# Machine Learning on Multiple Topological Materials Datasets

Yuqing He[1,2], Pierre-Paul De Breuck[1], Hongming Weng[2*], Matteo Giantomassi[1], Gian-Marco Rignanese[1,3,4*]

[1]Institute of Condensed Matter and Nanosciences, UCLouvain, Chemin des Étoiles 8, 1348 Louvain-la-Neuve, Belgium.
[2]Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing, China.
[3]WEL Research Institute, Avenue Pasteur 6, 1300 Wavre, Belgium.
[4]School of Materials Science and Engineering, Northwestern Polytechnical University, No. 127 Youyi West Road, Xi'an 710072 Shaanxi, China.

*Corresponding author(s). E-mail(s): hmweng@iphy.ac.cn; gian-marco.rignanese@uclouvain.be;

### Abstract

A dataset of 35,608 materials with their topological properties is constructed by combining the density functional theory (DFT) results of Materiae and the Topological Materials Database. Thanks to this, machine-learning approaches are developed to categorize materials into five distinct topological types, with the XGBoost model achieving an impressive 85.2% classification accuracy. By conducting generalization tests on different sub-datasets, differences are identified between the original datasets in terms of topological types, chemical elements, unknown magnetic compounds, and feature space coverage. Their impact on model performance is analyzed. Turning to the simpler binary classification between trivial insulators and nontrivial topological materials, three different approaches are also tested. Key characteristics influencing material topology are identified, with the maximum packing efficiency and the fraction of $p$ valence electrons being highlighted as critical features.

# 1 Introduction

In the past decade, topological electronic materials have drawn considerable attention due to their importance for both fundamental science and next-generation technological applications [1–3]. These materials exhibit unique topological configurations in their electronic band structures, resulting in peculiar electronic properties [2, 3]. A central and long-standing question in this field is how to determine whether a given material is topologically trivial or not. Thanks to advancements in symmetry indicator theory [4] and topological quantum chemistry theory [5] and to the use of first-principles calculations [6–8], it is now possible to categorize topologically non-trivial materials (NTMs) according to their symmetry indicators or elementary band representations (EBRs) and compatibility relations. One first distinguishes the topological semimetals (TSMs). These materials present electronic bands intersecting at discrete points or along lines in momentum space, that do not satisfy the compatibility relations, resulting in gapless states near the Fermi level. These topological nodes cannot be removed by symmetry-preserving perturbations. Depending on the position of these nodes identified according to the symmetry representations of the corresponding bands, TSMs can be further grouped into high-symmetry-point semimetals (HSPSMs) and high-symmetry-line semimetals (HSLSMs) [6]. Depending on the symmetry indicators, one then identifies the topological insulators (TIs) and the topological crystalline insulators (TCIs), which have a set of valence bands that satisfy the compatibility relations but cannot be decomposed into linear combination of EBRs [5]. Using high-throughput calculations integrating density functional theory (DFT) [9, 10] and topological quantum chemistry theory or symmetry indicator theory, tens of thousands of topological materials were detected by analyzing the symmetry of the wavefunctions of crystalline compounds from the Inorganic Crystal Structure Database [11] (ICSD), leading to the compilation of several databases [6–8]. Soon after, these two theories based on symmetry representations of wavefunctions have been extended to magnetic ordered materials [12–14].

However, from an experimental standpoint, the main characteristics of materials that may affect their topological properties, especially those which provide helpful chemical insights, are still unclear, hindering the design of new topological materials. Machine learning (ML) [15] techniques offer a novel approach to address these shortcomings. By exploring existing data, they can potentially identify the features that are critical for obtaining topological properties without the need to resort to heavy computations. In this framework, several studies have already been conducted. For example, neural networks and k-means clustering have been used to learn from the Hamiltonian [16–19]. Compressed-sensing [20], gradient boosted trees (GBTs) [21], and other methods have also been employed to learn from ab-initio data [22–28]. Claussen et al. [23] trained a GBT model with 35,009 symmetry-based entries from the Topological Materials Database (TMD) [7, 29]. By testing the effect of different features, their model reached an accuracy of 87.0% for classifying materials into 5 subclasses: Trivial Insulator (TrIs), two types of TSMs (Enforced Semimetals (ESs) or Enforced Semimetals with Fermi Degeneracy (ESFDs)), and two types of TIs (Split EBRs (SEBRs) or Not a Linear Combination of EBRs (NLC)). Andrejevic et al. [27] utilized 16,458 inorganic materials from TMD to develop a neural network classifier capable of distinguishing NTMs and TrIs based on X-ray absorption near-edge structure (XANES) spectra. It achieved a $F_1$ score of 89% for predicting NTMs from their XANES signatures.

In 2023, Ma et al. [28] introduced a parameter named *topogivity* for each chemical element measuring its tendency to form topological materials. They employed support vector machine (SVM) [30] to learn the topogivities of elements based on a dataset of 9,026 materials from Tang et al. [8]. This approach achieved an average accuracy of 82.7% in an 11×10-fold NCV procedure. Claussen et al. [23] found that the topology does not depend much on the particular positions of atoms in the crystal lattice. This is in contrast with the previous two findings indicating that the local environment of atoms in compounds is a decisive factor since both XANES and topogivity are sensitive to the elements and their local chemical environment. In addition, these previous studies used mainly generic ML algorithms, overlooking newly designed ones specifically tailored for materials science and which have demonstrated excellent performance [31–34]. Finally, we would like to point out that the outcomes of these studies are difficult to compare due to the diverse range of crystal systems considered (including specific classes of systems, 2D materials, or bulk 3D materials) and the variations in utilized databases, material classes, and types of features incorporated in the ML model construction.

In this paper, we conduct data curation to compile a comprehensive dataset consisting of 35,608 entries from Materiae [6] and TMD. This new dataset is then used to train models for classifying a material into five types as TrI, or NTM, specifically HSPSM, HSLSM, TI, and TCI. Using nested cross-validation (NCV) [35] on the data from Materiae, we first benchmark five different approaches: namely Random Forest (RF) [36], XGBoost [37] (an implementation of GBTs), Automatminer (AMM) [32], the Material Optimal Descriptor Network (MODNet) [33, 38], and the Materials Graph Network (MEGNet) [31]. We find that XGBoost performs the best. We then test this model on the complete dataset, achieving a mean NCV accuracy of 82.9% for the classification into the five types. We discuss the NCV results and the train-test procedures, highlighting differences between the datasets that affect the scores. Finally, we compare XGBoost with topogivity [28] and t-distributed Stochastic Neighbor Embedding (t-SNE) [39] for the binary classification between TrIs and NTMs. XGBoost performs the best with 92.4% accuracy. Additionally, we investigate the key factors influencing the topology of materials. Our analysis reveals that the maximum packing efficiency (MPE) and the fraction of $p$ valence electrons (FPV) are the factors that contribute the most to the distinction between TrIs and NTMs. It is noted that MPE is a structure-based feature that represents the maximum packing efficiency of atoms within a crystal lattice and FPV is a composition-based feature that indicates the fraction of $p$ electrons versus all valence electrons. We think this finding is reasonable and these features could be used as a heuristic for exploring new topological materials.

# 2 Results

## 2.1 Data Curation

Two datasets are constructed in this work, named $M$ and $T$. The dataset $M$ is extracted from Materiae following a thorough data curation procedure as described below, resulting in 25,683 compounds. Similarly, the dataset $T$ is constructed from the Topological Materials Database, resulting in 24,156 compounds after cleaning.

It should be noted that the names of the topological types in these two databases are different. Therefore, we here establish a correspondence between them during the curation process. Figure 1 depicts the type distribution for the two datasets, their intersection ($M \cap T$), and differences ($M \backslash T$ and $T \backslash M$). Roughly the same distribution is found, with a majority of TrIs and around 30% NTMs.

The data curation proceeds as follows. For the dataset $M$, we initially query Materiae, which includes 26,120 materials that are neither magnetic materials (i.e., for which the magnetic moment would be higher than 0.1 Bohr Magneton per unit cell according to the Materials Project [40] (MP) record) nor conventional metals (i.e., systems with an odd number of electrons per unit cell). By keeping only the results that were computed including spin-orbit coupling, an initial dataset of 25,895 materials (named $MAT$) is obtained including their topological properties. Subsequently, the dataset $M$ is constructed by removing those materials with labels conflicting with the dataset $T$ (see last paragraph). All the records in $MAT$ are indexed by their unique MP-ID.

For the dataset $T$, we start from the data available in the Topological Materials Database [29], which includes 73,234 compounds indexed by their ICSD-ID and grouped into 38,298 unique materials by common chemical formula, space group, and topological properties as determined from their calculated electronic structure. As some of the pre-assigned MP-IDs were found to be wrong, we decided to control them systematically with the *structure_matcher* of PYMATGEN [41] (using its default tolerance settings) and make our own MP-ID assignment. Given a set of compounds grouped as one unique material, we distinguish three cases to assign the MP-ID and the corresponding structure. First, when none of the compounds has an assigned MP-ID, one structure of the set is randomly selected and the corresponding MP-ID is indicated as not available. Second, when the compounds are associated with at most one MP-ID, one structure of the set is again randomly selected. We then check whether it matches the MP structure corresponding to the indicated MP-ID. If it does, the MP-ID is assigned to the structure. If not, the MP-ID is indicated as not available. Finally, when more than one MP-ID appears in the set, these MP-IDs are first ranked according to their energy above hull. Then, the different structures in the set are compared with the MP structures corresponding to these MP-IDs starting from the lowest in energy. In case of match, the corresponding MP-ID (i.e., the one with the lowest energy) is assigned to the structure. In case of absence of match with any of the ranked MP-IDs, the MP-ID is indicated as not available. At the end of the process, only one of the structures associated with the same MP-IDs is kept. The compounds are then sorted adopting the same classification as in Materiae. First, they undergoes the same curation as the one described for the dataset M: excluding magnetic materials and conventional metals. The materials containing rare-earth elements (Pr, Nd, Pm, Sm, Tb, Dy, Ho, Er, Tm, Yb, Lu, Sc) are also removed. This is done because, for these elements, the results of Materiae and the Topological Materials Database were obtained from calculations performed using pseudopotentials with a different number of valence electrons (typically odd in one case and even in the other). Furthermore, we label the resulting data according to Materiae's definition. For TSMs, the mapping is rather simple: ESFDs correspond to HSPSMs and ESs to HSLSMs. In contrasts, for TIs, the mapping is more comple. We label SEBRs and NLCs as TIs or TCIs as follows. The materials in the spacegroups 174, 187, 189, 188, or 190 are all labeled as TCIs. The others are labeled according to the parity of the last topological indices, odd ones as TIs while even ones as TCIs. The next curation step consists in removing the materials with duplicate MP-IDs, as well the 673 compounds with the same MP-ID but conflicting topological types. At the end, we were left with a total of 24,368 items with an assigned MP-IP (sometimes indicated as not available) and sorted according to the same classification as Materiae. Thanks to the curation performed, the compounds in the two datasets can easily be related based on their assigned MP-IDs. On this basis, we further removed 212 materials present in both datasets but with differing types, leaving 24,156 compounds in $T$.

At the end of the construction of the datasets $T$ and $M$, our global dataset $M \cup T$ contains a total of 35,608 materials while the intersection $M \cap T$ consists of 14,231 compounds, as shown in Fig. 1.

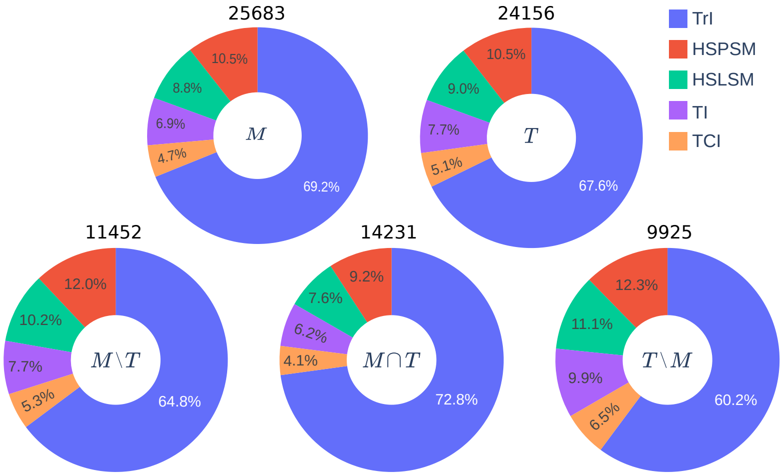

**Fig. 1** Composition in terms of the different topological types of the datasets $M$ (constructed from Materiae) and $T$ (originating from the Topological Materials Database) as well as of their differences ($M \backslash T$ and $T \backslash M$) and intersection ($M \cap T$).

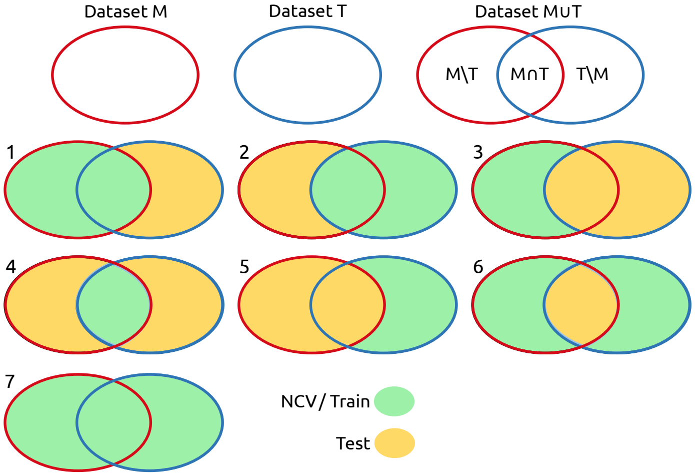

**Fig. 2** Diagram of the seven generalization tests. The datasets $M$ and $T$ are circled in red and blue, respectively. The union dataset $M \cup T$ can be split into three part: $M\backslash T$, $M \cap T$, and $T\backslash M$. In each of test, the dataset used for the training and the NCV of the ML model is filled in green, while the dataset used for testing it is filled in yellow.

## 2.2 Model

In order to select a model for further training and analysis, we first perform a benchmark on the $MAT$ dataset. Five different models are used: two generic ML algorithms (RF and XGBoost), and three well-developed algorithms in the field of material science (AMM, MODNet, and MEGNet). Moreover, for each method, two procedures are considered for the multiclass classification: either a direct multiclass classification (which gives the 5 possible labels as output) or a hierarchical binary classification (multiple models are trained following a tree such that each leaf represents a class). Figure 8 schematically represents these two procedures, with their respective accuracies. The highest accuracy (85.2%) is obtained with XGBoost using the direct multiclass classification. It is therefore used in the remainder of the work on all the data. More details about the benchmark are provided in Sec. 4.1.1.

## 2.3 Generalization tests

In principle, the NCV score should provide a comprehensive assessment of how the training model performs on new data. However, the model trained on the dataset $M$, which shows excellent performance (with a NCV accuracy of 85.7%), is found not to generalize well on the dataset $T\backslash M$ leading to an accuracy of only 71.8% (i.e., a decrease of 13.9%). Therefore, in order to further investigate the model performance, we perform a series of generalization tests by training on the different datasets at our disposal: $M$, $T$, their differences $M\backslash T$ and $T\backslash M$, their intersection $M \cap T$, their union $M \cup T$ as well as the union of their differences $(M\backslash T) \cup (T\backslash M)$.

**Table 1** Accuracy (in %) of the different nested cross-validation (NCV) and generalization tests depending on the training dataset. The results on (part of) the training dataset evaluated by NCV are indicated by a star.

| Test\Train | $M$ | $T$ | $M\backslash T$ | $M \cap T$ | $T\backslash M$ | $(M\backslash T) \cup (T\backslash M)$ | $M \cup T$ |
|---|---|---|---|---|---|---|---|
| NCV | 85.7 | 80.7 | 84.1 | 85.1 | 72.1 | 79.9 | 82.9 |
| $M\backslash T$ | 84.8* | 80.3 | 84.1* | 78.9 | 77.1 | 85.6* | 85.8* |
| $M \cap T$ | 86.1* | 86.0* | 82.0 | 85.1* | 81.6 | 83.6 | 86.6* |
| $T\backslash M$ | 71.8 | 73.2* | 69.8 | 70.2 | 72.1* | 73.3* | 74.3* |
| $M \cup T$ | 81.8 | 80.6 | 79.3 | 79.0 | 77.5 | 81.4 | 82.9 |

The seven tests are schematically represented in Fig. 2, where the datasets $M$ and $T$ are circled in red and blue, respectively. In each test, a ML model is first trained on the training set, depicted in green. A 5-fold NCV test is performed on the same data, followed by a generalization test on the test set depicted in yellow. The classification accuracy results obtained for each test are reported in Table 1, indicating the score obtained for the NCV, as well as on $M\backslash T$, $M \cap T$, $T\backslash M$, and $M \cup T$. Complementary metrics (i.e., $F_1$ score, precision, and recall) are reported in Table A1 in the Appendix.

As discussed below, the previously mentioned generalization issue is still present. For the generalization tests performed with $M$, $T$, $M\backslash T$, and $M \cap T$ as the training set (in green), the NCV accuracy is significantly higher than the test accuracy (i.e., for the corresponding datasets in yellow). When training on the dataset $M$, the NCV accuracy on the sub-dataset $M\backslash T$ (84.8%) and $M \cap T$ (86.1%) is also much larger than the test accuracy (71.8% for $T\backslash M$). When training on the dataset $T$, the results are more nuanced with the NCV on the sub-dataset $M \cap T$ (86.0%) being higher than the test accuracy (80.3% for $M\backslash T$). But that on the sub-dataset $T\backslash M$ (73.2%) is not.

It is worth noting that, when training on $T\backslash M$ and $(M\backslash T) \cup (T\backslash M)$, the NCV accuracy (72.1% and 79.9%) is smaller than all test accuracy values (81.6% and 83.6% for $M \cap T$, respectively; as well as 77.1% for $M\backslash T$ in the former case). Finally, the accuracy on $T\backslash M$ is the lowest one whatever the training set.

All the other metrics ($F_1$ score, precision, and recall) reported in Table A1 show the same trend. All these observations indicate that predicting the topological type on the materials of the dataset $T\backslash M$ seems to be more difficult than on those of the dataset $M$ (or its sub-datasets $M\backslash T$ and $M \cap T$). We propose four possible explanations for this bias (which are most probably combined).

The first reason is related to the distribution of the topological types in the datasets. As can be seen in Fig. 1, the proportion of TrIs is the lowest in $T\backslash M$, and the binary classification between TrIs and NTMs is much more accurate than the subsequent refined classifications of NTMs (see Fig. 8, Node 1 with respect to all the other nodes). Therefore, the proportion of TrIs affects the global accuracy.

The second rationalization is based on the distribution of the chemical elements in the datasets. Indeed, the accuracy of the model can be very low on compounds containing certain elements (e.g., as low as 37% on average for Gd), as illustrated in the Appendix (Fig. A1). In particular, the following elements with a low average accuracy are more present in $T\backslash M$ than in any other dataset: Ne, Mn, Fe, Eu, Gd, Po, Rn, Ra, Am. To test how this affects the global accuracy in each dataset, we recalculate the performance of the model when these elements are excluded. The corresponding accuracy, $F_1$ score, precision, and recall as well as the proportion of these materials are reported in the Appendix (Table A2). In general, the performance is smaller when including the elements above. This decrease is more important for the dataset $T \backslash M$ (3% compared to 0.5% for the other datasets). This could be expected as it contains a larger fraction of the elements above.

Following upon this observation, we search for possibly problematic elements in the dataset $M \cup T$. Their detection is based on more quantitative criteria. First, the number of materials containing such problematic element should be larger than 30, for statistical reasons. Second, the accuracy for the compounds containing this element should be lower than 75%. Finally, the recall for those materials should be lower than the one for those without that element. Applying these criteria, the following elements are identified: Cr, Mn, Fe, Cu, Tc, Eu, Os, Np. Table A3 contains the accuracies, $F_1$ score, precision and recall based on the presence of the previous elements.

A third potential cause of the bias for the dataset $T \backslash M$ is that about half of its compounds have an unknown magnetic type, since they could not be assigned an MP-ID. Table A4 investigates both the impact of elements and the presence of magnetic information. As can be seen, excluding the selected elements in the datasets $M \backslash T$, $M \cap T$ or $T \backslash M$ improves the accuracy by 5.3%. Excluding compounds with missing magnetic information further improves the score by 1.2%.

To analyze the cumulative effect of the above three explanations, we define the datasets $\widetilde{M \backslash T}$, $\widetilde{M \cap T}$, and $\widetilde{T \backslash M}$. These are formed by selecting the same number of compounds (3,372) in each original dataset ($M \backslash T$, $M \cap T$, and $T \backslash M$) adopting the same criteria as in Table A4 and in such a way that the distribution among the five different types is exactly the same (i.e., 2,339 TrIs, 315 HSPSMs, 279 HSLSMs, 271 TIs, and 168 TCIs). As can be seen in the NCV results reported in Table A5, the accuracy in the three datasets (79.2%, 79.9%, and 77.3%) is much more similar (the largest difference decreased to 2.6% from the previous 13.9%).

Finally, a fourth possible reason is related to the coverage of the feature space by the datasets. The ML model performance on a given test set obviously depends on how close its points are from those of the training dataset (interpolative predictions are better than extrapolative ones). To evaluate this effect, a heterogeneity metric is used, as explained in detail in Methods (see Eq. 5). It quantifies the similarity between the different datasets ($M \backslash T$, $M \cap T$, and $T \backslash M$), with a small heterogeneity leading in principle to a higher performance. The heterogeneity within each dataset (the diagonal part in Fig. 3) provides a reference value. Note that the heterogeneity in the dataset $T \backslash M$ is about 20% larger than in the others. This may explain the trend in the NCV accuracy for the models trained on the datasets $\widetilde{M \backslash T}$, $\widetilde{M \cap T}$, and $\widetilde{T \backslash M}$: as expected the lower the heterogeneity, the higher the NCV score. Furthermore, the heterogeneity increases significantly in the off-diagonal elements. This explains why a model trained on a given dataset tends not to generalize well to the other datasets.

## 2.4 Binary classification

In order to try to identify the main factors that influence the topology of a material, we turn to the binary classification between TrIs and NTMs on the whole dataset $M \cup T$. NTMs are considered as positive and TrIs as negative. Thus, the precision

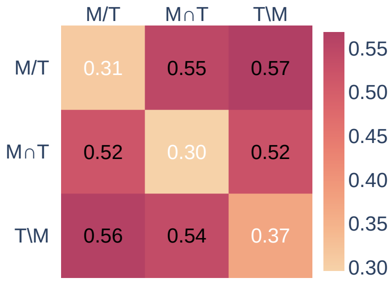

**Fig. 3** Heterogeneity metric between datasets: $M\backslash T$, $M\cap T$ and $T\backslash M$

**Table 2** Comparison of the NCV accuracy, $F_1$ score, precision, and recall (in %) of the XGBoost, topogivity, and t-SNE approaches. The Area Under the Receiver Operating Characteristic Curve (ROC_AUC) is also reported in %.

| | Accuracy | $F_1$ score | Precision | Recall | ROC_AUC |
|---|---|---|---|---|---|
| XGBoost | 92.4 | 88.5 | 89.5 | 87.5 | 97.5 |
| Topogivity | 87.2 | 79.5 | 85.6 | 74.1 | 90.6 |
| t-SNE | 75.7 | 70.8 | 59.0 | 88.3 | 89.1 |

measures the reliability of NTM predictions, and the recall measures the ability to detect all NTMs. The $F_1$ score, which is the harmonic mean of the precision and the recall, provides a balance between these two quantities (as they typically show an inverse relationship) and offers a better measure than the accuracy for an uneven class distribution.

Three approaches are considered here: the XGBoost model as above but for the binary classification; an existing heuristic model based on the *topogivity* of the elements [28] relying only on the composition of the compounds; and a generic dimension reduction method t-SNE [39] applied to the two most important features identified from XGBoost. All the details are available in Methods.

The results obtained on the dataset $M\cup T$ are provided in Table 2 and Fig. 4. Table 2 shows the results of the boolean predictions with the default threshold for each algorithm.

XGBoost shows the best performance with the highest accuracy, $F_1$ score, precision and ROC_AUC, thanks to its usage of a high-dimensional feature space to represent materials that fully describes the properties of materials. Figure 4 shows the trade-off of the scores as a function of the chosen threshold. XGBoost always has a better score. The topogivity and t-SNE approaches present an intersection point where they achieve the same scores. While their scores are lower than those of XGBoost, the topogivity and t-SNE approaches still provide reasonable results, and their advantage lies in their simplicity, making them easy to interpret.

The topogivity approach makes predictions based on a simple composition rule (see Eq. 1 in Methods) based on a single parameter, the elemental topogivity $\tau_E$ which approximately represents the inclination to form an NTM. Figure 5 shows a periodic

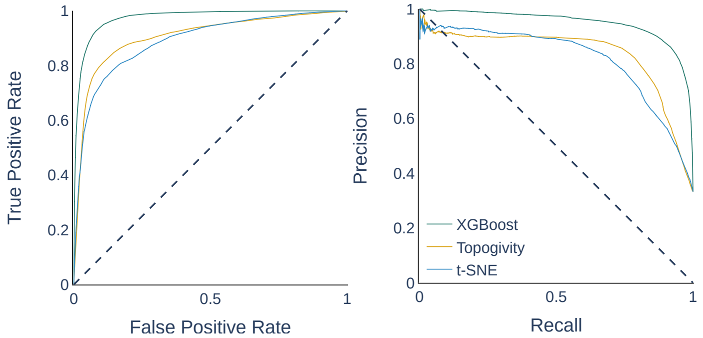

**Fig. 4** Receiver operating characteristic (ROC) and precision-recall curves for distinguishing nontrivial topological materials (NTMs) from trivial insulators (TrIs) on the dataset $M \cup T$.

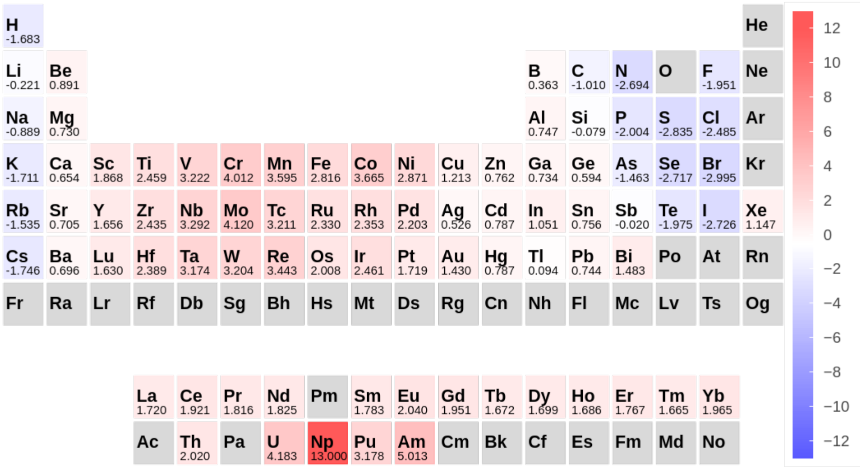

**Fig. 5** Periodic table of topogivities trained from dataset $M \cup T$. Existing topogivities are represented through numerical values with color coding, others are displayed in gray.

table with our newly trained topogivities for 83 elements (compared to 54 available previously).

The t-SNE approach developed here focuses on two features: the maximum packing efficiency in % (MPE) [42] and the fraction of $p$ valence electrons in % (FPV) [43, 44]. The points of the whole dataset are represented by two values representing their projections onto the t-SNE variables, as shown in Fig. 6. If the points are colored according to their type, a clear separation appears between NTMs and TrIs (in orange and blue, respectively). Taking the vertical line where t-SNE 1 is equal to zero as the

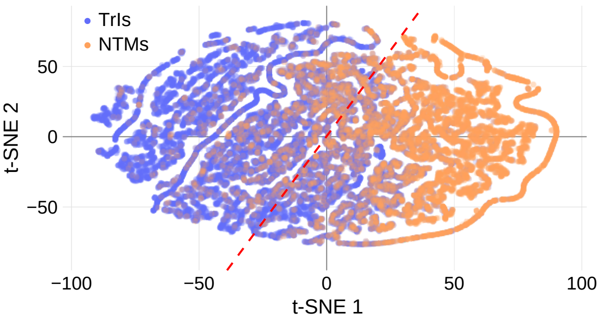

**Fig. 6** Visualization of the t-SNE results on the dataset $M \cup T$. Non-trivial materials are shown in blue, while trivial insulators are shown in orange. The red dashed line represents the decision boundary obtained using SVM.

splitting criterion between NTMs and TrIs, it is possible to predict 75.7% of materials correctly and to detect 88.3% of the NTMs. Furthermore, using a soft-margin linear SVM to identify the best frontier (dashed red line), the accuracy reaches 84.7%. This is still a bit lower than with the XGBoost and topogivity approaches, but it shows that even without using the target value (hence, in an unsupervised approach), the model can find the underlying relations between features and the topology of materials. The two selected features are clearly important to determine the topology of materials.

The distributions of their values in the dataset $M \cup T$ are displayed in Fig. 7. Panel (a) shows that the structures of NTMs are generally more closely packed than TrIs. This is consistent with our intuition that close-packing structures have stronger interatomic interactions, wider bands, and higher symmetry, thus promoting the appearance of nontrivial topological phases. Panel (b) demonstrates that NTMs tend to have a lower fraction of $p$ valence electrons. This can be rationalized as follows. Compounds with a higher fraction of $p$ valence electrons are mainly composed of elements of the top-right part of the periodic table which are more electronegative. These tend to form ionic or strongly covalent bonds with a large trivial band gap, hence to generate TrIs. This observation aligns well with the trends in the element topogivity, as depicted in Fig. 5. Elements located in the top-right part of the periodic table display negative topogivities, indicating their inclination to form TrIs.

## 3 Discussion

In this work, a dataset of 35,608 materials with their topological properties is constructed by combining the DFT results of Materiae and the Topological Materials Database, through a careful cleaning and curation process. The data from the two databases are found to be generally consistent with only 1% of the predictions which disagree. To the best of our knowledge, this is the first integration of materials from distinct data sources, a development that paves the way for more comprehensive and profound machine learning research. Using this newly created database, two research objectives were pursued.

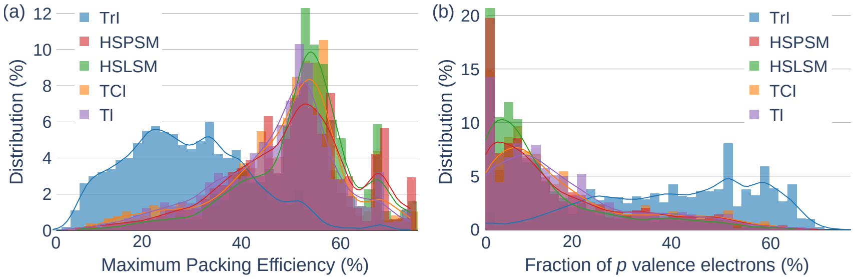

**Fig. 7** Distinction between trivial insulators (TrI) and nontrivial topological materials (NTM) based on (a) the maximum packing efficiency (%) and (b) the fraction of $p$ valence electrons in the dataset $M \cup T$. The NTM cannot be further discriminated into HSPSM, HSLSM, TCI, and TI.

First, machine-learning approaches were developed for categorizing materials into five distinct topological types (TrI, HSPSM, HSLSM, TI, and TCI). A thorough benchmark was performed on one of the databases to compare various machine learning approaches, obtained by combining five different models (MEGNet, Automatminer, MODNet, Random Forest, and XGBoost) with two possible procedures, namely one consisting of a series of binary-classification steps to obtain the final sorting, and the other being a direct multiclass categorization. The direct multiclass procedure relying on XGBoost was identified as the most promising approach, achieving an impressive 85.2% accuracy. A series of generalization tests were conducted that allowed for the identification of a series of differences between the two datasets (the distributions of the topological types and the chemical elements in the datasets, the presence of compounds of unknown magnetic type, as well as their coverage of the feature space). Their influence on the performance of the model was carefully analyzed.

Secondly, key factors influencing the material topology were identified by focusing on the binary classification between TrIs and NTMs. The previously developed approach relying on XGBoost performs even better on this simpler task achieving 92.4% accuracy. It was compared with two simpler methods, one relying on the use of the *topogivity* of the elements [28], and one being an unsupervised t-SNE. The latter only focuses on the two features identified as the most important, namely MPE and FPV. It demonstrates an accuracy of 84.7%. Such performance shows that the two features are very relevant to determining the topology of materials.

Upon analyzing the distribution of these features, we found notable disparities between NTMs and TrIs. These factors are compatible with our understanding and, together with topogivity, they offer heuristic intuitions for designing topological materials. This highlights the potential to discover critical features using machine learning approaches.

Prior to our work, Claussen et al. [23] had used a similar machine learning approach on TMD, also relying on XGBoost. They had found that the topology is mostly determined by the chemical composition and the crystal symmetry and that it does not depend much on the particular positions of atoms in the crystal lattice. In contrast, Andrejevic et al. [27] had suggested that XANES, a spectrum strongly related to the atomic type, the site, and the short-range interactions with surrounding atoms, could be used to identify NTMs and TrIs. Topogivity had also been introduced [28] as an intuitive chemical parameter related only to the elemental composition and had been found to work rather well. Therefore, it remained elusive whether or not local chemical environment and element-related atomic features are the important characteristics influencing the topology. In our study, we included new descriptors related to the crystal structure, the composition, and the atomic sites, in addition to those employed by Claussen et al. [23], enlarging the space to be explored. We observed that, rather than the crystal symmetry, the most important characteristic influencing the topology is MPE of atoms in the crystal lattice space. We rationalize this by the fact that the latter actually determines the hopping parameters of electrons and as a result influences the band width and the possibility of band inversion leading to non-trivial band topology. We found that the second most important characteristic is FPV. We think that the latter has an impact on the type of bonding. Indeed, many compounds showing essentially $p$ orbitals in the valence tend to be large band gap covalent insulators like diamond and silicon, or ionic insulators like NaCl and $CaF_2$.

Our study reveals the critical role of a comprehensive database in the ML research. The acquisition of a more extensive dataset, encompassing not only symmetry-indicator-based topological materials but also simulation results from Wilson loops, along with experimental data, holds immense importance for driving the further progress of this research.

# 4 Methods

## 4.1 ML Models

In this study, three main approaches have been considered for the classification of topological materials. The first one is a ML classifier into the five different types (TrI, HSPSM, HSLSM, TI, and TCI). For this approach, we tested five different models combined with two possible procedures. The other two can only separate the materials into two classes, namely TrIs and NTMs (the first approach can obviously also produce this simpler classification). The second one relies on the use of a previously developed heuristic model based on the *topogivity* of the elements [28]. The last one is an unsupervised ML approach relying on t-SNE.

### 4.1.1 ML classifier

For the ML classifier, we benchmark five different models (MEGNet, Automatminer, Random Forest, MODNet, and XGBoost) with two possible procedures. These are benchmarked on the dataset *MAT* to determine the best model.

**ML models**

MEGNet [31] is a graph neural network for machine learning molecules and crystals in materials science. MEGNet v1.2.9[1] is used in this work. For the multiclass classification (Tree 2, see below), the last layer is changed to softmax and the loss function to

[1] https://github.com/materialsvirtuallab/megnet

”categorical_crossentropy”. In the MEGNet model, the crystal is represented through a CrystalGraph which is truncated using a cutoff radius of 4Å for defining the neighbors of each atom. It is trained using 500 epochs. Given that structures containing isolated atoms cannot be handled by MEGNet, they are discarded from the training and test sets. The scores reported for MEGNet refer to the results obtained on the valid structures (i.e., without isolated atoms).

In all the other ML models, the crystal is represented using the features generated by Matminer [45]. This library transforms any crystal (based on their composition and structure) into a series of numerical descriptors with a physical and chemical meaning. It relies on various featurizers adapted from scientific publications.

Automatminer (AMM) [32] allows for the automatic creation of complete machine learning pipelines for materials science. Here, features are automatically generated with Matminer and then reduced. Using the Tree-based Pipeline Optimization Tool (TPOT) library [46], an AutoML stage is used to prototype and validate various internal ML pipelines. A customized ’express’ preset pipeline is performed on the training set (note that the ’EwaldEnergy’ is excluded from the AutoFeaturizer due to technical problems).

Random Forest (RF) [36], which is an ensemble learning method, constructs multiple decision trees. For the classification task, the final output is the result of majority voting. Here, we first use Matminer to extract Magpie_ElementProperty [43], SineCoulombMatrix [47], DensityFeatures and GlobalSymmetryFeatures from the input structures (the missing features are filled with the average of the known data). The hyperparameters are determined using a 5×3 NCV on the training data, where the stratified 3-fold inner cross-validation uses a grid search for determining the optimal values ({’max_features’ : [’auto’, ’sqrt’, ’log2’], ’criterion’ : [’gini’, ’entropy’]}) relying on the ’balanced_accuracy’ score.

MODNet [33, 38] is a framework based on feature selection and a feedforward neural network. The selection uses a relevance-redundancy score defined from the mutual information between the features and the target and between pairs of features. The framework is well suited for limited datasets. Here, we first use a predefined set of featurizers (DeBreuck2020Featurizer with accelerated oxidation state parameters) to generate features from the structures. Table 3 provides a complete list of all these Matminer featurizers. The missing features are filled with their default value (most are zero). The feature selection is performed on the training data only. The model hyperparameters are determined using a genetic algorithm.

XGBoost [37] is an ensemble of boosted trees. Here, the features generated by the MODNet approach are used as the input of the model. The hyperparameters are determined using a 5×5 NCV on the training data, where the stratified 5-fold inner cross-validation uses a grid search for determining the optimal values (see Table 4) relying on the $F_1$ score.

**Procedures**

Two different classification procedures are used, as shown in Fig. 8. The first approach (Tree 1) uses four one-vs-all binary-classification steps to obtain the final classification. The second approach (Tree 2) uses a direct multi-class classification for

**Table 3** The set of Matminer featurizers to generate the numerical descriptors in MODNet. They are used for algorithm MODNet and XGBoost.

| Composition | Structure | Site |
|---|---|---|
| AtomicOrbitals | BondFractions | AGNIFingerprints |
| AtomicPackingEfficiency | ChemicalOrdering | AverageBondAngle |
| BandCenter | CoulombMatrix | AverageBondLength |
| ElectronegativityDiff | DensityFeatures | BondOrientationalParameter |
| ElementFraction | EwaldEnergy | ChemEnvSiteFingerprint |
| ElementProperty ("magpie" preset) | GlobalSymmetryFeatures | CoordinationNumber |
| IonProperty | MaximumPackingEfficiency | CrystalNNFingerprint |
| Miedema | RadialDistributionFunction | GaussianSymmFunc |
| OxidationStates | SineCoulombMatrix | GeneralizedRadialDistributionFunction |
| Stoichiometry | StructuralHeterogeneity | LocalPropertyDifference |
| TMetalFraction | XRDPowderPattern | OPSiteFingerprint |
| ValenceOrbital | | VoronoiFingerprint |
| YangSolidSolution | | |

**Table 4** Hyperparameter grid searched for the XGBoost model.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Learning rate $\eta$ | 0.23 | $L^2$ regularization $\lambda$ | 1.33 |
| Maximal tree depth | [9, 10, 11] | Miniaml child weight | [0.1, 0.3, 0.5] |
| Column subsampling by tree | [0.75, 0.78] | Column subsampling by node | [0.75, 0.78] |

the five different types, using either majority voting for tree-based methods, or softmax activation for neural network based methods. The accuracy for both approaches are reported in the figure.

**Benchmark**

A consistent NCV testing procedure is used to evaluate the performance on the dataset *MAT* for the 10 different combinations of ML models and procedures. For all the ML models, an identical stratified 5-fold outer loop was used to determine the generalization error (test error).

The internal validation depends on the algorithm, as explained above.

In terms of the final classification, the results of the two procedures are very close with a ranking that depends on the ML model. The best result is achieved for the direct multiclass procedure relying on XGBoost which achieves an accuracy of 85.2%.

### 4.1.2 Topogivity

The topogivity $\tau_E$ of an element $E$ has been proposed as a measure of its tendency to form topological materials [28]. The $\tau_E$ values for 54 elements were originally obtained by using support vector machine (SVM) [30] model trained on a subset of 9,026 compounds from the database created by Tang *et al.* [8], with approximately one half classified as trivial and the other half as topological. The ML model is based on a heuristic chemical rule, which maps each material $M$ to a number $g(M)$ through the function

$$g(M) = \sum_E f_E(M)\tau_E \tag{1}$$

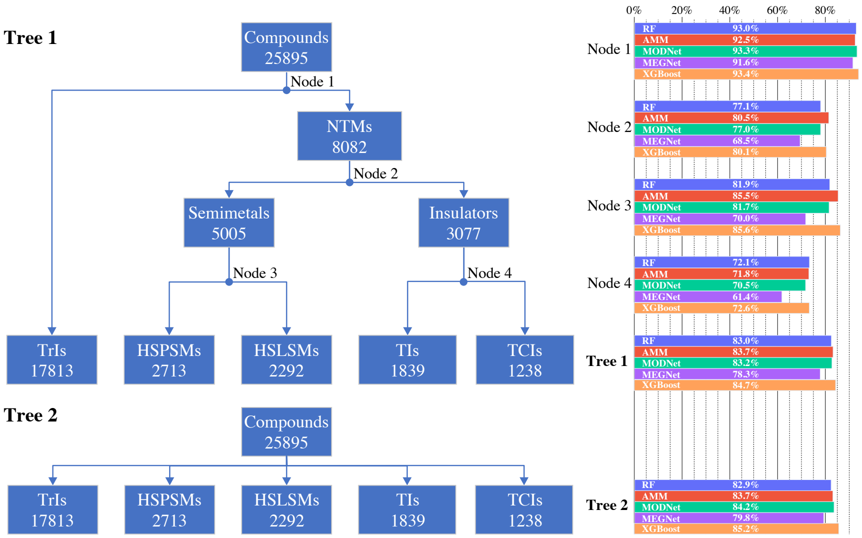

**Fig. 8** Illustration of the two different procedures used to categorize the materials according to their topological types. The Tree 1 relies on four binary-classification steps, while Tree is base on a direct multiclass approach. The accuracy achieved with the five ML models (RF, AMM, MODNet, MEGNet, and XGBoost) are reported both at the final stage of both procedures, as well as at each step in Tree 1.

where the summation runs over all elements $E$ in the chemical formula of material $M$, with $f_E(M)$ denoting the fraction of element $E$ within material $M$. A material $M$ is classified based on $g(M)$, as trivial (TrI) if negative and topological (NTM) if positive.

Here, we also build a new topogivity model trained on a subset from dataset $M \cup T$, which excludes the elements occurring less than 25 times. We construct a soft-margin linear SVM using the scikit-learn library (specifically, the sklearn.svm.SVC class). The hyperparameter $C$ of the model is determined through a grid search among 15 values evenly spaced on a log scale ranging from $10^4$ to $10^6$ relying on the $F_1$ score and adopting a 5-fold validation procedure.

The optimal value ($C = 3.73 \times 10^4$) is then used to train the final model on the whole dataset. This new topogivity model provides the $\tau_E$ value for 83 elements and covers 35,522 materials out of the 35,608 of the dataset $M \cup T$. These values are reported on the periodic table shown in Fig. 5.

For comparison, the previous model [28] gives the $\tau_E$ value for 54 elements and covers only 18,637 compounds in $M \cup T$. A quantitative comparison of the two models can thus only be performed on these 18,367 compounds. From the different scores reported in Table 5, it can be said that their performance is essentially similar.

**Table 5** NCV accuracy, $F_1$ score, precision and recall (in %) of the two different topogivity models using the 18,637 compounds of the $M \cup T$ dataset which only include the 54 elements for which $\tau_E$ is provided in Ref. [28].

|  | Accuracy | $F_1$ score | Precision | Recall |
|---|---|---|---|---|
| This work | 90.2 | 76.8 | 83.5 | 71.2 |
| Ref. [28] | 89.6 | 77.3 | 77.1 | 77.5 |

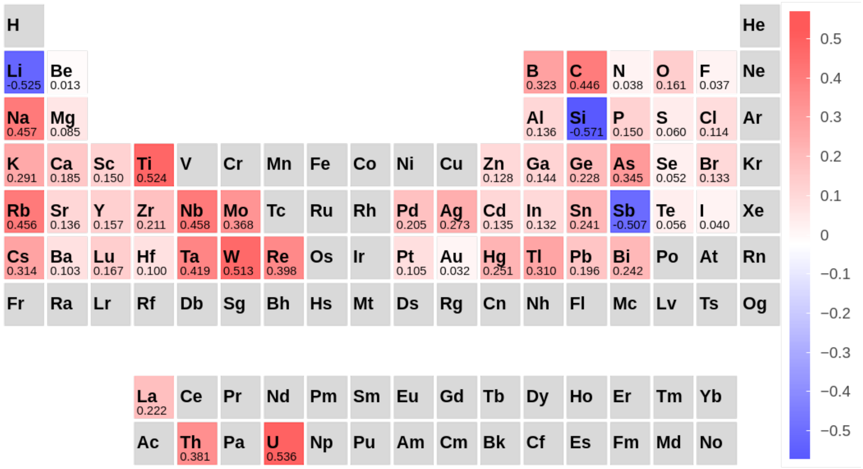

**Fig. 9** Comparison of the topogivities of Ref. [28] with those obtained here for 54 elements. The color code refers to $\Delta\tau_E$, a measure of their relative difference, as defined by Eq. (2). When the signs of the two topogivities are different, the value of $\Delta\tau_E$ is negative, and thus the element is highlighted in blue.

To compare the differences element by element, we use a form of relative difference defined by:

$$\Delta\tau_E = \frac{2\,\mathrm{sign}(\tau_E^p \tau_E^n)\,|\tau_E^p - \tau_E^n|}{2 + |\tau_E^p| + |\tau_E^n|}, \tag{2}$$

where $\tau_E^p$ and $\tau_E^n$ are the topogivities of the previous and new models, respectively. The values of $\Delta\tau_E$ are indicated on the periodic table shown in Fig. 9. They show that the two sets of results are essentially consistent with each other, except for three elements (Li, Si, and Sb), which exhibit different signs (as indicated by the blue color in the figure).

The final evaluation of the topogivity approach is performed using a 5×5 NCV. The results are presented in Table 2 and Fig. 4. It leads to an accuracy of 87.2%, a $F_1$ score of 79.5%, a precision of 85.6%, and a recall of 74.1%.

### 4.1.3 t-SNE

The last approach originates from the idea to identify the most important features to distinguish topological materials, which we then combine with the unsupervised algorithm t-SNE. All the Matminer features are first ranked in descending order of gain importance for training the XGBoost classification model between the 5 types for the datasets $M \backslash T$, $M \cap T$ and $T \backslash M$. The intersection of the top 15 most important features consists of 3 features: the maximum packing efficiency (MPE) in %, the fraction of $p$ valence electron (FPV) in %, and the formation enthalpy ($\Delta$H) in eV/atom as obtained from the semi-empirical Miedema model [48].

Based on this finding, the distribution of the compounds in $M \cup T$ according to the values of MPE, FPV, and $\Delta$H is analyzed to understand what helps the models to distinguish between TrIs from NTMs. The corresponding plots, in which the five different materials types have been separated, are reported in Fig. 7 for MPE and FPV (already commented in the Results section) and in Fig. 10(a) for $\Delta$H.

For the latter, it turns out that there is a concentration of TrIs around zero. The reason for this is, however, not physical but technical. In fact, the semi-empirical Miedema model [48] does not apply to all materials. For those compounds where it does not work, Matminer does not produce a value for the feature and, as commonly done in ML approaches, we replace these missing values with a zero. As it turns out, as shown in Fig. 10(c), the share of such missing values is much larger for TrIs (more than 87%) than NTMs, the ML model takes advantage of this flaw to identify them.

If the compounds without $\Delta$H value are removed, 12,518 entries are left in $M \cup T$, rather evenly sorted into 3,622 TrIs, 2,843 HSPSMs, 2,734 HSLSMs, 1,339 TCIs, and 1,980 TIs (see Fig. 10(d)). The corresponding distribution of the $\Delta$H values is reported in Fig. 10(b). It still shows a significant difference between TrIs (which mostly present positive values) and NTMs (which are mainly negative). The underlying reasons are still not completely clear to us. Given that value of the feature is missing for many materials, it can, however, not be used as a discriminator. Therefore, it is discarded from the subsequent analysis.

The dataset $M \cup T$ can now be simply visualized in 2D by representing each material by its values for MPE and FPV, as shown in Fig. 11. The plot reveals a rather clear distinction between TrIs and NTMs. And, if we apply t-SNE on the whole dataset, the distinction is even clearer. Finally, based on the t-SNE variables, a dividing line can be drawn using a soft-margin linear SVM in which $C$ is equal to 1.0.

## 4.2 Heterogeneity metric

To quantify the degree of diversity within a dataset or between two datasets, we adopted a heterogeneity metric defined as the k-nearest-neighbor (k-NN) distance in the space of the most important Matminer features. The most important features are defined based on their induced gain during the training of the XGBoost model. A total of 47 Matminer features (44 presenting continuous values and 3 discrete ones, respectively) was gathered by taking the union of the 20 most important ones for training on the sub-datasets $M \backslash T$, $M \cap T$, and $T \backslash M$. These originate from the five featurizers reported in Table 6.

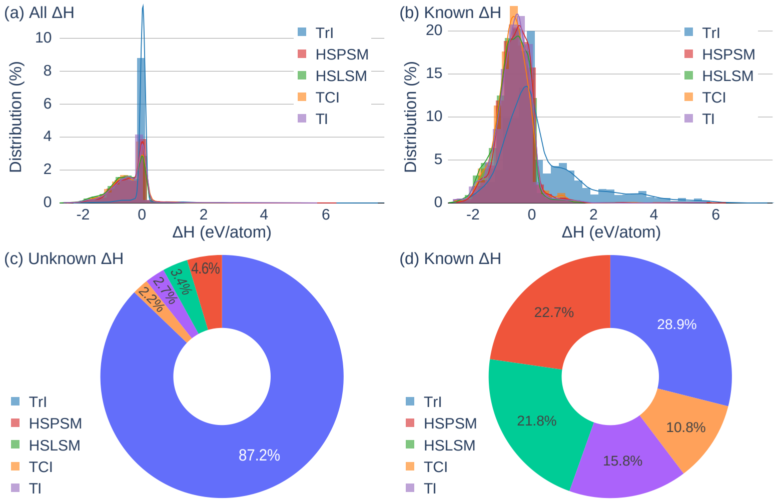

**Fig. 10** Distribution of ΔH values in the dataset $M \cup T$ according to the 5 topological types. Panel (a) shows the distribution of all the values, while panel (b) presents the distribution only for those compounds for which the value is known through the Miedema model [48]. Panels (c) and (d) show the pie-charts with distribution of the compounds among the 5 different topological types for those compounds with unknown and known values, respectively.

**Table 6** Matminer featurizers generating the features selected for defining the heterogeneity metric between datasets. A brief description and the source of the data is also provided.

| Matminer Featurizer | Description |
|---|---|
| • Miedema | Formation enthalpies of intermetallic compounds [48]. |
| • ElementProperty ("magpie" preset) | Weighted elemental statistics [43]. |
| • OxidationStates | Statistics about the oxidation states for each specie. Features are concentration-weighted statistics of the oxidation states. |
| • AtomicOrbitals | Estimation of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies based on the composition and the atomic orbital energies [49]. |
| • GlobalSymmetryFeatures | Symmetry information. |

All the continuous feature values are rescaled to range between 0 to 1 using the MinMaxScaler from scikit-learn on the dataset $M \cup T$. Then, the distance $d(x, y)$ between any two points $x$ and $y$ is defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n_c}(x_i - y_i)^2 + \sum_{j=n_c+1}^{n_c+n_d}(1/2(x_j == y_j))^2}\ , \tag{3}$$

where $n_c$=44 and $n_d$=3 are the numbers of features presenting continuous and discrete values, respectively. The distance $D(x, A)$ between a point $x$ and a dataset $A$ is defined

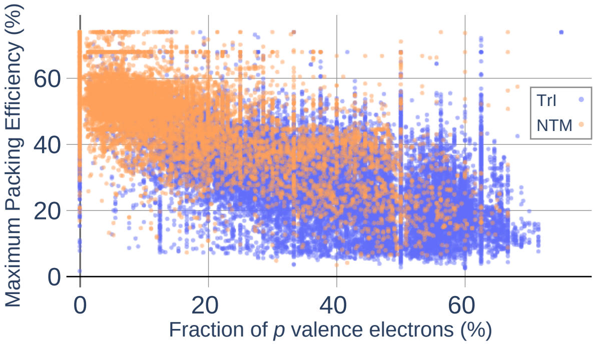

**Fig. 11** Visualization of the dataset $M \cup T$ according to the two most important features: the maximum packing efficiency and the fraction of $p$ valence electrons. This representation leads to a clear separation between TrIs and NTMs even though there is some overlap between the two types.

as the mean of the distances from point $x$ to the 5 NN-points ($q_j$ with $j = 1, \ldots, 5$) in dataset $A$:

$$D(x, A) = \frac{1}{5} \sum_{j=1}^{5} d(x, q_j). \tag{4}$$

Finally, the heterogeneity metric $H(A, B)$ between two datasets $A$ and $B$ (which can be the same dataset $A$) is defined based on the average distance between all the points of $A$ ($p_i$ with $i = 1, \ldots, N_A$) and the dataset $B$:

$$H(A, B) = \frac{1}{N_A} \sum_{i=1}^{N_A} D(p_i, B). \tag{5}$$

Note that with this definition $H(A, B)$ is an asymmetric measure (it follows from the fact that training and testing sets are not interchangeable).

# 5 Data availability

The data can be downloaded from the Materials Cloud Archive [50]: https://doi.org/10.24435/materialscloud:xx-xb. It can be viewed interactively (e.g., the plots corresponding to Figures 7 and 11) through the chemiscope visualization tool [51]. A Jupyter notebook is also available. All this information is also available here: https://topoclass.modl-uclouvain.org.

**Acknowledgments.** We thank N. Regnault for providing the data of TMD, which makes the data curation of the Materiae and TMD possible. Y. H. and H. W. acknowledge the funding from the National Key Research and Development Program of China (Grant No. 2022YFA1403800), the National Natural Science Foundation of China (Grants No. 12188101, 11925408), the Chinese Academy of Sciences (Grant No. XDB33000000). Y. H. was supported by China Scholarship Council (Grant No.

201904910878). H. W. is also supported by the New Cornerstone Science Foundation through the XPLORER PRIZE. Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

# References

[1] Kitaev, A. (ed.) *Periodic table for topological insulators and superconductors.* (ed. ) *AIP conference proceedings*, Vol. 1134, 22–30 (American Institute of Physics, 2009).

[2] Hasan, M. Z. & Kane, C. L. Colloquium: topological insulators. *Rev. Mod. Phys.* **82**, 3045 (2010).

[3] Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83**, 1057 (2011).

[4] Po, H. C., Vishwanath, A. & Watanabe, H. Symmetry-based indicators of band topology in the 230 space groups. *Nature communications* **8**, 50 (2017).

[5] Bradlyn, B. *et al.* Topological quantum chemistry. *Nature* **547**, 298–305 (2017).

[6] Zhang, T. *et al.* Catalogue of topological electronic materials. *Nature* **566**, 475–479 (2019).

[7] Vergniory, M. G. *et al.* A complete catalogue of high-quality topological materials. *Nature* **566**, 480–485 (2019).

[8] Tang, F., Po, H. C., Vishwanath, A. & Wan, X. Comprehensive search for topological materials using symmetry indicators. *Nature* **566**, 486–489 (2019).

[9] Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).

[10] Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).

[11] Hellenbrandt, M. The inorganic crystal structure database (icsd)—present and future. *Crystallogr. Rev.* **10**, 17–22 (2004).

[12] Watanabe, H., Po, H. C. & Vishwanath, A. Structure and topology of band structures in the 1651 magnetic space groups. *Science Advances* **4**, aat8685 (2018).

[13] Elcoro, L. *et al.* Magnetic topological quantum chemistry. *Nature communications* **12**, 5965 (2021).

[14] Peng, B., Jiang, Y., Fang, Z., Weng, H. & Fang, C. Topological classification and diagnosis in magnetically ordered electronic materials. *Phys. Rev. B* **105**, 235138 (2022). URL https://link.aps.org/doi/10.1103/PhysRevB.105.235138.

[15] Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **3**, 210–229 (1959).

[16] Zhang, Y. & Kim, E.-A. Quantum loop topography for machine learning. *Phys. Rev. Lett.* **118**, 216401 (2017).

[17] Zhang, P., Shen, H. & Zhai, H. Machine learning topological invariants with neural networks. *Phys. Rev. Lett.* **120**, 066401 (2018).

[18] Zhang, Y., Ginsparg, P. & Kim, E.-A. Interpreting machine learning of topological quantum phase transitions. *Phys. Rev. Res.* **2**, 023283 (2020).

[19] Scheurer, M. S. & Slager, R.-J. Unsupervised machine learning and band topology. *Phys. Rev. Lett.* **124**, 226401 (2020).

[20] Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory* **52**, 1289–1306 (2006).

[21] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).

[22] Acosta, C. M. *et al.* Analysis of topological transitions in two-dimensional materials by compressed sensing. *arXiv* 1805.10950 (2018).

[23] Claussen, N., Bernevig, B. A. & Regnault, N. Detection of topological materials with machine learning. *Phys. Rev. B* **101**, 245117 (2020).

[24] Cao, G. *et al.* Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites. *Phys. Rev. Mater.* **4**, 034204 (2020).

[25] Liu, J., Cao, G., Zhou, Z. & Liu, H. Screening potential topological insulators in half-heusler compounds via compressed-sensing. *J. Phys. Condens. Matter* **33**, 325501 (2021).

[26] Schleder, G. R., Focassio, B. & Fazzio, A. Machine learning for materials discovery: Two-dimensional topological insulators. *Appl. Phys. Rev.* **8**, 031409 (2021).

[27] Andrejevic, N. *et al.* Machine-learning spectral indicators of topology. *Adv. Mater.* **34**, 2204113 (2022).

[28] Ma, A. *et al.* Topogivity: A machine-learned chemical rule for discovering topological materials. *Nano Lett.* **23**, 772–778 (2023).

[29] Vergniory, M. G. *et al.* All topological bands of all nonmagnetic stoichiometric materials. *Science* **376**, eabg9094 (2022).

[30] Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

[31] Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).

[32] Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and automatminer reference algorithm. *npj Comput. Mater.* **6** (2020).

[33] De Breuck, P.-P., Hautier, G. & Rignanese, G.-M. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *npj Comput. Mater.* **7** (2021).

[34] Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).

[35] Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc., Ser. B, Methodol.* **36**, 111–133 (1974).

[36] Ho, T. K. (ed.) *Random decision forests.* (ed. ) *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1, 278–282 (IEEE, 1995).

[37] Chen, T. & Guestrin, C. (ed.) *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (ed. ) , KDD '16, 785–794 (ACM, 2016).

[38] De Breuck, P.-P., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on modnet. *J. Phys. Condens. Matter* **33**, 404002 (2021).

[39] Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** (2008).

[40] Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).

[41] Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

[42] Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017). URL http://link.aps.org/doi/10.1103/PhysRevB.96.024104.

[43] Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2** (2016).

[44] Deml, A. M., O'Hayre, R., Wolverton, C. & Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Phys. Rev. B* **93**, 085142 (2016). URL https://link.aps.org/doi/10.1103/PhysRevB.93.085142.

[45] Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).

[46] Olson, R. S. *et al.* *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I*, Ch. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, 123–137 (Springer International Publishing, 2016).

[47] Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).

[48] Zhang, R., Zhang, S., He, Z., Jing, J. & Sheng, S. Miedema calculator: A thermodynamic platform for predicting formation enthalpies of alloys within framework of miedema's theory. *Comput. Phys. Commun.* **209**, 58–69 (2016).

[49] Kotochigova, S., Levine, Z. H., Shirley, E. L., Stiles, M. D. & Clark, C. W. Local-density-functional calculations of the energy of atoms. *Phys. Rev. A* **55**, 191–199 (1997).

[50] Talirz, L. *et al.* Materials cloud, a platform for open computational science. *Scientific Data* **7** (2020). URL http://dx.doi.org/10.1038/s41597-020-00637-5.

[51] Fraux, G., Cersonsky, R. & Ceriotti, M. Chemiscope: interactive structure-property explorer for materials and molecules. *Journal of Open Source Software* **5**, 2117 (2020). URL http://dx.doi.org/10.21105/joss.02117.

# Appendix

**Table A1** $F_1$ score, precision, and recall (in %) of the different nested cross-validation (NCV) and generalization tests depending on the training dataset. The results on (part of) the training dataset evaluated by NCV are indicated by a star.

| Test \ Train | | $M$ | $T$ | $M\backslash T$ | $M \cap T$ | $T\backslash M$ | $(M\backslash T) \cup (T\backslash M)$ | $M \cup T$ |
|---|---|---|---|---|---|---|---|---|
| $F_1$ score | $M\backslash T$ | 94.9* | 93.4* | 94.4 | 94.0* | 90.6 | 95.2 | 94.9* |
| | $M \cap T$ | 92.7* | 92.7 | 87.6* | 92.9 | 88.0 | 88.9* | 92.6* |
| | $T\backslash M$ | 91.1 | 90.1* | 88.7 | 91.2 | 89.1* | 90.1* | 90.9* |
| | NCV | 93.7 | 91.6 | 94.4 | 92.7 | 89.4 | 92.6 | 92.9 |
| | $M \cup T$ | 93.0 | 92.2 | 90.3 | 92.8 | 89.2 | 91.4 | 92.9 |
| Precision | $M\backslash T$ | 83.7* | 81.7* | 84.0 | 77.4* | 82.5 | 86.1 | 85.3* |
| | $M \cap T$ | 85.3* | 84.8 | 84.9* | 82.1 | 83.5 | 86.4* | 86.7* |
| | $T\backslash M$ | 63.7 | 70.2* | 64.3 | 58.7 | 69.7* | 71.3* | 72.5* |
| | NCV | 84.4 | 77.4 | 84.0 | 81.9 | 69.6 | 78.8 | 81.2 |
| | $M \cup T$ | 77.5 | 78.8 | 77.7 | 72.7 | 78.5 | 81.3 | 81.5 |
| Recall | $M\backslash T$ | 88.9* | 87.2* | 88.9 | 84.9* | 86.4 | 90.4 | 89.8* |
| | $M \cap T$ | 88.8* | 88.6 | 86.2* | 87.2 | 85.7 | 87.6* | 89.5* |
| | $T\backslash M$ | 75.0 | 78.9* | 74.5 | 71.4 | 78.2* | 79.6* | 80.7* |
| | NCV | 88.8 | 83.9 | 88.9 | 87.0 | 78.3 | 85.1 | 86.7 |
| | $M \cup T$ | 84.6 | 85.0 | 83.5 | 81.5 | 83.5 | 86.0 | 86.8 |

**Table A2** Comparison of the accuracy, F1 score, precision, and recall (in %) on the datasets $M\backslash T$, $M \cap T$ and $T\backslash M$ with and without the materials containing at least one of the following elements: Ne, Mn, Fe, Eu, Gd, Po, Rn, Ra, Am. The first three columns show the NCV results on each sub-dataset, while the last three columns are the NCV results on the dataset $M \cup T$. The proportion of the problematic elements is also reported in the last row of the table for each dataset.

| | | NCV on each sub-dataset | | | NCV on $M \cup T$ | | |
|---|---|---|---|---|---|---|---|
| | | $M\backslash T$ | $M \cap T$ | $T\backslash M$ | $M\backslash T$ | $M \cap T$ | $T\backslash M$ |
| Accuracy | with | 84.1 | 85.4 | 72.0 | 85.9 | 86.6 | 74.2 |
| | without | 84.7 | 85.9 | 75.1 | 86.3 | 87.2 | 77.5 |
| $F_1$ score | with | 88.9 | 87.2 | 78.2 | 89.8 | 89.5 | 80.7 |
| | without | 89.0 | 87.2 | 78.8 | 89.7 | 89.6 | 81.8 |
| Precision | with | 94.4 | 92.9 | 89.1 | 94.9 | 92.6 | 90.9 |
| | without | 94.7 | 93.1 | 89.8 | 95.1 | 92.8 | 91.6 |
| Recall | with | 84.0 | 82.1 | 69.7 | 85.3 | 86.7 | 72.5 |
| | without | 84.0 | 82.0 | 70.1 | 85.0 | 86.7 | 73.9 |
| Element proportion (in %) | | 3.1 | 1.9 | 15.9 | 3.1 | 1.9 | 15.9 |

**Table A3** Comparison of the NCV accuracy, $F_1$ score, precision and recall (in %) on the dataset $M \cup T$ with and without materials containing Cr, Mn, Fe, Cu, Tc, Eu, Os, and Np.

| | | Cr | Mn | Fe | Cu | Tc | Eu | Os | Np |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | with | 69.3 | 61.4 | 60.6 | 74.3 | 68.1 | 69.2 | 70.0 | 70.2 |
| | without | 83.2 | 83.3 | 83.7 | 83.4 | 83.0 | 83.0 | 83.1 | 83.0 |
| $F_1$ score | with | 85.1 | 78.5 | 79.6 | 81.0 | 86.0 | 72.4 | 83.3 | 89.3 |
| | without | 86.7 | 87.0 | 87.1 | 87.1 | 86.7 | 86.8 | 86.7 | 86.7 |
| Precision | with | 92.6 | 90.6 | 88.0 | 94.4 | 93.0 | 90.5 | 89.1 | 100.0 |
| | without | 93.0 | 93.0 | 93.2 | 92.8 | 92.9 | 93.0 | 93.0 | 92.9 |
| Recall | with | 78.7 | 69.3 | 72.6 | 70.9 | 80.0 | 60.3 | 78.2 | 80.6 |
| | without | 81.3 | 81.7 | 81.7 | 82.1 | 81.2 | 81.4 | 81.3 | 81.2 |

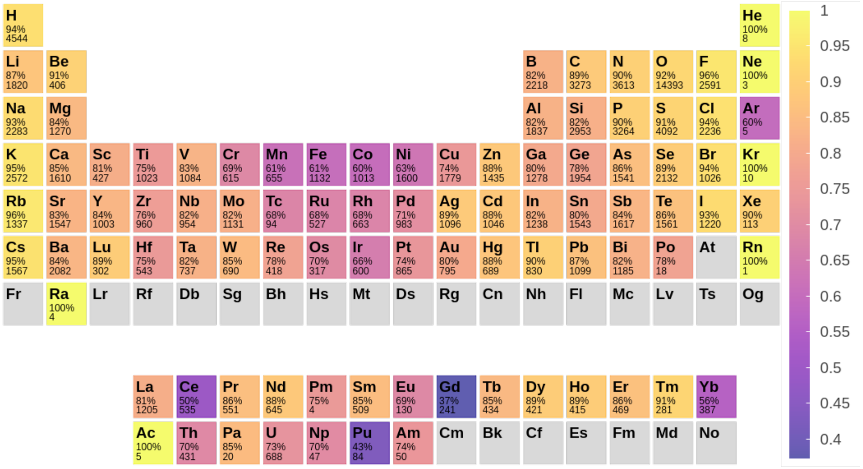

**Fig. A1** Periodic table with the average accuracy of the XGBoost model in the dataset $M \cup T$ per element. For each element, two numbers are reported. The first one (in %) shows the average accuracy for the compounds containing the element. This number is also reflected in the coloring of the cell in the periodic table. The second on illustrates the number of compounds containing this element.

**Table A4** Comparison of the NCV accuracy, $F_1$ score, precision and recall (in %) on the whole dataset $T\setminus M$ with the same results on the subsets 1 obtained by excluding the materials containing selected elements (Cr, Mn, Fe, Cu, Tc, Eu, Os or Np), and subset 2 constructed by further excluding materials without MP-ID.

| | Count | Accuracy | $F_1$ score | Precision | Recall |
|---|---|---|---|---|---|
| All data | 9,925 | 72.0 | 78.2 | 89.1 | 69.7 |
| Subset 1 | 7,364 | 77.3 | 79.9 | 90.1 | 71.8 |
| Subset 2 | 3,550 | 78.3 | 82.9 | 90.4 | 76.6 |

**Table A5** NCV accuracy, $F_1$ score, precision and recall (in %) on the datasets $\widetilde{M\setminus T}$, $\widetilde{M \cap T}$ and $\widetilde{T\setminus M}$.

| | Accuracy | $F_1$ score | Precision | Recall |
|---|---|---|---|---|
| $\widetilde{M\setminus T}$ | 79.2 | 82.8 | 92.8 | 74.7 |
| $\widetilde{M \cap T}$ | 79.9 | 84.4 | 90.7 | 78.8 |
| $\widetilde{T\setminus M}$ | 77.3 | 78.4 | 88.1 | 70.7 |

**Table A6** Comparison of the value of $g(M)$ (see Eq. eq:topog) based on the elemental topogivity $\tau_E$ from this work and from Ref. [28] for well-known nontrivial topological materials. Their type as determined experimentally is also reported. The "*" indicated for the type of TaAs refers to the fact that it is a Weyl semimetal which cannot be identified by symmetry-indicator theory nor TQC. CoSi is an example where Ref. [28] does not allow for a prediction because the topogivity of Co is not available.

| Compound | Type | This work | Ref. [28] |
|---|---|---|---|
| $Cd_2As_3$ | HSLSM | -2.815 | -5.375 |
| CoSi | HSPSM | 3.586 | — |
| $Na_3Bi$ | HSLSM | -1.184 | 1.606 |
| PtAl | HSPSM | 2.466 | 3.023 |
| SnTe | TCI | -1.219 | -0.574 |
| $TaAs_2$ | TCI | 0.248 | 0.410 |
| $Bi_2Se_3$ | TI | -5.185 | -4.411 |
| $Bi_2Te_3$ | TI | -2.959 | -1.102 |
| $Sb_2Te_3$ | TI | -5.965 | -4.118 |
| $ZrTe_5$ | TI | -7.440 | -5.816 |
| TaAs | * | 1.711 | 2.899 |