

# Machine-Learning Spectral Indicators of Topology

Nina Andrejevic,\* Jovana Andrejevic, B. Andrei Bernevig, Nicolas Regnault, Fei Han, Gilberto Fabbris, Thanh Nguyen, Nathan C. Drucker, Chris H. Rycroft,\* and Mingda Li\*

Topological materials discovery has emerged as an important frontier in condensed matter physics. While theoretical classification frameworks have been used to identify thousands of candidate topological materials, experimental determination of materials' topology often poses significant technical challenges. X-ray absorption spectroscopy (XAS) is a widely used materials characterization technique sensitive to atoms' local symmetry and chemical bonding, which are intimately linked to band topology by the theory of topological quantum chemistry (TQC). Moreover, as a local structural probe, XAS is known to have high quantitative agreement between experiment and calculation, suggesting that insights from computational spectra can effectively inform experiments. In this work, computed X-ray absorption near-edge structure (XANES) spectra of more than 10 000 inorganic materials to train a neural network (NN) classifier that predicts topological class directly from XANES signatures, achieving  $F_1$  scores of 89% and 93% for topological and trivial classes, respectively is leveraged. Given the simplicity of the XAS setup and its compatibility with multimodal sample environments, the proposed machine-learning-augmented XAS topological indicator has the potential to discover broader categories of topological materials, such as non-cleavable compounds and amorphous materials, and may further inform field-driven phenomena in situ, such as magnetic field-driven topological phase transitions.


## 1. Introduction

Topological materials are characterized by a topologically nontrivial electronic band structure from which they derive their exceptional transport properties.<sup>[1–6]</sup> The prospect of developing these exotic phases into useful applications has garnered widespread efforts to identify and catalogue candidate topological materials, evidenced by the emergence of numerous theoretical frameworks based on connectivity of electronic bands,<sup>[7–13]</sup> symmetry-based indicators,<sup>[7,14–21]</sup> electron-filling constraints,<sup>[7,22,23]</sup> and spin-orbit spillage.<sup>[24–26]</sup> These frameworks have facilitated the prediction of over 8000 topologically non-trivial phases,<sup>[27–34]</sup> a vast unexplored territory for experiments. This is strong motivation to develop complementary experimental techniques for high-throughput screening of candidate materials. Current state-of-the-art techniques such as angle-resolved photoemission spectroscopy (ARPES), scanning tunneling microscopy (STM), and

N. Andrejevic  
Center for Nanoscale Materials  
Argonne National Laboratory  
Lemont, IL 60439, USA  
E-mail: nandrejevic@alum.mit.edu; mingda@mit.edu

N. Andrejevic, F. Han, T. Nguyen, N. C. Drucker, M. Li  
Quantum Measurement Group  
Massachusetts Institute of Technology  
Cambridge, MA 02139, USA

N. Andrejevic  
Department of Materials Science and Engineering  
Massachusetts Institute of Technology  
Cambridge, MA 02139, USA

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adma.202204113>.

© 2022 The Authors. Advanced Materials published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

DOI: 10.1002/adma.202204113

J. Andrejevic  
Department of Physics  
University of Pennsylvania  
Philadelphia, PA 19104, USA  
E-mail: chr@seas.harvard.edu

J. Andrejevic, N. C. Drucker, C. H. Rycroft  
John A. Paulson School of Engineering and Applied Sciences  
Harvard University  
Cambridge, MA 02138, USA

B. A. Bernevig, N. Regnault  
Department of Physics  
Princeton University  
Princeton, NJ 08544, USA

B. A. Bernevig  
Donostia International Physics Center  
P. Manuel de Lardizabal 4, Donostia-San Sebastian 20018, Spain

B. A. Bernevig  
IKERBASQUE  
Basque Foundation for Science  
Plaza Euskadi 5, Bilbao 48009, Spain

quantum transport measurements are commonly used to detect topological signatures, but a few limitations remain: Methods like ARPES directly probe band topology but are surface-sensitive and thereby place strict requirements on sample preparation and the sample environment, limiting the range of experimentally accessible materials;<sup>[35,36]</sup> transport measurements, on the other hand, can be performed on more versatile samples but can be more difficult to interpret. Neither approach yet fully meets the demands of a high-throughput classification program.

Machine-learning methods are increasingly being adapted to materials research to accelerate materials discovery<sup>[37–44]</sup> and facilitate inverse design through high-throughput property prediction.<sup>[45–47]</sup> Several recent studies have proposed data-driven frameworks for predicting band topology from structural and compositional attributes<sup>[48–50]</sup> and quantum theoretical or simulated data.<sup>[51–54]</sup> At the same time, machine-learning methods are being adopted to automate and improve data analysis for a broad range of experimental techniques.<sup>[55–61]</sup> Importantly, machine learning presents a potential opportunity to not only accelerate data analysis, but to derive useful information from complex data in the absence of reliable theoretical models, or to extract new insights beyond traditional models.

In this work, we develop a data-driven classifier of electronic band topology using materials' X-ray absorption spectra. X-ray absorption spectroscopy (XAS) is widely used to characterize the chemical state and local atomic structure of atomic species in a material. This technique is suitable for the study of highly diverse samples and environments, including noncrystalline materials and extreme temperatures and pressures.<sup>[62]</sup> As a bulk probe, XAS also places few constraints on surface quality and sample preparation. The X-ray absorption near-edge structure (XANES), defined within  $\approx 50$  eV of an XAS absorption edge, provides a specie-specific fingerprint of the absorbing atom's local chemical environment, including coordination chemistry, orbital hybridization, and density of available electronic states. However, despite the rich electronic structural information contained in XANES spectra, the lack of a simple analytic description of XANES has compelled largely qualitative treatment of this energy regime, with individual spectral features attributed to properties of the electronic structure through empirical evidence and spectral matching.<sup>[63]</sup>

As a result, machine-learning methods have been introduced to automate the estimation of materials parameters such as coordination environments,<sup>[56,64–67]</sup> oxidation states,<sup>[64,67]</sup> and crystal-field splitting<sup>[68]</sup> from XANES and other core-level spectroscopies, and even enable direct prediction of XANES spectra from structural and atomic descriptors.<sup>[69–71]</sup> Here, we propose that machine-learning models can be used to extract other hidden electronic properties, namely the electronic band topology, from XANES signatures and thereby serve as a potentially useful diagnostic of topological character. The theory of topological quantum chemistry (TQC) has demonstrated the intimate link between a material's band topology and its local chemical bonding,<sup>[7]</sup> which motivates our inquiry into the unexplored connection between XANES spectra and band topology. In particular, we develop a machine-learning-enabled indicator of band topology based on K-edge XANES spectral inputs, which correspond to electronic transitions from the 1-s core shell states to unoccupied states above the Fermi energy. First, we summarize the data assembly procedure, which consists of labeling the database of computed XANES K-edge spectra<sup>[72]</sup> according to topological character using the catalogue of high-quality topological materials predicted by TQC.<sup>[27,34]</sup> We then conduct an exploratory analysis of topological indication for the K-edge XANES spectra of different elements based on principal component analysis (PCA) and *k*-means clustering. Finally, we develop a neural network (NN) classifier of topology that synthesizes insights from XANES signatures of all elements in a given compound. Our classifier achieves  $F_1$  scores of 89% and 93% for topological and trivial classes, respectively. Materials containing certain elements, including Be, Al, Si, Sc, Ti, Ga, Ag, and Hg, are predicted with  $F_1$  scores above 90% in both classes. Our work suggests the potential of machine learning to uncover topological character embedded in complex spectral features, especially when a mechanistic understanding is challenging to acquire.

## 2. Data Preparation and Pre-Processing

XAS data were obtained from the published database of computed K-edge XANES spectra<sup>[72]</sup> and additional examples distributed on the Materials Project,<sup>[73–76]</sup> which are computed using the FEFF9 program.<sup>[77]</sup> The materials from the XANES database were then labeled according to their classification in the database of topological materials,<sup>[27,34]</sup> which is based on the formalism of TQC.<sup>[7]</sup> The classifications in the TQC database are based on structures from the Inorganic Crystal Structure Database (ICSD),<sup>[78]</sup> and the ICSD identifier was used to associate topological class labels with entries in the XANES database. We note that the crystal structures in the two databases are not strictly identical, and ICSD identifiers are associated with structurally similar Materials Project entries according to pymatgen's StructureMatcher algorithm.<sup>[75,76]</sup> In rare cases, multiple ICSD identifiers corresponding to different topological classifications were associated with the same set of XANES spectra. Because small discrepancies between the ICSD and Materials Project structures could lead to different topological classification for some materials close to a phase transition, all multiply labeled examples were removed from the dataset.

F. Han, T. Nguyen, M. Li  
Department of Nuclear Science and Engineering  
Massachusetts Institute of Technology  
Cambridge, MA 02139, USA

G. Fabbris  
Advanced Photon Source  
Argonne National Laboratory  
Lemont, IL 60439, USA

C. H. Rycroft  
Department of Mathematics  
University of Wisconsin–Madison  
Madison, WI 53706, USA

C. H. Rycroft  
Computational Research Division  
Lawrence Berkeley Laboratory  
Berkeley, CA 94720, USA

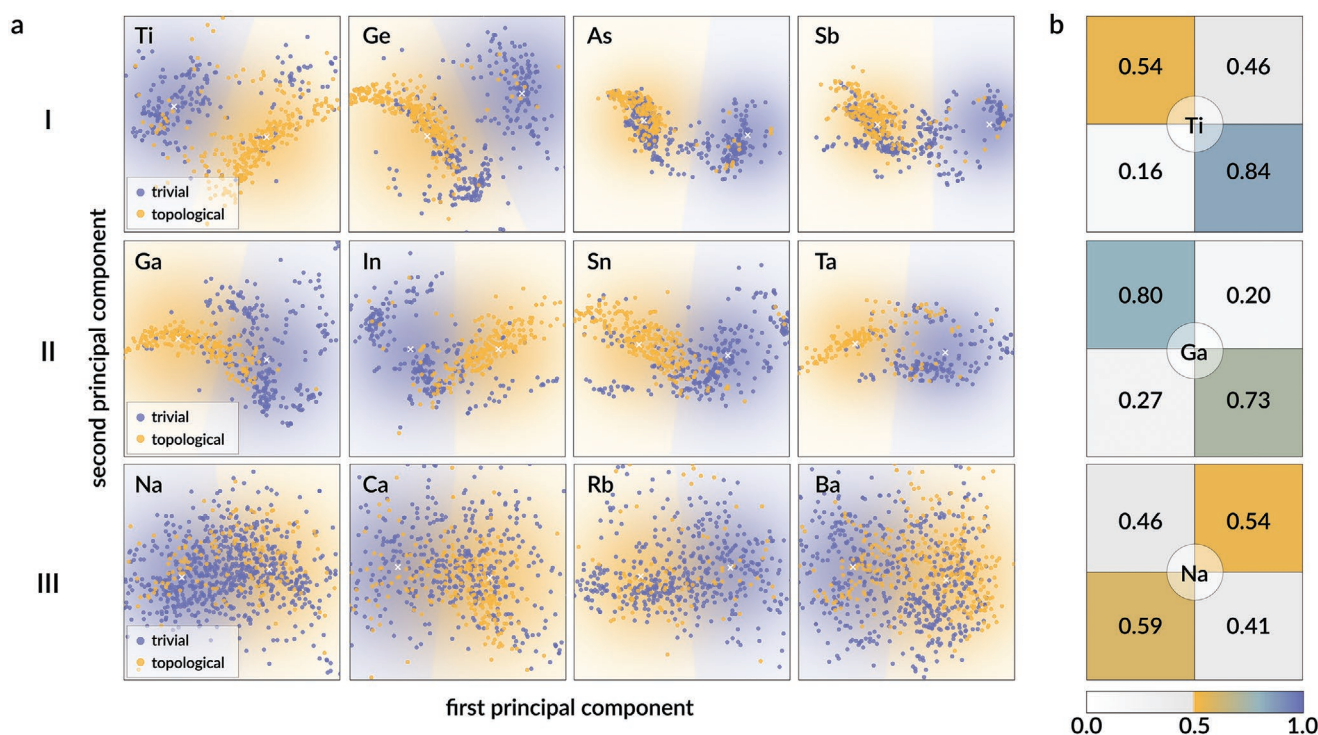
The materials data were further refined based on availability of both high-quality topological classification and spectral data, resulting in 13–151 total materials considered: 4957 topological ( $\approx 38\%$ ) and 8194 trivial ( $\approx 62\%$ ). Here, high-quality is defined following ref. [27], which considers only materials with well-determined structures and excludes alloys, magnetic compounds, and certain problematic  $f$ -electron atoms. Additionally, entries with spectra containing unphysical features such as large negative jumps were discarded. The materials in the final dataset are structurally and chemically diverse, representing 200 of 230 spacegroups and 63 different elements, with primitive unit cells ranging from 1 to 76 atoms and up to seven unique chemical species. The representation of different elements among topological and trivial examples is shown in Figure S1a,b, Supporting Information. Data were subdivided into training, validation, and test sets according to a 70/15/15% split. While samples were randomly distributed among the datasets, an assignment process was developed to ensure balanced representation of each absorbing element and topological class within each dataset. Specifically, the fraction of topological insulators (TI), topological semimetals (TSM), and topologically trivial materials represented in compounds containing a certain element was balanced as shown in Figure S1c, Supporting Information. For each example, the computed K-edge XANES spectra of each absorbing element were interpolated

and re-sampled at 200 evenly spaced energy values spanning an energy range of 56 eV surrounding the absorption edge. The spectra were standardized separately for different absorbing elements, which consisted of centering the mean of spectral intensities over each energy range, and scaling by the average intensity standard deviations.

### 3. Results

#### 3.1. Exploratory Analysis

Prior to training the NN classifier, we conducted an exploratory analysis of the assembled XANES spectra to estimate the separability by topological class exhibited by different elements. For all examples containing a given element, we performed a principal component analysis (PCA) on the high-dimensional spectra and subsequently carried out unsupervised  $k$ -means clustering on a subset of principal components of the training set. The number of retained principal components was selected to retain at least 80% of the explained variance of spectra for a given element. Results of the clustering analysis for a selection of elements are shown in Figure 1. The decision boundary between the two clusters identified by  $k$ -means clustering, projected along the first two principal components, lies at the

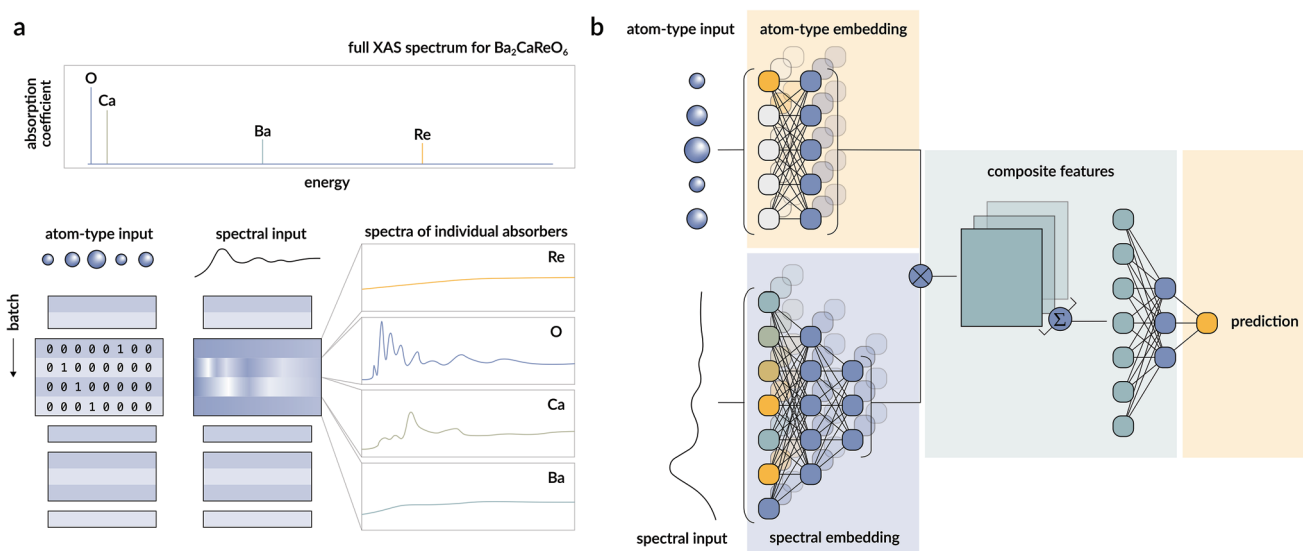


**Figure 1.** Exploratory analysis using principal components and  $k$ -means clustering. a) Decision boundary visualizations of classifications by unsupervised  $k$ -means clustering for selected elements. As detailed in the main text, the  $k$ -means clustering is performed on the subset of principal components accounting for at least 80% of the explained variance of spectra for a given element. The clusters are visualized along the first ( $x$ -axis) and second ( $y$ -axis) principal components in the scatter plots. Scattered points are colored according to their true class: topological (orange) or trivial (blue). The background is shaded according to the cluster-assigned class. The principal components exhibited three typical patterns: (row I) imbalanced classification in favor of topological examples, (row II) relatively balanced classification of topological and trivial examples, and (row III) no apparent clustering by class. b) Confusion matrices of representative examples in each of rows I, II, and III.

intersection of the blue (trivial) and orange (topological) shaded regions in Figure 1a. Since *k*-means clustering is not supervised by the true topological class of each example, cluster assignment was performed by solving an optimal matching problem that finds the pairing between clusters and topological classes that minimizes the number of misclassified examples, corrected for class imbalance. The examples from all three datasets (training, validation, and testing) are plotted as scattered points in the low-dimensional space and colored according to their known topological class. Additional visualizations are shown in Figure S2, Supporting Information. A quick survey of these results reveals a number of elements for which the classification accuracy of topological and trivial examples is imbalanced, and a few for which the classification accuracy is more balanced between the two classes. We correlated these observations with the decision boundary visualizations and noted three distinct patterns in the result of our unsupervised clustering. For some elements, nearly all topological examples were segregated within a single cluster (row I of Figure 1). This led to a strong score for topological examples but weaker score for trivial ones for elements such as Ti, Ge, As, and Sb. Other elements like Ga, In, Sn, and Ta exhibited more balanced classification accuracies between the two topological classes (row II of Figure 1). On the other hand, there were a number of unsuccessful examples of alkali and alkaline earth metals for which clustering of the data did not appear coincident with topological class (row III of Figure 1). Given that the feature transformations performed in our exploratory analysis were element-specific, the potential to discriminate data between the two classes is encouraging. This also suggests a possible advantage of synthesizing information of all constituent atom types in a given compound in order to improve prediction accuracy.

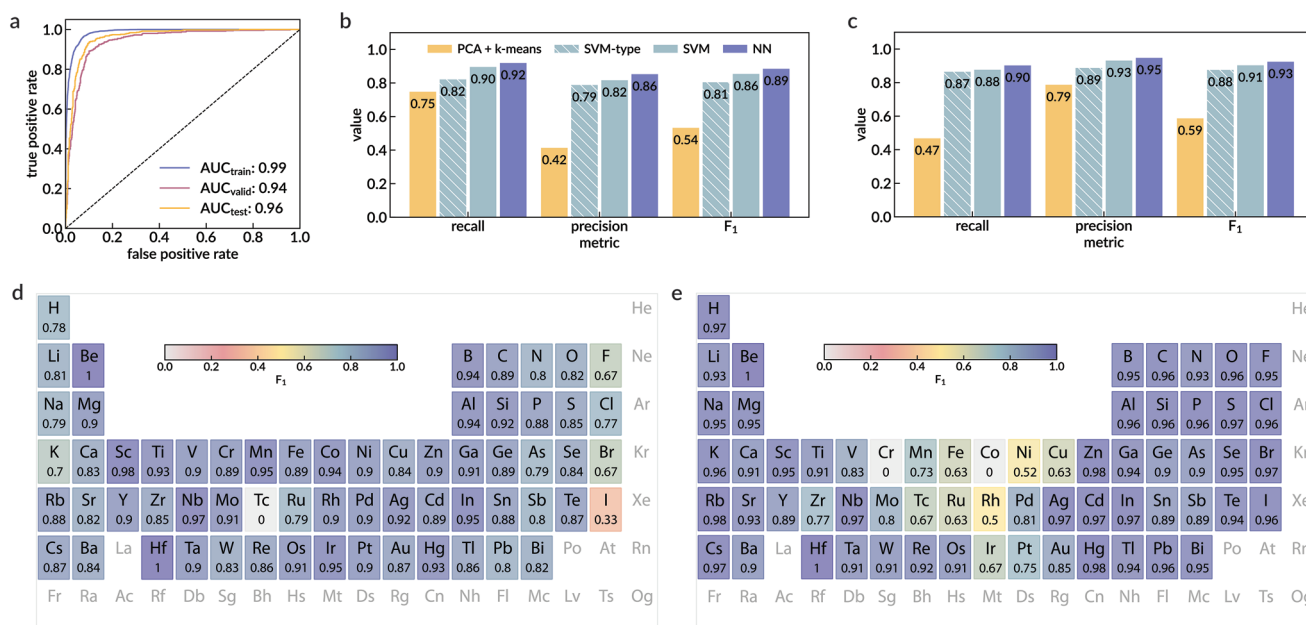
### 3.2. Network Architecture Optimization

The NN classifier inputs consist of the set of XANES spectra and atom types of each absorbing atom in a given material, as shown in Figure 2a, where atom types are encoded as one-hot feature vectors with a one at the index equaling the atomic number, and zeros elsewhere. The core-electron binding energy increases substantially with increasing atomic number, ranging from 284 eV for the C K-edge to 115 606 eV for the U K-edge,<sup>[79]</sup> and thus representing the XANES spectra of all absorbers on a continuous energy scale would be either poorly resolved or exceedingly high-dimensional (Figure 2a). Separating the spectral and atom type information at the input facilitates the construction of element-specific channels and allows us to retain the spectral energy resolution. In addition to enabling the synthesis of information from different absorbers, an NN comprises more complex, non-linear operations than PCA and thereby has the capability to learn more expressive representations of the input data. The network architecture is illustrated in Figure 2b. Fully connected layers first operate on each spectral and atom-type input to obtain intermediate representations, termed the spectral and atom-type embeddings, respectively. The embedded spectra are assigned to element-specific channels through a direct product with the corresponding atom-type embedding. These composite features are subsequently added for a given material and flattened to a single array, which is passed to another series of fully connected layers and activations that output the predicted binary topological class. Due to moderate class imbalance, samples were weighted to add greater penalty to the misclassification of topological examples.



**Figure 2.** Data structure and model architecture. a) A schematic of the full XANES spectrum for a representative sample in the dataset, showing the signatures from different absorbing elements on an absolute energy scale. For a given material, the inputs to the NN classifier consist of one-hot encoded atom types (left) and XANES spectra (right) for all absorbing atoms. b) Schematic of the NN architecture predicting the (binary) topological class using spectral and atom-type inputs. Spectral and atom-type inputs are individually embedded by fully connected layers before performing a direct product between corresponding spectral and atomic channels. These composite features are aggregated for a given material and passed to a final fully connected block to predict the topological class.





**Figure 3.** NN classifier performance. a) The receiver operating characteristic (ROC) curve showing the tradeoff between true and false positive rates for the NN model. The area under the curve (AUC) for each dataset is noted in the legend. b,c) Comparative plots of the overall recall, precision, and  $F_1$  scores for topological (b) and trivial (c) examples obtained using different methods discussed in the main text. d,e) Element-specific  $F_1$  scores for topological (d) and trivial (e) examples. Each element's entry lists its atomic number, atomic symbol, and  $F_1$  score. Elements with no score listed were not present in the dataset.

### 3.3. Machine-Learning Model Performance

**Figure 3** summarizes the performance of the trained NN classifier. The receiver operating characteristic (ROC) curve, which indicates the tradeoff between true and false positive rates, is shown in Figure 3a. We use three different metrics in assessing the quality of prediction: recall, precision, and  $F_1$  score. These metrics are defined as

$$\text{recall} : r = \frac{t_p}{t_p + f_n} \quad (1a)$$

$$\text{precision} : p = \frac{t_p}{t_p + f_p} \quad (1b)$$

$$F_1 \text{ score} : F_1 = 2 \frac{pr}{p+r} \quad (1c)$$

where  $t_p$  and  $t_n$  denote the number of true positive and true negative predictions, and  $f_p$  and  $f_n$  denote the number of false positive and false negative predictions of a given class, respectively. The NN classifier achieved  $F_1$  scores of 89% and 93% for topological and trivial classes, respectively. We compare these results to the performance of a traditional support vector machine (SVM) operating on one-hot encoded atom types only (denoted SVM-type) and on a concatenated array of spectra for all atom types (denoted SVM), as shown in Figure 3b,c. The average performance of the PCA and  $k$ -means clustering approach across all elements is also included for reference. Note that the concatenated feature vector input to the SVM contains zeros in place of spectra corresponding to elements not contained in the compound. We find that both the NN and SVM classifiers based on XANES spectral inputs outperform

the baseline model relying on atom types alone, suggesting that XANES spectral features provide meaningful insight to topological indication. To maintain the same number of neurons between SVM-type and SVM models, the SVM-type inputs were copied 200 times (the length of the spectral inputs) to construct the input features, which led to a combined increase of 5% in the  $F_1$  scores compared to a minimal SVM-type model reported in Figure S5a, Supporting Information, for comparison. The NN further improves upon the SVM model predictions, particularly in the precision of topological classification which increased by 4%. We note that the NN with both spectral and atom-type inputs achieves a combined improvement of  $\approx 7\%$  in the  $F_1$  scores compared to a NN model of similar size operating on atom-type inputs alone (Figure S5a, Supporting Information). Additional details about the reference models are provided in the Supporting Information. We also assess the sensitivity to the spectral energy resolution in Figure S7, Supporting Information. While the main results of this work are obtained for spectra sampled at intervals of  $\approx 0.28$  eV, we see that a sampling of  $\approx 5$  eV is sufficient for comparable performance. Finally, we compute the average metric scores obtained by the NN classifier individually for each absorbing element, shown in Figure 3d,e for topological and trivial examples, respectively. Corresponding results for the SVM model and additional plots for the NN classifier are shown in Figures S4 and S6, Supporting Information, respectively.

### 3.4. Application to Experimental Spectra

While we are unable to include experimental spectra in our training set due to limited availability, we present a preliminary

**Table 1.** Predictions on corresponding experimental and computational spectra.

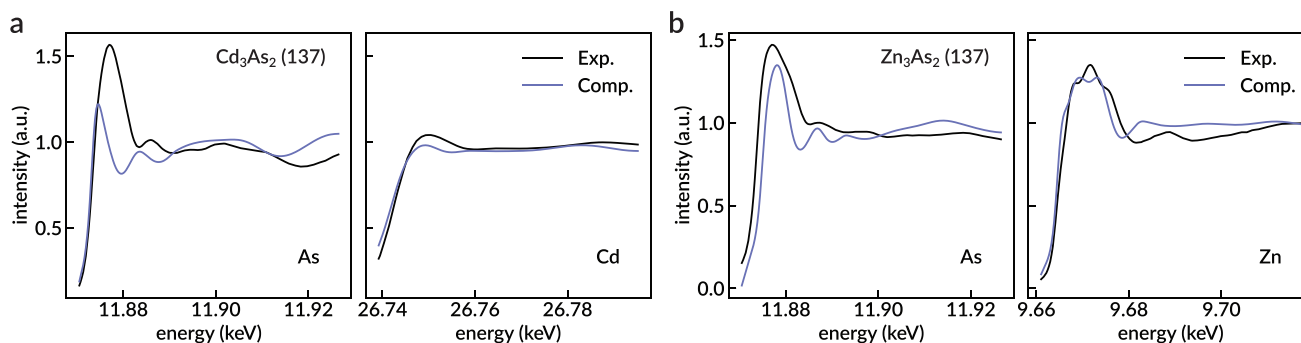
Material	Spacegroup	Class		
		True	Pred. (Exp.)	Pred. (Comp.)
NbAs	109	Topo.	Topo.	Topo.
LaAlGe <sup>a)</sup>	109	Topo.	Topo.	Topo.
Zn <sub>3</sub> As <sub>2</sub>	137	Trivial	Trivial	Trivial
Cd <sub>3</sub> As <sub>2</sub>	137	Topo.	Topo.	Topo.
CdGeYb <sup>b)</sup>	189	Topo.	Topo.	~
MoSe <sub>2</sub>	194	Trivial	Topo. <sup>d)</sup>	Topo. <sup>d)</sup>
CdTe	*c)	Trivial	Trivial	Trivial

<sup>a)</sup>Al K-edge was not measured; <sup>b)</sup>Yb K-edge was not measured; <sup>c)</sup>The same classifications are obtained for all computed spacegroups: 63, 152, 186, 216, and 225; <sup>d)</sup>Incorrectly predicted.

effort by making predictions on a small set of seven experimental XAS spectra and their computational counterparts, where available. The XAS experiments were performed at the 4-ID-D beamline of the advanced photon source (APS) and include measurements of both topological and trivial compounds listed in **Table 1**. The predictions obtained using the experimental and computational spectra were all consistent with one another, though in one instance (MoSe<sub>2</sub>) both are incorrectly classified, as shown in **Table 1**. Specifically, within this set of examples, the NN correctly classifies six of the seven sets of experimental spectra, and five of the six sets of computational spectra (computational spectra were unavailable for one of the seven compounds). Additionally, we note that for the two ternary compounds, LaAlGe and CdGeYb, one of the three absorption edges could not be measured at this time; in these cases, the two available experimental spectra were used to make a prediction. As an example, **Figure 4** shows the experimental and computational XAS spectra of the topological semimetal Cd<sub>3</sub>As<sub>2</sub> (**Figure 4a**) and the isostructural trivial compound Zn<sub>3</sub>As<sub>2</sub> (**Figure 4b**). While there is some misalignment of the experimental spectra relative to the computed ones, many of the key qualitative features are preserved. We expect that a certain tolerance in the misalignment is admissible, further reinforced by the results of the sensitivity analysis discussed in the previous section. Spectra for the remaining experimental examples are provided in **Figure S9**, Supporting Information.

## 4. Discussion

Our results indicate that the NN classifier enables higher and more balanced predictive accuracy over the PCA and *k*-means clustering approach for a majority of elements, including significant improvement for alkali metals. Certain elements are better indicators of one class over another; for instance, the alkali metals and halogens appear to serve as somewhat poor indicators of topological samples but are well-predicted in trivial compounds. A possible explanation for this is that the elements in these columns rarely contribute to frontier orbitals (valence and conduction bands) in materials, and are thereby poor indicators of topology. Certain transition-metal elements, such as Cr, Co, Ni, Tc, and Rh, also exhibit imbalanced accuracy in the prediction of trivial and topological classes. This is most likely due to the over-representation of topological examples containing Cr, Co, Ni, and Rh (**Figure S5c**, Supporting Information), since accurate prediction of topological compounds is prioritized during training. Tc is the least abundant element in the dataset (**Figure S1a,b**, Supporting Information), which accounts for the model's weak performance on Tc-containing compounds. However, further investigation of the relevant spectroscopic features—whether pre-edge, edge, or post-edge—in connection with the corresponding electronic transitions (e.g., 1s → 3d) may be useful to better understand performance barriers for transition metals. Finally, we comment on the



**Figure 4.** a,b) Comparison between experimental and computational XAS spectra. Experimental (black) and computational (blue) K-edge XANES spectra of As and Cd in Cd<sub>3</sub>As<sub>2</sub> (topological) (a) and As and Zn in Zn<sub>3</sub>As<sub>2</sub> (trivial) (b). The spacegroup of each structure is indicated in parentheses. Both experimental and computational inputs in (a,b) are correctly classified.

**Table 2.** Predictions on mislabeled Weyl semimetals.

Material	Spacegroup	Predicted class
TaAs	109	Topological
NbAs	109	Topological
NbP	109	Topological
WTe <sub>2</sub>	31	Topological
Ag <sub>2</sub> Se	17	Trivial
LaAlGe	109	Topological
Ba <sub>7</sub> Al <sub>4</sub> Ge <sub>9</sub>	42	Topological
Cu <sub>2</sub> SnTe <sub>3</sub>	44	Topological
BiTeI	143	Trivial
Al <sub>4</sub> Mo	8	Topological
KOs <sub>2</sub> O <sub>6</sub>	216	Topological
Zn <sub>2</sub> In <sub>2</sub> S <sub>5</sub>	186	Trivial

comparatively low precision obtained for topological over trivial examples, 86% and 95%, respectively. While the higher false positive rate of topological materials may suggest additional model improvements are needed, it may also indicate missed topological candidates. In fact, since the TQC formalism considers only the characters of electronic bands at high-symmetry points, it may incorrectly classify certain Weyl semimetals with topological singularities at arbitrary  $k$ -points.<sup>[27]</sup> In particular, we identified 12 experimentally verified<sup>[5]</sup> or theoretically predicted Weyl semimetals<sup>[80]</sup> that are labeled as trivial in the TQC database, nine of which we correctly predict as topological using our NN classifier (Table 2). Thus, the potential presence of topological singularities not considered in the TQC formalism might account for some loss of precision in the classification of topological examples. In addition, we summarize in Table S1, Supporting Information, the top 100 predicted topological materials from a collection of 459 samples not represented in the TQC database. These are the top candidates predicted by our model that may contain topological singularities. We do note that the success of the NN classifier can be attributed significantly to the presence of particular elements; further work is being pursued to more accurately decouple this contribution from that of more subtle variations in the XAS spectral features for a given absorbing element.

## 5. Conclusion

We explored the predictive power of XAS as a potential discriminant of topological character by training and evaluating a NN classifier on more than 10–000 examples of computed XANES spectra<sup>[72]</sup> labeled according to the largest catalogue of topological materials.<sup>[27,34]</sup> A number of important extensions are envisioned for this work, such as its application to experimental XANES data, incorporation of a multi-fidelity approach to favor experimentally validated examples,<sup>[81]</sup> expansion of the energy range to the extended X-ray absorption fine structure regime, and inquiry into the detailed contribution from spectral features for individual elements. The theoretical connections between band topology and the local chemical environment encoded in

XANES spectra has not yet been established, and we envision data-driven methods as a possible tool in aiding this theoretical development. Our current results demonstrate a promising pathway to develop robust experimental protocols for high-throughput screening of candidate topological materials aided by machine-learning methods. Additionally, the flexibility of the XAS sample environment can further enable the study of materials whose topological phases emerge when driven by electric, magnetic, or strain fields, and even present the opportunity to study topology with strong disorder and topology in amorphous materials.<sup>[82,83]</sup> Thus, machine-learning-empowered XAS may be poised to become a simple but powerful experimental tool for topological classification.

## 6. Experimental Section

**Data Processing:** The computed XANES spectra of each absorbing atom were interpolated and re-sampled at 200 evenly spaced energy values. Each XANES spectrum spanned an energy range of 56 eV, and spectra from the same absorbing atom were co-aligned using the calculated absolute energy scale. Spectra of the same absorbing atom were standardized by centering the mean of the average intensities over the sampled energy range, and scaling by the mean of the standard deviations in intensity values.

**Machine Learning:** Principal component analysis and SVM model implementation and training were carried out using the scikit-learn Python library.<sup>[84]</sup> The NN models presented in this work were implemented in Python using the PyTorch<sup>[85]</sup> and PyTorch Geometric<sup>[86]</sup> libraries. The atom-type embeddings were obtained using a single fully connected layer with 93 input and output neurons. The spectral embeddings of the original 200-feature spectra were obtained using a series of two fully connected layers with 256 and 64 output neurons, respectively, each followed by a dropout layer with a rate of 0.5 and a rectified linear unit (ReLU) activation. The composite embedded features had dimensions of 5952 and were passed to a second series of two fully connected layers with 256 and 64 output neurons, respectively, each followed by a dropout layer with a rate of 0.5 and a ReLU activation. A final, sigmoid-activated, fully connected layer was then used to output the scalar prediction. The models were trained on a Quadro RTX 6000 graphics processing unit (GPU) with 24GB of random access memory. Optimization was performed using the Adam optimizer to minimize the binary cross-entropy loss.

**Sample Preparation:** NbAs and CdTe crystals were grown using chemical vapor transport while LaAlGe, CdGeYb, Cd<sub>3</sub>As<sub>2</sub>, Zn<sub>3</sub>As<sub>2</sub>, and MoSe<sub>2</sub> crystals were grown using the flux method as described in the literature. The samples exhibited clear, lustrous surfaces with demarcated straight edges indicating the orientation of the crystal axes. The samples were not polished.

**X-ray Absorption Spectroscopy:** XAS experiments were performed at the 4-ID-D beamline of the Advanced Photon Source, Argonne National Laboratory. The X-ray energy was selected using a Si (111) double crystal monochromator, which was detuned to reject harmonics. Measurements were recorded near the  $K$  absorption edge for each element. For absorption edges below 23 keV, a Pd mirror was employed to further reject harmonics. Measurements were done at room temperature and in transmission mode using N<sub>2</sub> and Ar filled ion chambers to detect both incident and transmitted intensities, respectively. Prior to making predictions, experimental spectra were pre-processed as follows. First, a linear background was fit to the pre-edge region and subtracted. The resulting spectra were fit with an arctangent function of the form  $a_1(1 + 2\tan^{-1}(a_2(E - a_3))/\pi)/2$  with fitting parameters  $\{a_i\}$  and measured energies  $E$ , and subsequently scaled by  $1/a_1$ . This ensured that experimental intensities were scaled consistently with computational ones, which were 0 at energies well below the absorption edge and approach 1 at energies well above the absorption edge. Finally, the

experimental spectra were shifted in energy so that the area under the fitted arctangent matched that of the arctangent fit to the average computational spectrum for each absorbing atom. Examples of these pre-processing steps are shown in Figure S10, Supporting Information. Finally, experimental spectra were interpolated and scaled according to the means and standard deviations of the computational spectra as described in the Data Processing section.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

N.A. and J.A. contributed equally to this work. N.A. acknowledges National Science Foundation GRFP support under Grant No. 1122374. J.A. acknowledges National Science Foundation GRFP support under Grant No. DGE-1745303. N.A. and M.L. acknowledge the support from the U.S. Department of Energy (DOE), Office of Science (SC), Basic Energy Sciences (BES), Award No. DE-SC0021940. F.H., T.N., and M.L. acknowledge the support from the DOE Award No. DE-SC0020148. M.L. is partially supported by NSF DMR-2118448, the Norman C. Rasmussen Career Development Chair, and the Class of 1947 Career Development Chair, and acknowledges the support from Dr. R. Wachnik. B.A.B. and N.R. gratefully acknowledge financial support from the Schmidt DataX Fund at Princeton University made possible through a major gift from the Schmidt Futures Foundation, NSF-MRSEC Grant No. DMR-2011750 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 101020833). C.H.R. was partially supported by the Applied Mathematics Program of the U.S. DOE Office of Science Advanced Scientific Computing Research under Contract No. DE-AC02-05CH11231. Work performed at the Center for Nanoscale Materials, a U.S. Department of Energy Office of Science User Facility, was supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. This material is based, in part, upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Keywords

machine learning, topological materials, X-ray absorption spectroscopy

Received: May 6, 2022

Revised: July 18, 2022

Published online: October 31, 2022

- [1] M. Z. Hasan, C. L. Kane, *Rev. Mod. Phys.* **2010**, *82*, 3045.
- [2] X.-L. Qi, S.-C. Zhang, *Rev. Mod. Phys.* **2011**, *83*, 1057.
- [3] B. Yan, S.-C. Zhang, *Rep. Prog. Phys.* **2012**, *75*, 096501.
- [4] A. Bansil, H. Lin, T. Das, *Rev. Mod. Phys.* **2016**, *88*, 021004.
- [5] B. Yan, C. Felser, *Annu. Rev. Condens. Matter Phys.* **2017**, *8*, 337.
- [6] N. P. Armitage, E. J. Mele, A. Vishwanath, *Rev. Mod. Phys.* **2018**, *90*, 015001.
- [7] B. Bradlyn, L. Elcoro, J. Cano, M. Vergniory, Z. Wang, C. Felser, M. Aroyo, B. A. Bernevig, *Nature* **2017**, *547*, 298.
- [8] J. Kruthoff, J. De Boer, J. Van Wezel, C. L. Kane, R.-J. Slager, *Phys. Rev. X* **2017**, *7*, 041069.
- [9] J. Cano, B. Bradlyn, Z. Wang, L. Elcoro, M. Vergniory, C. Felser, M. Aroyo, B. A. Bernevig, *Phys. Rev. B* **2018**, *97*, 035139.
- [10] L. Elcoro, Z. Song, B. A. Bernevig, *Phys. Rev. B* **2020**, *102*, 035110.
- [11] B. J. Wieder, B. Bradlyn, J. Cano, Z. Wang, M. G. Vergniory, L. Elcoro, A. A. Soluyanov, C. Felser, T. Neupert, N. Regnault, B. A. Bernevig, *Nat. Rev. Mater.* **2022**, *7*, 196.
- [12] A. Bouhon, G. F. Lange, R.-J. Slager, *Phys. Rev. B* **2021**, *103*, 245127.
- [13] D. Călugăru, A. Chew, L. Elcoro, Y. Xu, N. Regnault, Z.-D. Song, B. A. Bernevig, *Nat. Phys.* **2022**, *18*, 185.
- [14] R.-J. Slager, A. Mesaros, V. Juričić, J. Zaanen, *Nat. Phys.* **2013**, *9*, 98.
- [15] P. Jadaun, D. Xiao, Q. Niu, S. K. Banerjee, *Phys. Rev. B* **2013**, *88*, 085110.
- [16] C.-K. Chiu, J. C. Teo, A. P. Schnyder, S. Ryu, *Rev. Mod. Phys.* **2016**, *88*, 035005.
- [17] H. C. Po, A. Vishwanath, H. Watanabe, *Nat. Commun.* **2017**, *8*, 50.
- [18] Z. Song, T. Zhang, Z. Fang, C. Fang, *Nat. Commun.* **2018**, *9*, 3530.
- [19] Z. Song, S.-J. Huang, Y. Qi, C. Fang, M. Hermele, *Sci. Adv.* **2019**, *5*, eaax2007.
- [20] H. C. Po, *J. Phys.: Condens. Matter* **2020**, *32*, 263001.
- [21] B. Peng, Y. Jiang, Z. Fang, H. Weng, C. Fang, *Phys. Rev. B* **2022**, *105*, 235138.
- [22] R. Chen, H. C. Po, J. B. Neaton, A. Vishwanath, *Nat. Phys.* **2018**, *14*, 55.
- [23] H. Watanabe, H. C. Po, A. Vishwanath, *Sci. Adv.* **2018**, *4*, eaat8685.
- [24] K. Choudhary, K. F. Garrity, F. Tavazza, *Sci. Rep.* **2019**, *9*, 8534.
- [25] K. Choudhary, K. F. Garrity, J. Jiang, R. Pachter, F. Tavazza, *npj Comput. Mater.* **2020**, *6*, 49.
- [26] K. Choudhary, K. F. Garrity, N. J. Ghimire, N. Anand, F. Tavazza, *Phys. Rev. B* **2021**, *103*, 155131.
- [27] M. Vergniory, L. Elcoro, C. Felser, N. Regnault, B. A. Bernevig, Z. Wang, *Nature* **2019**, *566*, 480.
- [28] T. Zhang, Y. Jiang, Z. Song, H. Huang, Y. He, Z. Fang, H. Weng, C. Fang, *Nature* **2019**, *566*, 475.
- [29] F. Tang, H. C. Po, A. Vishwanath, X. Wan, *Nature* **2019**, *566*, 486.
- [30] F. Tang, H. C. Po, A. Vishwanath, X. Wan, *Sci. Adv.* **2019**, *5*, eaau8725.
- [31] F. Tang, H. C. Po, A. Vishwanath, X. Wan, *Nat. Phys.* **2019**, *15*, 470.
- [32] D. Wang, F. Tang, J. Ji, W. Zhang, A. Vishwanath, H. C. Po, X. Wan, *Phys. Rev. B* **2019**, *100*, 195108.
- [33] Y. Xu, L. Elcoro, Z.-D. Song, B. J. Wieder, M. Vergniory, N. Regnault, Y. Chen, C. Felser, B. A. Bernevig, *Nature* **2020**, *586*, 702.
- [34] M. G. Vergniory, B. J. Wieder, L. Elcoro, S. S. Parkin, C. Felser, B. A. Bernevig, N. Regnault, *Science* **2022**, *376*, eabg9094.
- [35] S. Suga, A. Sekiyama, *Photoelectron Spectroscopy: Bulk and Surface Electronic Structures*, Vol. 176, Springer, Berlin, Germany **2013**.
- [36] B. Lv, T. Qian, H. Ding, *Nat. Rev. Phys.* **2019**, *1*, 609.
- [37] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.
- [38] Y. Liu, T. Zhao, W. Ju, S. Shi, *J. Mater. Sci.* **2017**, *3*, 159.
- [39] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.
- [40] H. Zhang, K. Hippalgaonkar, T. Buonassisi, O. M. Løvvik, E. Sagvolden, D. Ding, *ES Energy Environ.* **2018**, *2*, 1.



- [41] P. Mikulskis, M. R. Alexander, D. A. Winkler, *Adv. Intell. Syst.* **2019**, 1, 1900045.
- [42] Y. Juan, Y. Dai, Y. Yang, J. Zhang, *J. Mater. Sci. Technol.* **2020**, 79, 178.
- [43] A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta, I. Takeuchi, *Nat. Commun.* **2020**, 11, 5966.
- [44] A. Mannodi-Kanakkithodi, M. K. Chan, *Trends Chem.* **2021**, 3, 79.
- [45] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, 3, 2810.
- [46] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2016**, 2, 16028.
- [47] J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, *Phys. Rev. X* **2014**, 4, 011019.
- [48] N. Claussen, B. A. Bernevig, N. Regnault, *Phys. Rev. B* **2020**, 101, 245117.
- [49] J. F. Rodriguez-Nieva, M. S. Scheurer, *Nat. Phys.* **2019**, 15, 790.
- [50] A. Ma, Y. Zhang, T. Christensen, H. C. Po, L. Jing, L. Fu, M. Soljačić, *arXiv: 2202.05255*, **2022**.
- [51] Y. Zhang, E.-A. Kim, *Phys. Rev. Lett.* **2017**, 118, 216401.
- [52] W. Lian, S.-T. Wang, S. Lu, Y. Huang, F. Wang, X. Yuan, W. Zhang, X. Ouyang, X. Wang, X. Huang, L. He, X. Chang, D.-L. Deng, L. Duan, *Phys. Rev. Lett.* **2019**, 122, 210503.
- [53] M. S. Scheurer, R.-J. Slager, *Phys. Rev. Lett.* **2020**, 124, 226401.
- [54] P. Zhang, H. Shen, H. Zhai, *Phys. Rev. Lett.* **2018**, 120, 066401.
- [55] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, *Rev. Mod. Phys.* **2019**, 91, 045002.
- [56] M. R. Carbone, S. Yoo, M. Topsakal, D. Lu, *Phys. Rev. Mater.* **2019**, 3, 033604.
- [57] A. Cui, K. Jiang, M. Jiang, L. Shang, L. Zhu, Z. Hu, G. Xu, J. Chu, *Phys. Rev. Appl.* **2019**, 12, 054049.
- [58] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, J. I.-J. Wang, Q. Ma, Y. Cao, D. Rodan-Legrain, Y.-Q. Bie, E. Navarro-Moratalla, D. Klein, D. MacNeill, S. Wu, H. Kitadai, X. Ling, P. Jarillo-Herrero, J. Kong, J. Yin, T. Palacios, *Adv. Mater.* **2020**, 32, 2000953.
- [59] A. M. Samarakoon, K. Barros, Y. W. Li, M. Eisenbach, Q. Zhang, F. Ye, V. Sharma, Z. Dun, H. Zhou, S. A. Grigera, C. D. Batista, D. A. Tennant, *Nat. Commun.* **2020**, 11, 892.
- [60] Y. Zhang, A. Mesaros, K. Fujita, S. Ekins, M. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. S. Davis, E. Khatami, E.-A. Kim, *Nature* **2019**, 570, 484.
- [61] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, C. Weitenberg, *Nat. Phys.* **2019**, 15, 917.
- [62] M. Newville, *Rev. Mineral. Geochem.* **2014**, 78, 33.
- [63] A. Gaur, B. Shrivastava, *Rev. J. Chem.* **2015**, 5, 361.
- [64] S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, L. Hung, *npj Comput. Mater.* **2020**, 6, 109.
- [65] C. Zheng, C. Chen, Y. Chen, S. P. Ong, *Patterns* **2020**, 1, 100013.
- [66] S. Kiyohara, T. Miyata, K. Tsuda, T. Mizoguchi, *Sci. Rep.* **2018**, 8, 13548.
- [67] A. Guda, S. Guda, A. Martini, A. Kravtsova, A. Algasov, A. Bugaev, S. Kubrin, L. Guda, P. Šot, J. van Bokhoven, C. Copéret, A. Soldatov, *npj Comput. Mater.* **2021**, 7, 203.
- [68] Y. Suzuki, H. Hino, M. Kotsugi, K. Ono, *npj Comput. Mater.* **2019**, 5, 39.
- [69] M. R. Carbone, M. Topsakal, D. Lu, S. Yoo, *Phys. Rev. Lett.* **2020**, 124, 156401.
- [70] C. D. Rankine, M. M. Madkhali, T. J. Penfold, *J. Phys. Chem. A* **2020**, 124, 4263.
- [71] J. Lüder, *Phys. Rev. B* **2021**, 103, 045140.
- [72] K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong, K. A. Persson, *Sci. Data* **2018**, 5, 180151.
- [73] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, 1, 011002.
- [74] 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, S. P. Ong, *npj Comput. Mater.* **2018**, 4, 12.
- [75] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, 68, 314.
- [76] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K. A. Persson, *Comput. Mater. Sci.* **2015**, 97, 209.
- [77] J. J. Rehr, J. J. Kas, F. D. Vila, M. P. Prange, K. Jorissen, *Phys. Chem. Chem. Phys.* **2010**, 12, 5503.
- [78] G. Bergerhoff, I. Brown, F. H. Allen, *Crystallographic Databases*, International Union of Crystallography, Chester, UK **1987**.
- [79] J. E. Penner-Hahn, *Compr. Coord. Chem. II* **2003**, 2, 159.
- [80] Q. Xu, Y. Zhang, K. Koepf, W. Shi, J. van den Brink, C. Felser, Y. Sun, *npj Comput. Mater.* **2020**, 6, 32.
- [81] X. Meng, G. E. Karniadakis, *J. Comput. Phys.* **2020**, 401, 109020.
- [82] A. Agarwala, V. B. Shenoy, *Phys. Rev. Lett.* **2017**, 118, 236402.
- [83] E. Prodan, *J. Phys. A: Math. Theor.* **2011**, 44, 113001.
- [84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, 12, 2825.
- [85] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in *Advances in Neural Information Processing Systems 32*, (Eds: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett), Curran Associates, Inc., Red Hook, NY, USA **2019**, pp. 8024–8035.
- [86] M. Fey, J. E. Lenssen, *arXiv: 1903.02428*, **2019**.