

Lawrence Berkeley National Laboratory

LBL Publications

Title

Machine learning spectral indicators of topology

Permalink

<https://escholarship.org/uc/item/01m864fh>

Authors

Andrejevic, Nina
Andrejevic, Jovana
Bernevig, B Andrei
[et al.](#)

Publication Date

2020-03-02

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Machine learning spectral indicators of topology

Nina Andrejevic,^{1,2,3,*} Jovana Andrejevic,^{4,*} B. Andrei Bernevig,^{5,6,7}
Nicolas Regnault,⁵ Chris H. Rycroft,^{4,8,†} and Mingda Li^{1,9,‡}

¹*Quantum Measurement Group, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

²*Department of Materials Science and Engineering,
Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

³*Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, USA*

⁴*John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, MA 02138, USA*

⁵*Department of Physics, Princeton University, Princeton, NJ 08544, USA*

⁶*Donostia International Physics Center, P. Manuel de Lardizabal 4, 20018 Donostia-San Sebastian, Spain*

⁷*IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*

⁸*Computational Research Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720, USA*

⁹*Department of Nuclear Science and Engineering,
Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

(Dated: June 21, 2024)

Topological materials discovery has emerged as an important frontier in condensed matter physics. While theoretical classification frameworks such as topological quantum chemistry 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 environment; thus, it may indirectly encode signatures of materials' topology. 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, we show that XAS can potentially uncover materials' topology when augmented by machine learning. Using the computed X-ray absorption near-edge structure (XANES) spectra of more than 10,000 inorganic materials, we train a neural network classifier that predicts topological class directly from XANES signatures with F_1 scores of 89% and 93% for topological and trivial classes, respectively, and achieves F_1 scores above 90% for materials containing certain elements. Given the simplicity of the XAS setup and its compatibility with multimodal sample environments, the proposed machine learning-empowered 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 a variety of field-driven phenomena *in situ*, such as magnetic field-driven topological phase transitions.

Keywords: machine learning, topological materials, X-ray absorption spectroscopy

I. INTRODUCTION

Topological materials are characterized by a topologically nontrivial electronic band structure from which they derive their exceptional transport properties [1–6]. 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 [7–13], symmetry-based indicators [7, 14–21], electron-filling constraints [7, 22, 23], and spin-orbit spillage [24–26]. These frameworks have facilitated the prediction of over 8,000 topologically non-trivial phases [27–34], 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 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 [35, 36]; 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 [37–44] and facilitate inverse design through high-throughput property prediction [45–47]. Several recent studies have proposed data-driven frameworks for predicting band topology from structural and compositional attributes [48, 49] and quantum theoretical or simulated data [50–53]. At the same time, machine learn-

* These authors contributed equally to this work.

Corresponding author: nandrejevic@alum.mit.edu

† Corresponding author: chr@seas.harvard.edu

‡ Corresponding author: mingda@mit.edu

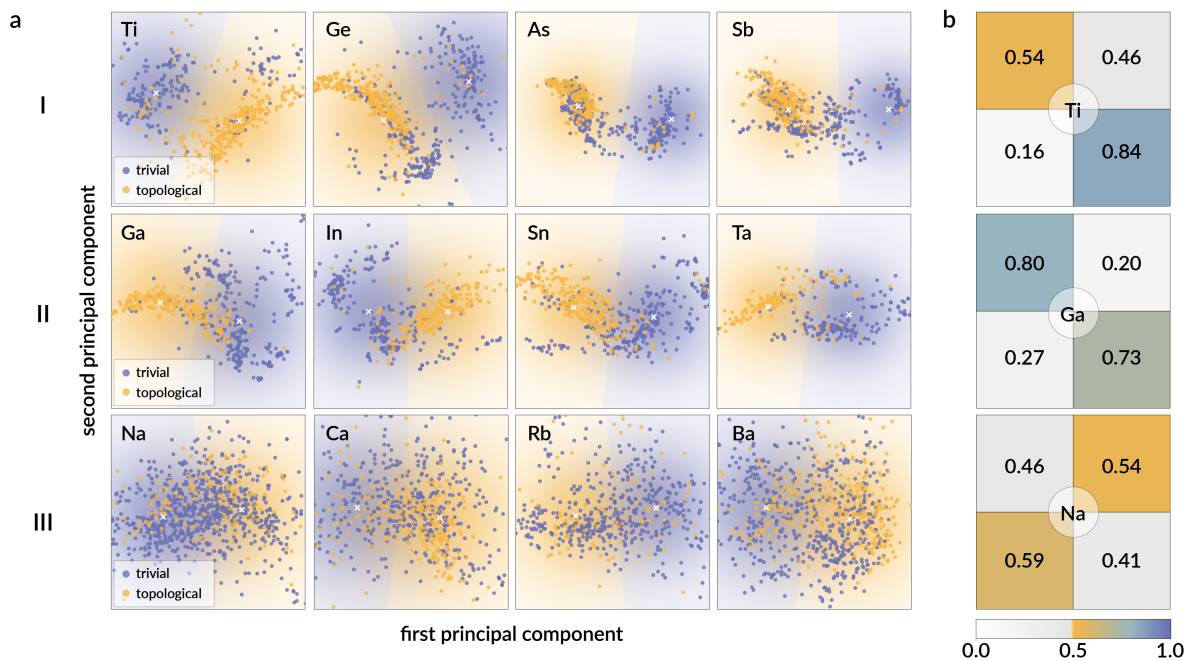


FIG. 1. **Exploratory analysis using principal components and k -means clustering.** (a) Decision boundary visualizations of classifications by unsupervised k -means clustering for selected elements. As detailed in the main text, the k -means clustering is performed on the subset of principal components 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.

ing methods are being adopted to automate and improve data analysis for a broad range of experimental techniques [54–60]. 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 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 [61]. 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 approximately 50 eV of an XAS absorption edge, provides a specie-specific fingerprint of 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 [62]. As a result, machine learning methods have been introduced to automate the estimation of materials parameters such as coordination environments [55, 63–66], oxidation states, [63, 66], and crystal-field splitting [67] from XANES and other core-level spectroscopies, and even enable direct prediction of XANES spectra from structural and atomic descriptors [68–70]. 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. 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 $1s$ 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 [71] according to topological character using the catalogue of high-quality topological materials predicted by topological quantum chemistry (TQC) [27, 34]. 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 signa-

tures 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.

II. DATA PREPARATION AND PRE-PROCESSING

XAS data were obtained from the published database of computed K-edge XANES spectra [71] and additional examples distributed on the Materials Project [72–75], which are computed using the FEFF9 program [76]. The materials from the XANES database were then labeled according to their classification in the database of topological materials [27, 34], which is based on the formalism of TQC [7]. The classifications in the TQC database are based on structures from the Inorganic Crystal Structure Database (ICSD) [77], 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 Structure-Matcher algorithm [74, 75]. 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. 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: 4,957 topological ($\sim 38\%$) and 8,194 trivial ($\sim 62\%$). 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 7 unique chemical species. The representation of different elements among topological and trivial examples is shown in **Fig. S1(a-b)**. 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 **Fig. S1(c)**. 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.

III. RESULTS

A. Exploratory analysis

Prior to training the neural network 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 **Fig. 1**. The decision boundary between the two clusters identified by k -means clustering, projected along the first two principal components, lies at the intersection of the blue (trivial) and orange (topological) shaded regions in **Fig. 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 **Fig. S2**. 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 **Fig. 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 **Fig. 1**). On the other hand, there were a number of unsuccessful examples of alkali and alkaline earth metals

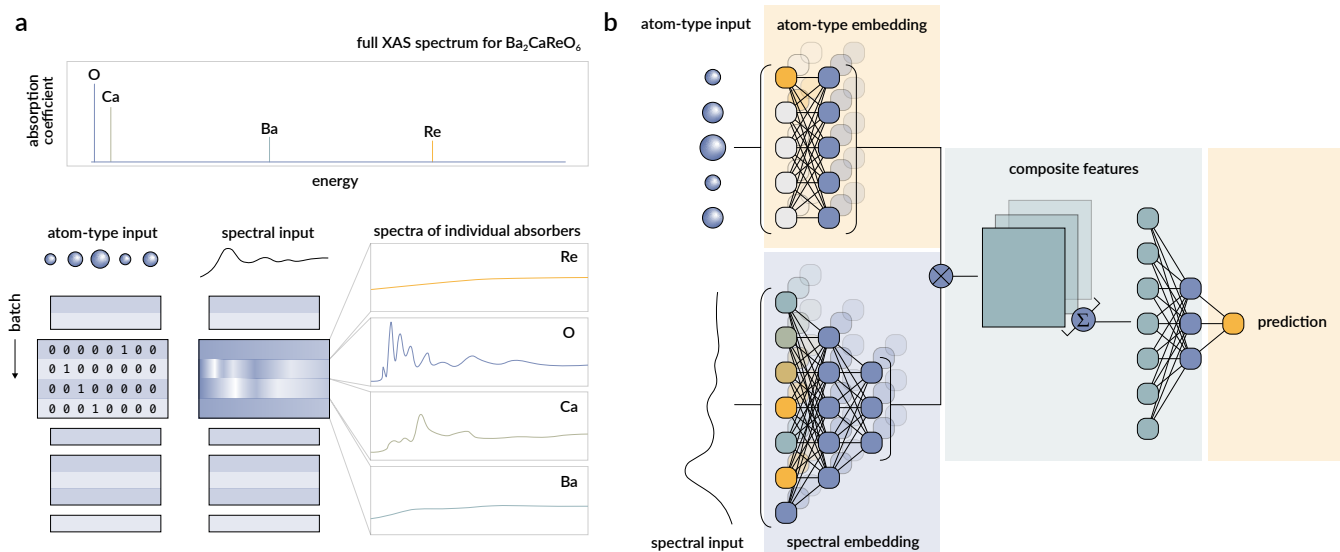


FIG. 2. **Data structure and model architecture.** (a) A schematic of the full XANES spectrum for a representative sample in the dataset, showing the signatures from different absorbing elements on an absolute energy scale. For a given material, the inputs to the NN classifier consist of one-hot encoded atom types (left) and XANES spectra (right) for all absorbing atoms. (b) Schematic of the neural network 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.

for which clustering of the data did not appear coincident with topological class (row III of **Fig. 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.

B. 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 **Fig. 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 [78], and thus representing the XANES spectra of all absorbers on a continuous energy scale would be either poorly resolved or exceedingly high-dimensional (**Fig. 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, a neural network 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 **Fig. 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 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.

C. Machine learning model performance

Fig. 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 **Fig. 3a**. We use three different metrics in assessing the quality of prediction: recall, precision, and F_1 score. These metrics are defined

as

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

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

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

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 **Fig. 3b** and **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 **Fig. S4a** 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 $\sim 7\%$ in the F_1 scores compared to a NN model of similar size operating on atom-type inputs alone (**Fig. S4a**). Additional details about the reference models are provided in the Supplementary Information. Finally, we compute the average metric scores obtained by the NN classifier individually for each absorbing element, shown in **Fig. 3d** and **e** for topological and trivial examples, respectively. Corresponding results for the SVM model and additional plots for the NN classifier are shown in **Fig. S4** and **Fig. S6**, respectively.

IV. 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

TABLE I. Predictions on mislabeled Weyl semimetals

Material	Spacegroup	Predicted class
TaAs	109	Topological
NbAs	109	Topological
NbP	109	Topological
WTe2	31	Topological
Ag2Se	17	Trivial
LaAlGe	109	Topological
Ba7Al4Ge9	42	Topological
Cu2SnTe3	44	Topological
BiTeI	143	Trivial
Al4Mo	8	Topological
KOs2O6	216	Topological
Zn2In2S5	186	Trivial

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 overrepresentation of topological examples containing Cr, Co, Ni, and Rh (**Fig. S5c**), since accurate prediction of topological compounds is prioritized during training. Tc is the least abundant element in the dataset (**Fig. S1a** and **b**), 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. $1s \rightarrow 3d$) may be useful to better understand performance barriers for transition metals. Finally, we comment on the 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 [27]. In particular, we identified 12 experimentally-verified [5] or theoretically-predicted Weyl semimetals [79] that are labeled as trivial in the TQC database, 9 of which we correctly predict as topological using our NN classifier (Table I). 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. We do note that the success of the neural network 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 varia-

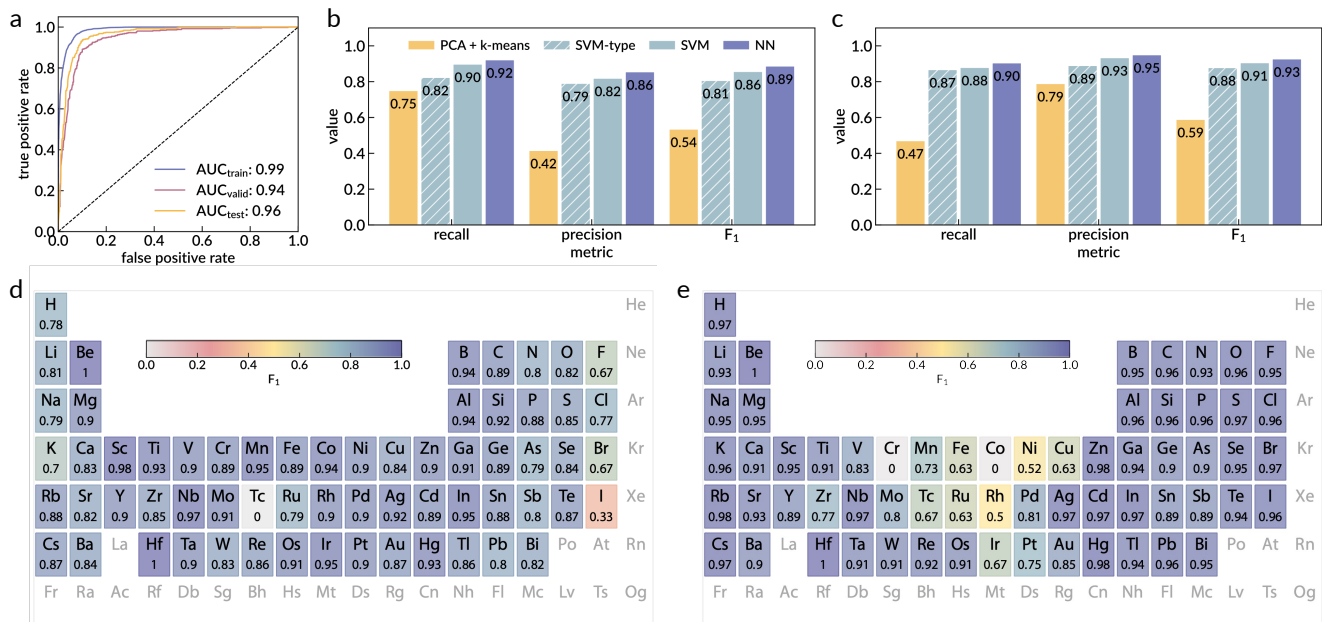


FIG. 3. **Neural network (NN) classifier performance.** (a) The receiver operating characteristic (ROC) curve showing the tradeoff between true and false positive rates for the NN model. The area under the curve (AUC) for each dataset is noted in the legend. (b-c) Comparative plots of the overall recall, precision, and F₁ scores for (b) topological and (c) trivial examples obtained using different methods discussed in the main text. (d-e) Element-specific F₁ scores for (d) topological and (e) trivial examples. Each element’s entry lists its atomic number, atomic symbol, and F₁ score. Elements with no score listed were not present in the dataset.

tions in the XAS spectral features for a given absorbing element.

V. CONCLUSION

We explored the predictive power of XAS as a potential discriminant of topological character by training and evaluating a neural network classifier on more than 10,000 examples of computed XANES spectra [71] labeled according to the largest catalogue of topological materials [27, 34]. 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 [80], expansion of the energy range to the extended X-ray absorption fine structure (EXAFS) regime, and inquiry into the detailed contribution from spectral features for individual elements. Our 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 presents the opportunity to study topology with strong disorder and topology in amorphous materials [81, 82]. Thus, machine learning-empowered XAS may

be poised to become a simple but powerful experimental tool for topological classification.

VI. METHODS

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 [83]. The neural network models presented in this work were implemented in Python using the PyTorch [84] and PyTorch Geometric [85] 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 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 24 GB of random access memory (RAM). Optimization was performed using the Adam optimizer to minimize the binary cross-entropy loss.

ACKNOWLEDGMENTS

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-SC0020148. M.L. acknowledges the support from the U.S. Department of Energy (DOE), Office of Science (SC), Award No. DE-SC0021940. 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.

Competing interests The authors declare no competing interests.

Additional information Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Nina Andrejevic (nandrejevic@alum.mit.edu), Mingda Li (mingda@mit.edu), or Chris H. Rycroft (chr@seas.harvard.edu).

Data availability All the data and code supporting the findings are available from the corresponding authors upon reasonable request.

-
- [1] M. Z. Hasan and C. L. Kane, Colloquium: Topological insulators, *Rev. Mod. Phys.* **82**, 3045 (2010).
- [2] X.-L. Qi and S.-C. Zhang, Topological insulators and superconductors, *Rev. Mod. Phys.* **83**, 1057 (2011).
- [3] B. Yan and S.-C. Zhang, Topological materials, *Reports on Progress in Physics* **75**, 096501 (2012).
- [4] A. Bansil, H. Lin, and T. Das, Colloquium: Topological band theory, *Rev. Mod. Phys.* **88**, 021004 (2016).
- [5] B. Yan and C. Felser, Topological materials: Weyl semimetals, *Annual Review of Condensed Matter Physics* **8**, 337 (2017).
- [6] N. P. Armitage, E. J. Mele, and A. Vishwanath, Weyl and Dirac semimetals in three-dimensional solids, *Rev. Mod. Phys.* **90**, 015001 (2018).
- [7] B. Bradlyn, L. Elcoro, J. Cano, M. Vergniory, Z. Wang, C. Felser, M. Aroyo, and B. A. Bernevig, Topological quantum chemistry, *Nature* **547**, 298 (2017).
- [8] J. Kruthoff, J. De Boer, J. Van Wezel, C. L. Kane, and R.-J. Slager, Topological classification of crystalline insulators through band structure combinatorics, *Physical Review X* **7**, 041069 (2017).
- [9] J. Cano, B. Bradlyn, Z. Wang, L. Elcoro, M. Vergniory, C. Felser, M. Aroyo, and B. A. Bernevig, Building blocks of topological quantum chemistry: Elementary band representations, *Physical Review B* **97**, 035139 (2018).
- [10] L. Elcoro, Z. Song, and B. A. Bernevig, Application of induction procedure and smith decomposition in calculation and topological classification of electronic band structures in the 230 space groups, *Physical Review B* **102**, 035110 (2020).
- [11] B. J. Wieder, B. Bradlyn, J. Cano, Z. Wang, M. G. Vergniory, L. Elcoro, A. A. Soluyanov, C. Felser, T. Neupert, N. Regnault, *et al.*, Topological materials discovery from nonmagnetic crystal symmetry, arXiv preprint arXiv:2106.00709 (2021).
- [12] A. Bouhon, G. F. Lange, and R.-J. Slager, Topological correspondence between magnetic space group representations and subdimensions, *Physical Review B* **103**, 245127 (2021).
- [13] D. Călugăru, A. Chew, L. Elcoro, N. Regnault, Z.-D. Song, and B. A. Bernevig, General construction and topological classification of all magnetic and non-magnetic flat bands, arXiv preprint arXiv:2106.05272 (2021).
- [14] R.-J. Slager, A. Mesaros, V. Juričić, and J. Zaanen, The space group classification of topological band-insulators, *Nature Physics* **9**, 98 (2013).
- [15] P. Jadaun, D. Xiao, Q. Niu, and S. K. Banerjee, Topological classification of crystalline insulators with space group symmetry, *Physical Review B* **88**, 085110 (2013).
- [16] C.-K. Chiu, J. C. Teo, A. P. Schnyder, and S. Ryu, Classification of topological quantum matter with symmetries, *Reviews of Modern Physics* **88**, 035005 (2016).
- [17] H. C. Po, A. Vishwanath, and H. Watanabe, Symmetry-based indicators of band topology in the 230 space groups, *Nature Communications* **8**, 1 (2017).
- [18] Z. Song, T. Zhang, Z. Fang, and C. Fang, Quantitative

- mappings between symmetry and topology in solids, *Nature Communications* **9**, 1 (2018).
- [19] Z. Song, S.-J. Huang, Y. Qi, C. Fang, and M. Hermele, Topological states from topological crystals, *Science Advances* **5**, eaax2007 (2019).
- [20] H. C. Po, Symmetry indicators of band topology, *Journal of Physics: Condensed Matter* **32**, 263001 (2020).
- [21] B. Peng, Y. Jiang, Z. Fang, H. Weng, and C. Fang, Topological classification and diagnosis in magnetically ordered electronic materials, *arXiv preprint arXiv:2102.12645* (2021).
- [22] R. Chen, H. C. Po, J. B. Neaton, and A. Vishwanath, Topological materials discovery using electron filling constraints, *Nature Physics* **14**, 55 (2018).
- [23] H. Watanabe, H. C. Po, and A. Vishwanath, Structure and topology of band structures in the 1651 magnetic space groups, *Science Advances* **4**, eaat8685 (2018).
- [24] K. Choudhary, K. F. Garrity, and F. Tavazza, High-throughput discovery of topologically non-trivial materials using spin-orbit spillage, *Scientific Reports* **9**, 1 (2019).
- [25] K. Choudhary, K. F. Garrity, J. Jiang, R. Pachter, and F. Tavazza, Computational search for magnetic and non-magnetic 2d topological materials using unified spin-orbit spillage screening, *NPJ Computational Materials* **6**, 1 (2020).
- [26] K. Choudhary, K. F. Garrity, N. J. Ghimire, N. Anand, and F. Tavazza, High-throughput search for magnetic topological materials using spin-orbit spillage, machine learning, and experiments, *Physical Review B* **103**, 155131 (2021).
- [27] M. Vergniory, L. Elcoro, C. Felser, N. Regnault, B. A. Bernevig, and Z. Wang, A complete catalogue of high-quality topological materials, *Nature* **566**, 480 (2019).
- [28] T. Zhang, Y. Jiang, Z. Song, H. Huang, Y. He, Z. Fang, H. Weng, and C. Fang, Catalogue of topological electronic materials, *Nature* **566**, 475 (2019).
- [29] F. Tang, H. C. Po, A. Vishwanath, and X. Wan, Comprehensive search for topological materials using symmetry indicators, *Nature* **566**, 486 (2019).
- [30] F. Tang, H. C. Po, A. Vishwanath, and X. Wan, Topological materials discovery by large-order symmetry indicators, *Science Advances* **5**, eaau8725 (2019).
- [31] F. Tang, H. C. Po, A. Vishwanath, and X. Wan, Efficient topological materials discovery using symmetry indicators, *Nature Physics* **15**, 470 (2019).
- [32] D. Wang, F. Tang, J. Ji, W. Zhang, A. Vishwanath, H. C. Po, and X. Wan, Two-dimensional topological materials discovery by symmetry-indicator method, *Physical Review B* **100**, 195108 (2019).
- [33] Y. Xu, L. Elcoro, Z.-D. Song, B. J. Wieder, M. Vergniory, N. Regnault, Y. Chen, C. Felser, and B. A. Bernevig, High-throughput calculations of magnetic topological materials, *Nature* **586**, 702 (2020).
- [34] M. G. Vergniory, B. J. Wieder, L. Elcoro, S. S. Parkin, C. Felser, B. A. Bernevig, and N. Regnault, All topological bands of all stoichiometric materials, *arXiv preprint arXiv:2105.09954* (2021).
- [35] S. Suga and A. Sekiyama, *Photoelectron Spectroscopy: Bulk and Surface Electronic Structures*, Vol. 176 (Springer, 2013).
- [36] B. Lv, T. Qian, and H. Ding, Angle-resolved photoemission spectroscopy and its application to topological materials, *Nature Reviews Physics* **1**, 609 (2019).
- [37] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments, *Nature* **533**, 73 (2016).
- [38] Y. Liu, T. Zhao, W. Ju, and S. Shi, Materials discovery and design using machine learning, *Journal of Materiomics* **3**, 159 (2017).
- [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, and A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, *ACS Central Science* **4**, 268 (2018).
- [40] H. Zhang, K. Hippalgaonkar, T. Buonassisi, O. M. Løvvik, E. Sagvolden, and D. Ding, Machine learning for novel thermal-materials discovery: early successes, opportunities, and challenges, *arXiv preprint arXiv:1901.05801* (2019).
- [41] P. Mikulskis, M. R. Alexander, and D. A. Winkler, Toward interpretable machine learning models for materials discovery, *Advanced Intelligent Systems* **1**, 1900045 (2019).
- [42] Y. Juan, Y. Dai, Y. Yang, and J. Zhang, Accelerating materials discovery using machine learning, *Journal of Materials Science & Technology* (2020).
- [43] A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, *et al.*, On-the-fly closed-loop materials discovery via bayesian active learning, *Nature Communications* **11**, 1 (2020).
- [44] A. Mannodi-Kanakkithodi and M. K. Chan, Computational data-driven materials discovery, *Trends in Chemistry* (2021).
- [45] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Accelerating materials property predictions using machine learning, *Scientific reports* **3**, 1 (2013).
- [46] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Computational Materials* **2**, 16028 (2016).
- [47] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling, *Physical Review X* **4**, 011019 (2014).
- [48] N. Claussen, B. A. Bernevig, and N. Regnault, Detection of topological materials with machine learning, *arXiv preprint arXiv:1910.10161* (2019).
- [49] J. F. Rodríguez-Nieva and M. S. Scheurer, Identifying topological order through unsupervised machine learning, *Nature Physics* **15**, 790 (2019).
- [50] Y. Zhang and E.-A. Kim, Quantum loop topography for machine learning, *Phys. Rev. Lett.* **118**, 216401 (2017).
- [51] 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, and L. Duan, Machine learning topological phases with a solid-state quantum simulator, *Phys. Rev. Lett.* **122**, 210503 (2019).
- [52] M. S. Scheurer and R.-J. Slager, Unsupervised machine learning and band topology, *arXiv preprint arXiv:2001.01711* (2020).
- [53] P. Zhang, H. Shen, and H. Zhai, Machine learning topological invariants with neural networks, *Physical review letters* **120**, 066401 (2018).
- [54] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**,

- 045002 (2019).
- [55] M. R. Carbone, S. Yoo, M. Topsakal, and D. Lu, Classification of local chemical environments from X-ray absorption spectra using supervised machine learning, *Phys. Rev. Materials* **3**, 033604 (2019).
- [56] A. Cui, K. Jiang, M. Jiang, L. Shang, L. Zhu, Z. Hu, G. Xu, and J. Chu, Decoding phases of matter by machine-learning Raman spectroscopy, *Phys. Rev. Applied* **12**, 054049 (2019).
- [57] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, V. Fatemi, L. Zhou, J. I.-J. Wang, Q. Ma, *et al.*, Deep learning enabled fast optical characterization of two-dimensional materials, arXiv preprint arXiv:1906.11220 (2019).
- [58] A. M. Samarakoon, K. Barros, Y. W. Li, M. Eisenbach, Q. Zhang, F. Ye, Z. Dun, H. Zhou, S. A. Grigera, C. D. Batista, *et al.*, Machine learning assisted insight to spin ice dy2ti2o7, arXiv preprint arXiv:1906.11275 (2019).
- [59] Y. Zhang, A. Mesaros, K. Fujita, S. Edkins, M. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. S. Davis, E. Khatami, *et al.*, Machine learning in electronic-quantum-matter imaging experiments, *Nature* **570**, 484 (2019).
- [60] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, and C. Weitenberg, Identifying quantum phase transitions using artificial neural networks on experimental data, *Nature Physics* **15**, 917 (2019).
- [61] M. Newville, Fundamentals of xafs, consortium for advanced radiation sources, University of Chicago, Chicago, IL (2004).
- [62] A. Gaur and B. Shrivastava, Speciation using x-ray absorption fine structure (xafs), *Review Journal of Chemistry* **5**, 361 (2015).
- [63] S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, and L. Hung, Random forest machine learning models for interpretable x-ray absorption near-edge structure spectrum-property relationships, *npj Computational Materials* **6**, 1 (2020).
- [64] C. Zheng, C. Chen, Y. Chen, and S. P. Ong, Random forest models for accurate identification of coordination environments from x-ray absorption near-edge structure, *Patterns* **1**, 100013 (2020).
- [65] S. Kiyohara, T. Miyata, K. Tsuda, and T. Mizoguchi, Data-driven approach for the prediction and interpretation of core-electron loss spectroscopy, *Scientific reports* **8**, 1 (2018).
- [66] A. Guda, S. Guda, A. Martini, A. Kravtsova, A. Algasov, A. Bugaev, S. Kubrin, L. Guda, P. Šot, J. van Bokhoven, *et al.*, Understanding x-ray absorption spectra by means of descriptors and machine learning algorithms, *npj Computational Materials* **7**, 1 (2021).
- [67] Y. Suzuki, H. Hino, M. Kotsugi, and K. Ono, Automated estimation of materials parameter from x-ray absorption and electron energy-loss spectra with similarity measures, *Npj Computational Materials* **5**, 1 (2019).
- [68] M. R. Carbone, M. Topsakal, D. Lu, and S. Yoo, Machine-learning x-ray absorption spectra to quantitative accuracy, *Physical review letters* **124**, 156401 (2020).
- [69] C. D. Rankine, M. M. Madkhali, and T. J. Penfold, A deep neural network for the rapid prediction of x-ray absorption spectra, *The Journal of Physical Chemistry A* **124**, 4263 (2020).
- [70] J. Lueder, A machine learning approach to predict l-edge x-ray absorption spectra of light transition metal ion compounds, arXiv preprint arXiv:2107.13149 (2021).
- [71] 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, *Scientific data* **5**, 180151 (2018).
- [72] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, The Materials Project: A materials genome approach to accelerating materials innovation, *APL Materials* **1**, 011002 (2013).
- [73] 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 Computational Materials* **4**, 12 (2018), <https://doi.org/10.1038/s41524-018-0067-x>.
- [74] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Computational Materials Science* **68**, 314 (2013).
- [75] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, *Computational Materials Science* **97**, 209 (2015).
- [76] J. J. Rehr, J. J. Kas, F. D. Vila, M. P. Prange, and K. Jorissen, Parameter-free calculations of X-ray spectra with FEFF9, *Physical Chemistry Chemical Physics* **12**, 5503 (2010).
- [77] G. Bergerhoff and I. Brown, Crystallographic databases. FH Allen *et al.* (Hrsg.) Chester, International Union of Crystallography (1987).
- [78] J. E. Penner-Hahn *et al.*, X-ray absorption spectroscopy, *Comprehensive Coordination Chemistry II* **2**, 159 (2003).
- [79] Q. Xu, Y. Zhang, K. Koepf, W. Shi, J. van den Brink, C. Felser, and Y. Sun, Comprehensive scan for nonmagnetic weyl semimetals with nonlinear optical response, *npj Computational Materials* **6**, 1 (2020).
- [80] X. Meng and G. E. Karniadakis, A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems, *Journal of Computational Physics* **401**, 109020 (2020).
- [81] A. Agarwala and V. B. Shenoy, Topological insulators in amorphous systems, *Phys. Rev. Lett.* **118**, 236402 (2017).
- [82] E. Prodan, Disordered topological insulators: a non-commutative geometry perspective, *Journal of Physics A: Mathematical and Theoretical* **44**, 113001 (2011).
- [83] 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, and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011).
- [84] 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, and S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–

8035.

- [85] M. Fey and J. E. Lenssen, Fast graph representation learning with PyTorch Geometric, in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

Supplemental material for “Machine learning spectral indicators of topology”

Nina Andrejevic,^{1,2,3,*} Jovana Andrejevic,^{4,*} B. Andrei Bernevig,^{5,6,7}

Nicolas Regnault,⁵ Chris H. Rycroft,^{4,8,†} and Mingda Li^{1,9,‡}

¹*Quantum Measurement Group, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

²*Department of Materials Science and Engineering,*

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

³*Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, USA*

⁴*John A. Paulson School of Engineering and Applied Sciences,*

Harvard University, Cambridge, MA 02138, USA

⁵*Department of Physics, Princeton University, Princeton, NJ 08544, USA*

⁶*Donostia International Physics Center, P. Manuel de Lardizabal 4, 20018 Donostia-San Sebastian, Spain*

⁷*IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*

⁸*Computational Research Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720, USA*

⁹*Department of Nuclear Science and Engineering,*

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

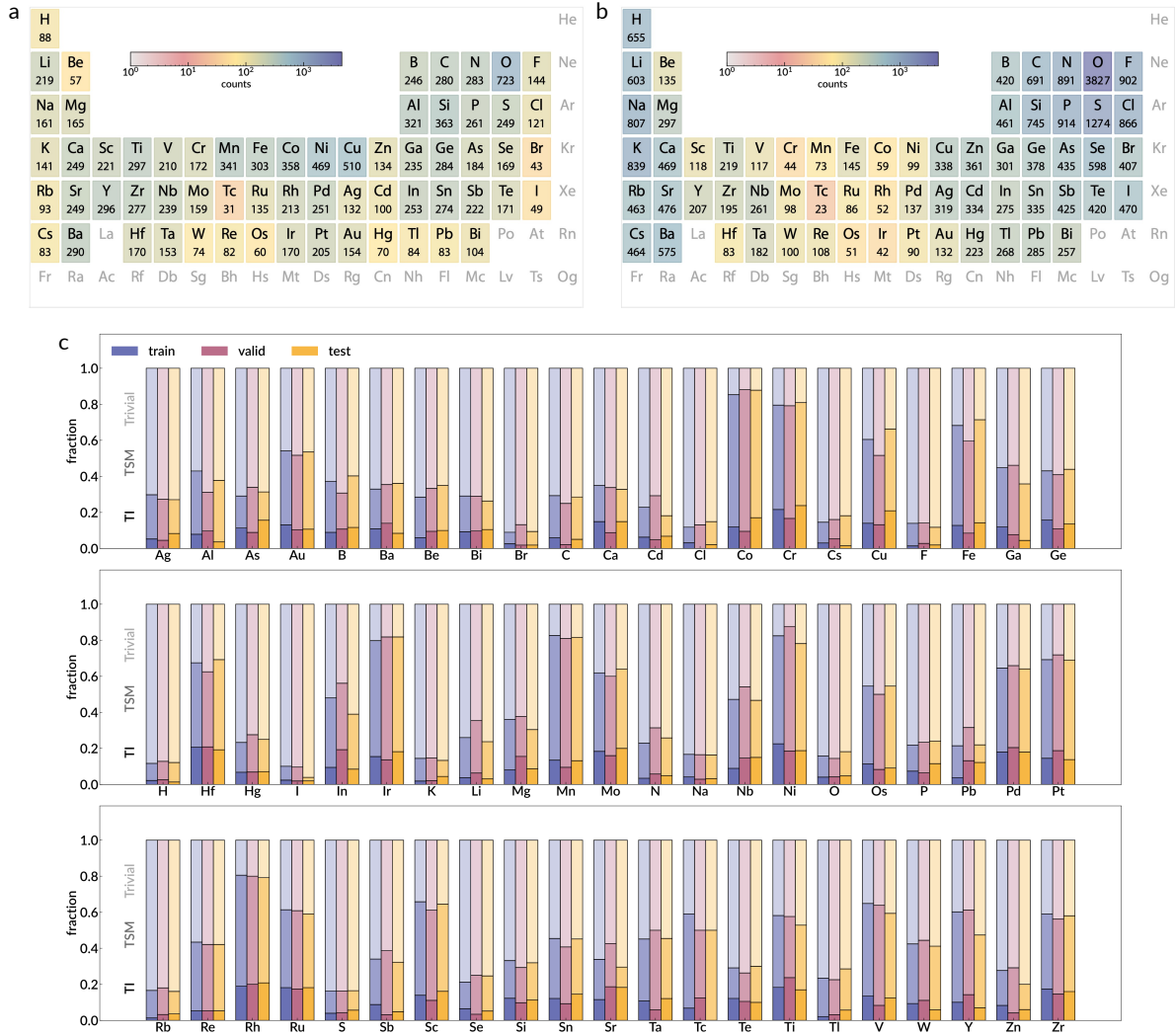


FIG. S1. Element and topological class frequencies in the dataset. (a) The total number of topological samples across training, validation, and testing data containing each element. Each element's entry includes its atomic number, atomic symbol, and number of samples, and is colored by the number of samples. (b) The total number of trivial samples by element. Elements with no counts listed were not present in the dataset. (c) The fraction of topological and trivial samples, by element, in the training, validation, and testing sets. The data subdivision reflects a balanced representation of absorbing elements and topological class across the datasets. TI and TSM denote topological insulators and topological semimetals, respectively.

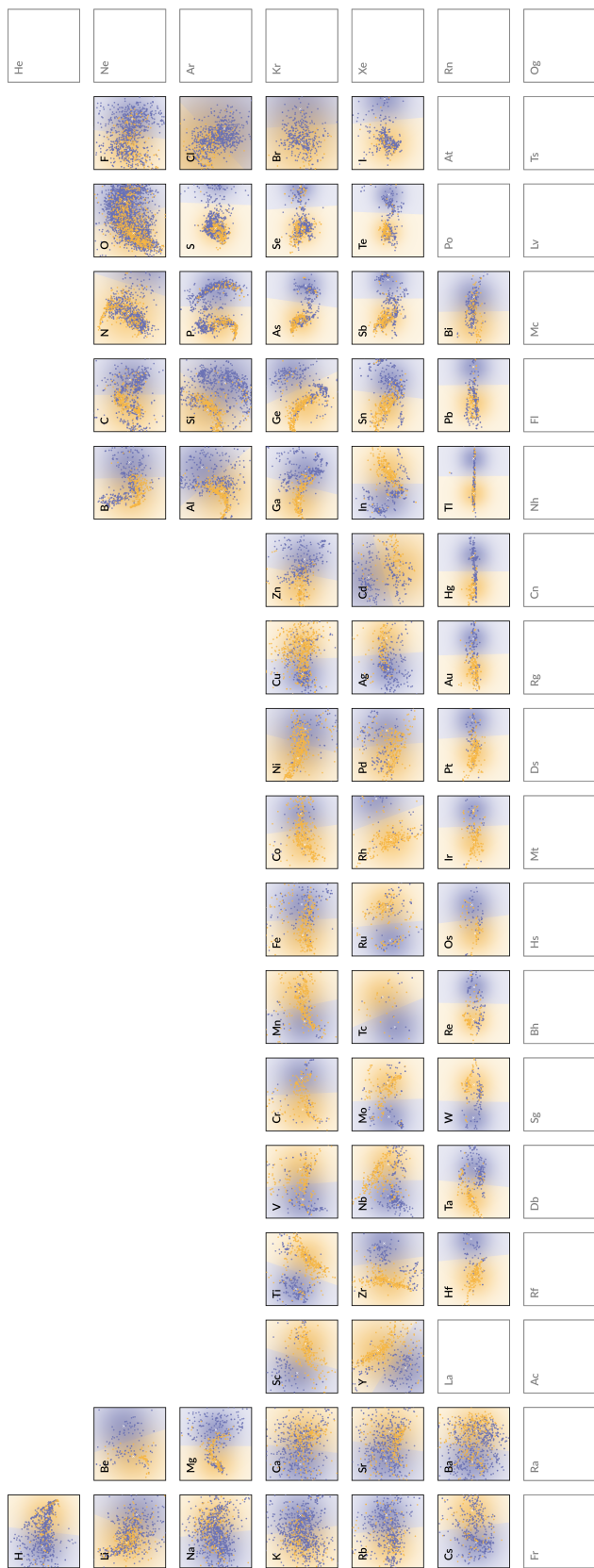


FIG. S2. **Complete results of PCA and k -means clustering.** Decision boundary visualizations of classifications by unsupervised k -means clustering for all elements. The clusters are visualized along the first (x -axis) and second (y -axis) principal components in the scatter plots. Scattered points are colored according to their true class: topological (orange) or trivial (blue). The background is shaded according to the cluster-assigned class.



FIG. S3. **Average XANES K-edge spectra by element and topological class.** Visualization of the average XANES spectra for each absorbing atom present in the dataset used for this work, separated by topological (orange) and trivial (blue) class. Shaded regions correspond to one standard deviation.

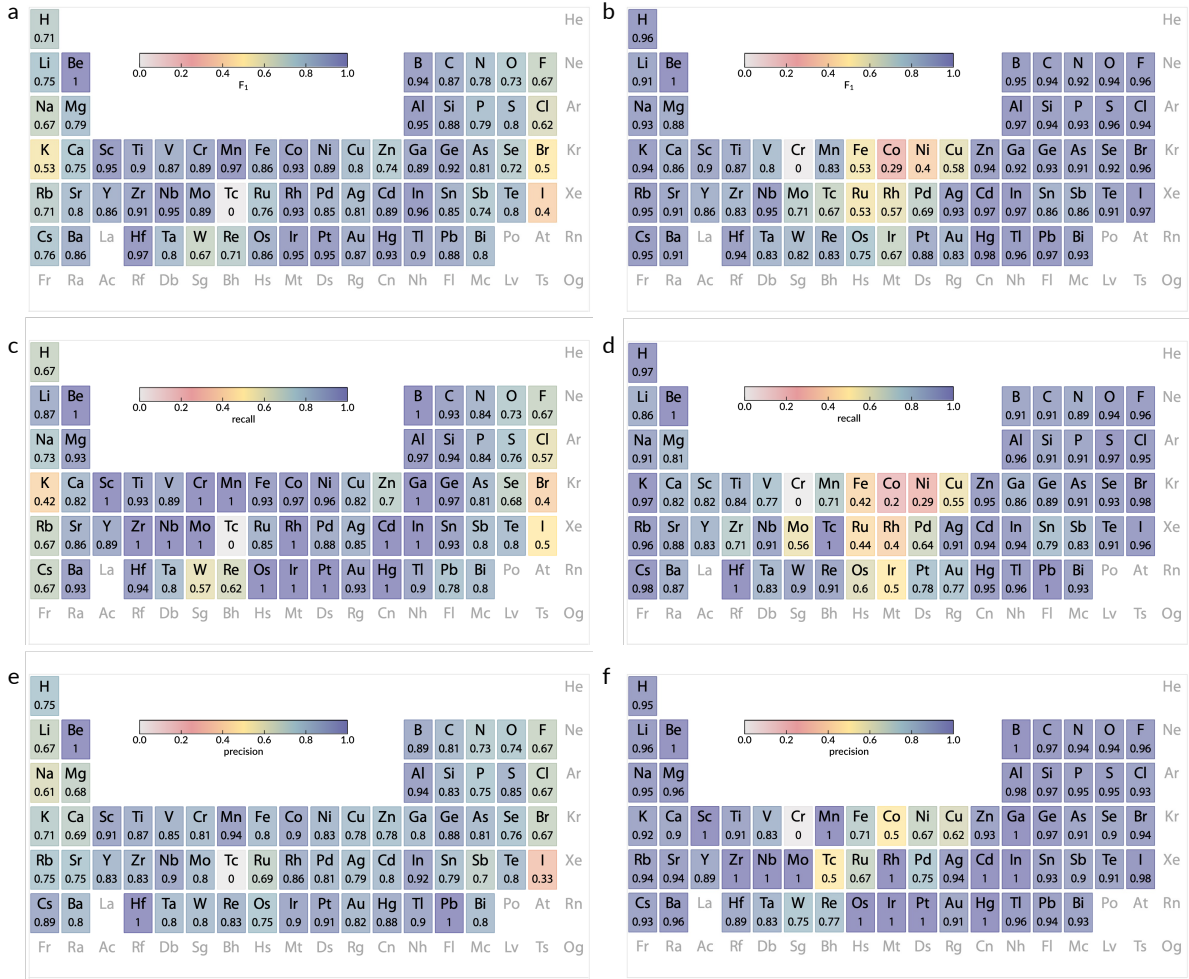


FIG. S4. **Support vector machine (SVM) model performance.** Element specific (a-b) F_1 scores, (c-d) recall, and (e-f) precision for topological (left column) and trivial (right column) examples, respectively. Elements with no score listed were not present in the dataset.

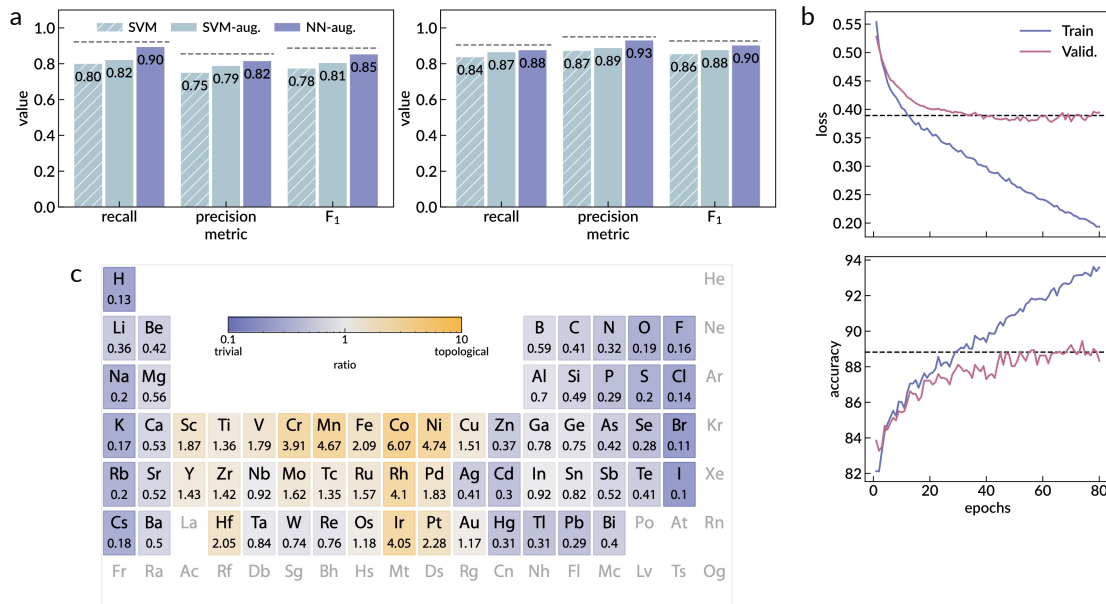


FIG. S5. **Additional model performance metrics.** (a) Comparative plots of the overall recall, precision, and F₁ scores for (left) topological and (right) trivial examples obtained using only atom-type inputs. As noted in the main text, augmented inputs consisting of 200 copies of the one-hot encoded atom types were passed to the SVM to maintain the same number of neurons between SVM-type and SVM models. In the barplots shown, SVM and SVM-aug. refer to SVM models operating on the original and augmented atom-type inputs, respectively. NN-aug. refers to the neural network model without the spectral embedding layers; instead, the direct product is performed between the atom-type embedding and an array of ones equal in length to the spectral embedding vector. Dashed gray lines indicate the scores of the full NN model reported in the main text. (b) Representative training history of the full NN model, indicating the loss and accuracy at the early stopping point determined using the validation set (black dashed lines). (c) Ratios of topological to trivial examples present in the dataset for each absorbing atom, highlighting the overrepresentation of topological examples containing certain transition metals.

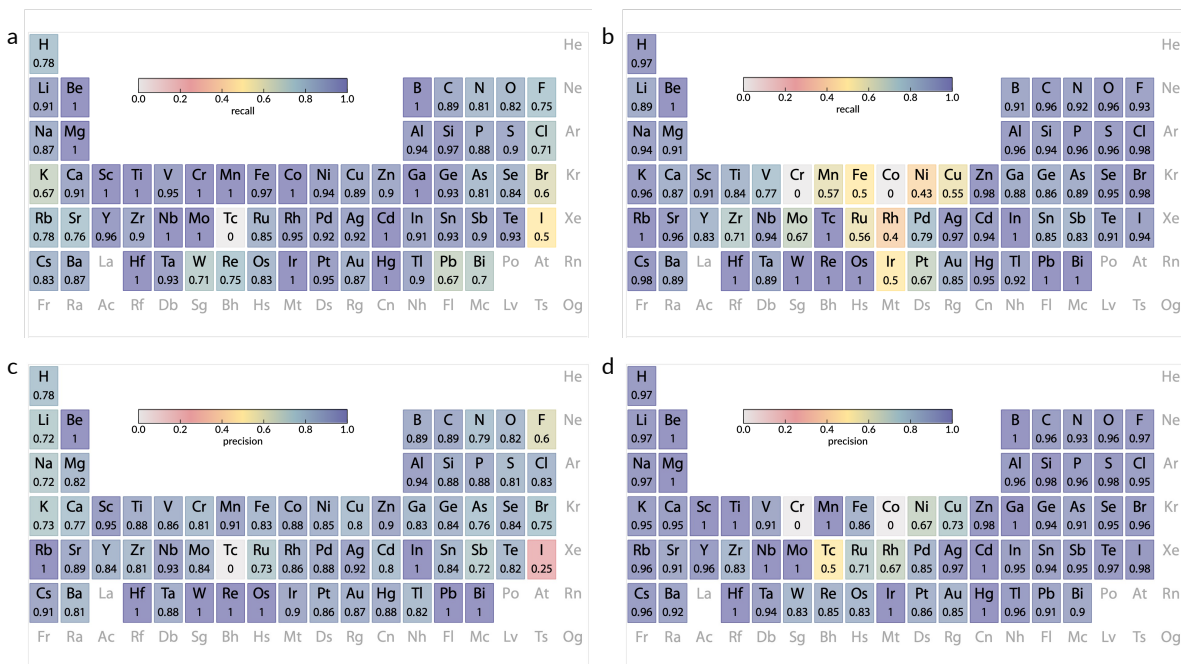


FIG. S6. **Neural network (NN) classifier recall and precision.** Element specific (a-b) recall and (c-d) precision for topological (left column) and trivial (right column) examples, respectively. Elements with no score listed were not present in the dataset.