed.* **2004**, *43* (6), 749–752.

(7) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76.

(8) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133–A1138.

(9) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136* (3B), B864–B871.

(10) Ong, S. P. Accelerating Materials Science with High-Throughput Computations and Machine Learning. *Comput. Mater. Sci.* **2019**, *161* (February), 143–150.

(11) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; et al. FireWorks: A Dynamic Workflow System Designed for High-Throughput Applications. *Concurr. Comput. Pract. Exp.* **2015**, *27* (17), 5037–5059.

(12) 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. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(13) Zhu, Z.; Chu, I.-H.; Ong, S. P. Li3Y(PS4)2 and Li5PS4Cl2: New Lithium Superionic Conductors Predicted from Silver Thiophosphates Using Efficiently Tiered Ab Initio Molecular Dynamics Simulations. *Chem. Mater.* **2017**, *29* (6), 2474–2484.

(14) Deng, Z.; Zhu, Z.; Chu, I.-H.; Ong, S. P. Data-Driven First-Principles Methods for the Study and Design of Alkali Superionic Conductors. *Chem. Mater.* **2017**, *29* (1), 281–288.

(15) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65* (11), 1501–1509.

(16) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; et al. AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012**, *58*, 227–235.

(17) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 011002.

(18) de Jong, M.; Chen, W.; Angsten, T.; Jain, A.; Notestine, R.; Gamst, A.; Sluiter, M.; Krishna Ande, C.; van der Zwaag, S.; Plata, J. J.; et al. Charting the Complete Elastic Properties of Inorganic Crystalline Compounds. *Sci. Data* **2015**, *2*, 150009.

(19) Tran, R.; Xu, Z.; Radhakrishnan, B.; Winston, D.; Sun, W.; Persson, K. A.; Ong, S. P. Surface Energies of Elemental Crystals. *Sci. Data* **2016**, *3*, 160080.

(20) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *Npj Comput. Mater.* **2017**, *3* (1), 1–13.

(21) Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349* (6245), 255–260.

(22) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* **2014**, *89* (9), 094104.

(23) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *Npj Comput. Mater.* **2016**, *2*, 16028.

(24) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31* (9), 3564–3572.

(25) Ye, W.; Chen, C.; Wang, Z.; Chu, I. H.; Ong, S. P. Deep Neural Networks for Accurate Predictions of Crystal Stability. *Nat. Commun.* **2018**, *9* (1), 1–6.

(26) de Jong, M.; Chen, W.; Notestine, R.; Persson, K.; Ceder, G.; Jain, A.; Asta, M.; Gamst, A. A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-Nary Inorganic Polycrystalline Compounds. *Sci Rep* **2016**, *6* (June), 34256.

(27) Yabuuchi, N.; Kubota, K.; Dahbi, M.; Komaba, S. Research Development on Sodium-Ion Batteries. *Chem. Rev.* **2014**, *114* (23), 11636–11682.

(28) Shimono, T.; Tanabe, D.; Kobayashi, W.; Moritomo, Y. Structural Response of P2-Type $\mathrm{Na}_x\mathrm{MnO}_2$ against Na^+ Intercalation. *J. Phys. Soc. Jpn.* **2013**, *82* (8), 083601.

(29) Ong, S. P.; Chevrier, V. L.; Hautier, G.; Jain, A.; Moore, C.; Kim, S.; Ma, X.; Ceder, G. Voltage, Stability and Diffusion Barrier Differences between Sodium-Ion and Lithium-Ion Intercalation Materials. *Energy Environ. Sci.* **2011**, *4* (9), 3680.

(30) Yabuuchi, N.; Hara, R.; Kubota, K.; Paulsen, J.; Kumakura, S.; Komaba, S. A New Electrode Material for Rechargeable Sodium Batteries: P2-Type Na2/3[Mg0.28Mn 0.72]O2 with Anomalously High Reversible Capacity. *J Mater Chem A* **2014**, *2*, 16851–16855.

(31) Yabuuchi, N.; Kajiyama, M.; Iwatate, J.; Nishikawa, H.; Hitomi, S.; Okuyama, R.; Usui, R.; Yamada, Y.; Komaba, S. P2-Type Na(x)[Fe(1/2)Mn(1/2)]O2 Made from Earth-Abundant Elements for Rechargeable Na Batteries. *Nat. Mater.* **2012**, *11* (6), 512–517.

(32) Wang, X.; Tamaru, M.; Okubo, M.; Yamada, A. Electrode Properties of P2–Na2/3MnyCo1–YO2 as Cathode Materials for Sodium-Ion Batteries. *J. Phys. Chem. C* **2013**, *117* (30), 15545–15551.

(33) Zhao, J.; Xu, J.; Lee, D. H.; Dimov, N.; Meng, Y. S.; Okada, S. Electrochemical and Thermal Properties of P2-Type Na2/3Fe1/3Mn2/3O2 for Na-Ion Batteries. *J. Power Sources* **2014**, *264*, 235–239.

(34) Delmas, C.; Fouassier, C.; Hagenmuller, P. Structural Classification and Properties of the Layered Oxides. *Phys. BC* **1980**, *99* (1–4), 81–85.

(35) Berthelot, R.; Carlier, D.; Delmas, C. Electrochemical Investigation of the P2–NaxCoO2 Phase Diagram. *Nat. Mater.* **2011**, *10* (1), 74–80.

(36) Caballero, a.; Hernán, L.; Morales, J.; Sánchez, L.; Santos Peña, J.; Aranda, M. a. G. Synthesis and Characterization of High-Temperature Hexagonal P2-Na0.6 MnO2 and Its Electrochemical Behaviour as Cathode in Sodium Cells. *J. Mater. Chem.* **2002**, *12* (4), 1142–1147.

(37) Thorne, J. S.; Dunlap, R. A.; Obrovac, M. N. Investigation of P2-Na2/3Mn1/3Fe1/3Co1/3O2 for Na-Ion Battery Positive Electrodes. *J. Electrochem. Soc.* **2014**, *161* (14), A2232–A2236.

(38) Kim, D.; Kang, S.-H.; Slater, M.; Rood, S.; Vaughey, J. T.; Karan, N.; Balasubramanian, M.; Johnson, C. S. Enabling Sodium Batteries Using Lithium-Substituted Sodium Layered Transition Metal Oxide Cathodes. *Adv. Energy Mater.* **2011**, *1* (3), 333–336.

(39) Meng, Y. S.; Hinuma, Y.; Ceder, G. An Investigation of the Sodium Patterning in NaxCoO2 (0.5⩽x⩽1) by Density Functional Theory Methods. *J. Chem. Phys.* **2008**, *128* (10), 104708.

(40) Wang, Y.; Ni, J. Ground State Structure of Sodium Ions in NaxCoO2: A Combined Monte Carlo and First-Principles Approach. *Phys. Rev. B* **2007**, *76* (9), 1–5.

(41) Hinuma, Y.; Meng, Y. S.; Ceder, G. Temperature-Concentration Phase Diagram of P2 $-$ $\mathrm{Na}_x\mathrm{CoO}_2$ from First-Principles Calculations. *Phys. Rev. B* **2008**, *77* (22), 224111.

(42) Nelson, L. J.; Ozoliņš, V.; Reese, C. S.; Zhou, F.; Hart, G. L. W. Cluster Expansion Made Easy with Bayesian Compressive Sensing. *Phys. Rev. B* **2013**, *88* (15), 155105.

(43) Capelle, K. A Bird's-Eye View of Density-Functional Theory. **2002**, 69.

(44) Koningsberger, D. C.; Prins, R. *X-Ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS, and XANES*; Wiley-Interscience, New York, 1988.

(45) Bunker, G. *Introduction to XAFS: A Practical Guide to X-Ray Absorption Fine Structure Spectroscopy*; Cambridge University Press, 2010.

(46) Lin, Y.-C.; Wen, B.; Wiaderek, K. M.; Sallis, S.; Liu, H.; Lapidus, S. H.; Borkiewicz, O. J.; Quackenbush, N. F.; Chernova, N. A.; Karki, K.; et al. Thermodynamics, Kinetics and Structural Evolution of ε-LiVOPO 4 over Multiple Lithium Intercalation. *Chem. Mater.* **2016**, *28* (6), 1794–1805.

(47) Quackenbush, N. F.; Paik, H.; Woicik, J. C.; Arena, D. A.; Schlom, D. G.; Piper, L. F. J.; Source-ii, N. S. L. X-Ray Spectroscopy of Ultra-Thin Oxide/Oxide Heteroepitaxial Films: A Case Study of Single-Nanometer VO2/TiO2. **2015**, *2* (100), 5452–5466.

(48) Cheng, J.-H.; Pan, C.-J.; Lee, J.-F.; Chen, J.-M.; Guignard, M.; Delmas, C.; Carlier, D.; Hwang, B.-J. Simultaneous Reduction of Co3+ and Mn4+ in P2-Na2/3Co2/3Mn1/3O2 As Evidenced by X-Ray Absorption Spectroscopy during Electrochemical Sodium Intercalation. *Chem. Mater.* **2014**, *26* (2), 1219–1225.

(49) Newville, M. Fundamentals of XAFS. *Rev. Mineral. Geochem.* **2014**, *78* (1), 33–74.

(50) Jung, Y. H.; Christiansen, A. S.; Johnsen, R. E.; Norby, P.; Kim, D. K. In Situ X-Ray Diffraction Studies on Structural Changes of a P2 Layered Material during Electrochemical Desodiation/Sodiation. *Adv. Funct. Mater.* **2015,** *25* (21), 3227–3237.

(51) Manceau, A.; Marcus, M.; Lenoir, T. Estimating the Number of Pure Chemical Components in a Mixture by X-Ray Absorption Spectroscopy. *J. Synchrotron Radiat.* **2014**, *21* (5), 1140–1147.

(52) Fay, M. J.; Proctor, A.; Hoffmann, D. P.; Houalla, M.; Hercules, D. M. Determination of the Mo Surface Environment of Mo/TiO2 Catalysts by EXAFS, XANES and PCA. *Mikrochim. Acta* **1992**, *109* (5–6), 281–293.

(53) Ravel, B.; Newville, M. ATHENA, ARTEMIS, HEPHAESTUS: Data Analysis for X-Ray Absorption Spectroscopy Using IFEFFIT. *J. Synchrotron Radiat.* **2005**, *12* (4), 537–541.

(54) Ewels, P.; Sikora, T.; Serin, V.; Ewels, C. P.; Lajaunie, L. A Complete Overhaul of the Electron Energy-Loss Spectroscopy and X-Ray Absorption Spectroscopy Database: Eelsdb.Eu. *Microsc. Microanal.* **2016**, *22* (03), 717–724.

(55) Suplee, C. XCOM: Photon Cross Sections Database https://www.nist.gov/pml/xcom-photon-cross-sections-database (accessed Jun 24, 2019).

(56) XAS Spectra Library https://cars.uchicago.edu/xaslib (accessed Jun 10, 2019).

(57) Chang, J. Y.; Lin, B. N.; Hsu, Y. Y.; Ku, H. C. Co K-Edge XANES and Spin-State Transition of RCoO3(R = La, Eu). *Phys. B Condens. Matter* **2003**, *329–333* (II), 826–828.

(58) Farges, E. C. Æ. F.; Jr, Æ. G. E. B.; Xanes, M. Á. G. Á. A Pre-Edge Analysis of Mn K-Edge XANES Spectra to Help Determine the Speciation of Manganese in Minerals and Glasses. **2009**, 111–126.

(59) Sano, M.; Komorita, S.; Yamatera, H. XANES Spectra of Copper(II) Complexes: Correlation of the Intensity of the 1s .Fwdarw. 3d Transition and the Shape of the Complex. *Inorg. Chem.* **1992**, *31* (3), 459–463.

(60) Le Fèvre, P.; Magnan, H.; Chandesris, D.; Jupille, J.; Bourgeois, S.; Barbier, A.; Drube, W.; Uozumi, T.; Kotani, A. Hard X-Ray Resonant Electronic Spectroscopy in Transition Metal Oxides. *Nucl. Instrum. Methods Phys. Res. Sect. Accel. Spectrometers Detect. Assoc. Equip.* **2005**, *547* (1), 176–186.

(61) 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.* **2008**, *37* (6), 572–584.

(62) 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.* **2017**, *8* (20), 5091–5098.

(63) Carbone, M. R.; Yoo, S.; Topsakal, M.; Lu, D. Classification of Local Chemical Environments from X-Ray Absorption Spectra Using Supervised Machine Learning. *Phys. Rev. Mater.* **2019**, *3* (3), 033604.

(64) 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.* **2010**, *12* (21), 5503.

(65) Zimmermann, N. E. R.; Horton, M. K.; Jain, A.; Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Front. Mater.* **2017**, *4* (November), 1–13.

(66) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(67) Belgiu, M.; Drăguţ, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31.

(68) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer Texts in Statistics; Springer-Verlag: New York, 2013.

(69) Qing Song; Wenjie Hu; Wenfang Xie. Robust Support Vector Machine with Bullet Hole Image Classification. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2002**, *32* (4), 440–448.

(70) Chatzidakis, M.; Botton, G. A. Towards Calibration-Invariant Spectroscopy Using Deep Learning. *Sci. Rep.* **2019**, *9* (1), 1–10.

(71) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444.

(72) Kubota, K.; Komaba, S. Review—Practical Issues and Future Perspective for Na-Ion Batteries. *J. Electrochem. Soc.* **2015**, *162* (14), A2538–A2550.

(73) Hayashi, A.; Noi, K.; Tanibata, N.; Nagao, M.; Tatsumisago, M. High Sodium Ion Conductivity of Glass–Ceramic Electrolytes with Cubic Na3PS4. *J. Power Sources* **2014**, *258*, 420–423.

(74) Zhu, Z.; Chu, I.-H.; Deng, Z.; Ong, S. P. Role of Na+ Interstitials and Dopants in Enhancing the Na+ Conductivity of the Cubic Na3PS4 Superionic Conductor. *Chem. Mater.* **2015**, *27* (24), 8318–8325.

(75) Nanjundaswamy, K. Synthesis, Redox Potential Evaluation and Electrochemical Characteristics of NASICON-Related-3D Framework Compounds. *Solid State Ion.* **1996**, *92* (1–2), 1–10.

(76) Chu, I.-H.; Christopher, S. K.; Han, N.; Zhu, Z.; Sunny, H.; Deng, Z.; Meng, Y. S.; Ong, S. P. Room-Temperature All-Solid-State Rechargeable Sodium-Ion Batteries with a Cl-Doped Na3PS4 Superionic Conductor. *Sci. Rep.* **2016**, *6*, 33733.

(77) Hasa, I.; Passerini, S.; Hassoun, J. A Rechargeable Sodium-Ion Battery Using a Nanostructured Sb–C Anode and P2-Type Layered Na0.6Ni0.22Fe0.11Mn0.66O2 Cathode. *RSC Adv.* **2015**, *5* (60), 48928–48934.

(78) Yoshida, H.; Yabuuchi, N.; Kubota, K.; Ikeuchi, I.; Garsuch, A.; Schulz-Dobrick, M.; Komaba, S. P2-Type Na2/3Ni1/3Mn2/3−xTixO2 as a New Positive Electrode for Higher Energy Na-Ion Batteries. *Chem. Commun.* **2014**, *50* (28), 3677.

(79) Yabuuchi, N.; Yano, M.; Yoshida, H.; Kuze, S.; Komaba, S. Synthesis and Electrode Performance of O3-Type NaFeO2-NaNi1/2Mn1/2O2 Solid Solution for Rechargeable Sodium Batteries. *J. Electrochem. Soc.* **2013**, *160* (5), A3131–A3137.

(80) Thorne, J. S.; Dunlap, R. A.; Obrovac, M. N. Structure and Electrochemistry of NaxFex Mn1-xO2 (1.0 ≤ x ≤ 0.5) for Na-Ion Battery Positive Electrodes. **2013**, *160* (2), 361–367.

(81) Park, K.; Han, D.; Kim, H.; Chang, W.; Choi, B.; Anass, B.; Lee, S. Characterization of a P2-Type Chelating-Agent-Assisted Na2/3Fe1/2Mn1/2O2 Cathode Material for Sodium-Ion Batteries. *RSC Adv.* **2014**, *4* (43), 22798.

(82) Talaie, E.; Duffort, V.; Smith, H. L.; Fultz, B.; Nazar, L. F. Structure of the High Voltage Phase of Layered P2-Na2/3−z[Mn1/2Fe1/2]O2 and the Positive Effect of Ni Substitution on Its Stability. *Energy Env. Sci* **2015**, *8* (8), 2512–2523.

(83) Li, Z.; Gao, R.; Sun, L.; Hu, Z.; Liu, X. Designing an Advanced P2-Na0.67Mn0.65Ni0.2Co0.15O2 Layered Cathode Material for Na-Ion Batteries. *J Mater Chem A* **2015**, *3*, 16272–16278.

(84) Mortemard de Boisse, B.; Cheng, J.-H.; Carlier, D.; Guignard, M.; Pan, C.-J.; Bordère, S.; Filimonov, D.; Drathen, C.; Suard, E.; Hwang, B.-J.; et al. O3–NaxMn1/3Fe2/3O2 as a Positive Electrode Material for Na-Ion Batteries: Structural Evolutions and Redox Mechanisms upon Na+ (de)Intercalation. *J Mater Chem A* **2015**, *3* (20), 10976–10989.

(85) Wu, X.; Guo, J.; Wang, D.; Zhong, G.; McDonald, M. J.; Yang, Y. P2-Type Na0.66Ni0.33–XZnxMn0.67O2 as New High-Voltage Cathode Materials for Sodium-Ion Batteries. *J. Power Sources* **2014**.

(86) Ding, J. J.; Zhou, Y. N.; Sun, Q.; Yu, X. Q.; Yang, X. Q.; Fu, Z. W. Electrochemical Properties of P2-Phase Na0.74CoO2 Compounds as Cathode Material for Rechargeable Sodium-Ion Batteries. *Electrochimica Acta* **2013**, *87*, 388–393.

(87) Huang, Q.; Foo, M. L.; Lynn, J. W.; Zandbergen, H. W.; Lawes, G.; Wang, Y.; Toby, B. H.; Ramirez, A. P.; Ong, N. P.; Cava, R. J. Low Temperature Phase Transitions and Crystal Structure of Na0.5CoO2. *J. Phys. Condens. Matter* **2004**, *16* (32), 5803.

(88) Meng, Y.; Van der Ven, a.; Chan, M.; Ceder, G. Ab Initio Study of Sodium Ordering in Na0.75CoO2 and Its Relation to Co3+⁄Co4+ Charge Ordering. *Phys. Rev. B* **2005**, *72* (17), 172103.

(89) Zhang, P.; Capaz, R.; Cohen, M.; Louie, S. Theory of Sodium Ordering in NaxCoO2. *Phys. Rev. B* **2005**, *71* (15), 153102.

(90) Shu, G.; Chou, F. Sodium-Ion Diffusion and Ordering in Single-Crystal P2-NaxCoO2. *Phys. Rev. B* **2008**, *78* (5), 052101.

(91) Xu, J.; Lee, D. H.; Clément, R. J.; Yu, X.; Leskes, M.; Pell, A. J.; Pintacuda, G.; Yang, X.-Q.; Grey, C. P.; Meng, Y. S. Identifying the Critical Role of Li Substitution in P2–Na

x[LiyNizMn1– y – z]O 2 (0 < x , y , z < 1) Intercalation Cathode Materials for High-Energy Na-Ion Batteries. *Chem. Mater.* **2014**, *26* (2), 1260–1269.

(92) Liu, L.; Li, X.; Bo, S.-H.; Wang, Y.; Chen, H.; Twu, N.; Wu, D.; Ceder, G. High-Performance P2-Type Na2/3(Mn1/2Fe1/4Co1/4)O2 Cathode Material with Superior Rate Capability for Na-Ion Batteries. *Adv. Energy Mater.* **2015**, *5* (22), 1500944.

(93) Carlier, D.; Cheng, J. H.; Berthelot, R.; Guignard, M.; Yoncheva, M.; Stoyanova, R.; Hwang, B. J.; Delmas, C. The P2-Na(2/3)Co(2/3)Mn(1/3)O2 Phase: Structure, Physical Properties and Electrochemical Behavior as Positive Electrode in Sodium Battery. *Dalton Trans. Camb. Engl. 2003* **2011**, *40* (36), 9306–9312.

(94) Clément, R. J.; Bruce, P. G.; Grey, C. P. Review—Manganese-Based P2-Type Transition Metal Oxides as Sodium-Ion Battery Cathode Materials. *J. Electrochem. Soc.* **2015**, *162* (14), A2589–A2604.

(95) Moradabadi, A.; Kaghazchi, P. Mechanism of Li Intercalation/Deintercalation into/from the Surface of LiCoO2. *Phys Chem Chem Phys* **2015**, *17* (35), 22917–22922.

(96) Mo, Y.; Ong, S. P.; Ceder, G. Insights into Diffusion Mechanisms in P2 Layered Oxide Materials by First-Principles Calculations. *Chem. Mater.* **2014**, *26* (18), 5208–5214.

(97) Guo, S.; Sun, Y.; Yi, J.; Zhu, K.; Liu, P.; Zhu, Y.; Zhu, G.; Chen, M.; Ishida, M.; Zhou, H. Understanding Sodium-Ion Diffusion in Layered P2 and P3 Oxides via Experiments and First-Principles Calculations: A Bridge between Crystal Structure and Electrochemical Performance. *NPG Asia Mater.* **2016**, *8* (4), e266.

(98) Hasa, I.; Buchholz, D.; Passerini, S.; Hassoun, J. A Comparative Study of Layered Transition Metal Oxide Cathodes for Application in Sodium-Ion Battery. *ACS Appl. Mater. Interfaces* **2015**, *7* (9), 5206–5212.

(99) Kuo, L.-Y.; Moradabadi, A.; Huang, H.-F.; Hwang, B.-J.; Kaghazchi, P. Structure and Ionic Conductivity of the Solid Electrolyte Interphase Layer on Tin Anodes in Na-Ion Batteries. *J. Power Sources* **2017**, *341*, 107–113.

(100) Hart, G. L. W.; Forcade, R. W. Generating Derivative Structures from Multilattices: Algorithm and Application to Hcp Alloys. *Phys. Rev. B* **2009**, *80* (1), 014120.

(101) Toumar, A. J.; Ong, S. P.; Richards, W. D.; Dacek, S.; Ceder, G. Vacancy Ordering in O3 -Type Layered Metal Oxide Sodium-Ion Battery Cathodes. *Phys. Rev. Appl.* **2015**, *4* (6), 1–9.

(102) Lu, Z.; Donaberger, R. a.; Dahn, J. R. Superlattice Ordering of Mn, Ni, and Co in Layered Alkali Transition Metal Oxides with P2, P3, and O3 Structures. *Chem. Mater.* **2000**, *12* (12), 3583–3590.

(103) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.

(104) Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, a. P. Electron-Energy-Loss Spectra and the Structural Stability of Nickel Oxide: An LSDA+U Study. *Phys. Rev. B* **1998**, *57* (3), 1505–1509.

(105) Wang, L.; Maxisch, T.; Ceder, G. Oxidation Energies of Transition Metal Oxides within the GGA+U Framework. *Phys. Rev. B* **2006**, *73* (19), 195107.

(106) Jain, A.; Hautier, G.; Ong, S. P.; Moore, C. J.; Fischer, C. C.; Persson, K. a.; Ceder, G. Formation Enthalpies by Mixing GGA and GGA + U Calculations. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2011**, *84*, 1–10.

(107) Su, J.; Pei, Y.; Yang, Z.; Wang, X. First-Principles Investigation on the Structural, Electronic Properties and Diffusion Barriers of Mg/Al Doped NaCoO 2 as the Cathode Material of Rechargeable Sodium Batteries. *RSC Adv* **2015**, *5* (35), 27229–27234.

(108) Ong, S. P.; Wang, L.; Kang, B.; Ceder, G. Li - Fe - P - O2 Phase Diagram from First Principles Calculations. *Chem. Mater.* **2008**, *20* (4), 1798–1807.

(109) Aydinol, M.; Kohan, a.; Ceder, G.; Cho, K.; Joannopoulos, J. Ab Initio Study of Lithium Intercalation in Metal Oxides and Metal Dichalcogenides. *Phys. Rev. B* **1997**, *56* (3), 1354–1365.

(110) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81* (1), 511–519.

(111) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

(112) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113* (22), 9901.

(113) Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113* (22), 9978.

(114) Zandbergen, H. W.; Foo, M.; Xu, Q.; Kumar, V.; Cava, R. J. Sodium Ion Ordering in NaxCoO2: Electron Diffraction Study. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2004**, *70* (2), 1–8.

(115) Zhou, F.; Cococcioni, M.; Marianetti, C.; Morgan, D.; Ceder, G. First-Principles Prediction of Redox Potentials in Transition-Metal Compounds with LDA+U. *Phys. Rev. B* **2004**, *70* (23), 235121.

(116) Li, X.; Ma, X.; Su, D.; Liu, L.; Chisnell, R.; Ong, S. P.; Chen, H.; Toumar, A.; Idrobo, J.-C.; Lei, Y.; et al. Direct Visualization of the Jahn–Teller Effect Coupled to Na Ordering in $Na_{5/8}MnO_2$. *Nat. Mater.* **2014**, *13* (6), 586–592.

(117) Ouyang, C. Y.; Shi, S. Q.; Lei, M. S. Jahn–Teller Distortion and Electronic Structure of LiMn2O4. *J. Alloys Compd.* **2009**, *474* (1–2), 370–374.

(118) Ma, X.; Chen, H.; Ceder, G. Electrochemical Properties of Monoclinic NaMnO2. *J. Electrochem. Soc.* **2011**, *158* (12), A1307.

(119) Mendiboure, a.; Delmas, C.; Hagenmuller, P. Electrochemical Intercalation and Deintercalation of NaxMnO2 Bronzes. *J. Solid State Chem.* **1985**, *57* (3), 323–331.

(120) Sanchez, J. M.; Ducastelle, F.; Gratias, D. Generalized Cluster Description of Multicomponent Systems. *Phys. Stat. Mech. Its Appl.* **1984**, *128* (1), 334–350.

(121) Van der Ven, A. Lithium Diffusion in Layered LixCoO2. *Electrochem. Solid-State Lett.* **1999**, *3* (7), 301.

(122) Sykes, M. F.; Essam, J. W. Exact Critical Percolation Probabilities for Site and Bond Problems in Two Dimensions. *J. Math. Phys.* **1964**, *5* (8), 1117–1127.

(123) Doubaji, S.; Valvo, M.; Saadoune, I.; Dahbi, M.; Edström, K. Synthesis and Characterization of a New Layered Cathode Material for Sodium Ion Batteries. *J. Power Sources* **2014**, *266*, 275–281.

(124) Yuan, D.; He, W.; Pei, F.; Wu, F.; Wu, Y.; Qian, J.; Cao, Y.; Ai, X.; Yang, H. Synthesis and Electrochemical Behaviors of Layered Na0.67[Mn0.65Co0.2Ni0.15]O2 Microflakes as a Stable Cathode Material for Sodium-Ion Batteries. *J. Mater. Chem. A* **2013**, *1* (12), 3895–3899.

(125) Buchholz, D.; Chagas, L. G.; Winter, M.; Passerini, S. P2-Type Layered Na0.45Ni0.22Co0.11Mn0.66O2 as Intercalation Host Material for Lithium and Sodium Batteries. *Electrochimica Acta* **2013**, *110*, 208–213.

(126) Yoshida, J.; Guerin, E.; Arnault, M.; Constantin, C.; Mortemard de Boisse, B.; Carlier, D.; Guignard, M.; Delmas, C. New P2-Na0.70Mn0.60Ni0.30Co0.10O2 Layered Oxide as Electrode Material for Na-Ion Batteries. *J. Electrochem. Soc.* **2014**, *161* (14), A1987–A1991.

(127) Wang, H.; Yang, B.; Liao, X.-Z.; Xu, J.; Yang, D.; He, Y.-S.; Ma, Z.-F. Electrochemical Properties of P2-Na2/3[Ni1/3Mn2/3]O2 Cathode Material for Sodium Ion Batteries When Cycled in Different Voltage Ranges. *Electrochimica Acta* **2013**, *113*, 200–204.

(128) Shibata, T.; Fukuzumi, Y.; Kobayashi, W.; Moritomo, Y. Fast Discharge Process of Layered Cobalt Oxides Due to High Na+ Diffusion. *Sci. Rep.* **2015**, *5*, 9006.

(129) Kang, K.; Meng, Y. S.; Bréger, J.; Grey, C. P.; Ceder, G. Electrodes with High Power and High Capacity for Rechargeable Lithium Batteries. *Science* **2006**, *311*, 977–980.

(130) Van der Ven, a.; Ceder, G.; Asta, M.; Tepesch, P. D. First-Principles Theory of Ionic Diffusion with Nondilute Carriers. *Phys. Rev. B* **2001**, *64* (18), 184307.

(131) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 011002.

(132) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31* (9), 3564–3572.

(133) Zheng, C.; Radhakrishnan, B.; Chu, I.-H.; Wang, Z.; Ong, S. P. Effects of Transition-Metal Mixing on Na Ordering and Kinetics in Layered P2 Oxides. *Phys. Rev. Appl.* **2017**, *7* (6), 064003.

(134) Seidler, G. T.; Mortensen, D. R.; Remesnik, A. J.; Pacold, J. I.; Ball, N. A.; Barry, N.; Styczinski, M.; Hoidn, O. R. A Laboratory-Based Hard x-Ray Monochromator for High-Resolution x-Ray Emission Spectroscopy and x-Ray Absorption near Edge Structure Measurements. *Rev. Sci. Instrum.* **2014**, *85* (11), 113906.

(135) Bilderback, D. H.; Elleaume, P.; Weckert, E. Review of Third and next Generation Synchrotron Light Sources. *J. Phys. B At. Mol. Opt. Phys.* **2005**, *38* (9), S773–S797.

(136) Ankudinov, A. L.; Takimoto, Y.; Rehr, J. J. Combined Bethe-Saltpeter Equations and Time-Dependent Density-Functional Theory Approach for x-Ray Absorption Calculations. *Phys. Rev. B* **2005**, *71* (16), 165110.

(137) Tanaka, I.; Mizoguchi, T. First-Principles Calculations of x-Ray Absorption near Edge Structure and Energy Loss near Edge Structure: Present and Future. *J. Phys. Condens. Matter* **2009**, *21* (10), 104201. https://doi.org/10.1088/0953-8984/21/10/104201.

(138) Laskowski, R.; Blaha, P. Understanding the $L_{2,3}$ x-Ray Absorption Spectra of Early $3d$ Transition Elements. *Phys. Rev. B* **2010**, *82* (20), 205104.

(139) Asakura, K.; Ikemoto, I.; Kuroda, H.; Kobayashi, T.; Shirakawa, H. Dopant Structure in FeCl3-Doped Polyacetylene Studied by X-Ray Absorption Spectroscopy and X-Ray Photoelectron Spectroscopy. *Bull. Chem. Soc. Jpn.* **1985**, *58* (7), 2113–2120.

(140) Newville, M.; Carroll, S. A.; O'Day, P. A.; Waychunas, G.; Ebert, M. A Web-Based Library of XAFS Data on Model Compounds. *J. Synchrotron Radiat.* **1999**, *6* (3), 276–277.

(141) Zheng, C.; Mathew, K.; Chen, C.; Chen, Y.; Tang, H.; Dozier, A.; Kas, J. J.; Vila, F. D.; Rehr, J. J.; Piper, L. F. J.; et al. Automated Generation and Ensemble-Learned Matching of X-Ray Absorption Spectra. *npj Comput Mater* **2018**, *4* (1), 1–9.

(142) Mathew, K.; Montoya, J. H.; Faghaninia, A.; Dwarakanath, S.; Aykol, M.; Tang, H.; Chu, I.; Smidt, T.; Bocklund, B.; Horton, M.; et al. Atomate: A High-Level Interface to Generate, Execute, and Analyze Computational Materials Science Workflows. *Comput. Mater. Sci.* **2017**, *139*, 140–152.

(143) Takigawa, I.; Shimizu, K.; Tsuda, K.; Takakusagi, S. Machine-Learning Prediction of the d-Band Center for Metals and Bimetals. *RSC Adv.* **2016**, *6* (58), 52587–52595.

(144) Zheng, C.; Mathew, K.; Chen, C.; Chen, Y.; Tang, H.; Dozier, A.; Kas, J. J.; Vila, F. D.; Rehr, J. J.; Piper, L. F. J.; et al. Automated Generation and Ensemble-Learned Matching of X-Ray Absorption Spectra. *Npj Comput. Mater.* **2018**, *4* (1), 12.

(145) Rehr, J. J. Theoretical Approaches to X-Ray Absorption Fine Structure. *Rev. Mod. Phys.* **2000**, *72* (3), 621–654.

(146) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. a. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based on REpresentational State Transfer (REST) Principles. *Comput. Mater. Sci.* **2015**, *97*, 209–215.

(147) Ravel, B.; Hester, J. R.; Solé, V. A.; Newville, M. Towards Data Format Standardization for X-Ray Absorption Spectroscopy. *J. Synchrotron Radiat.* **2012**, *19* (6), 869–874.

(148) Ravel, B.; Newville, M. XAFS Data Interchange: A Single Spectrum XAFS Data File Format. *J. Phys. Conf. Ser.* **2016**, *712* (1), 1–5.

(149) Rana, J.; Glatthaar, S.; Gesswein, H.; Sharma, N.; Binder, J. R.; Chernikov, R.; Schumacher, G.; Banhart, J. Local Structural Changes in LiMn1.5Ni0.5O4 Spinel Cathode Material for Lithium-Ion Batteries. *J. Power Sources* **2014**, *255*, 439–449.

(150) Rana, J.; Kloepsch, R.; Li, J.; Scherb, T.; Schumacher, G.; Winter, M.; Banhart, J. On the Structural Integrity and Electrochemical Activity of a 0.5Li2MnO3·0.5LiCoO2 Cathode Material for Lithium-Ion Batteries. *J. Mater. Chem. A* **2014**, *2* (24), 9099.

(151) Moreno, M. S.; Jorissen, K.; Rehr, J. J. Practical Aspects of Electron Energy-Loss Spectroscopy (EELS) Calculations Using FEFF8. *Micron* **2007**, *38* (1), 1–11.

(152) Rehr, J. J.; Soininen, J. A.; Shirley, E. L. Final-State Rule vs the Bethe-Salpeter Equation for Deep-Core x-Ray Absorption Spectra. *Phys. Scr.* **2005**, *2005* (T115), 207.

(153) Vinson, J.; Rehr, J. J. Ab Initio Bethe-Salpeter Calculations of the x-Ray Absorption Spectra of Transition Metals at the L-Shell Edges. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2012**, *86* (19), 1–6.

(154) Xu, J.; Krüger, P.; Natoli, C. R.; Hayakawa, K.; Wu, Z.; Hatada, K. X-Ray Absorption Spectra of Graphene and Graphene Oxide by Full-Potential Multiple Scattering Calculations with Self-Consistent Charge Density. *Phys. Rev. B* **2015**, *92* (12), 125408.

(155) Alperovich, I.; Smolentsev, G.; Moonshiram, D.; Jurss, J. W.; Concepcion, J. J.; Meyer, T. J.; Soldatov, A.; Pushkar, Y. Understanding the Electronic Structure of 4d Metal Complexes: From Molecular Spinors to L-Edge Spectra of a Di-Ru Catalyst. *J. Am. Chem. Soc.* **2011**, *133* (39), 15786–15794.

(156) Newville, M. EXAFS Analysis Using FEFF and FEFFIT. *J. Synchrotron Radiat.* **2001**, *8* (2), 96–100.

(157) Shishido, T.; Asakura, H.; Yamazoe, S.; Teramura, K.; Tanaka, T. Structural Analysis of Group V, VI, VII Metal Compounds by XAFS and DFT Calculation. *J. Phys. Conf. Ser.* **2009**, *190*, 012073.

(158) Newville, M. Fundamentals of XAFS. *Rev. Mineral. Geochem.* **2014**, *78* (1), 33–74.

(159) Ravel, B.; Newville, M. ATHENA , ARTEMIS , HEPHAESTUS : Data Analysis for X-Ray Absorption Spectroscopy Using IFEFFIT. *J. Synchrotron Radiat.* **2005**, *12* (4), 537–541.

(160) Lin, Y.-C.; Wen, B.; Wiaderek, K. M.; Sallis, S.; Liu, H.; Lapidus, S. H.; Borkiewicz, O. J.; Quackenbush, N. F.; Chernova, N. A.; Karki, K.; et al. Thermodynamics, Kinetics and Structural Evolution of ε-LiVOPO4 over Multiple Lithium Intercalation. *Chem. Mater.* **2016**, *28* (6), 1794–1805.

(161) Yu, X.; Wang, Q.; Zhou, Y.; Li, H.; Yang, X.-Q.; Nam, K.-W.; Ehrlich, S. N.; Khalid, S.; Meng, Y. S. High Rate Delithiation Behaviour of LiFePO4 Studied by Quick X-Ray Absorption Spectroscopyw. *Chem Commun Chem Commun* **2012**, *48* (48), 11537–11539.

(162) Cheng, J.-H.; Pan, C.-J.; Lee, J.-F.; Chen, J.-M.; Guignard, M.; Delmas, C.; Carlier, D.; Hwang, B.-J. Simultaneous Reduction of Co3+ and Mn4+ in P2-Na2/3Co2/3Mn1/3O2 As Evidenced by X-Ray Absorption Spectroscopy during Electrochemical Sodium Intercalation. *Chem. Mater.* **2014**, *26* (2), 1219–1225.

(163) Koningsberger, D. C.; Prins, R. *X-Ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS, and XANES*; Wiley-Interscience, New York, 1988.

(164) Bunker, G. *Introduction to XAFS: A Practical Guide to X-Ray Absorption Fine Structure Spectroscopy*; Cambridge University Press, 2010.

(165) Newville, M. Fundamentals of XAFS. *Rev. Mineral. Geochem.* **2014**, *78* (1), 33–74.

(166) Ravel, B.; Newville, M. ATHENA, ARTEMIS, HEPHAESTUS: Data Analysis for X-Ray Absorption Spectroscopy Using IFEFFIT. *J. Synchrotron Radiat.* **2005**, *12* (4), 537–541.

(167) Ewels, P.; Sikora, T.; Serin, V.; Ewels, C. P.; Lajaunie, L. A Complete Overhaul of the Electron Energy-Loss Spectroscopy and X-Ray Absorption Spectroscopy Database: Eelsdb.Eu. *Microsc. Microanal.* **2016**, *22* (03), 717–724.

(168) Egerton, R. F. *Electron Energy-Loss Spectroscopy in the Electron Microscope*, 3rd ed.; Springer US: Boston, MA, 2011.

(169) 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.* **2010**, *12* (21), 5503.

(170) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 011002.

(171) 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. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(172) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; et al. FireWorks: A Dynamic Workflow System Designed for High-Throughput Applications. *Concurr. Comput. Pract. Exp.* **2015**, *27* (17), 5037–5059.

(173) Wang, Z.; Lee, J. Z.; Xin, H. L.; Han, L.; Grillon, N.; Guy-Bouyssou, D.; Bouyssou, E.; Proust, M.; Meng, Y. S. Effects of Cathode Electrolyte Interfacial (CEI) Layer on Long Term Cycling of All-Solid-State Thin-Film Batteries. *J. Power Sources* **2016**, *324*, 342–348.

(174) Jia, Q.; Ramaswamy, N.; Hafiz, H.; Tylus, U.; Strickland, K.; Wu, G.; Barbiellini, B.; Bansil, A.; Holby, E. F.; Zelenay, P.; et al. Experimental Observation of Redox-Induced Fe–N Switching Behavior as a Determinant Role for Oxygen Reduction Activity. *ACS Nano* **2015**, *9* (12), 12496–12505.

(175) Behafarid, F.; Matos, J.; Hong, S.; Zhang, L.; Rahman, T. S.; Roldan Cuenya, B. Structural and Electronic Properties of Micellar Au Nanoparticles: Size and Ligand Effects. *ACS Nano* **2014**, *8* (7), 6671–6681.

(176) Jorissen, K.; Rehr, J. J. Calculations of Electron Energy Loss and X-Ray Absorption Spectra in Periodic Systems without a Supercell. *Phys Rev B* **2010**, *81* (24), 245124.

(177) Vinson, J.; Rehr, J. J. Ab Initio Bethe-Salpeter Calculations of the x-Ray Absorption Spectra of Transition Metals at the L-Shell Edges. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2012**, *86* (19), 1–6.

(178) Wu, Z.; Cohen, R. E. More Accurate Generalized Gradient Approximation for Solids. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2006**, *73* (23), 2–7.

(179) Kresse, G.; Harl, J. Accurate Bulk Properties from Approximate Many-Body Techniques. *Phys. Rev. Lett.* **2009**, *103* (5), 4–7.

(180) Haas, P.; Tran, F.; Blaha, P. Calculation of the Lattice Constant of Solids with Semilocal Functionals. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2009**, *79* (8), 1–10.

(181) Klimeš, J.; Bowler, D. R.; Michaelides, A. Van Der Waals Density Functionals Applied to Solids. *Phys. Rev. B* **2011**, *83* (19), 195131.

(182) Alkauskas, A.; Pasquarello, A. Band-Edge Problem in the Theoretical Determination of Defect Energy Levels: The O Vacancy in ZnO as a Benchmark Case. *Phys. Rev. B* **2011**, *84* (12), 125206.

(183) Perdew, J.; Ruzsinszky, A.; Csonka, G.; Vydrov, O.; Scuseria, G.; Constantin, L.; Zhou, X.; Burke, K. Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys. Rev. Lett.* **2008**, *100* (13), 136406.

(184) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid Functionals Based on a Screened Coulomb Potential. *J. Chem. Phys.* **2003**, *118* (18), 8207–8215.

(185) Paier, J.; Asahi, R.; Nagoya, A.; Kresse, G. Cu2 ZnSnS4 as a Potential Photovoltaic Material: A Hybrid Hartree-Fock Density Functional Theory Study. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2009**, *79* (11), 1–8.

(186) Da Silva, J. L. F.; Ganduglia-Pirovano, M. V.; Sauer, J.; Bayer, V.; Kresse, G. Hybrid Functionals Applied to Rare-Earth Oxides: The Example of Ceria. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2007**, *75* (4), 19–24.

(187) Wróbel, J.; Kurzydłowski, K. J.; Hummer, K.; Kresse, G.; Piechota, J. Calculations of ZnO Properties Using the Heyd-Scuseria-Ernzerhof Screened Hybrid Density Functional. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2009**, *80* (15), 1–8.

(188) Ong, S. P.; Mo, Y.; Ceder, G. Low Hole Polaron Migration Barrier in Lithium Peroxide. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2012**, *85* (8), 2–5.

(189) Carey, C.; Boucher, T.; Mahadevan, S.; Bartholomew, P.; Dyar, M. D. Machine Learning Tools Formineral Recognition and Classification from Raman Spectroscopy. *J. Raman Spectrosc.* **2015**, *46* (10), 894–903.

(190) Liu, J.; Bell, A. W.; Bergeron, J. J. M.; Yanofsky, C. M.; Carrillo, B.; Beaudrie, C. E. H.; Kearney, R. E. Methods for Peptide Identification by Spectral Comparison. *Proteome Sci.* **2007**, *5*, 3.

(191) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (9), 859–866.

(192) BEARDEN, J. A.; BURR, A. F. Reevaluation of X-Ray Atomic Energy Levels. *Rev. Mod. Phys.* **1967**, *39* (1), 125–142.

(193) Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W. D.; Gamst, A. C.; Persson, K. A.; Ceder, G. The Thermodynamic Scale of Inorganic Crystalline Metastability. *Sci. Adv.* **2016**, *2* (11), e1600225–e1600225.

(194) Xu, J.; Krüger, P.; Natoli, C. R.; Hayakawa, K.; Wu, Z.; Hatada, K. X-Ray Absorption Spectra of Graphene and Graphene Oxide by Full-Potential Multiple Scattering Calculations with Self-Consistent Charge Density. *Phys. Rev. B* **2015**, *92* (12), 125408.

(195) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. a. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based on REpresentational State Transfer (REST) Principles. *Comput. Mater. Sci.* **2015**, *97*, 209–215.

(196) Ravel, B. A Practical Introduction to Multiple Scattering Theory. *J. Alloys Compd.* **2005**, *401* (1–2), 118–126.

(197) Zoubir, Arnaud. *Raman Imaging*; Zoubir, A., Ed.; Springer Series in Optical Sciences; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; Vol. 168.

(198) Hansen, M. E.; Smedsgaard, J. A New Matching Algorithm for High Resolution Mass Spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (8), 1173–1180.

(199) Hernández-Rivera, E.; Coleman, S. P.; Tschopp, M. A. Using Similarity Metrics to Quantify Differences in High-Throughput Data Sets: Application to X-Ray Diffraction Patterns. *ACS Comb. Sci.* **2017**, *19* (1), 25–36.

(200) Deza, M. M.; Deza, E. *Encyclopedia of Distances*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013.

(201) Ho, T. K.; Hull, J. J.; Srihari, S. N. Decision Combination in Multiple Classifier Systems. *IEEE Trans Pattern Anal Mach Intell* **1994**, *16* (1), 66–75.

(202) Black, D. *The Theory of Committees and Elections*; Springer Netherlands: Dordrecht, 1986.

(203) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. 2014.

(204) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13* (2), 22–30.

(205) O'Day, P. A.; Newville, M.; Neuhoff, P. S.; Sahai, N.; Carroll, S. A. X-Ray Absorption Spectroscopy of Strontium(II) Coordination. *J. Colloid Interface Sci.* **2000**, *222* (2), 184–197.

(206) Chaurand, P.; Rose, J.; Briois, V.; Salome, M.; Proux, O.; Nassif, V.; Olivi, L.; Susini, J.; Hazemann, J.; Bottero, J. New Methodological Approach for the Vanadium K-Edge X-Ray Absorption Near-Edge Structure Interpretation: Application to the Speciation of Vanadium in Oxide Phases from Steel Slag. *J. Phys. Chem. B* **2007**, *111* (19), 5101–5110.

(207) Silversmit, G.; Bokhoven, J. A. V.; Poelman, H.; Eerden, A. M. J. V. D.; Marin, G. B. The Structure of Supported and Unsupported Vanadium Oxide under Calcination , Reduction and Oxidation Determined with XAS. **2005**, *285*, 151–162.

(208) Farges, F.; Brown, G. E.; Rehr, J. J. Ti *K*-Edge XANES Studies of Ti Coordination and Disorder in Oxide Compounds: Comparison between Theory and Experiment. *Phys. Rev. B* **1997**, *56* (4), 1809–1819.

(209) Farges, F.; Brown, G. E.; Petit, P.-E.; Munoz, M. Transition Elements in Water-Bearing Silicate Glasses/Melts. Part I. a High-Resolution and Anharmonic Analysis of Ni Coordination Environments in Crystals, Glasses, and Melts. *Geochim. Cosmochim. Acta* **2001**, *65* (10), 1665–1678.

(210) DeBeer George, S.; Brant, P.; Solomon, E. I. Metal and Ligand K-Edge XAS of Organotitanium Complexes: Metal 4p and 3d Contributions to Pre-Edge Intensity and Their Contributions to Bonding. *J. Am. Chem. Soc.* **2005**, *127* (2), 667–674.

(211) Westre, T. E.; Kennepohl, P.; DeWitt, J. G.; Hedman, B.; Hodgson, K. O.; Solomon, E. I. A Multiplet Analysis of Fe K-Edge 1s → 3d Pre-Edge Features of Iron Complexes. *J. Am. Chem. Soc.* **1997**, *119* (27), 6297–6314.

(212) 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.* **2008**, *37* (6), 572–584.

(213) Sano, M.; Komorita, S.; Yamatera, H. XANES Spectra of Copper(II) Complexes: Correlation of the Intensity of the 1s .Fwdarw. 3d Transition and the Shape of the Complex. *Inorg. Chem.* **1992**, *31* (3), 459–463.

(214) Fernandez-Garcia, M.; Marquez Alvarez, C.; Haller, G. L. XANES-TPR Study of Cu-Pd Bimetallic Catalysts: Application of Factor Analysis. *J. Phys. Chem.* **1995**, *99* (33), 12565–12569.

(215) Manceau, A.; Marcus, M.; Lenoir, T. Estimating the Number of Pure Chemical Components in a Mixture by X-Ray Absorption Spectroscopy. *J. Synchrotron Radiat.* **2014**, *21* (5), 1140–1147.

(216) Fay, M. J.; Proctor, A.; Hoffmann, D. P.; Houalla, M.; Hercules, D. M. Determination of the Mo Surface Environment of Mo/TiO2 Catalysts by EXAFS, XANES and PCA. *Mikrochim. Acta* **1992**, *109* (5–6), 281–293.

(217) Beauchemin, S.; Hesterberg, D.; Beauchemin, M. Principal Component Analysis Approach for Modeling Sulfur K-XANES Spectra of Humic Acids. *Soil Sci. Soc. Am. J.* **2002**, *66* (1), 83.

(218) Bajt, S.; Sutton, S. R.; Delaney, J. S. X-Ray Microprobe Analysis of Iron Oxidation States in Silicates and Oxides Using X-Ray Absorption near Edge Structure (XANES). *Geochim. Cosmochim. Acta* **1994**, *58* (23), 5209–5214.

(219) Tanaka, I.; Mizoguchi, T. First-Principles Calculations of x-Ray Absorption near Edge Structure and Energy Loss near Edge Structure: Present and Future. *J. Phys. Condens. Matter* **2009**, *21* (10), 104201.

(220) Laskowski, R.; Blaha, P. Understanding the $L_{2,3}$ x-Ray Absorption Spectra of Early $3d$ Transition Elements. *Phys. Rev. B* **2010**, *82* (20), 205104.

(221) 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. *Sci. Data* **2018**, *5*, 180151.

(222) Kiyohara, S.; Miyata, T.; Tsuda, K.; Mizoguchi, T. Data-Driven Approach for the Prediction and Interpretation of Core-Electron Loss Spectroscopy. *Sci. Rep.* **2018**, *8* (1), 13548.

(223) Suzuki, Y.; Hino, H.; Kotsugi, M.; Ono, K. Automated Estimation of Materials Parameter from X-Ray Absorption and Electron Energy-Loss Spectra with Similarity Measures. *Npj Comput. Mater.* **2019**, *5* (1), 39.

(224) Ankudinov, A. L.; Ravel, B.; Rehr, J. J.; Conradson, S. D. Real-Space Multiple-Scattering Calculation and Interpretation of x-Ray-Absorption near-Edge Structure. *Phys. Rev. B* **1998**, *58* (12), 7565–7576.

(225) Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier Chains for Multi-Label Classification. *Mach. Learn.* **2011**, *85* (3), 333–359.

(226) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst Technol* **2011**, *2* (3), 27:1–27:27.

(227) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(228) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* **2011**, *12*, 2825–2830.

(229) LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional Networks and Applications in Vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*; 2010; pp 253–256.

(230) 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. *Sci. Data* **2018**, *5*, 180151.

(231) Gray, R. M. *Entropy and Information Theory*; 2011.

(232) Prado, R. J.; M.Flank, A. Sodium K Edge XANES Calculation in NaCl Type Structures. *Phys. Scr.* **2005**, 165.

(233) Nakanishi, K.; Ohta, T. Verification of the FEFF Simulations to K-Edge XANES Spectra of the Third Row Elements. *J. Phys. Condens. Matter Inst. Phys. J.* **2009**, *21* (10), 104214.

(234) Strobl, C.; Boulesteix, A. L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* **2007**, *8*.

(235) Yokoyama, T.; Kosugi, N.; Kuroda, H. Polarized Xanes Spectra of CuCl2 · 2H2O. Further Evidence for Shake-down Phenomena. *Chem. Phys.* **1986**, *103* (1), 101–109.

(236) Asakura, K.; Iwasawa, Y. Metal Oxide Catalysts. In *X-Ray Absorption Fine Structure for Catalysts and Surfaces*; Series on Synchrotron Radiation Techniques and Applications; WORLD SCIENTIFIC, 1996; Vol. Volume 2, pp 192–215.

(237) Wong, J.; Lytle, F. W.; Messmer, R. P.; Maylotte, D. H. *K*-Edge Absorption Spectra of Selected Vanadium Compounds. *Phys. Rev. B* **1984**, *30* (10), 5596–5610.

(238) Rana, J.; Glatthaar, S.; Gesswein, H.; Sharma, N.; Binder, J. R.; Chernikov, R.; Schumacher, G.; Banhart, J. Local Structural Changes in LiMn1.5Ni0.5O4 Spinel Cathode Material for Lithium-Ion Batteries. *J. Power Sources* **2014**, *255*, 439–449.

(239) Rana, J.; Kloepsch, R.; Li, J.; Scherb, T.; Schumacher, G.; Winter, M.; Banhart, J. On the Structural Integrity and Electrochemical Activity of a 0.5Li2MnO3·0.5LiCoO2 Cathode Material for Lithium-Ion Batteries. *J. Mater. Chem. A* **2014**, *2* (24), 9099.

(240) Weng, T. C.; Waldo, G. S.; Penner-Hahn, J. E. A Method for Normalization of X-Ray Absorption Spectra. *J. Synchrotron Radiat.* **2005**, *12* (4), 506–510.

(241) Wu, Z.; Cohen, R. E. More Accurate Generalized Gradient Approximation for Solids. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2006**, *73* (23), 2–7.

(242) Cotton, F. A.; Ballhausen, C. J. Soft X-Ray Absorption Edges of Metal Ions in Complexes. I. Theoretical Considerations. *J. Chem. Phys.* **2005**, *25* (4), 617–619.

(243) Zimmermann, N. E. R.; Vorselaars, B.; Quigley, D.; Peters, B. Nucleation of NaCl from Aqueous Solution: Critical Sizes, Ion-Attachment Kinetics, and Rates. *J. Am. Chem. Soc.* **2015**, *137* (41), 13352–13361.

(244) Tsoumakas, G.; Katakis, I. Multi-Label Classification. *Int. J. Data Warehous. Min.* **2007**, *3* (3), 1–13.