A Machine Learning Approach for Understanding and Discovering Topological Materials

by

Andrew Ma

A.B. Physics
Princeton University (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

© Massachusetts Institute of Technology 2021. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
August 27, 2021

Certified by .....
Marin Soljačić
Professor of Physics
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

A Machine Learning Approach for Understanding and Discovering Topological Materials

by

Andrew Ma

Submitted to the Department of Electrical Engineering and Computer Science
on August 27, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Topological materials are of significant interest for both basic science and next-generation technological applications due to their unconventional electronic properties. The majority of currently-known topological materials have been discovered using methods that involve symmetry-based analysis of the quantum mechanical wavefunction. Here we use machine learning to develop a heuristic chemical rule, which diagnoses whether a material is topological using only its chemical formula. It is based on a notion that we term topogivity, which is a learned numerical value for each element that loosely captures the tendency of an element to form topological materials. Topogivities provide chemical insights for understanding topological materials. We implement a high-throughput procedure for discovering topological materials that are not diagnosable by symmetry indicators. The procedure is based on heuristic rule prediction followed by ab initio validation. The concept of topogivity represents a fundamentally new approach to the study of topological materials, and opens up new directions of research at the intersection of chemistry, machine learning, and band topology.

Thesis Supervisor: Marin Soljačić
Title: Professor of Physics

Acknowledgments

First and foremost, I would like to thank my advisor, Marin Soljačić, for his invaluable help and guidance throughout my PhD. I would like to thank Yang Zhang, Thomas Christensen, Hoi Chun Po, Li Jing, Liang Fu, and Marin Soljačić for collaboration on this work.

I would like to thank Pawan Goyal for assisting preliminary work, and thank Samuel Kim, Peter Lu, Rumen Dangovski, and Edward Zhang for helpful discussions about this work.

I would like to acknowledge the National Science Foundation Graduate Research Fellowship (Grant No. 1745302) and the MIT EECS Alan L. McWhorter Fellowship for helping fund my graduate studies.

Research was sponsored in part by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This material is also based upon work supported in part by the Air Force Office of Scientific Research under the award number FA9550-20-1-0115, as well as in part by the US Office of Naval Research (ONR) Multidisciplinary University Research Initiative (MURI) grant N00014-20-1-2325 on Robust Photonic Materials with High-Order Topological Protection. This material is also based upon work supported in part by the U. S. Army Research Office through the Institute for Soldier Nanotechnologies at MIT, under Collaborative Agreement Number W911NF-18-2-0048. This work is also supported in part by the the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, `http://iaifi.org/`).

The work described in this thesis forms a portion of the following manuscript:

Andrew Ma*, Yang Zhang*, Thomas Christensen, Hoi Chun Po, Li Jing, Liang Fu, and Marin Soljačić. Topogivity: A machine-learned chemical rule for discovering topological materials. *In preparation*, 2021.

* Equal contributions

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Topological materials are an exotic class of materials that have been of major interest in recent years. Their unconventional properties – such as the protected conducting surfaces of topological insulators – make them important for both fundamental science and next-generation technological applications [1, 2]. There are a wide variety of topological materials, including strong and weak topological insulators [3], topological crystalline insulators [4, 5], higher-order topological insulators [6], Weyl semimetals [7, 8], Dirac semimetals [9], nodal-line semimetals [10], and multifold fermion semimetals [11].

Since the genesis of the field, a central and enduring question has been how to determine whether a given electronic material is topological. Finding answers to this question is important both for deepening our fundamental understanding of topological materials, and also in the search for ideal candidates for experimental realization and realisitic device applications.

Efforts to answer this question have predominately taken a first-principles, quantum-mechanical approach [12]. In particular, a recurring theme has been to exploit symmetry to simplify the diagnosis problem, as exemplified by the Fu-Kane criterion for inversion-symmetric materials [13]. Recently-developed theories known as symmetry indicators [14] and topological quantum chemistry [15] enable the diagnosis of a wide range of topological materials using symmetry-based analysis of the wavefunction. These symmetry-based approaches are broadly-applicable and require rel-

atively little computational cost [14, 15, 16], and they have enabled high-throughput computational searches for topological materials [17, 18, 19, 20, 21]. Despite these successes, symmetry-based approaches provide only limited intuition regarding why a given material is topological. Moreover, some topological materials are not diagnosable by symmetry indicators [14] (we will refer to such materials as non-symmetry-diagnosable topological materials). For example, the first experimentally-observed Weyl semimetal, tantalum arsenide (TaAs) [7], is a non-symmetry-diagnosable topological material [18].

Many physical phenomena can be understood both at a quantum-mechanical level and at a chemical heuristic level. A well-known example is bonding, which can be understood using quantum-mechanical approaches such as molecular orbital theory [22], as well as using heuristics such as the difference of element electronegativities. While the former approach is more detailed and accurate, the latter gives an immediate answer and provides valuable and easily transferable intuition. An intriguing question is whether topological materials could also be understood using a chemical heuristic approach. Already, there has been research on connections between chemistry and electronic band topology [23], e.g. finding evidence that certain elements are more likely to form topological materials than others [24, 25] and using chemical intuition and heuristics to help find topological materials [26, 27]. However, these approaches are not as broadly applicable as wavefunction-based approaches.

In the field of materials science, one topic of growing interest is the use of machine learning [28], which can reduce computational costs [29, 30], yield interpretable models [31, 32], guide the discovery of new materials [33, 34, 35], and enable the discovery of novel scientific understandings and insights [36, 37, 38]. At the intersection of chemistry and machine learning, there have been recent applications in developing chemical heuristics for many problems in materials science [39]. In the area of topological materials, one direction has been in the study of toy models of electronic band topology [40, 41, 42, 43]. Another direction – enabled by ab initio data – has been the study of topology in real electronic materials [44, 45, 46, 47, 48, 49]. Outside of the area of topological materials, there have also been a number of works on the use

of machine learning for topology in other physical systems [50, 51, 52, 53, 54, 55, 56].

In our work, we (i) use machine learning to develop a broadly-applicable heuristic chemical rule that diagnoses whether or not a material is topological, and (ii) apply this rule to predict topological materials that are subsequently validated using ab initio calculations. This SM thesis will focus primarily on the machine learning portion of this collaboration. Key parts of the ab initio calculations (which were performed by Yang Zhang) will also be described. Full results will be presented in our manuscript in preparation [57].

The heuristic rule is based on the notion of a learned parameter for each element that loosely captures the tendency of an element to form topological materials, which we term an element's *topogivity*. The heuristic rule is simple, hand-calculable, and interpretable: a given material is diagnosed as topologically nontrivial (trivial) if the weighted average of its elements' topogivities is positive (negative). A key highlight is that the heuristic rule can generalize beyond the scope of existing symmetry-based diagnosis methods [14, 15]. This is important because current first-principles approaches for non-symmetry-diagnosable topological materials (e.g., using Wilson loops [12]) typically involve significant computational cost. We integrate the heuristic rule into a high-throughput procedure for discovering non-symmetry-diagnosable topological materials, which involves screening using the heuristic rule followed by density functional theory (DFT) validation.

Chapter 2

Classes of Materials and Dataset

2.1 Modeling Goals

Within the broad umbrella of research on determining whether a real material is topological, much of the previous literature has focused on issues of diagnosing specific kinds of topology. For example, the Fu-Kane criterion for inversion-symmetric materials is a method for diagnosing strong and weak topological insulators [13]. Similarly, symmetry indicators often indicate a material's topological invariant modulo an integer (e.g., Chern numbers via rotation eigenvalues) [58]. In contrast, in this work we are interested in a coarser-grained distinction between topological materials (i.e., topologically nontrivial materials) and trivial materials (i.e., topologically trivial materials). Topological materials include both topological insulators (Fig. 2-1, left) and topological semimetals (Fig. 2-1, right). Note that in this work, our criterion for what constitutes a topological semimetal will focus more on energetics (specifically, we will require that the protected band-degeneracies be near or at the Fermi level).

We aim to develop a machine learning model – which takes the form of a heuristic chemical rule – that can distinguish between topological materials and trivial materials. For our supervised learning approach to fitting this model, we require one set of materials labeled as "topological" (i.e., positive labels) as well as another set of materials labeled as "trivial" (i.e., negative labels). Additionally, we also need to identify a third set of materials that will be used as materials to screen in the

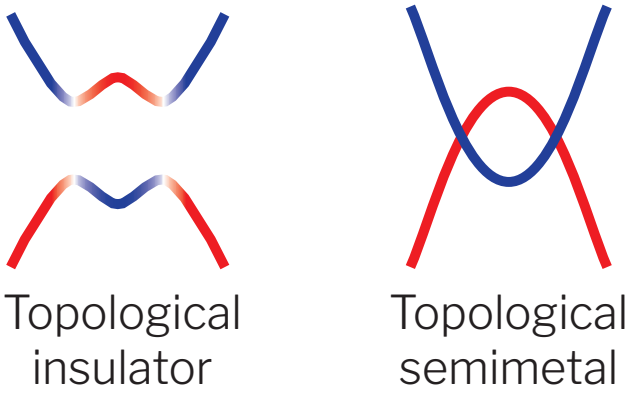

Figure 2-1: **Band topology schematics.** Pictorial visualizations of example band structures of topological insulators and topological semimetals, emphasizing the role of band-inversions.

high-throughput materials discovery process. We identify materials for each of these three sets by choosing suitable subsets of an ab initio dataset that was generated using symmetry indicators by [18] (note that one could also use one of the other similar datasets [19, 20, 21] for this purpose). The dataset consists of stoichiometric, non-magnetic, three-dimensional materials treated with spin-orbital coupling (our modeling consequently applies to this setting). We do basic pre-processing of the dataset (e.g., to remove materials that contain rare elements).

2.2 Symmetry Indicators Dataset

The symmetry indicators method uses symmetry information to try to determine whether a material can be distinguished from an atomic insulator – the diagnoses are obtained by making use of DFT that is performed at high-symmetry points [14]. We will make use of the fact that the fine-grained classifications produced by symmetry indicators can be coarse-grained into two categories: Undiagnosable by Symmetry Indicators (USI) and Not an Atomic Insulator (NAI). In the terminology of [17] and [18], USI corresponds to "case 1" and NAI corresponds to "case 2" and "case 3". Many USI materials are trivial and many NAI materials are topological. However, a USI mate-

rial is not guaranteed to be trivial since it could also be a non-symmetry-diagnosable topological material, and an NAI material is not guaranteed to be topological (under our criterion) since it could also be a "trivial" metal (given that the symmetry indicator diagnosis ignores energetic aspects of the electronic bands). For example, the Weyl semimetal TaAs [7] is categorized by symmetry indicators as USI [18].

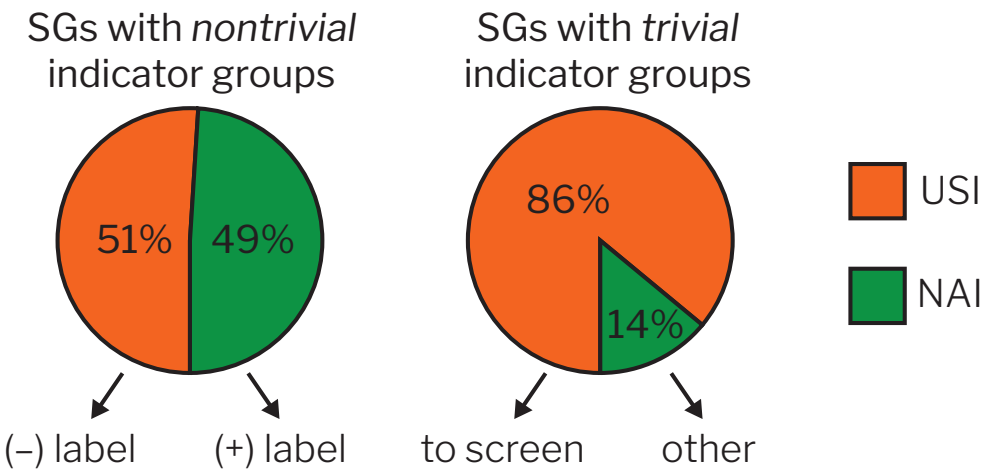

Figure 2-2: **Materials dataset.** The symmetry-indicator-generated ab initio dataset is partitioned into two sets of materials based on SG. Materials that have an SG with a nontrivial indicator group are in the first set and the remaining materials are in the second set. The two sets have substantially different ratios of NAI (Not an Atomic Insulator) to USI (Undiagnosable by Symmetry Indicators). The first set is used as labeled data for learning the values of the topogivities, and the USI portion of the second set are the materials to screen in the high-throughput materials discovery process.

To facilitate our goal of selecting suitable subsets, we partition the dataset's materials into two sets based on space group (SG) (see Fig. 2-2). All materials that have an SG with a nontrivial indicator group [14] are placed in the first set, and the rest are placed in the second set. Symmetry indicators are more successful in searching for topological materials in the first set in the sense that: (i) 49% of the materials in the first set are NAI compared to only 14% in the second set, and (ii) symmetry indicators can diagnose both gapped and gapless NAI materials in the first set but can only diagnosis gapless NAI materials in the second set [14]. Motivated by this, we use the first set's NAI materials as the "topological" labeled set and the first set's USI materials as the "trivial" labeled set; collectively, we refer to these two sets as

the labeled data. From a machine learning perspective, this defines a dataset with noisy labels. One source of noise is the fact that the distinction between NAI and USI is mathematically not the same as the distinction between topological and trivial. There are also other sources of noise, such as inaccuracies in the DFT itself [59]. We use the second set's USI materials as the materials to screen in the high-throughput materials discovery process. A useful consequence of this choice is that any material from the materials to screen that turns out to be topological must be a non-symmetry-diagnosable topological material as long as the symmetry indicator based calculation in the original dataset [18] was correct. The labeled data consists of 9,026 materials and the materials to screen consists of 1,433 materials.

2.3 Frequencies

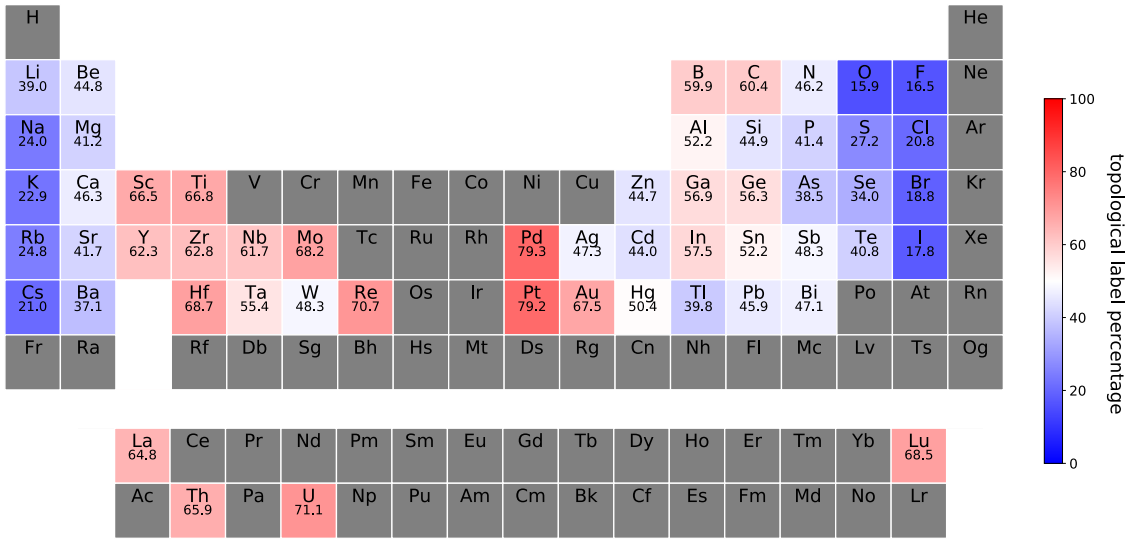

Figure 2-3: **Topological label percentage for each element.** Percentages are shown by color-coding and in values. These are calculated for the labeled data. It represents the percentage of materials with the "topological" label among the materials that contain a given element.

In Fig. 2-3, we show the topological label percentage for each element. For each element, this represents the percentage of materials that have the "topological label" among all of the labeled data materials that contain that element. From this figure,

we can see that the frequency of topological materials varies substantially across different elements.

Chapter 3

Machine Learning Model

3.1 Formulation

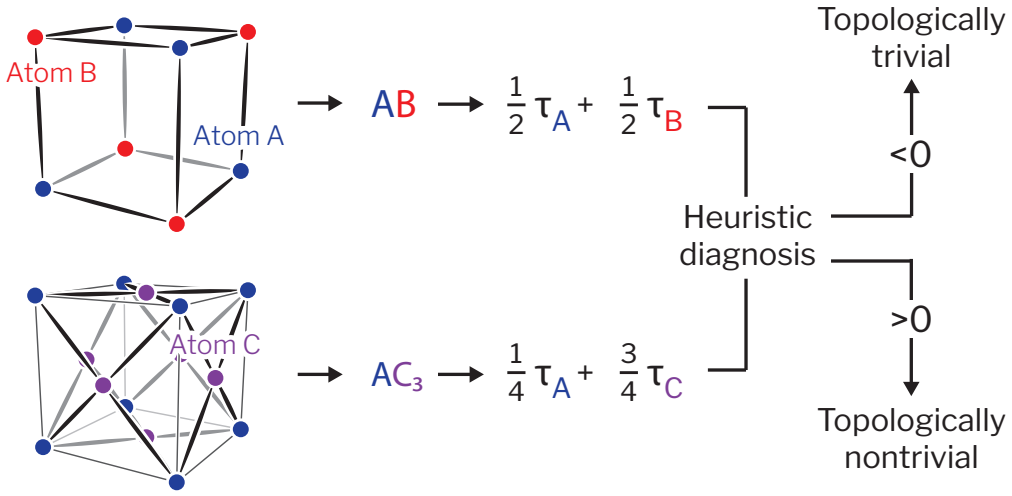

Figure 3-1: **Topogivity-based diagnosis of materials.** Given a stoichiometric material, the topogivity-based heuristic diagnosis is evaluated by simply weighting the material's elements' (atoms A, B, C) topogivities (τ_A, τ_B, τ_C) by their relative abundance in the unit cell, or equivalently, in the chemical formula (AB and AC_3). The sign indicates the topological classification and the magnitude indicates roughly how confident we are in this classification. Each element's topogivity is a machine-learned parameter, which loosely captures the element's tendency to form topological materials.

We represent the heuristic chemical rule as a parameterized machine learning model (Fig. 3-1). More specifically, the model maps each material M to a real number

$g(M)$ according to the function

$$g(M) = \sum_E f_E(M)\tau_E, \tag{3.1}$$

where the summation runs over the elements present in the chemical formula of material M, τ_E is a learned parameter for each element E, and $f_E(M)$ is the element fraction for the element E in material M (e.g., for a chemical formula $A_xB_yC_z$, $f_A(M) = \frac{x}{x+y+z}$, $f_B(M) = \frac{y}{x+y+z}$, and $f_C(M) = \frac{z}{x+y+z}$). Classification decisions are made according to the sign of $g(M)$: classify as topological if positive and classify as trivial if negative. A greater magnitude of $g(M)$ roughly corresponds to a more confident classification decision. The model is a heuristic *chemical* rule in the sense that all the information required for obtaining a diagnosis is contained in the material's chemical formula. Nevertheless, it is possible that symmetry or other spatial information is used implicitly, e.g. due to relationships between chemical composition and crystal structure [60].

The optimized parameters $\{\tau_E\}$ are obtained by fitting the model as a parameterized binary classifier to the data. Rather than directly optimizing the parameters, we use an approach based on featurizing the materials as vectors of element fractions with one element dropped, learning a linear function, and then mapping the weights to the corresponding topogivities. The linear function is trained using soft-margin linear SVM [61], which we implemented using the Scikit-learn package [62].

For each element E, we refer to the optimized parameter τ_E as its *topogivity*. For a given material M, $g(M)$ is simply the weighted average of its elements' topogivities, where the weighting is with respect to each element's relative abundance, as identifiable from the material's chemical formula. Conceptually, an element's topogivity loosely captures its tendency to form topological materials – greater topogivity roughly corresponds to greater tendency.

3.2 Performance Evaluation

Accuracy (%)	Recall (%)	Precision (%)	F1 Score (%)
82.7 ± 1.0	78.0 ± 1.8	85.6 ± 1.9	81.6 ± 1.1

Table 3.1: **Results from nested cross validation**. Results are shown in the format mean ± standard deviation. Note that the "topological" label corresponds to the positive class.

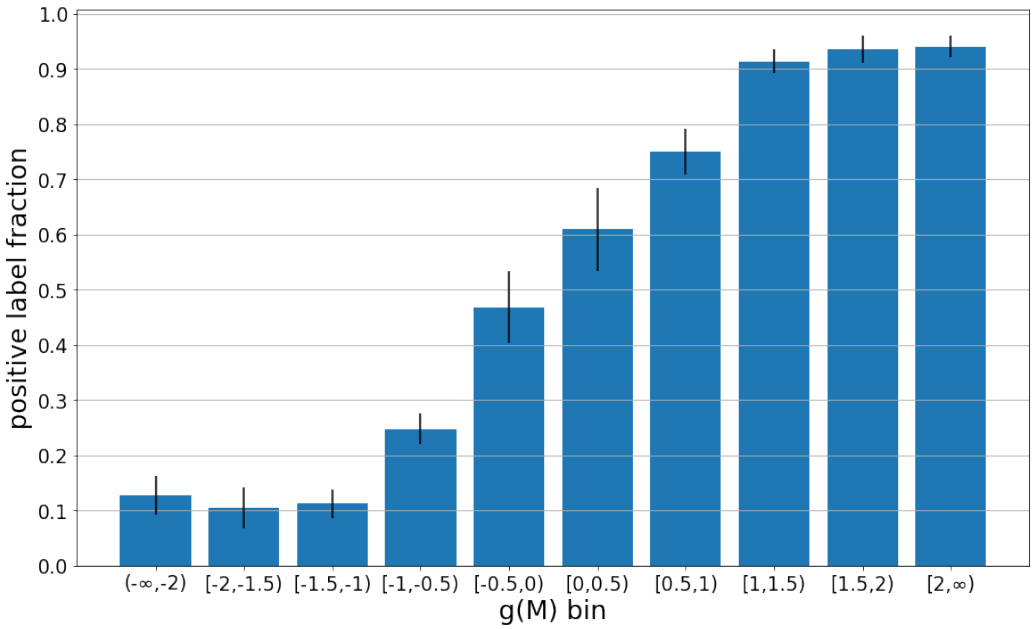

Figure 3-2: **Positive label fraction vs. $g(M)$ bin.** For each bin of $g(M)$ values, we calculate the fraction of test set materials that have a positive label (i.e., topological label). Shown are nested cross validation means, with error bars indicating plus/minus one standard deviation.

We use a nested cross validation procedure, in which the inner loop is used for hyperparameter selection and the outer loop is used for performance evaluation. The outer loop is 11-fold, and the inner loop is 10-fold. The dataset splitting is stratified. We find a test accuracy of 82.7±1.0% (mean ± standard deviation). Additional results are shown in Table 3.1. Additionally, we find empirical evidence that as the magnitude of $g(M)$ is increased, the fraction of correctly-classified materials first increases and

then plateaus, with the plateau beginning around $|g(M)| \approx 1$, as shown in Fig. 3-2. We heuristically set a threshold of 1.0 for a high-confidence positive classification and observe that that $93.0 \pm 1.2\%$ of materials with $g(M) \geq 1.0$ are correctly classified. Choosing a threshold in this way means we are prioritizing precision over recall – this choice is motivated by the fact that the DFT computations will be computationally expensive and so we want our screening success rate to be reasonably high, but other choices of threshold may also be reasonable depending on the purpose. Finally, we train on all of the materials in the labeled data to obtain the final model, which is what we will use for our high-throughput discovery process.

3.3 Learned Topogivities and Analysis

We visualize the final model's learned topogivities in Fig. 3-3. This periodic table of topogivities enables an immediate heuristic diagnosis of any stoichiometric material whose elements are featured in the table. This is illustrated with examples in Fig. 3-3 for the trivial insulator NaCl and the Dirac semimetal Na_3Bi [9]. The Weyl semimetal TaAs [7] is also worth highlighting: TaAs is non-symmetry-diagnosable [18] and does not appear in the labeled data, but is successfully diagnosed as topological by the topogivity approach: $g(\text{TaAs}) = \frac{1}{2}\tau_{\text{Ta}} + \frac{1}{2}\tau_{\text{As}} = 1.450$.

The simplicity of our model enables us to readily extract chemical insights from the periodic table of topogivities. First, we observe that elements that are near each other in the periodic table tend to have similar topogivities, which is consistent with intuition. Second, we observe that the elements with negative topogivities (i.e., that exhibit a tendency to form trivial materials) are located in two clusters respectively in the top right and bottom left parts of the periodic table. This is also consistent with intuition: elements in the top right and bottom left are respectively highly electronegative and highly electropositive, and thus tend to form highly ionic compounds – such compounds tend to have large trivial band gaps. Third, considering column 15 of the periodic table (the pnictogens), we observe that while the topogivities of Ni, P, and As are negative (and Sb is slightly positive), the topogivity of Bi is positive with

a relatively large magnitude. This is consistent with the intuition arising from the experimental fact that Bi is featured in many well-known topological materials [25]. Finally, we observe a region of high topogivities in the early transition metals – future work could attempt to understand the reasons for this (note that there is a chance that this is partially an artifact of typical oxidation states). Overall, while the element topogivities are parameters whose specific learned values are affected by dataset issues and modeling limitations (e.g., the effects of correlated features), the fact that we can extract chemical insights that are consistent with intuition is evidence that the topogivity picture is a meaningful way to study topological materials.

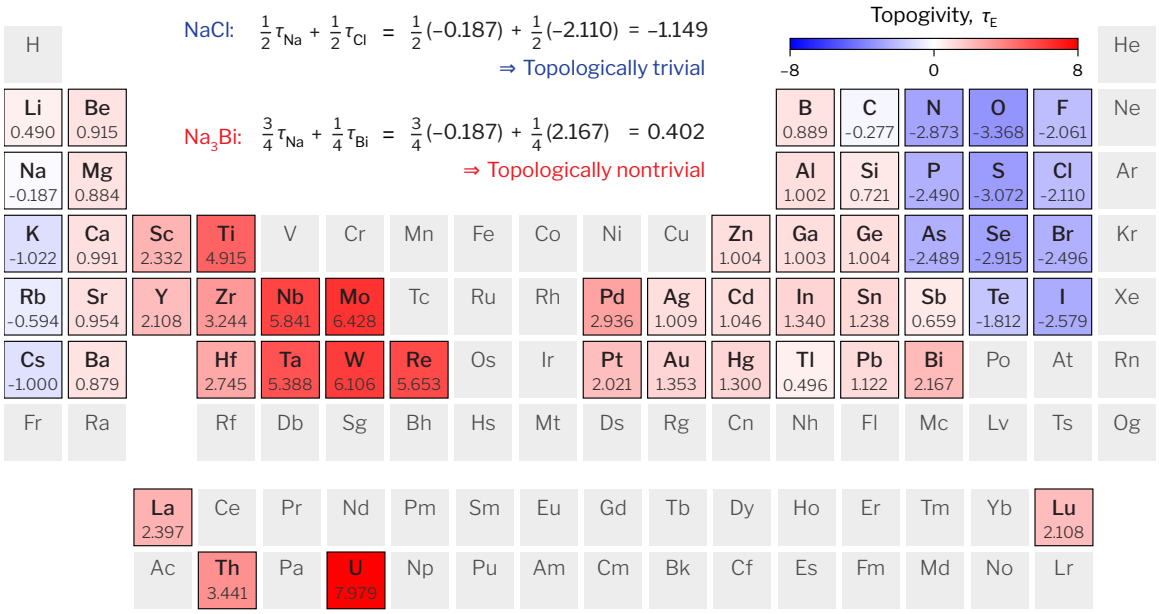

Figure 3-3: **Periodic table of topogivities.** Machine-learned topogivities τ_E are shown by color-coding and in values. Elements that do not appear in any material in the post-processed dataset are shown in gray. Example applications of the topogivity-based heuristic chemical rule is shown for two materials, NaCl (trivial) and Na_3Bi (nontrivial).

Chapter 4

High-Throughput Topological Materials Discovery

Some of the results described in this chapter are preliminary results. The finalized results will be included in [57]. Additionally, since this SM thesis focuses on the machine learning component of the collaboration, we will not elaborate on the DFT methodological details, which will be described in detail in [57].

4.1 Topogivity-Based Screening Approach

To identify topological materials using the learned topogivities, we compute $g(M)$ for each of the 1,433 materials to screen (Fig. 4-1). We restrict our attention to the materials that have a $g(M)$ value corresponding to a high-confidence positive classification (i.e., $g(M) \geq 1.0$), of which there are 73 materials (a repeated structure and a material that does not exist in ICSD were removed). Additionally, since it is difficult to obtain accurate DFT calculations for f electron materials, we exclude materials containing 4f and 5f electrons, which eliminates 5 materials and thus leaves us with 68 materials for ab initio validation. Note that even though SG is not used explicitly by the model, we consider two materials to be distinct if they have different SGs (one consequence is that some of these 68 materials have chemical formulas that also appeared in training the model).

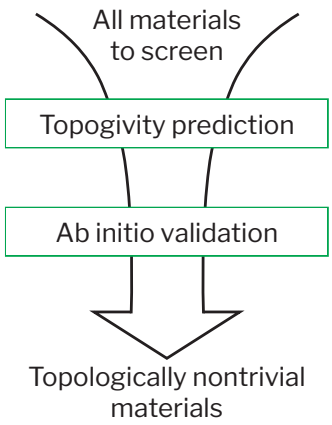

Figure 4-1: **Topogivity-based discovery of topological materials.** We leverage our framework to perform high-throughput topological materials discovery by first screening through a suitable collection of materials using topogivities – identifying candidate topological materials rapidly – and subsequently performing ab initio validation using DFT.

4.2 Ab Initio Calculations

For each of the 68 materials, we perform DFT using the Full-Potential Local-Orbital program (FPLO) [63] within the generalized gradient approximation (GGA) [64]. Our calculations can detect many forms of nontrivial topology, but we note that our DFT does not check for all types of nontrivial topology, so in principle it is possible that some topological materials are missed. Of the 68 materials, we find that 56 are topological, corresponding to a success rate of 82.4%. All 56 of these topological materials are topological semimetals. Some of these 56 topological materials have previously been predicted in the literature and a smaller portion have also already been experimentally observed, e.g., TaAs [7]. Beyond these, our DFT calculations also turn up multiple new topological materials that to our knowledge have not previously been identified.

The band structures of four newly-discovered topological materials that are promising for experimental observation are shown in Fig. 4-2. The band crossings are circled in red – we can see that they are near or at the Fermi level. We emphasize that the reason these four materials are non-symmetry-diagnosable is that their topological nodes are all within the valence band manifold or conduction band manifold. Such

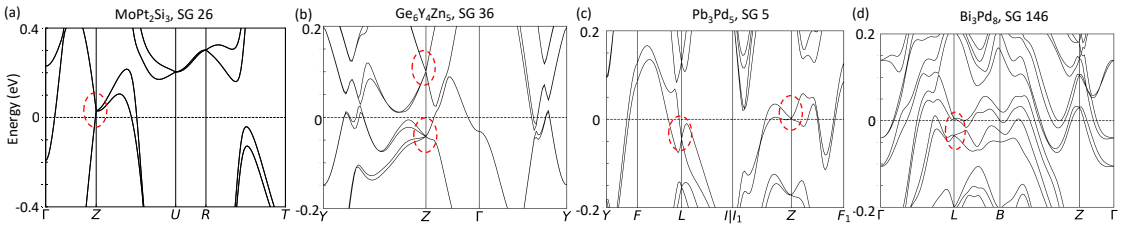

Figure 4-2: **Example of newly-discovered topological materials.** These materials are not diagnosable using symmetry indicators, but were successfully discovered using our topogivity-based approach. The band structures were computed using DFT. All four are topological semimetals.

nodes cannot be diagnosed by the symmetry indicators method, which is formulated based on the electron filling and therefore cannot target band degeneracies that are not *between* the valence and conduction bands [14].

Chapter 5

Discussion and Outlook

The topogivity approach developed here represents a broadly-applicable, systematic approach for diagnosing topological materials. An important feature of the approach is that it is not restricted to the types of topological materials that are diagnosable using symmetry indicators. Strikingly, a topogivity-based diagnosis uses only the chemical formula and requires merely a handful of arithmetic operations to evaluate. Furthermore, the periodic table of topogivities provides simple intuition for a complex, exotic phenomena.

These merits of the topogivity approach are enabled in part by the heuristic nature of the chemical rule and, accordingly, come at the cost of rigorous classification: separate validation of a topogivity-based prediction requires either ab initio calculations – as done here – or experiments. In addition to the diagnoses being heuristic, the topogivities themselves are not unambiguously defined, and depend e.g., on the weighted average formulation as well as the choice of machine learning algorithm. Moreover, the topogivity approach provides only a coarse-grained topological classification – nontrivial or trivial – without the fine-grained detail of ab initio approaches (such as symmetry indicators [14] or Wilson loops [12]).

The machine-learned topogivities enabled us both to extract chemical insights and to discover new topological materials. One promising future direction is to look for a more complete understanding of the underlying reasons for the values of the elements' topogivities, which may in turn shed new light on the fundamental question

of why some materials are topological while others are not. Another promising future direction is to perform more comprehensive searches for new topological materials using topogivity-based materials discovery strategies. More broadly, the notion of topogivity represents a new paradigm in the field, which lays the groundwork for further interdisciplinary research on both understanding and discovering topological materials.

Bibliography

[1] Arun Bansil, Hsin Lin, and Tanmoy Das. Colloquium: Topological band theory. *Rev. Mod. Phys.*, 88(2):021004, 2016.

[2] M Zahid Hasan and Charles L Kane. Colloquium: topological insulators. *Rev. Mod. Phys.*, 82(4):3045, 2010.

[3] Liang Fu, Charles L Kane, and Eugene J Mele. Topological insulators in three dimensions. *Phys. Rev. Lett.*, 98(10):106803, 2007.

[4] Liang Fu. Topological crystalline insulators. *Phys. Rev. Lett.*, 106(10):106802, 2011.

[5] Timothy H Hsieh, Hsin Lin, Junwei Liu, Wenhui Duan, Arun Bansil, and Liang Fu. Topological crystalline insulators in the snte material class. *Nat. Commun.*, 3(1):1–7, 2012.

[6] Wladimir A Benalcazar, B Andrei Bernevig, and Taylor L Hughes. Quantized electric multipole insulators. *Science*, 357(6346):61–66, 2017.

[7] Su-Yang Xu, Ilya Belopolski, Nasser Alidoust, Madhab Neupane, Guang Bian, Chenglong Zhang, Raman Sankar, Guoqing Chang, Zhujun Yuan, Chi-Cheng Lee, et al. Discovery of a weyl fermion semimetal and topological fermi arcs. *Science*, 349(6248):613–617, 2015.

[8] Alexey A Soluyanov, Dominik Gresch, Zhijun Wang, QuanSheng Wu, Matthias Troyer, Xi Dai, and B Andrei Bernevig. Type-ii weyl semimetals. *Nature*, 527(7579):495–498, 2015.

[9] ZK Liu, Bo Zhou, Yong Zhang, ZJ Wang, HM Weng, Dharmalingam Prabhakaran, S-K Mo, ZX Shen, Zhong Fang, Xi Dai, et al. Discovery of a three-dimensional topological dirac semimetal, na3bi. *Science*, 343(6173):864–867, 2014.

[10] AA Burkov, MD Hook, and Leon Balents. Topological nodal semimetals. *Phys. Rev. B*, 84(23):235126, 2011.

[11] Barry Bradlyn, Jennifer Cano, Zhijun Wang, MG Vergniory, C Felser, Robert Joseph Cava, and B Andrei Bernevig. Beyond dirac and weyl fermions:

Unconventional quasiparticles in conventional crystals. *Science*, 353(6299):558–564, 2016.

[12] Jiewen Xiao and Binghai Yan. First-principles calculations for topological quantum materials. *Nat. Rev. Phys.*, pages 1–15, 2021.

[13] Liang Fu and Charles L Kane. Topological insulators with inversion symmetry. *Phys. Rev. B*, 76(4):045302, 2007.

[14] Hoi Chun Po, Ashvin Vishwanath, and Haruki Watanabe. Symmetry-based indicators of band topology in the 230 space groups. *Nat. Commun.*, 8(1):1–9, 2017.

[15] Barry Bradlyn, L Elcoro, Jennifer Cano, MG Vergniory, Zhijun Wang, C Felser, MI Aroyo, and B Andrei Bernevig. Topological quantum chemistry. *Nature*, 547(7663):298–305, 2017.

[16] Jorrit Kruthoff, Jan de Boer, Jasper van Wezel, Charles L. Kane, and Robert-Jan Slager. Topological classification of crystalline insulators through band structure combinatorics. *Phys. Rev. X*, 7:041069, 2017.

[17] Feng Tang, Hoi Chun Po, Ashvin Vishwanath, and Xiangang Wan. Efficient topological materials discovery using symmetry indicators. *Nat. Phys.*, 15(5):470–476, 2019.

[18] Feng Tang, Hoi Chun Po, Ashvin Vishwanath, and Xiangang Wan. Comprehensive search for topological materials using symmetry indicators. *Nature*, 566(7745):486–489, 2019.

[19] Tiantian Zhang, Yi Jiang, Zhida Song, He Huang, Yuqing He, Zhong Fang, Hongming Weng, and Chen Fang. Catalogue of topological electronic materials. *Nature*, 566(7745):475–479, 2019.

[20] MG Vergniory, L Elcoro, Claudia Felser, Nicolas Regnault, B Andrei Bernevig, and Zhijun Wang. A complete catalogue of high-quality topological materials. *Nature*, 566(7745):480–485, 2019.

[21] Maia G Vergniory, Benjamin J Wieder, Luis Elcoro, Stuart SP Parkin, Claudia Felser, B Andrei Bernevig, and Nicolas Regnault. All topological bands of all stoichiometric materials. *arXiv:2105.09954*, 2021.

[22] Warren J Hehre. Ab initio molecular orbital theory. *Acc. Chem. Res.*, 9(11):399–406, 1976.

[23] Nitesh Kumar, Satya N Guin, Kaustuv Manna, Chandra Shekhar, and Claudia Felser. Topological quantum materials from the viewpoint of chemistry. *Chem. Rev.*, 121(5):2780–2815, 2021.

[24] QD Gibson, Leslie Mareike Schoop, L Muechler, LS Xie, M Hirschberger, Nai Phuan Ong, Roberto Car, and Robert Joseph Cava. Three-dimensional dirac semimetals: Design principles and predictions of new materials. *Phys. Rev. B*, 91(20):205128, 2015.

[25] Anna Isaeva and Michael Ruck. Crystal chemistry and bonding patterns of bismuth-based topological insulators. *Inorg. Chem.*, 59(6):3437–3451, 2020.

[26] Sebastian Klemenz, Aurland K Hay, Samuel ML Teicher, Andreas Topp, Jennifer Cano, and Leslie M Schoop. The role of delocalized chemical bonding in square-net-based topological semimetals. *J. Am. Chem. Soc.*, 142(13):6350–6359, 2020.

[27] Xin Gui, Ivo Pletikosic, Huibo Cao, Hung-Ju Tien, Xitong Xu, Ruidan Zhong, Guangqiang Wang, Tay-Rong Chang, Shuang Jia, Tonica Valla, et al. A new magnetic topological quantum material candidate by design. *ACS Cent. Sci.*, 5(5):900–910, 2019.

[28] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[29] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. rev. lett.*, 98(14):146401, 2007.

[30] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. rev. lett.*, 120(14):143001, 2018.

[31] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120(14):145301, 2018.

[32] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. commun.*, 8(1):1–12, 2017.

[33] Jesús Carrete, Wu Li, Natalio Mingo, Shidong Wang, and Stefano Curtarolo. Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X*, 4(1):011019, 2014.

[34] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. mat.*, 15(10):1120–1127, 2016.

[35] Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, et al. Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183, 2020.

[36] Wenhao Sun, Christopher J Bartel, Elisabetta Arca, Sage R Bauers, Bethany Matthews, Bernardo Orvañanos, Bor-Rong Chen, Michael F Toney, Laura T Schelhas, William Tumas, et al. A map of the inorganic ternary metal nitrides. *Nat. Mater.*, 18(7):732–739, 2019.

[37] Yi Zhang, A Mesaros, Kazuhiro Fujita, SD Edkins, MH Hamidian, K Ch'ng, H Eisaki, S Uchida, JC Séamus Davis, Ehsan Khatami, et al. Machine learning in electronic-quantum-matter imaging experiments. *Nature*, 570(7762):484–490, 2019.

[38] Volker L Deringer, Noam Bernstein, Gábor Csányi, Chiheb Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A Drabold, and Stephen R Elliott. Origins of structural and electronic transitions in disordered silicon. *Nature*, 589(7840):59–64, 2021.

[39] Janine George and Geoffroy Hautier. Chemist versus machine: Traditional knowledge versus machine learning techniques. *Trends Chem.*, 3(2):86–95, 2020.

[40] Yi Zhang and Eun-Ah Kim. Quantum loop topography for machine learning. *Phys. Rev. Lett.*, 118(21):216401, 2017.

[41] Pengfei Zhang, Huitao Shen, and Hui Zhai. Machine learning topological invariants with neural networks. *Phys. Rev. Lett.*, 120(6):066401, 2018.

[42] Mathias S Scheurer and Robert-Jan Slager. Unsupervised machine learning and band topology. *Phys. Rev. Lett.*, 124(22):226401, 2020.

[43] Yi Zhang, Paul Ginsparg, and Eun-Ah Kim. Interpreting machine learning of topological quantum phase transitions. *Physical Review Research*, 2(2):023283, 2020.

[44] Carlos Mera Acosta, Runhai Ouyang, Adalberto Fazzio, Matthias Scheffler, Luca M Ghiringhelli, and Christian Carbogno. Analysis of topological transitions in two-dimensional materials by compressed sensing. *arXiv:1805.10950*, 2018.

[45] Nikolas Claussen, B Andrei Bernevig, and Nicolas Regnault. Detection of topological materials with machine learning. *Phys. Rev. B*, 101(24):245117, 2020.

[46] Nina Andrejevic, Jovana Andrejevic, Chris H Rycroft, and Mingda Li. Machine learning spectral indicators of topology. *arXiv:2003.00994*, 2020.

[47] Guohua Cao, Runhai Ouyang, Luca M Ghiringhelli, Matthias Scheffler, Huijun Liu, Christian Carbogno, and Zhenyu Zhang. Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites. *Phys. Rev. Mater.*, 4(3):034204, 2020.

[48] Huijun Liu, GH Cao, ZZ Zhou, and Jianghui Liu. Screening potential topological insulators in half-heusler compounds via compressed-sensing. *J. Phys.: Condens. Matter*, 33(32):325501, 2020.

[49] Gabriel R Schleder, Bruno Focassio, and Adalberto Fazzio. Machine learning for materials discovery: Two-dimensional topological insulators. *Appl. Phys. Rev.*, 8(3):031409, 2021.

[50] Evert PL Van Nieuwenburg, Ye-Hua Liu, and Sebastian D Huber. Learning phase transitions by confusion. *Nat. Phys.*, 13(5):435–439, 2017.

[51] Juan Carrasquilla and Roger G Melko. Machine learning phases of matter. *Nature Physics*, 13(5):431–434, 2017.

[52] Ivan Glasser, Nicola Pancotti, Moritz August, Ivan D Rodriguez, and J Ignacio Cirac. Neural-network quantum states, string-bond states, and chiral topological states. *Phys. Rev. X*, 8(1):011006, 2018.

[53] Joaquin F Rodriguez-Nieva and Mathias S Scheurer. Identifying topological order through unsupervised machine learning. *Nat. Phys.*, 15(8):790–795, 2019.

[54] Benno S Rem, Niklas Käming, Matthias Tarnowski, Luca Asteria, Nick Fläschner, Christoph Becker, Klaus Sengstock, and Christof Weitenberg. Identifying quantum phase transitions using artificial neural networks on experimental data. *Nat. Phys.*, 15(9):917–920, 2019.

[55] Eran Lustig, Or Yair, Ronen Talmon, and Mordechai Segev. Identifying topological phase transitions in experiments using manifold learning. *Phys. Rev. Lett.*, 125(12):127401, 2020.

[56] T Mendes-Santos, X Turkeshi, M Dalmonte, and Alex Rodriguez. Unsupervised learning universal critical behavior via the intrinsic dimension. *Phys. Rev. X*, 11(1):011040, 2021.

[57] Andrew Ma, Yang Zhang, Thomas Christensen, Hoi Chun Po, Li Jing, Liang Fu, and Marin Soljačić. Topogivity: A machine-learned chemical rule for discovering topological materials. *In preparation*, 2021.

[58] Zhida Song, Tiantian Zhang, Zhong Fang, and Chen Fang. Quantitative mappings between symmetry and topology in solids. *Nat. Commun.*, 9(1):1–7, 2018.

[59] Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E Castelli, Stewart J Clark, Andrea Dal Corso, et al. Reproducibility in density functional theory calculations of solids. *Science*, 351(6280), 2016.

[60] VM Goldschmidt. Crystal structure and chemical constitution. *Trans. Faraday Soc.*, 25:253–283, 1929.

[61] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[62] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[63] Klaus Koepernik and Helmut Eschrig. Full-potential nonorthogonal local-orbital minimum-basis band-structure scheme. *Phys. Rev. B*, 59(3):1743, 1999.

[64] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865, 1996.