MACHINE LEARNING TOPOLOGICAL CHARACTERISTICS FROM MULTIPLE ELECTRONIC MATERIALS DATABASES

YUQING HE

Supervisors:

Gian-Marco Rignanese

(Co-Supervisor) Matteo Giantomassi

Members of the jury:

(President) Xavier Urbain
Xavier Gonze
Jean-Christophe Charlier
Hongming Weng
Xiaotong Liu

UCLouvain
Institute of Condensed Matter and Nanosciences

December 2023

ABSTRACT

Significant advances have been made in predicting new topological materials using high-throughput *ab–initio* and symmetry-based indicator calculations. To date, thousands of materials have been identified to be topologically nontrivial by scanning the existing databases. However, this approach is severely limited to non-magnetic systems with well-defined symmetries, leaving a much larger materials space unexplored. Additionally, the existing results have not been able to link the topology of materials to some basic factors that are applicable in the experimental field, for example, the difference in electronegativities of the chemical composition. The machine learning method is an effective way to overcome the above shortcomings by exploring the topological materials data. A large dataset including 35,608 entries is obtained by merging two symmetry-indicator-based databases: Materiae and Topological Materials Database. We mainly address two tasks: 5-types classification for categorizing materials into the following classes - trivial insulator, high-symmetry-point semimetal, high-symmetry-line semimetal, topological insulator, and topological crystalline insulator; and a binary classification to distinguish trivial insulators from topologically nontrivial materials. The first task aims to train a model to explore new materials. After a benchmark with an identical nested cross-validation procedure, the method adopting gradient boosted trees reaches the highest accuracy, 85.2%. We further apply this method in a group of tests to analyze how well this model would generalize to new, unseen data in practical applications. Then a robust model is trained on our full dataset, and it is used to predict the topological type of 906 materials. The second task is to study the essential factors that influence the topology of materials. We interpret some important features used in our model and combine them with an unsupervised algorithm to analyze their effectiveness.

ACKNOWLEDGMENTS

I would like to extend my profound gratitude to the following individuals and institutions whose valuable guidance and unwavering support made this thesis possible.

First and foremost, my sincere appreciation goes to my supervisor, Prof. Gian-Marco Rignanese. His expertise, guidance and invaluable feedback have not only shaped the direction of this research but have also been a constant source of inspiration. His scientific curiosity and dedication to academics are truly commendable.

I am equally indebted to my co-supervisor, Dr. Matteo Giantomassi, whose solutions to various challenges were always effective. His broad and in-depth knowledge deeply inspired me.

To my esteemed Committee Members, Prof. Xavier Urbain, Prof. Xavier Gonze, Prof. Jean-Christophe Charlier, Prof. Hongming Weng, and Prof. Xiaotong Liu, I express my sincere anticipation for your forthcoming insights and constructive criticisms, which undoubtedly enhance the quality of this thesis.

I would like to extend my gratitude to my collaborators, Prof. Hongming Weng for his unremitting participation and Yi Jiang for his assistance in the early stages of the dataset construction. Particular thanks to Pierre Paul De Breuck for his valuable contributions and suggestions during a range of discussions.

My heartfelt thanks go to my colleagues, including Vinciane Gandibleux, Jiaqi Zhou, Maryam Azizi, Guillaume Brunin, Ionel-Bogdan Guster, Wei Chen, and many more. You are my friends at the same time. Thank you for your patience in addressing various problems, both physical questions in research and issues in my life. Your help encourage me to overcome all kinds of difficulties.

To my family, my mom Jie, my dad Jun, and my grandma Yalan, I am profoundly grateful for your unwavering belief in my abilities, which has been a driving force throughout this journey.

To my friends, Mark, Yanru, Min, Zehan, Yao, Yinghua, and others, your steadfast friendship, care, and support have been a constant source of inspiration and joy. A special mention goes to my boyfriend, Mark, whose encouragement and support were invaluable during the completion of this thesis.

I would like to acknowledge the China Scholarship Council (Grant No. 201904910878) for their financial support during my graduate studies.

Finally, my heartfelt thanks extend to all the participants who generously shared their time and insights during conferences and workshops. This thesis is a culmination of the collective efforts of many individuals and entities, and I am deeply appreciative of the role each has played in its completion.

Thank you all for your invaluable contributions.

CONTENTS

ACRONYMS

Automatminer AutoML + Matminer

deltaH Miedema deltaH_inter

DFT density functional theory

ES enforced semimetal

ESFD enforced semimetal with fermi degeneracy

FN false negative

FP false positive

FPV fraction p valence electrons

GBTs gradient boosted trees

HOMO highest occupied molecular orbital

HSLSM high-symmetry-line semimetal

HSPSM high-symmetry-point semimetal

ICSD Inorganic Crystal Structure Database

k-NN k-nearest-neighbor

KS Kohn-Sham

LUMO lowest unoccupied molecular orbital

MEGNet Materials graph network

ML machine learning

MODNet Material optimal descriptor network

MP Material Project

MPE max packing efficiency

NCV nested cross-validation

NLC not equal to a Linear combination

NMI normalized mutual information

NTM topologically nontrivial material

PCA principal component analysis

RF random forests

ROC receiver operating characteristic

ROC-AUC the area under the receiver operating characteristic curve

RR relevance and redundancy

SEBR split extended brillouin zone representation

SISSO Sure Independence Screening and Sparsifying Operator

SVM support vector machine

t-SNE t-distributed Stochastic Neighbor Embedding

TCI topological crystalline insulator

TI topological insulator

TMD Topological Materials Database

TN true negative

TP true positive

TPOT Tree-based Pipeline Optimization Tool atomic packing efficiency

TQC Topological Quantum Chemistry theory

TrI trivial insulator

TSM topological semimetal

VASP Vienna ab initio simulation package

INTRODUCTION

Since the 20th century, quantum theory [1] has revolutionized our understanding of the world by revealing that all matter can exhibit wave properties under a suitable energy scale. Condensed matter physics is a field focusing on the utilization of quantum theory and statistical principles in the study of solids to understand their physical properties and behavior. A standard approach to this problem is to formulate the system's Hamiltonian, solve the many-body Schrödinger equation, and subsequently compute the observables. However, given that real-world systems involve an astronomically large number of particles, on the order of 10^{23}, attempting to compute the many-body wave function within such a huge Hilbert space is unrealistic. In this context, physicists introduced the idea of "more is different": Although everything in condensed matter physics always can be explained by basic theories (quantum mechanics and statistical physics), complex real-world periodic structures will emerge to an effective theory that is quite different from the basic theory [2]. Remarkably, even in the face of intricate complexity within actual systems, these effective theories can exhibit a remarkable degree of simplicity and generality. Various formalisms, such as Monte Carlo methods [3], molecular dynamics [4], and density functional theory (DFT) [5, 6], have been developed to study crystalline structures, phase transitions, and electronic properties. Moreover, in recent years, a series of research to apply machine learning (ML) methods to simulation or experimental results have also progressed significantly [7, 8].

Materials are at the heart of condensed matter physics. In recent years, topological material is an exotic class of materials that have been of major interest due to their importance for both fundamental science and next-generation technological applications. [9–11] They have nontrivial topological configurations in the geometry of their electronic band structures, thus exhibiting unique electronic properties and unconventional electromagnetic activity. [12–14]. Figure 1 shows a schematic of topology of the band structures. [15] Due to the existence of a finite band gap, ordinary insulators are non-conductive. Trivial semimetals exhibit zero band gaps at the critical point. In the case where the conduction and valence bands are inverted, it results in an intrinsic topological insulator (TI). At the critical point during the transition from a TI to an ordinary insulator with symmetry protection, the crossing of the conduction and the valence bands leads to a topological semimetal (TSM). Determining whether a given material is topological is a central and enduring question in this domain. Topological band theory [9–11, 14] is a simple yet effective approach to the problem. In topological band theory, the interaction of quasiparticles is treated to the mean field level, the band topology is considered to be a solvable problem. Combined with first-principle calculations, by analyzing the symmetry of the non-interacting wave functions and their electromagnetic response, the diagnosis of a wide range of topological materials is made possible [16–19]. This symmetry-based approach has the advantages of wide applicability and relatively low computational cost [20–25]. Thanks to high-throughput calculations performed on the crystal structures from the Inorganic Crystal Structure Database (ICSD) [26, 27], tens of thousands of topological materials were detected, and several databases were compiled [28–31].

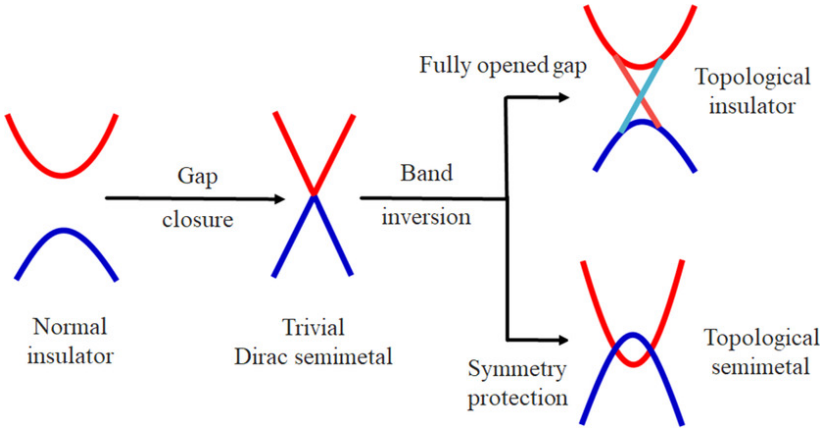


Figure 1: Band topology schematics. Image taken from Ref. [15]

Unfortunately despite its efficiency, this method is not suitable for certain materials, particularly magnetic materials, and structures lacking well-defined symmetries, and only restricted insight about why a certain material is topological is given. These limitations constrain the exploration and design of "ideal" topological materials. However, the existence of large amounts of data paves the way for adopting ML techniques. ML in material science is an active field of research that leverages data-driven algorithms to advance materials research and discovery. By analyzing large databases of known materials with the associated properties, ML models can predict a wide range of quantities, such as thermal conductivity [32–34], vibrational properties [35, 36], and electronic band structures [37, 38]. A major advantage of ML is the reduction of computational cost [39, 40], thus enabling the rapid screening and accelerating the discovery of novel materials with specific properties [41–43]. Moreover, ML is capable of providing new physical insights by identifying patterns and relationships within the data [44–46]. Over the past few years, there has been a rapid development of frameworks specifically tailored for adopting ML in material science, including tools for feature extraction [47, 48] and algorithms for model generation. The algorithms are universal for all properties, some have the capacity to directly learn from materials (e.g., crystal structures or molecules) [36, 49–51], conversely, others require numerical descriptors for their operation [52, 53]. According to the benchmark results [50, 54] obtained on a series of datasets with different properties, relatively accurate predictions can be made, with each algorithm exhibiting its own advantages in various aspects.

In this work, we adopt ML techniques with existing data to explore some insightful models for predicting topological materials. Our objective is to classify materials into distinct categories, namely trivial insulator (TrI) or topologically nontrivial material (NTM), specifically high-symmetry-point semimetal (HSPSM), high-symmetry-line semimetal (HSLSM), topological insulator (TI) or topological crystalline insulator (TCI). Chapter 1 introduces some basic concepts about ML techniques and provides a concise review of the literature regarding applications in the realm of topological materials. Chapter 2 discusses the process of data curation on two symmetry-based DFT-computed topological materials databases: Materiae [28] and Topological Materials Database (TMD) [30, 31]. Chapter 3 presents an overview of the learning process and parameter settings. Chapter 4 discusses detailed results for classifying materials into five distinct types and presents the predictions made by our final model. Chapter 5 considers the results of diagnosing NTM from TrI, and studies the critical factors that influence topological properties.

1 STATE OF THE ART

1.1 MACHINE LEARNING METHODS

The essence of ML is to make predictions or decisions not through explicit programs, but rather by models autonomously discovered from the data. ML includes three basic paradigms, supervised learning, unsupervised learning, and reinforcement learning. Supervised learning consists in finding a general rule that maps inputs to the target outputs by learning from labeled examples. Both inputs and their desired outputs are included in the training data. Two common examples of supervised learning tasks include classification, which involves discrete output variables, and regression, which deals with continuous output variables. Unsupervised learning aims to identify patterns from unlabeled data, thereby revealing the underlying structure inherent in the given data. This is achieved through various techniques, such as clustering to group similar instances and dimensionality reduction to project high-dimensional data onto a more compact subspace. Reinforcement learning [55] focuses on what to do, learning how to map situations to actions so that agents can get the maximum numerical reward signal in a dynamic environment, and has found notable applications in domains like robotics and game playing. In this work, we focus on solving a classification task with supervised learning, and at the end we use unsupervised learning to study some essential features. In the upcoming sections, I will provide a concise introduction to fundamental ML concepts, such as the process of completing a task and the learning algorithms that can be used. To facilitate the comprehension, some of these concepts are illustrated with a simple example of a binary classification task where the objective is to distinguish between two classes: "spam" and "non-spam" emails.

The task of filtering out unwanted or potentially harmful messages is essential for ensuring a better user experience and improved email security. The goal is to develop a ML model capable of analyzing email content and accurately classifying incoming emails based on certain features. The informative attributes extracted from the data that aid in the task are called features. Several features commonly utilized in this task are listed below. The sender's email address, if unusual or randomly generated information exists in the name, and verifying the domain name matches the claimed organization. Identifying embedded links and URLs within the email and cross-referencing them with known blacklists or reputable sources. Detection of suspicious attachment files. Analyzing the content and structure of email, spam emails contain attention-grabbing subject lines and promotional keywords, they are poorly constructed with spelling and grammar errors. Each data point can be encoded into a sequence of numerical variables according to these features, which are subsequently inputted into the relevant algorithm for analysis and modeling.

1.1.1 *Learning algorithms*

Extensive research has been conducted on a multitude of general algorithms that demonstrate applicability across diverse learning tasks. Here, I will briefly introduce several fundamental algorithms: decision tree [56], support vector machine (SVM) [57], artificial

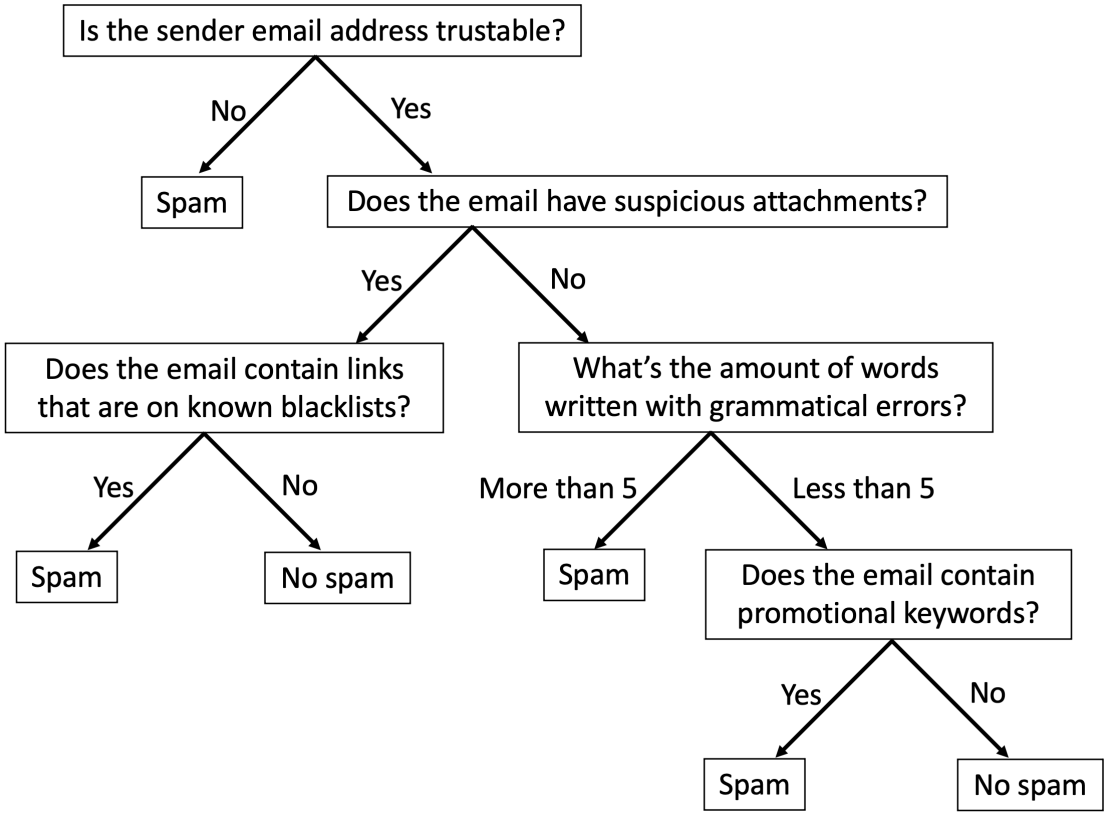


Figure 1.1: An example of decision tree.

neural networks [58], and principal component analysis (PCA) [59, 60]. The intention of this section is not to present all existing models, but to provide some brief examples. They either participate as a small step or serve as fundamental constituents for other well-designed algorithms in our research.

A decision tree [56] is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from the root to the leaf represent classification rules. Figure 1.1 shows an example of a decision tree for distinguishing the spam email. A decision tree is simple to understand and interpret, but often relatively inaccurate compared to other models. This can be remedied by replacing a single decision tree with an assembly of decision trees at the cost of losing some interpretability. The random forests (RF) [61] and gradient boosted trees (GBTs) [62] are two common ensemble learning methods built upon decision trees taking bagging and boosting approach.

SVM [57] aims to find optimal hyperplanes that effectively separate data points belonging to distinct classes within a given feature space. Suppose the sets to discriminate are linearly separable in that space. In that case, a logical selection for the optimal hyperplane is one that represents the largest separation, or margin, between the two classes, the so-called maximum-margin hyperplane. It divides the input variable space such that the width of the gap between different classes is maximized. Predictions of new examples are made based on which side they fall when mapping into that same space. Figure 1.2 shows a simple example in 2D space, it classifies a pair of (x1,x2) coordinates into the category

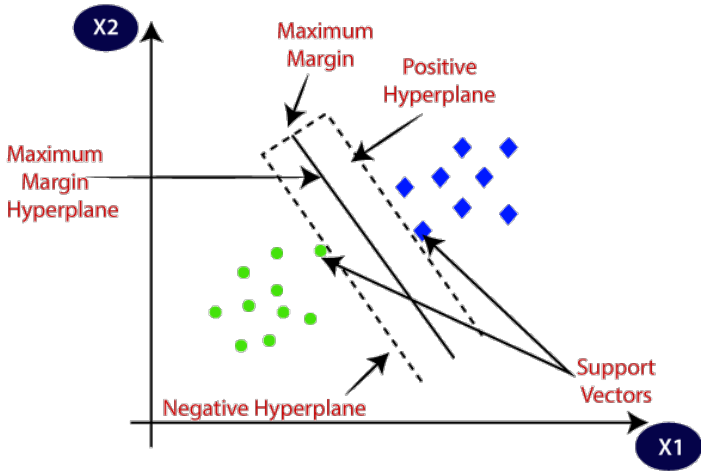


Figure 1.2: An example of a support vector machine. Image taken from Ref. [64]

red or blue, the best hyperplane (here a line) is the one that maximizes the margins, the distance between the hyperplane and the nearest data point of the two categories. These points are called support vectors. For non-linearly separable cases, a proposed solution [63] is to transform the original data into a higher-dimensional space where classes presumably become more separable.

Artificial neural networks [58], usually simply called neural networks, are computing systems inspired by the biological neural networks that constitute animal brains. A neural network is based on a collection of connected units or nodes called artificial neurons. Each neuron can transmit a signal, a real number to other neurons, and the output of each neuron is computed by a function of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds by backward propagation. The weight increases or decreases the strength of the signal at a connection. Typically, neurons are aggregated into different layers and perform different transformations. Signals travel from the first layer (the input layer) to the last layer (the output layer) possibly after traversing several layers. Figure 1.3 shows a simple schematic of a neural network with two hidden layers. With our example, the input can be a three-dimensional vector that represents each data point based on the following features: "Is the email address trustable?", "Are suspicious links or attachments present?", "The amount of spelling and grammar errors". The output can be a vector [0, 1] or [1, 0] which represents "spam" or "non-spam" email.

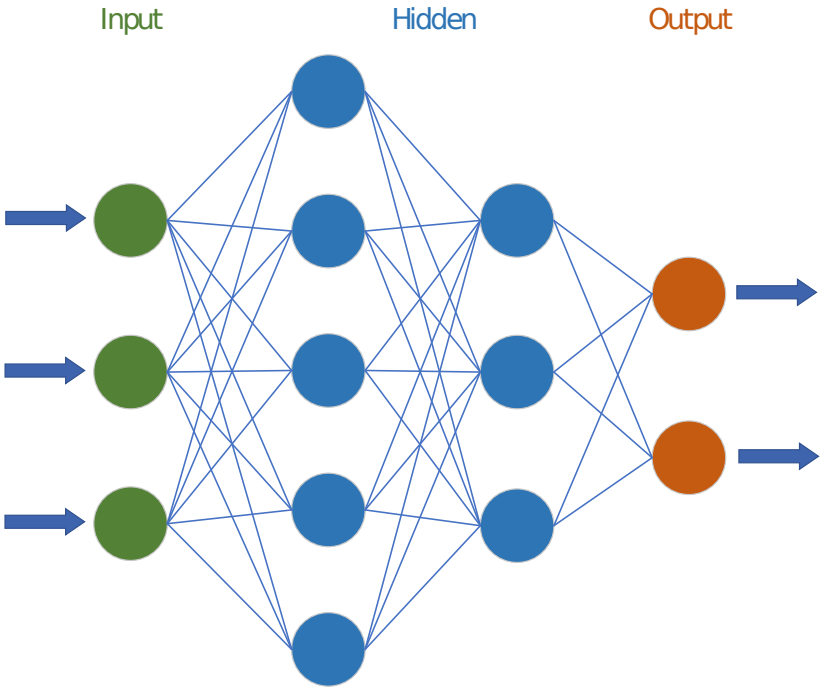


Figure 1.3: Schematic of a neural network. Here, each circular node represents an artificial neuron while each arrow represents a connection from the output of one artificial neuron to the input of another. Image adjusted from Ref. [65]

PCA [59, 60] is a fundamental dimensionality reduction technique widely used in data analysis and machine learning. Its primary aim is to transform high-dimensional data into a lower-dimensional subspace while preserving the essential variability and patterns present in the original data. The key idea behind PCA is to identify the principal components, which are orthogonal vectors that represent the directions of maximum variance in the data. Through diagonalizing the covariance matrix and sorting eigenvectors according to their eigenvalues descendingly, the principal components are sequentially assigned. By projecting the original features onto a small set of uncorrelated principal components, PCA reduces the dimensionality while retaining the most significant information. PCA is commonly used for data visualization of high-dimensional data in lower dimensions, and noise reduction or feature extraction by identifying dominant patterns in the data. One of the primary advantages of PCA is that it is an unsupervised learning technique, thus it does not require labeled data. However, its limitation lies in the fact that it may not be suitable for describing non-linear relationships present in the data. For such cases, nonlinear dimensionality reduction techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) [66] may be more appropriate.

1.1.2 *Learning Process*

Completing a ML project involves a structured series of interdependent stages that guide the entire process, from problem definition to model deployment. A typical Machine Learning process is shown in Figure 1.4. The initial phase involves precisely defining the problem while understanding its scope and objectives. Data collection and preparation ensues, where relevant datasets are gathered, cleaned, and transformed in order to be usable. Subsequently, a model is developed on the training data. This model is then tested

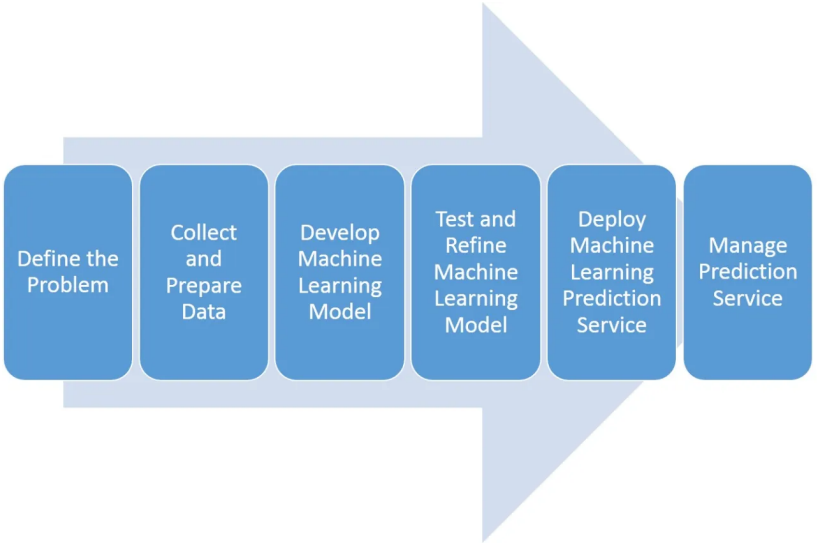


Figure 1.4: The diagram of a typical Machine Learning process. These steps are iterative to refine the Model. Image taken from Ref. [67]

on unseen data followed by a possible refinement. In the end, analyzing and deploying the best model for prediction. Though the final model is imperfect, the iterative nature of ML projects dictates continual refinement and enhancement of the model, thereby facilitating the realization of effective solutions. Each of these steps is critical for the project to succeed, but generally, extra attention for data curation and model development can take up most of the time and effort.

Data curation is an essential part of any ML solution and requires a great deal of human intervention. Effective data curation is a fundamental prerequisite for accurate and reliable model development, and contributes to the generalization and adaptability of models. The quantity of data ensures model robustness and effective generalization, a substantial dataset providing representative samples resulting in more accurate and adaptable predictions across diverse scenarios. High-quality data is the bedrock upon which accurate and effective ML models are constructed. Flawed data can indeed lead to biased or unreliable predictions, undermining the credibility and usefulness of the model's outcomes. Data curation involves three key sub-steps: sourcing, cleaning, and preprocessing. Sourcing data entails identifying and consolidating relevant data sources, joining multiple data sources, and rationalizing them into one dataset. Cleaning data is a crucial step due to the potential for missing, erroneous, or duplicated values, strategies for rectifying missing values encompass data source re-evaluation, removal of affected rows, or replacement with other values. The subsequent stage encompasses data preprocessing to transform data into an optimal format for machine learning models, a set of quantitative attributes. One of the key components is feature engineering [68], that is extracting a group of efficient machine-understandable representations from the original data by adding in domain knowledge. These meaningful attributes of data are referred to as features. In situations where specific data points lack certain features, feature imputation [69] becomes a valuable tool, it entails the systematic replacement or estimation of absent values. Methods encompass both single imputation strategies, such as mean or median replacement, or straightforward zero filling. Alternatively, the approach extends to multiple imputations, exemplified by predicting missing values based on modeling with existing data points. Another vital process involves

the encoding of categorical variables, for example, the target of the classification task. Two techniques are commonly used: ordinal encoding converts labels into distinct numeric values while one-hot encoding transforms unique values into binary arrays where the target value is represented as 1 while the other values are represented as zeros.

Model development refers to the process of creating, training, and refining a predictive or analytical model based on input data. Suitable algorithms are chosen depending on the problem type and data, rigorous training and tuning optimize the model, while evaluation and testing provide a clear understanding on how well the model generalizes to new, unseen data. The process typically involves the following steps. In the initial phase, one chooses choosing an appropriate machine learning algorithm or model architecture based on the nature of the problem and the characteristics of the data. Different algorithms may be more suitable for specific types of tasks and data. For instance, in a classification task, algorithms such as logistic regression [70], decision trees [56], SVM [57], or neural networks [58] come into consideration. Subsequent to algorithm selection, the training data is input into the chosen algorithm to teach the model. This training process entails the dynamic adjustment of the model's parameters, orchestrated with the objective of minimizing a predetermined loss function. Following this phase, is a key step in model development, particularly for supervised models, the model evaluation. It assesses the model's performance on a separate validation dataset using appropriate evaluation metrics. In tandem, model refinement follows to optimize the model's performance, through architectural modifications and tuning of hyperparameters. These hyperparameters represent attributes set prior to training that influence the learning process, such as learning rate and loss function. Techniques like grid search or random search are often used to find the best hyperparameters aiming for a cautious equilibrium. Overly complex models might overfit the training data and perform poorly on new data, while overly simple models might underfit and not capture the underlying patterns, proper model refinement can lead to improved predictive accuracy and a more reliable machine learning solution. Culminating, scoring the final model on an independent test dataset, its performance on test data is used to estimate how well the model can generalize to unseen data, and hence how well it might perform once deployed. Model development constitutes an iterative cycle of experimentation, validation, and refinement to create a model that effectively addresses the problem and provides valuable insights or predictions. It involves a trade-off between various factors, such as interpretability, computational complexity, and bias-variance (to strike a balance between a model that fits the training data well and generalizes well to new, unseen data).

1.1.3 *Scoring functions*

For learning tasks with the target output, scoring metrics serve as quantitative measures to assess different aspects of the performance and effectiveness of ML algorithms. They provide a comprehensive framework for comparing the model's predictions against the actual values within the labeled dataset. By distilling intricate performance nuances into quantifiable values, these metrics offer valuable insights into various dimensions of the model, enable comparisons between different algorithms, and aid in selecting the best suitable algorithm or hyperparameters for a specific task or problem. These metrics encompass a range of statistical measures that collectively illuminate the model's strengths and limitations. For regression tasks, scoring functions commonly include metrics such as mean squared error, mean absolute error, and R-squared [71]. For classification tasks,

Table 1.1: The confusion matrix of binary classification. It indicates different cases of predictions.

	Positive-Predicted	Negative-Predicted
Positive-Actual	True Positive (TP)	False Negative (FN)
Negative-Actual	False Positive (FP)	True Negative (TN)

on the contrary, confusion matrix [72], accuracy, precision, recall, F1-score [73], area under the receiver operating characteristic (ROC-AUC) curve [74], and the precision-recall curve are commonly used. Each of these metrics provides distinct insights into the model's capabilities, collectively offering a comprehensive framework for model evaluation. Some details about the metrics used in our work are introduced below.

The confusion matrix serves as a fundamental tool to summarize the model's predictions and their alignment with actual outcomes. For binary prediction, four possible cases are illustrated in Table 1.1. Considering our example task, the predictions for 20 emails are listed in Table 1.2, for the type of an email, 0 represents "non-spam" and 1 represents "spam". These predictions fall into four categories: 4 TP (correctly predicted "spam"), 2 FP (incorrectly predicted as "spam" when actually "non-spam"), 12 TN (correctly predicted "non-spam"), and 1 FN (incorrectly predicted as "non-spam" when actually "spam").

Accuracy quantifies the proportion of correctly classified instances out of the total predictions made by the model:

$$Accuracy = \frac{Correct\ classifications}{All\ classifications} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1.1}$$

So Precision, also called positive predictive value, focuses on the ratio of true positive predictions to the total number of positive predictions:

$$Precision = \frac{True\ positive}{Predicted\ positive} = \frac{TP}{TP + FP} \tag{1.2}$$

It highlights the model's ability to avoid false positive errors. On the other hand, recall, also known as sensitivity or true positive rate, calculates the ratio of true positive predictions to the total actual positives:

$$Recall = \frac{True\ positive}{Actual\ positive} = \frac{TP}{TP + FN} \tag{1.3}$$

It showcases the model's effectiveness in capturing all relevant positive instances among the actual positives. For the predictions listed in Table 1.2, the accuracy, precision and recall are respectively 0.85, 0.67 and 0.8. To distinguish "spam" from "non-spam", a high precision indicates that when the model labels an email as "spam," it is indeed likely to be a genuine spam email, minimizing the occurrence of false positives to avoid misclassification of important emails as spam, which potentially cause users to miss important communications. A high recall indicates that the model is proficient in identifying a substantial portion of genuine spam emails, minimizing the inconvenience of false negatives that leave actual spam emails undetected.

The F_1 score combines the precision and recall into a single value, defined in Eq. 1.4. It is an overall measure of a model's effectiveness that strikes a balance between correctly identifying positive instances while also minimizing false positives and false negatives.

Table 1.2: An example of predictions for 20 emails. In column "Actual type" and "Predicted type", 0 represents "non-spam" and 1 represents "spam". The predictions are made according to the default threshold, 0.5.

Email-id	Actual type	Predicted type	Predict-Prob0	Predict-Prob1
01	0	0	0.82	0.18
02	0	0	0.92	0.08
03	0	0	0.96	0.04
04	0	0	0.75	0.25
05	1	1	0.18	0.82
06	0	0	0.97	0.03
07	1	1	0.40	0.60
08	0	0	0.98	0.02
09	0	0	0.99	0.01
10	0	0	0.98	0.02
11	0	0	0.76	0.24
12	1	1	0.39	0.61
13	0	1	0.44	0.56
14	0	0	0.82	0.18
15	0	0	0.75	0.25
16	0	0	0.63	0.37
17	0	1	0.23	0.77
18	0	0	0.51	0.49
19	1	0	0.92	0.08
20	1	1	0.39	0.61

In other words, a good F1 score means that one is correctly identifying real threats and is not disturbed by false alarms, an F_1 score is considered perfect when it is 1, while the model is a total failure when it is 0. This balanced metric is particularly valuable when the consequences of false positives and false negatives are both significant. It can be a better measure than accuracy for an uneven class distribution.

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{1.4}$$

Additionally, when the probability scores or confidence values are provided for the predictions from the model, the receiver operating characteristic (ROC) curve and the precision-recall curve provide a visual representation of a model's performance across different probability thresholds. The ROC curve plots the true positive rate (sensitivity, recall) against the false positive rate at various probability thresholds. These thresholds are set based on the model's confidence in its predictions. For example, in the context of email spam detection, the model assigns a probability of 0.82 to email 05 being "spam" which means the model is 82% confident that the email is indeed spam. By varying the probability threshold, the ROC curve illustrates how the true positive rate and false positive rate change, allowing you to assess the model's performance across different levels of confidence. The ROC analysis offers a comprehensive view of a model's performance, enabling making informed decisions about its effectiveness and trade-offs between true positive and false positive rates. It is a valuable tool for evaluating the discriminative capabilities of classification models and selecting the most suitable threshold for the specific task at hand. The area under the ROC curve (ROC-AUC) is a numerical summary of the ROC curve's performance. It quantifies the overall discriminatory power of the model across all possible thresholds, a higher ROC-AUC indicates better performance. In the case of email spam detection, a higher ROC-AUC score indicates that the model is effective at predicting "spam" emails correctly. Complementing this, the precision-recall curve artfully captures the interplay between precision and recall, when the threshold is higher, precision tends to increase, but recall may decrease, conversely, when the threshold is lowered, recall may improve, but precision might decrease. It is especially valuable in addressing imbalanced datasets or when false positives and false negatives have different impacts on the problem. The ROC curve and precision-recall curve of the example predictions in Table 1.2 are illustrated in Figure 1.5.

Collectively, these metrics empower practitioners to comprehensively evaluate the performance of binary classification models. It is essential to understand the nuances of each metric and choose the one that aligns best with the desired outcomes. The most relevant metric for a given task enables informed decisions in selecting, optimizing, and refining models to achieve desired outcomes in real-world scenarios.

1.1.4 *Nested cross-validation*

Train-test split is a fundamental technique in ML used to evaluate the performance of a model. The primary goal of this approach is to simulate how the model is likely to perform on new, previously unseen data, thus assessing its ability to generalize on new instances. Typically, the dataset is partitioned into a larger training set, used to teach the model the underlying patterns and relationships in the data, and a smaller test set reserved for assessment. The test set serves as a proxy for evaluating the model's real-world

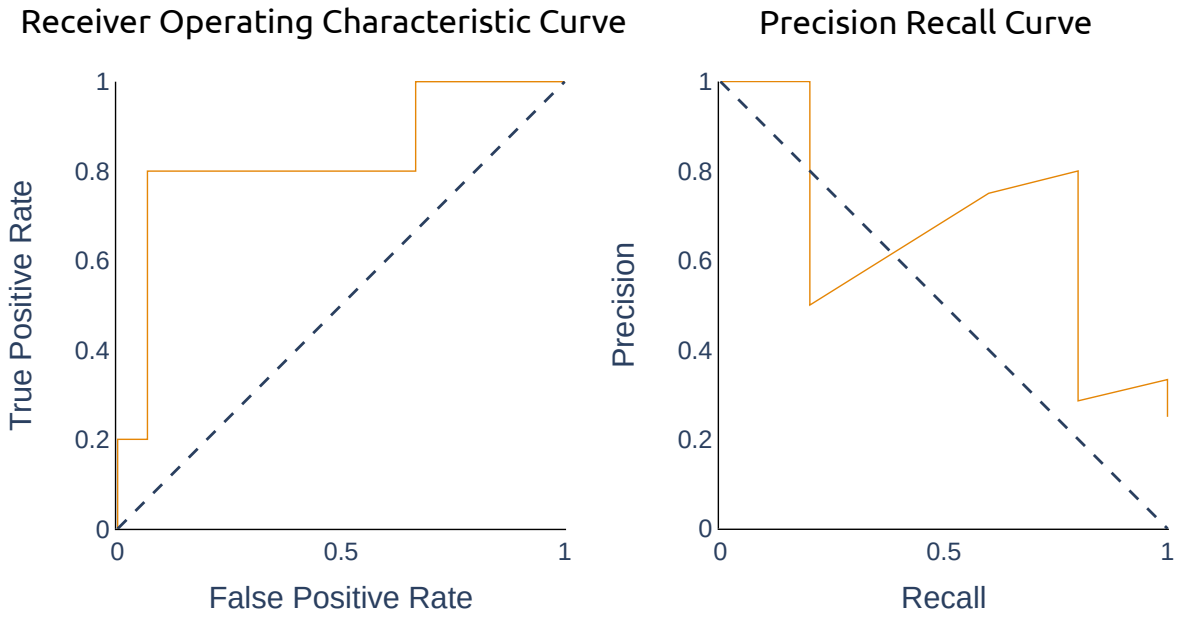


Figure 1.5: Receiver operating characteristic (ROC) curve and precision-recall curve of the prediction in Table 1.2.

performance, providing insights into its effectiveness and potential shortcomings. This is a crucial step in model development as it helps to reliably estimate the quality of the model and prevent the issue of overfitting, where the model memorizes the training data, even noise and random fluctuations, rather than learning general patterns. It is important to note that the train-test split should be carefully chosen to ensure a representative distribution of data and maintain the integrity of the evaluation process. The splitting ratio depends on factors such as the dataset size, the complexity of the problem, and the desired level of confidence in the model's performance.

Stratified train-test split is a common method to use when dealing with classification tasks on imbalanced datasets. In a stratified train-test split, the data is divided into training and testing subsets while ensuring that the class distribution (the proportion of different categories) remains consistent across both subsets.

Train-validation-test split takes the process a step further, considering the hyper-parameter tuning of the model, a validation set is split from the training set. Figure 1.6 shows the procedure. Multiple models, each configured with different parameters, are trained on the remaining training set and scored on the validation part. Subsequently, the set of hyper-parameters that behaved best on the validation part is applied to retrain the model on the entire training dataset. Finally, this model is evaluated on the test set to assess its ultimate performance.

However, there is a risk that this result may depend on a particular random choice for the pair of train(-validation)-test sets. Cross-validation [75] is a solution to erase the impact of the choice for the test set, since one repeats the train-test procedure on different parts of the dataset and averages the score. The most common form of cross-validation is k-fold cross-validation. In this approach, the dataset is divided into k equal-sized subsets (folds), with one fold reserved as the test set and the remaining k-1 folds used as the training set. This process is repeated k times, with each fold serving as the test set exactly once. The

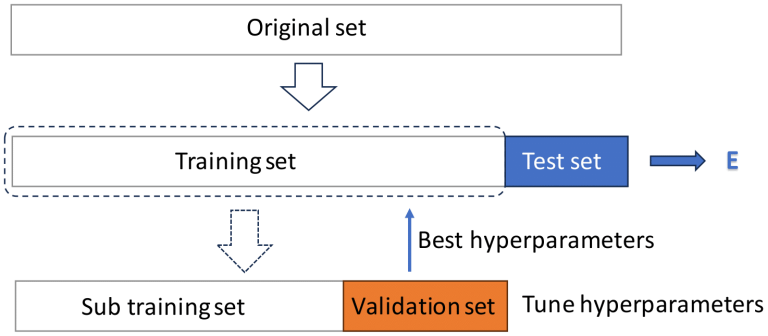


Figure 1.6: Train-validation-test split. The optimal combination of hyperparameters is selected on the validation set, and the generalization error of the model is estimated on the test set.

performance metrics obtained from these iterations are then averaged to provide a more reliable estimate of the model's performance.

Nested cross-validation (NCV) is used for further reducing the influence of the validation set, and mitigate the risk of overfitting during hyperparameter tuning, where a model performs well on the training data but poorly on new data. The method involves a dual-layered cross-validation process. The outer loop follows the traditional k-fold cross-validation approach. The inner loop, nested within the outer loop, focuses on hyperparameter tuning. For each iteration of the outer loop, the training data is further divided into m folds. Various combinations of hyperparameters are tested on these inner folds, and the model's performance is assessed using appropriate metrics. The best hyperparameter configuration is selected based on the performance results within the inner loop. Figure 1.7 shows an example of a 5×3-fold NCV procedure. This nested process ensures that each hyperparameter configuration is evaluated across different training and validation subsets, providing a more robust assessment of the model's performance and its sensitivity to hyperparameter changes. The final performance estimate is obtained by aggregating the results from the outer loop, capturing the model's ability to generalize to new, unseen data. While NCV is more computationally intensive compared to standard cross-validation, it provides a more reliable and unbiased estimate of a model's performance. Using multiple validation sets and multiple rounds of testing reduces the likelihood of selecting hyperparameters that perform well by chance on a particular validation set. Instead, it aids in selecting models that consistently lead to better generalization across different data distributions, providing a more robust evaluation of its performance in real-world applications.

1.2 Machine learning with topological electronic materials

1.2.1 *Topological electronic materials*

The key characteristic of topological materials is the presence of topological invariants [77, 78], which are mathematical quantities that remain unchanged even under continuous deformations of the electronic structure. For example, Thouless et al. [77] in 1982 found that the electronic wave function of electrons in solids has a topological invariant called "Chern number". In a two-dimensional material, the Berry curvature describes how quantum states evolve as a parameter (typically the crystal momentum) is varied. The Chern number is calculated by integrating the Berry curvature over the Brillouin zone. It is an integer value

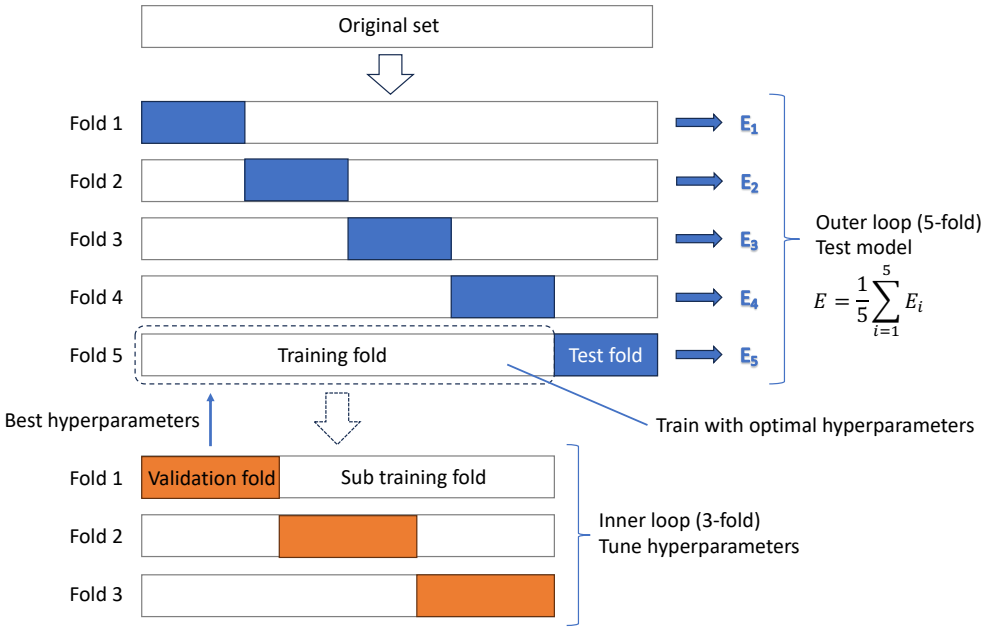
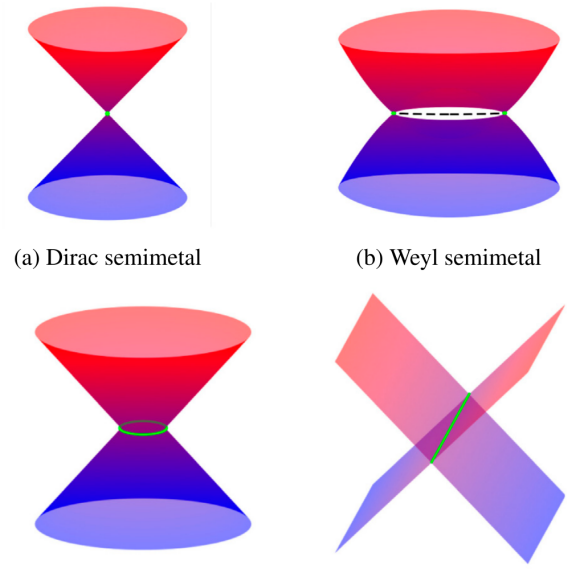


Figure 1.7: Nested 5×3-fold cross-validation for algorithm selection. The optimal combination of hyperparameters is selected during the inner 3-fold cross-validation loop, and the generalization error of the model is estimated on the outer 5-fold cross-validation loop. Image adjusted from Ref. [76]

that describes the quantization of the Hall conductance and reflects the number of chiral edge states in these materials [79, 80]. If the Chern number is non-zero, then the system is a Chern insulator. Since 2005 [78], in a series of theoretical work, researchers have gradually realized that in addition to the Chern number, almost any symmetry of the crystal, such as time-reversal invariance, mirror reflection, etc., may have its corresponding topological invariants [81–84]. They provide insight into the underlying topology of the electronic bands, enabling the identification of materials with unique electronic properties [12]. The existence of topological invariants protects their electronic states from scattering or backscattering, thus making them robust against certain types of defects and disorder. This ensures that the conducting edge or surface states of these materials remain stable even in the presence of perturbations.

TSMs and TIs are two main classes of topological materials. TSMs have protected electronic states, such as Weyl or Dirac fermions [85–87], at specific points or lines in their electronic band structure. Figure 1.8 shows a schema of three types TSMs. In Weyl semimetals, the band structure features Weyl nodes, which are points in the Brillouin zone where the electronic bands touch linearly. Dirac semimetals, in contrast, exhibit Dirac nodes characterized by linear band crossings, resulting in the formation of Dirac fermions. Nodal-line semimetals are characterized by the presence of one-dimensional lines in the Brillouin zone where the energy bands touch. Unlike conventional semimetals, where band crossings can be unstable and easily gapped out, topological semimetals have protected band degeneracies due to specific symmetries or topology. These protected crossings are stable against small perturbations, making them robust and distinctively different from ordinary semimetals. TIs are materials with insulating bulk properties, characterized by a finite band gap resulting from strong spin-orbit coupling and time-reversal symmetry [10, 11].



(a) Dirac semimetal (b) Weyl semimetal

(c) Nodal line semimetals

Figure 1.8: Schematic illustration of band crossing in momentum space of 3D topological semimetals. Image taken from Ref. [92].

However, TIs exhibit unique conducting surface or edge states, known as topological surface states or Dirac surface states. These surface states are protected by global, nontrivial topological invariants, ensuring that they remain immune to backscattering, allowing for the efficient transport of charge or spin currents, and are highly robust against impurities, defects, and disorder. Figure 1.9 shows a schema of the spin-polarized edge channels in HgTe [88], a two-dimensional TI. This remarkable feature sets TIs apart from ordinary insulators and holds significant potential for applications in various fields of research and technology. According to various symmetries, topological invariants, and the specific nature of protected surface or edge states in these materials, TIs are categorized into various types. For instance, time-reversal invariant TIs obey time-reversal symmetry and are characterized by a nontrivial Z2 topological order; Z2 TIs are characterized by a Z2 topological invariant, determining whether the material is in a topologically nontrivial or trivial phase; TCIs exhibit protected surface states due to the presence of crystal symmetries, unlike the Z2 TIs, TCIs rely on specific crystalline symmetries to protect their topological surface states. Accompanied by theoretical predictions, experimental research has successfully identified and characterized topological materials in various systems. For instance, materials like cadmium arsenide (Cd_3As_2) [89, 90] and tantalum arsenide (TaAs) [91] have been identified as Weyl semimetals, where Weyl nodes appear at points in their band structure. Bismuth selenide (Bi_2Se_3) and bismuth telluride (Bi_2Te_3) [16, 17, 19] are well-studied TIs, and one was among the first materials where the topological insulating behavior was experimentally observed.

In the field of topological materials, many computational physicists and materials scientists are actively searching for new and improved topological materials, including TIs [16, 17, 19, 82, 88, 93], TCIs [83, 94], Weyl semimetals [89–91, 95, 96], Dirac semimetals [97], and nodal-line semimetals [98]. Substantial advancements have been made, yet significant challenges persist, preventing the realization of "ideal" topological materials.

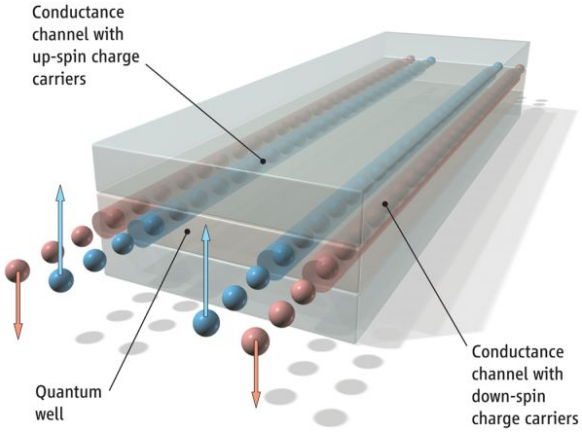


Figure 1.9: Spin-polarized edge channels in HgTe quantum well. Image taken from Ref. [88].

While some topological materials have been identified, many topological materials are synthesized under specific conditions, which might not always be stable or feasible for practical applications. Some topological effects are prominent at low temperatures or require extreme conditions, making them challenging to harness for practical applications. Designing topological materials that exhibit the desired topological properties while satisfying additional material requirements (e.g., mechanical, thermal, optical properties) adds additional complexity to the process.

Despite all these challenges, ongoing research continues to pave the way for practical applications of topological materials, and the detection and realization of topological materials remain an active area of research. Topological quantum chemistry theory (TQC) draws upon the principles of both condensed matter physics and quantum chemistry for analyzing the connectivity of electronic energy bands in the Brillouin zone to identify topological phase transitions and predict novel materials with specific electronic properties [21, 23–25, 99]. It provides a unified framework for treating all topological phases arising from crystalline symmetries for nonmagnetic materials. Relying on the notion of elementary band representations and compatibility relations, how bands can connect across the Brillouin zone is constrained.

In 2019, adopting similar high-throughput algorithms that combined DFT and TQC, three independent but generally consistent databases were published together. They filtered crystal structures from ICSD by different criteria, then computed and analyzed their band structure including spin–orbit coupling. Vergniory et al. [30] and Zhang et al. [28] calculated more than 25,000 nonmagnetic materials and found that approximately 30% of them were topological, whereas Tang et al. [29] found several hundreds of topological materials from thousands of candidates. Zhang et al. swept through a total of 39,519 materials that are simultaneously registered in the Material Project (MP) [100] and the ICSD and displayed the topological properties of 26,120 materials in Materiae. The band structures are computed using the Vienna ab initio simulation package (VASP) [101, 102]. What is special about Zhang et al. is that they also provided results without spin–orbit coupling. Because light atoms usually exhibit negligible spin-orbit coupling, for certain materials composed of light atoms, evaluating their topology without spin-orbit coupling is more physically relevant. Whereas Vergniory et al. picked the 'high-quality' ICSD entries (materials for which the atomic positions and structure are measured very accurately) and filtered with 'stoichiometric', 'few atoms (<30)', 'not alloy', and 'not-usual magnetic atoms' to find

26,938 stoichiometric compounds. They used VASP for DFT calculations and analyzed the topological properties when the symmetry was lowered into subgroups and displayed their results on TMD. In 2022, they extended it to include 73,234 entries with converged electronic structures resulting from all 96,196 processable ICSD entries. Results both with and without incorporating the effects of spin-orbit coupling are displayed [31]. Tang et al. [29] picked 17,258 materials with relatively clean Fermi surfaces (i.e. systems where the size of the trivial pockets is smaller than approximately 10% of the Brillouin zone volume) and completed the ab initio calculations using WIEN2K [103].

DFT is one of the most popular and versatile modeling methods to investigate materials in a theoretical way. Guided by the Hohenberg-Kohn theorems [5], which assert that the ground-state properties of a many-electron system are uniquely determined by its electron density n, DFT investigates the electronic structure of many-body systems by focusing on the electron density instead of the much more complicated many-electron wavefunction. In the Kohn-Sham (KS) [6] formulation of DFT, one assumes the existence of a fictitious system of non-interacting electrons governed by an effective mean-field Hamiltonian whose ground state reproduces the density of the many-body system. The KS Hamiltonian consists of the kinetic energy of the independent electrons, the Hartree potential v_H describing the classical electrostatic repulsion plus the so-called exchange-correlation potential v_{xc} that includes all the remaining quantum mechanical effects that are not accounted for by the first two terms. Very schematically the KS equations can be summarized as follows:

$$\left[-\frac{1}{2}\Delta + v_{ext}(r) + v_H[n](r) + v_{xc}[n](r)\right]\psi_i(r) = \epsilon_i\psi_i(r) \tag{1.5}$$

where $\psi_i(r)$ are the single-particle KS orbitals with the associated eigenvalues ϵ_i while the electronic density is given by

$$n(r) = \sum_{i}^{occ} |\psi_i(r)|^2. \tag{1.6}$$

The obtained equations have the form of independent-particle equations, but as the Hartree and exchange-correlation potential depends on the density the solution must be found self-consistently.

While the above symmetry-based methods combined with DFT calculations have achieved notable successes, they are limited to nonmagnetic crystal structures with well-defined symmetry, leaving a large space undetected. Meanwhile, symmetry indicators are not capable of diagnosing topology beyond eigenvalues, so a material that has zero indicators can still possibly be an NTM [21, 23], for instance, the first experimentally observed Weyl semimetal, tantalum arsenide (TaAs) [95]. (Wilson loop [104] helps to characterize the topology of electronic band structures in these materials but presents computational challenges.)

Furthermore, our intuition into the material's topological nature remains limited, in the experimental field, the essential factors of materials that may affect the topological properties are still unclear, which hinders the process of designing topological materials. ML techniques, a method that has developed rapidly in recent years, provide a novel view to use existing data to extract knowledge from them and explore the hidden information. It could bypass the heavy computation and potentially infer the quantities that are essential to the topology of a material, potentially overcoming the above shortcomings.

1.2.2 *Review of the literature*

In the realm of topological materials, several studies have been conducted to explore the application of ML algorithms. Parts of the research are about learning from the Hamiltonian. In 2017, Zhang et al. [105] introduced a method called Quantum Loop Topography that transforms Hamiltonian or wave function into a multidimensional image. It is then utilized as input data for a neural network model to distinguish the Chern insulator and the fractional Chern insulator from TrI. In 2020 they further built three simple neural network models learning transitions between topologically trivial states and the Chern insulator, Z2 TI, or Z2 quantum spin liquid to explore the interpretability in these cases [106]. Another research in 2018 [107] trained a neural network for predicting topological winding numbers from Hamiltonians of one-dimensional insulators. They found the model did learn the discrete version of the winding number formula. An unsupervised approach has been explored in 2020 [108], where the authors use k-means clustering to identify paths of adiabatic deformations between Hamiltonians thereby allowing the diagnosing of trivial and topological phases by comparing with a generally designated set of TrIs.

There are also works on learning from ab initio data. Utilizing compressed sensing [109], a study [110] investigated 220 functionalized 2D honeycomb-lattice materials isoelectronic to graphene. They identified a descriptor for the Z2 invariant to distinguish metals, TrIs, and 2D TIs. Furthermore, an approach developed based on the compressed-sensing technique, SISSO (Sure Independence Screening and Sparsifying Operator) [52] was applied for TI/TrI classification. One [111] study considered 243 alloyed tetradymites to learn a 2D descriptor. It solely relies on the atomic number and electronegativity of the constituent atoms. Another [112] work focused on 116 ternary half-Heusler compounds to learn a 2D descriptor defined by the atomic number, the number of valence electrons, and the Pauli electronegativity of the constituent atoms. Several works used a larger dataset with general materials for training. Claussen et al. [113] trained a model of GBTs with 35,009 symmetry-based entries on TMD. They tested the effect of different features and built a full model that reached an accuracy of 87.0% for classifying materials into 5 subclasses: TrIs, two types of TSMs: Enforced Semimetal (ES) or Enforced Semimetal with Fermi Degeneracy (ESFD), and two types of TIs: Split Extended Brillouin Zone Representation (SEBR) or Not Equal to a Linear Combination (NLC). In 2021, a method [114] combined SISSO to construct a feature space and then applied GBTs to predict the topological classification of insulating materials has been developed based on a dataset of TIs and TrIs collected from 2D materials databases [115–119]. Andrejevic et al. [120] utilize 16,458 inorganic materials with computed X-ray absorption near-edge structure (XANES) spectra from ICSD and topological class from TMD to develop a neural network classifier capable of distinguishing topological class NTM and TrI based on XANES signatures. In 2023, Ma et al. [121] introduced a technique named "Topogivity" to categorize NTM and TrI based on a dataset of 9,026 materials from Tang [29]. They employed SVM to learn the topogivities of elements and constructed a quantity that relies on the chemical formula of the material, indicating information about its topological class. This approach achieved an average accuracy of 82.7% in an 11×10-fold NCV procedure.

Table 1.3: Atomic, bond, and state attributes used for crystals. Table adapted from Ref. [49]

Level	Attributes name	Description
Atom	Z	The atomic number of element (1-94)
Bond	Spatial distance	Distance r valued on Gaussian basis, $\exp(-(r-r_0)^2/\sigma^2)$, where r_0 takes values at 100 locations linearly placed between 0 and 5 and the width $\sigma = 0.5$.
State	Two zeros	Placeholder for global information exchange.

1.2.3 *Adopted Algorithms*

However, in the above studies, only generic ML algorithms were employed, while some newly designed methods [36, 49–51] which have demonstrated better performance, were overlooked. Moreover, the outcomes from these studies lack comparability due to the diverse range of crystal systems considered (encompassing specific classes of systems, 2D materials, or bulk 3D materials) and variations in the utilized databases, specifically: the defined classes of materials and the types of features incorporated in model construction. To address these issues, we compared several ML techniques on a unified dataset for predicting the DFT-based topological type from a crystal structure. Eight algorithms are used in this work: three well-established algorithms in the field of material science: Materials graph network (MEGNet) [49], AutoML + Matminer (Automatminer) [50], and Material optimal descriptor network (MODNet) [36]; four generic ML models: SISSO, RF, GBTs, and t-SNE; and a special method developed for topological materials: Topogivity. The following sections will give a brief introduction to these methods.

MEGNet

MEGNet [49] is an implementation of DeepMind's graph networks [122] for machine learning molecules and crystals in materials science. It harnesses the principles of graph neural networks to model complex interatomic interactions in materials. Materials are represented as graphs, where atoms are nodes and their interactions are edges. The initial graph is constructed by the set of atomic attributes $V = \{v_i\}_{i=1:N}^{v}$, bond attributes $E = \{(e_k, r_k, s_k)\}_{k=1:N}^{e}$ and global state attributes $\boldsymbol{u}$. Table 1.3 summarizes the full set of atomic, bond, and state attributes used for crystals. A graph network module (Figure 1.10) contains a series of update operations that map an input graph $G = (E, V, \boldsymbol{u})$ to an output graph $G' = (E', V', \boldsymbol{u}')$. This graph-based representation enables MEGNet to capture both short-range and long-range interactions, making it particularly effective for materials with diverse bonding patterns. The MEGNet architecture integrates atomic features and their relationships, allowing it to learn intricate material behaviors, such as electronic structures and chemical reactions.

According to the result of the benchmark test with various tasks done in MatBench [50], MEGNet performs well when the database is larger (more than 10,000 items), but the accuracy crucially depends on the quantity of the available data. Another limitation is that not all the materials can be processed, MEGNet requires a structure that does not contain isolated atoms as input to construct the initial graph.

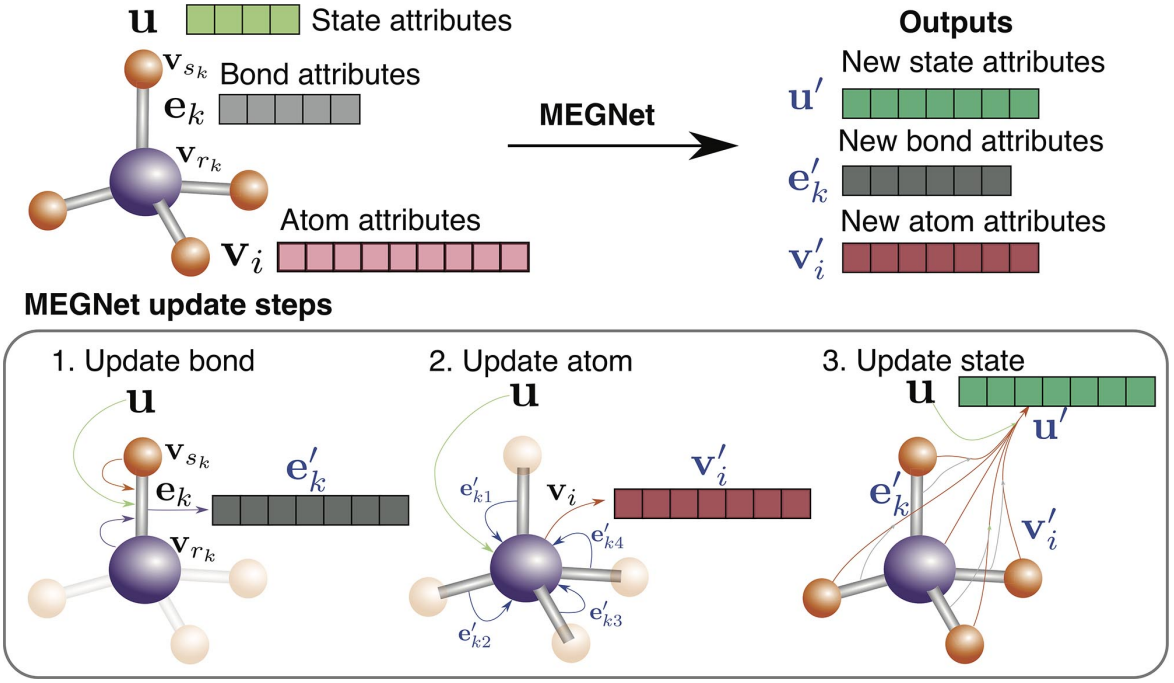


Figure 1.10: Overview of a MEGNet module. From the initial graph, it first updates the bond attributes, this process involves a flow of information originating from the atoms composing the bond, the state attributes, and the preceding bond attributes to the new bond attributes. Analogously, the subsequent two steps update the atomic and global state attributes. Ultimately, these iterative steps culminate in a new graph representation. Image taken from Ref. [49]

Matminer

Matminer [47] is utilized in conjunction with the next 6 algorithms: Automatminer, MODNet, SISSO, RF, GBTs, and t-SNE. An overview of it is displayed in Figure 1.11. It is a package for data mining the properties of materials containing three main components for accomplishing the following tasks: (i) obtaining ready-made datasets or retrieving materials properties from various repositories, (ii) featurizing meaningful numerical descriptors from complex materials attributes (such as crystal structure and chemical composition), and (iii) visualizing materials by the outcomes of data mining. Matminer does not provide ML routines itself but serves to facilitate the integration between materials data and various external machine learning libraries such as scikit-learn [123] and Keras [124].

In this work, functions of feature extraction are mainly used. A single crystal structure is transformed into a set of numerical descriptors to input into algorithms for further training. For example, given a crystal structure of face-centered cubic palladium, the GlobalSymmetryFeatures featurizer can give its ['spacegroup_num', 'crystal_system', 'crystal_system_int', 'is_centrosymmetric'] as [225, 'cubic', 1, True]. As shown in Figure 1.12, a total of 47 featurizers adapted from scientific publications are provided by Matminer, grouped into five submodules according to the object to generate features from: chemical compositions, crystal sites and structures, electronic bandstructure and density of states (DOS). Some brief introduction about different features used in this work are explained below.

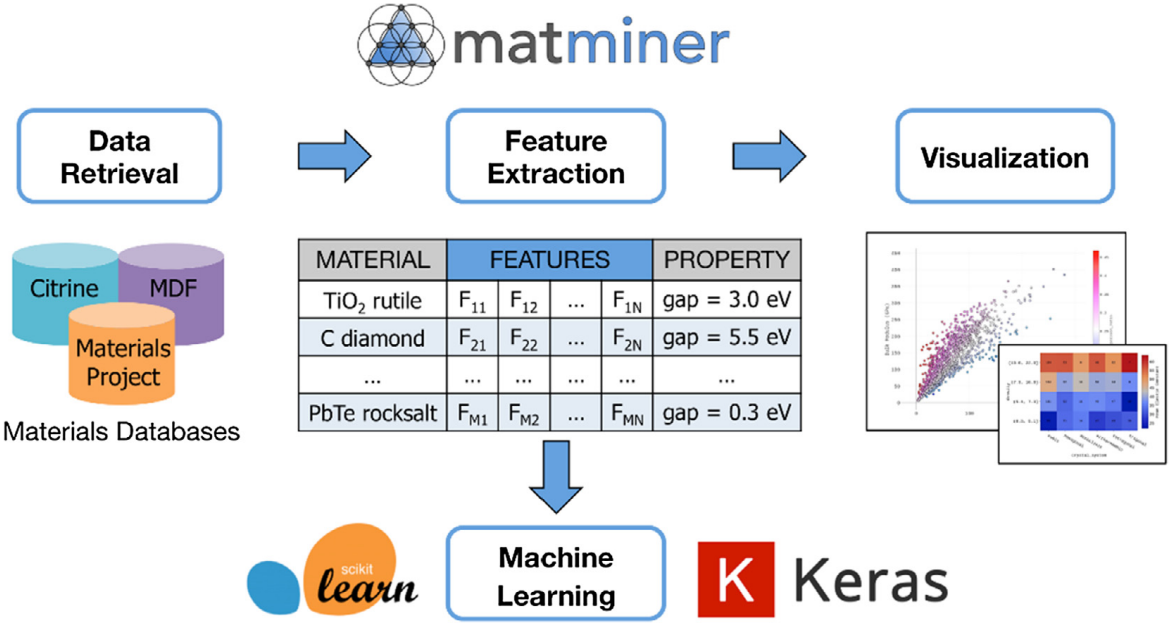


Figure 1.11: Overview of Matminer. It includes three main capabilities: (i)data retrieval from materials databases, (ii)feature extraction from materials, (iii)feature visualization. Image taken from Ref. [47]

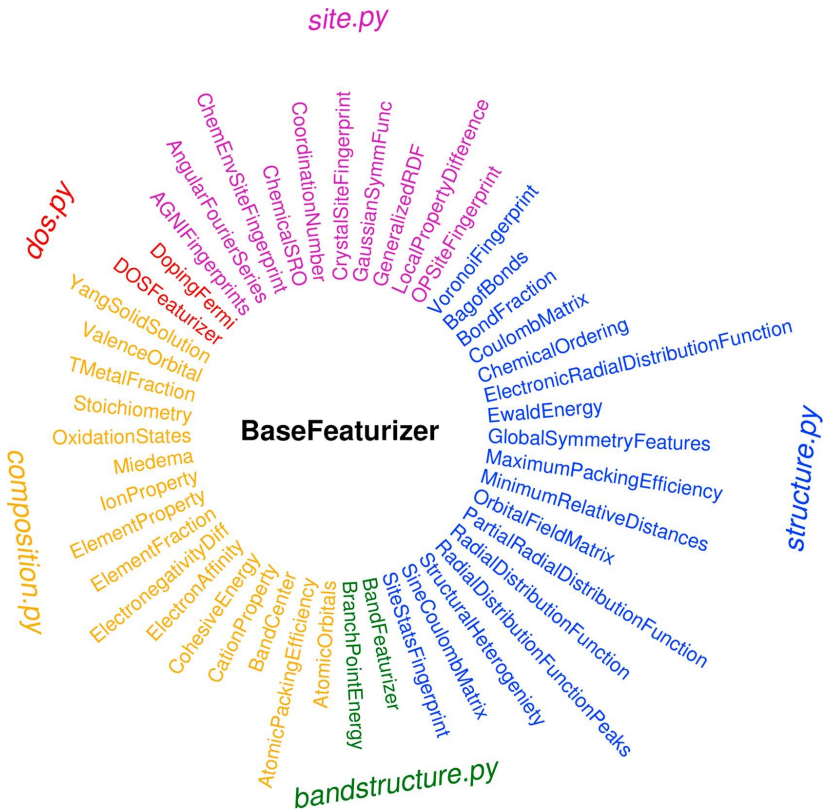


Figure 1.12: Overview of featurizers in Matminer. They transform a material attribute into a group of numerical descriptors. Image taken from Ref. [47]

From a composition object, the following composition-based features mainly create elements-related descriptors, the value of the properties or their different statistics, such as the mean atomic number of the constituent elements, the maximum oxidation state.

ElementFraction computes the atomic fraction of each element in the composition. A full vector indexed following the atomic number order, with these fractions and zero for the remaining atoms is generated.

ElementProperty generates a list of statistics concerning the elemental properties of the constituent elements, such as the maximum atomic number (or row, column), and the mean number of electrons in the s, p, d, or f-orbitals. Four presets are provided: "matminer" preset [125], "deml" preset [126], “matscholar” preset [127], and “magpie” preset [128].

Stoichiometry represents a composition as various L^p norms of the elemental fractions vector V=[V_1, ..., V_n], where V_i goes through the constituent elements. [128]

AtomicOrbitals according to Ref. [129]. The atomic orbital energies of the composition can lead to an estimation of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). It provides the HOMO/LUMO character (s, p, d, or f) and the respective HOMO/LUMO element.

AtomicPackingEfficiency delineated in Ref. [130] provides two features: atomic packing efficiency and L_2 distance. Atomic packing efficiency quantifies the difference between the ratio of radii from the central atom to neighboring atoms and the ideal ratio of other atoms with optimal packing efficiency. L_2 distance is computed between the material and a k-cluster whose packing efficiency is less than the ideal.

BandCenter [131] estimates the absolute position of the band center found by a fundamental geometric mean calculation involving electronegativities.

CohesiveEnergy [132] is the amount of energy (quantified per unit cell) necessary for the formation of a crystal from the elementary constituents.

ElectronAffinity [126] is defined as $E_{aff} = a_i E_{aff,i} \cdot Q_{anion,i}$, where a_i is the molar fraction of the composing anions, $E_{aff,i}$ is the affinity of anion i, and $Q_{anion,i}$ is the formal charge.

ElectronegativityDiff [126] computes the difference between the electronegativity of each cation i and the sum of the electronegativity of all anions, then generate statistics of them as the final features.

IonProperty is introduced in Ref. [128], providing three functions. The first attribute is a boolean that indicates if it is possible to form a neutral ionic compound. Then based on the difference in electronegativity between each pair of atoms, computes the maximum and average ionic character of the composing atomic bonds.

Miedema provides the formation enthalpy (eV/atom) of intermetallic compounds, solid solutions, and amorphous phases calculated based on the semi-empirical Miedema model (refer to Ref. [133] for comprehensive details).

OxidationStates is a simple function that gathers the oxidation states for each element of the composition and computes a variety of their statistics, such as mean and variance.

ValenceOrbital [128] evaluates the valence electrons and their orbitals of each constituent element. A range of statistical analyses is then performed on them, to calculate metrics about these valence electrons, such as the maximum of p-electrons, and the fraction of d-electrons.

From a structure object, in addition to these composition-based features, some spatial information about the crystal structure can also be extracted.

GlobalSymmetryFeatures describes three features related to the crystal symmetry: the spacegroup number, the crystal system, and a boolean signifying the presence or absence of an inversion symmetry operation in the crystal structure.

DensityFeatures determines three features associated with material density characteristics. The first feature concerns the density of the material, the mass per unit volume. The second is the volume allocated per atom, the total unit cell volume divided by the number of constituent sites within it. The third is the packing fraction, the summation of the atomic volumes, estimated through the formula $\frac{4}{3}\pi r_i^3$, with r_i representing the radius of each species, divided by the total volume of the unit cell.

MaximumPackingEfficiency, detailed in Ref. [134], is a feature that quantifies the maximum possible packing efficiency within this crystal lattice. It uses a Voronoi tessellation to determine the largest radius each atom can have before touching any one of its neighbors, thus assess the largest volume fraction occupied by the atoms within the unit cell.

CoulombMatrix captures intricate electrostatic Coulomb interactions within a structure or a molecule, presented as a matrix M. It was initially introduced in Ref. [135] for organic molecules. With the nuclear charge Z_i and the position R_i of atom i, matrix M is defined by $M_{ij,i\neq j} = \frac{Z_i \cdot Z_j}{||R_i - R_j||}$ as off-diagonal elements and $M_{ii} = 0.5 Z_i^{2.4}$ as the diagonal elements. To employ it effectively in ML algorithms, it is transformed into a fixed-length vector which consists of the eigenvalues of the Coulomb matrix in descending order, and zero-padding to the end.

SineCoulombMatrix is introduced in Ref. [136]. It expands the Coulomb Matrix in modeling organic molecules to encompass periodic crystals.

The introduction of the above features is not intended to provide an exhaustive list of the features we employed but rather to offer some physical explanations. Several of these features will be elaborated upon during the interpretation of the results. A group of example materials featurized with some mentioned attributes are listed in Appendix A.

Automatminer

Automatminer [50] is an automatic tool developed from Matminer for creating complete ML pipelines for materials science. It can be employed across datasets with materials in the form of either chemical composition, crystal structure, or electronic bandstructure information. The routine composed of four general stages is shown in Figure 1.13. First is the automatic featurization with Matminer to create and assign potentially relevant features for each sample. To ensure the given featurizers are able to produce non-null descriptors, a precheck functionality is used to ensure that only a featurizer valid for more than 90% input objects will be utilized. Next is the data cleaning stage, which handles errors in the dataset (e.g., imputing unknown values) and encodes categorical features to prepare the input for ML. The third step involves feature reduction, which employs one or more dimensionality reduction algorithms to eliminate redundant or linearly dependent feature sets. Finally, the AutoML stage is applied to search for optimal configurations in ML pipelines. The TPOT (Tree-based Pipeline Optimization Tool) library [137] is used for prototyping and validating internal ML pipelines. Thus the best ML model will be provided for predictions. Automatminer is designed to be easy to use, it provides several preset pipelines that can be considered as black boxes to perform typical training steps and produce a machine learning model, such as the "debug" preset for quick predictions, the "express" preset with a moderate degree of accuracy, and the "heavy" preset that involves more featurization and

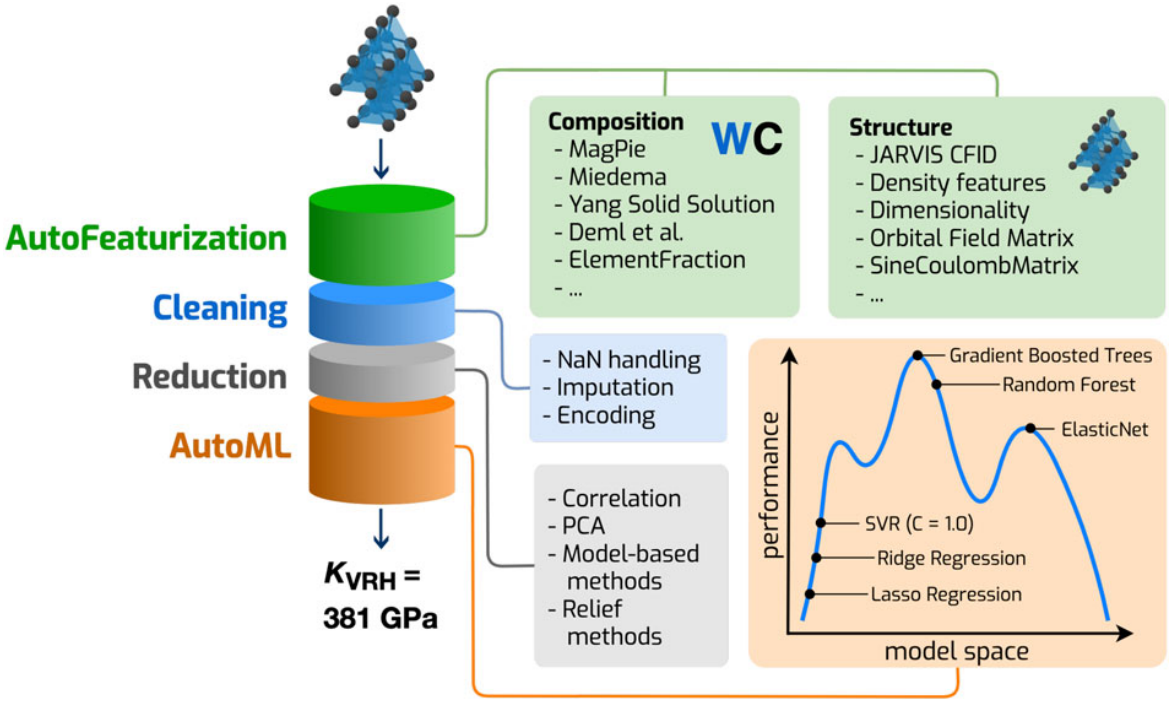


Figure 1.13: The Automatminer pipeline. Associated to the regression task for learning the bulk modulus. The procedure proceeds with the following steps: autofeaturization to extract features from the composition and structure of the material, data cleaning, and feature reduction stages to prepare the input matrices, and searching and building the optimal model for prediction. Image taken from Ref. [50]

longer training times. To facilitate end-user needs, customization of each stage to modify a preset pipeline or design a new pipeline is also supported.

MODNet

MODNet [36] is a supervised machine learning framework based on a feedforward neural network. It is designed to simultaneously learn multiple material properties using joint learning (single property learning is also supported). The ordinary routine of MODNet is to start with primitive materials, the composition or crystal structure, employing one of the predefined sets of featurizers from Matminer for feature extraction. Then on a bunch of numerical descriptors, MODNet provides a feature reduction method based on the normalized mutual information (NMI) and relevance and redundancy (RR) score. This helps to identify and retain the most informative and relevant features. Finally, a simple model or an ensemble model can be used to complete training and predicting, where an estimate of uncertainty is provided by the ensemble model. However, MODNet also allows users to construct a MODData object directly using other descriptors, such as custom-designed ones, and continue to further steps, either processing feature reduction or training the model. The architecture of the MODNet model is depicted in Figure 1.14.

A brief introduction of the NMI and RR score are given below. The information entropy $H(X)$ of a continuous random variable X is defined as:

$$\mathbf{H(X) = -\int P_X(x) log(P_X(x)) dx} \tag{1.7}$$

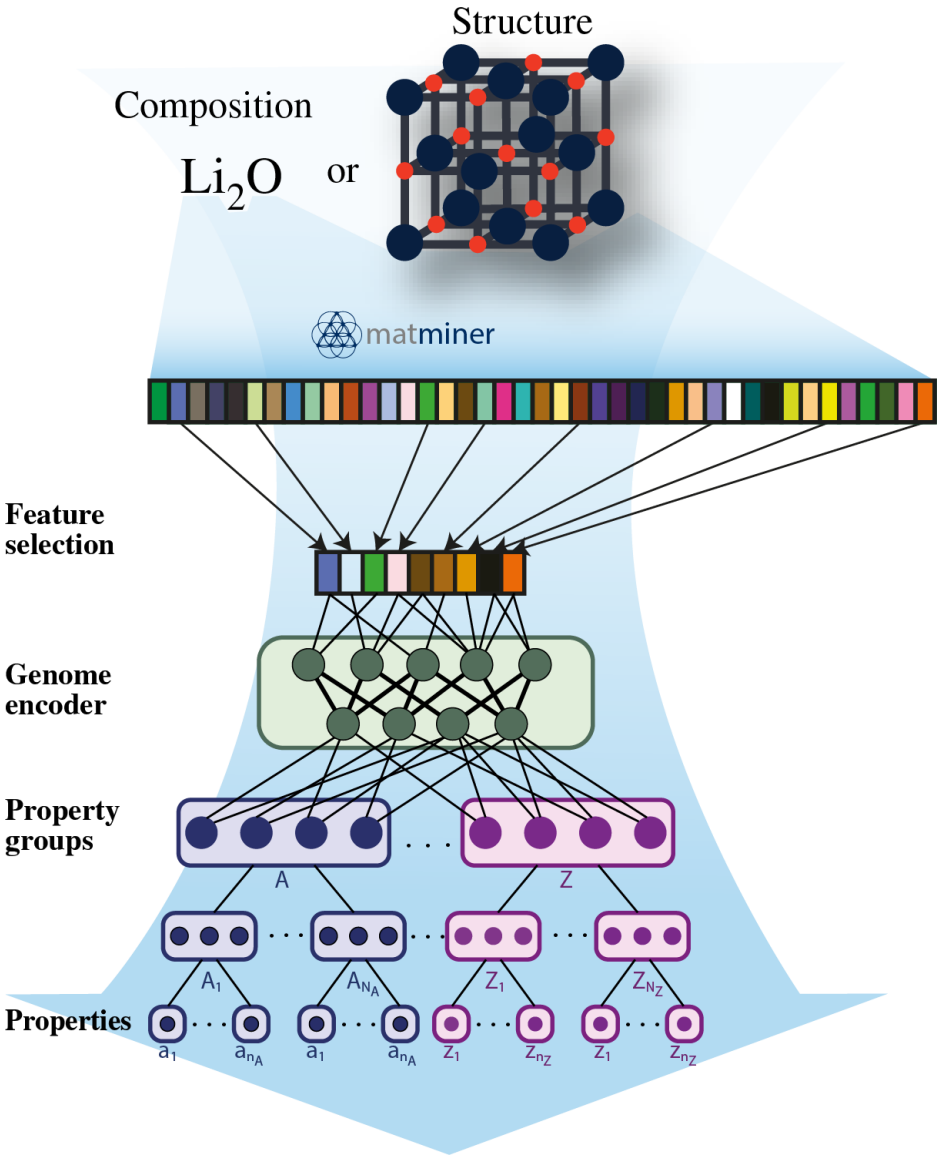


Figure 1.14: Schematic of the MODNet model when employed to learn multiple properties. Following the feature engineering, a hierarchical neural network is used. The initial green block of the neural network encodes the material into an all-round vector. Various properties are organized into groups from A to Z according to similar nature (e.g., mechanical, optical, etc.), and further blocks re-encode to get a more target-specific representation. Finally, learning is done for each property. Image taken from Ref. [36]

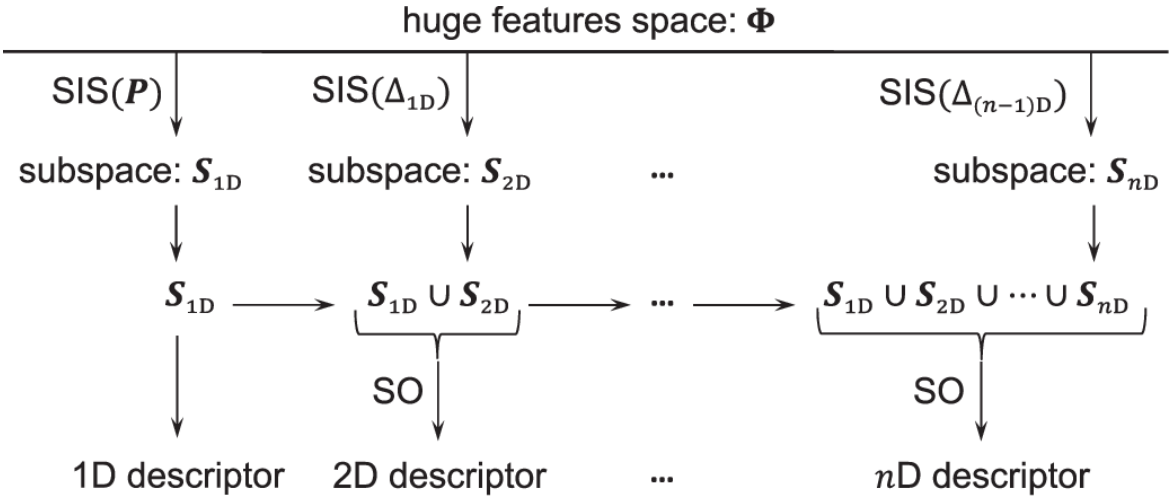


Figure 1.15: Overview of SISSO. From a huge feature space Φ, it generates unified subspaces characterized by the highest correlation with the residual errors $\mathbf{\Delta}$ (or the target material property $\boldsymbol{P}$) through sure independence screening and merges them with the sparsifying operator to further extract the optimal descriptor. Image taken from Ref. [52]

in which $P_X(x)$ is the probability density function of X. The mutual information $MI(X,Y)$ measures the mutual dependence between variables X and Y. It is then scaled and transformed into a common range as $NMI(X,Y)$:

$$\mathbf{NMI}(\mathbf{X},\mathbf{Y}) = \frac{\mathbf{MI}(\mathbf{X},\mathbf{Y})}{(\mathbf{H}(\mathbf{X})+\mathbf{H}(\mathbf{Y}))/\mathbf{2}} \tag{1.8}$$

Based on this, one constructs a feature space $\mathcal{F}_s$ beginning from the feature with the highest NMI between the target, then interactively adds feature f_x into F that has the highest RR score defined as:

$$\mathbf{RR}(\mathbf{f}) = \frac{\mathbf{NMI}(\mathbf{f},\mathbf{y})}{[\mathbf{max}_{\mathbf{f_s}\in\mathcal{F}_s}(\mathbf{NMI}(\mathbf{f},\mathbf{f_s}))]^{\mathbf{p}}+\mathbf{c}} \tag{1.9}$$

Benefiting from feature selection by a method to dynamically optimize the RR ratio, this framework is well suited for limited datasets [54]. Furthermore, the NMI and RR scores saved in the object MODData provide a simple way to glance at the dataset from a statistical perspective and facilitate the interpretation of the model in terms of dependent physical variables.

SISSO

SISSO [52] is an algorithm to identify the best low-dimensional descriptor out of huge and possibly strongly correlated feature spaces within the framework of compressed-sensing-based dimensionality reduction. Figure 1.15 illustrates the architecture of it.

First, SISSO constructs feature spaces recursively as

$$\mathbf{\Phi_n} \equiv \bigcup_{i=1}^{n} \hat{\boldsymbol{H}}^{(m)} \left[\phi_1,\phi_2\right], \forall \phi_1,\phi_2 \in \Phi_{i-1} \tag{1.10}$$

with the primary feature space (all inputs features) Φ_0 and a defined operators set

$$\hat{\boldsymbol{H}}^{(m)} \equiv \{I, +, -, \times, \backslash, exp, log \cdots\} [\phi_1, \phi_2] . \tag{1.11}$$

Then a sparse-solution algorithm by combining sparsifying operators with sure independence screening is used to find the final descriptor. Sure independence screening [138] uses correlation magnitude (i.e., the absolute of inner product between the target property and a feature) to assess each feature. It retains only the highest-ranked features based on this metric. After this reduction, the optimal n-dimensional descriptor is identified by sparsifying operators. The goal is to seek the best outcome that minimizes dimensionality, This involves progressively testing larger values of n until the remaining residual error aligns with the expected level of quality.

SISSO is designed to solve tasks when only relatively small training sets are available, but is also applicable to larger datasets with fewer features (considering the memory issue caused by the size of features space Φ_n increases very fast with Φ_0). It yields a mathematical model represented by explicit, analytical functions of fundamental input physical quantities. This function is not only useful for the prediction of new data but may also provide a hint to reveal the relation between features and the target.

Random forest

RF [61], also called random decision forests is an ensemble learning method that operates by constructing a multitude of decision trees at training time. Each tree is trained on bootstrapped subsets of the data with randomly selected features at each split point. It supports classification and regression tasks. For classification, the output of RF is the class selected by most trees (as shown in Figure 1.16). For regression, the average prediction of the individual trees is returned. By bagging, RF produces relatively accurate predictions while reducing overfitting and enhancing robustness to noisy data with the cost of losing the interpretability of a single decision tree. However, insights into feature importance (such as GINI importance and permutation importance) are offered, making it valuable for feature selection and interpretation. It is a well-known algorithm that is widely used across diverse domains, thanks to its simplicity and reliable ability for handling complex data sets in various tasks.

GINI importance and permutation importance are methods used to gauge the importance of features in machine learning models. GINI importance is specifically designed for classification tasks, primarily employed in ensemble decision tree-based models like RF and GBTs. It is based on the Gini impurity criterion to quantify feature importance by assessing how frequently it is used and the reduction in impurity (or increase in purity) results from it when used to split data into classes. On the other hand, permutation importance is a versatile, model-agnostic technique applicable to both classification and regression tasks. It evaluates feature importance by evaluating the drop in model performance when the values of a specific feature are randomly shuffled, thus the relationship between the feature and the target is disrupted. Although GINI importance indicates which features are more influential in making predictions and is relatively simple to interpret, it is susceptible to overfitting and exhibits a strong bias towards favoring high cardinality features (features with many unique values, typically numerical features) over low cardinality features (features with few unique values, such as categorical variables with a limited number of distinct categories.) Permutation importance prevents the affection from overfitting and biases related to feature type and model structure since it can be computed on unseen data with performance

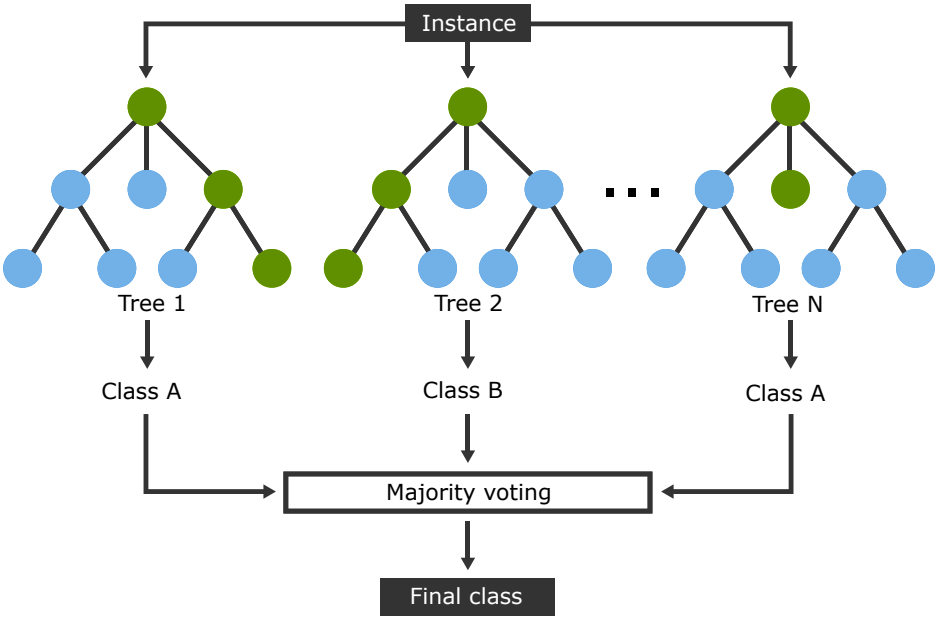


Figure 1.16: Diagram of a random forest for classification.

metrics (such as accuracy and mean squared error) for predictions of the model. It has the limitation that results can be influenced by random permutations, and it assumes that features are independent, while numerous studies have found that the importance is highly misleading when significant correlations exist between features. [139]

Gradient boost trees

GBTs [62] is another robust ensemble learning method based on decision trees. Unlike building independent trees in RF, it is a boosting algorithm that constructs a collection of decision trees sequentially, with each new tree focusing on rectifying the errors made by the preceding ones. GBT's strength lies in its use of gradient descent optimization, where it identifies and assigns higher weights to data points where the current model struggles the most, gradually refining predictions. It excels at improving model accuracy and is particularly effective when dealing with complex, nonlinear relationships in the data. To prevent overfitting, regularization techniques such as shrinkage (learning rate) and subsampling are used for balancing a trade-off between bias and variance. Thanks to the ability of highly accurate prediction, it has applications in a wide range of domains, such as natural language processing, image recognition, and weather forecasting.

Gain importance is a measurement commonly used in XGBoost, calculated by evaluating the improvement in the model's loss function (e.g., accuracy for classification task or mean squared error for regression task) when a particular feature is used for splitting data points. This metric provides a relative evaluation of feature importance within the model, and its value is intimately linked to the model's predictive accuracy.

Topogivity

Topogivity [121] approach defines a framework to find a heuristic chemical rule for diagnosing NTMs based on the chemical formula of the material, aiming at learning insights

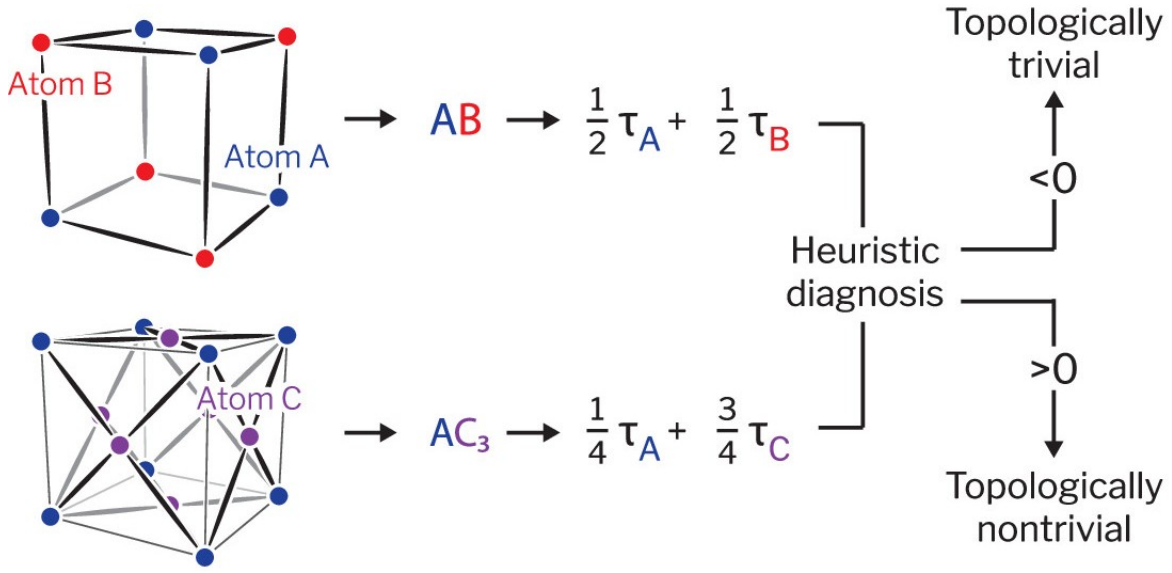


Figure 1.17: Chemical rule of topogivity. Given a stoichiometric material (e.g., AB and AC_3), its topological type is diagnosed by weighting the topogivities (τ_A, τ_B, τ_C) with the element fraction. The sign indicates the topological classification, while their magnitude shows confidence. Image taken from Ref. [121]

about electronic topology. As shown in Figure 1.17, it maps each material M to a number $g(M)$ with the function

$$g(M) = \sum_E f_E(M)\tau_E \tag{1.12}$$

where the summation runs over the elements E in the chemical formula of material M, $f_E(M)$ is the element fraction for the element E in material M, and τ_E is a learned parameter for each element E, referred as topogivity.

The value of each τ_E is learned by constructing a soft-margin linear SVM model, where the input is the element fraction of the material, and the target is the topological type (TrI or NTM)

Their final model is displayed in Figure 1.18. This heuristic chemical rule is widely applicable as it solely requires the chemical formula. The periodic table of topogivities provides an intuitive understanding: for certain element E, the topogivity τ_E roughly captures its tendency to form an NTM. Nevertheless, only a broad classification (TrI or NTM) can be provided, rigorous classification requires further detection (e.g., ab initio calculations). Additionally, the definition of topogivities lacks clarity and can be influenced by factors such as the training dataset and the parameters of the SVM algorithm.

T-distributed Stochastic Neighbor Embedding

The t-SNE [66] is a nonlinear dimensionality reduction method widely used for visualizing and exploring high-dimensional data. Without taking the target into account, it is particularly effective in revealing the underlying structure and patterns in complex data by preserving local pairwise similarities. The primary objective of t-SNE is to project high-dimensional data points into a lower-dimensional space, typically 2-dimension or 3-dimensions, while preserving their pairwise similarities more faithfully.

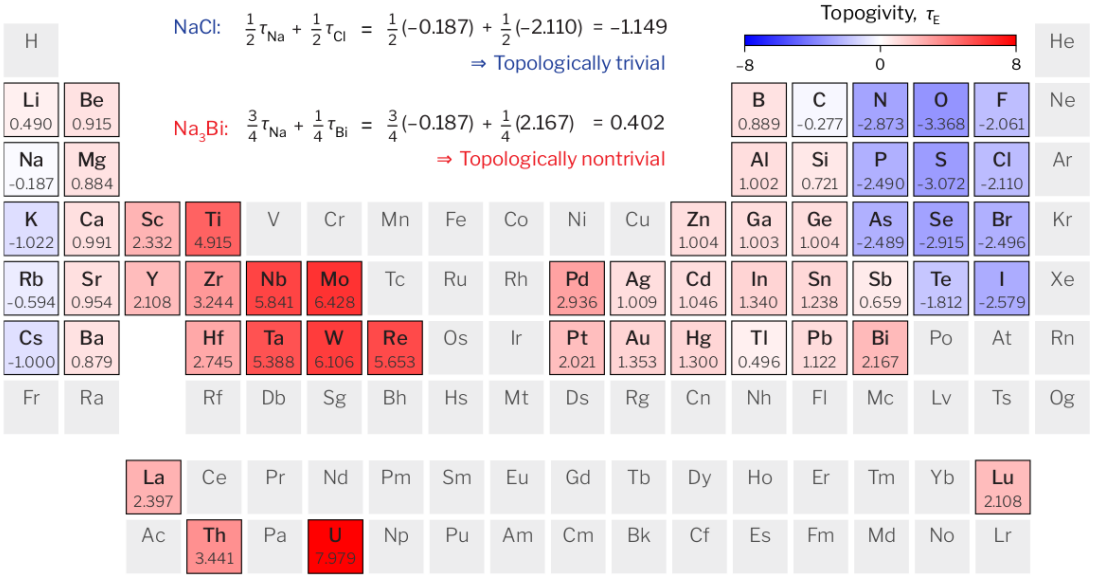


Figure 1.18: Periodic table of topogivities from Ref. [121]. The topogivities τ_E are learned with a symmetry-based dataset from Ref. [29]. Existing topogivities are represented through numerical values with color coding, others are displayed in gray. Two example materials: $NaCl$ and Na_3Bi are given.

The process of t-SNE begins by computing pairwise similarities between data points in a high-dimensional space. These similarities represent the conditional probability of two points being neighbors if sampled from a Gaussian distribution. Next, the algorithm initializes lower-dimensional representations of data points randomly and iteratively refines them to minimize differences in pairwise similarities between the original and lower-dimensional spaces. The optimization process in t-SNE involves gradient descent, where the algorithm tries to find the optimal lower-dimensional representation that best captures the local relationships in the original data. The employment of t-distribution in the lower-dimensional space preserves better pairwise similarities.

2 DATASETS

This chapter introduces the process of collecting and cleaning the data from Materiae and TMD to build our dataset. Two datasets are built from them, named M and T. 35,608 samples are kept for training our final model, they are indexed either with the MP-ID or the ICSD-ID, and each sample includes the corresponding pymatgen [125] structure object and the topological type from the DFT calculation and TQC analysis when the presence of spin-orbit coupling is considered.

2.1 DATASET MAT

Dataset *MAT* is constructed from the Materiae. In Materiae, the topological properties of 26,120 materials are displayed, these are results computed with VASP and SymTopo [140] following the workflow shown in Figure 2.1.

There are three things to notice in their final database: a) magnetic materials and conventional metals are not included; b) the structures are checked and processed with PHONOPY [141]; c) materials in spacegroup 1 are excluded.

They selected 39,519 materials that are concurrently registered in both the MP and ICSD databases. Before the DFT calculation, the magnetic materials and conventional metals are first labeled, since the further analysis tool, SymTopo is not suitable for magnetic materials and conventional metals are not potential topological material candidates. These materials are not included in their final database. The magnetic materials are labeled for those whose magnetic moment is higher than 0.1 μ_B per unit cell based on the MP records. Materials are categorized as conventional metals when they have an odd number of electrons per unit cell, as determined by the employed pseudopotential (the generalized gradient approximation of the Perdew–Burke–Ernzerhof type exchange-correlation potential [142] from the VASP software package, for each material, listed at http://materiae.iphy.ac.cn/).

The initial structures are retrieved from the MP database. As these structures are not in a unified convention, such as the orientation of the primitive cell is arbitrary and the choice of origin point can vary, their symmetries are then checked with PHONOPY. To enable automated high-throughput calculations and ensure accurate symmetry recognition by VASP, we utilize PHONOPY to standardize and symmetrize the atomic positions following the loading of lattice parameters and atomic coordinates. During this process, 166 materials were excluded due to disparities between the spacegroups identified by PHONOPY and those listed on the MP, with discrepancies tolerated up to an error of 0.1 Å in atomic positions.

The last thing is that all the materials in spacegroup 1 are also not included, as in that case, based on TQC, all of them are classified as TrIs.

Subsequently, the remaining materials are adopted for further DFT calculations. For TSMs, a set of their bands below the Fermi level are degenerate, thus can not be separated from other bands due to a failure of satisfying the compatibility relations. Their topological nodes cannot be removed by symmetry-preserving perturbations. Depending on the degeneracy and dimensionality of their nodes, a topological semimetal is classified into

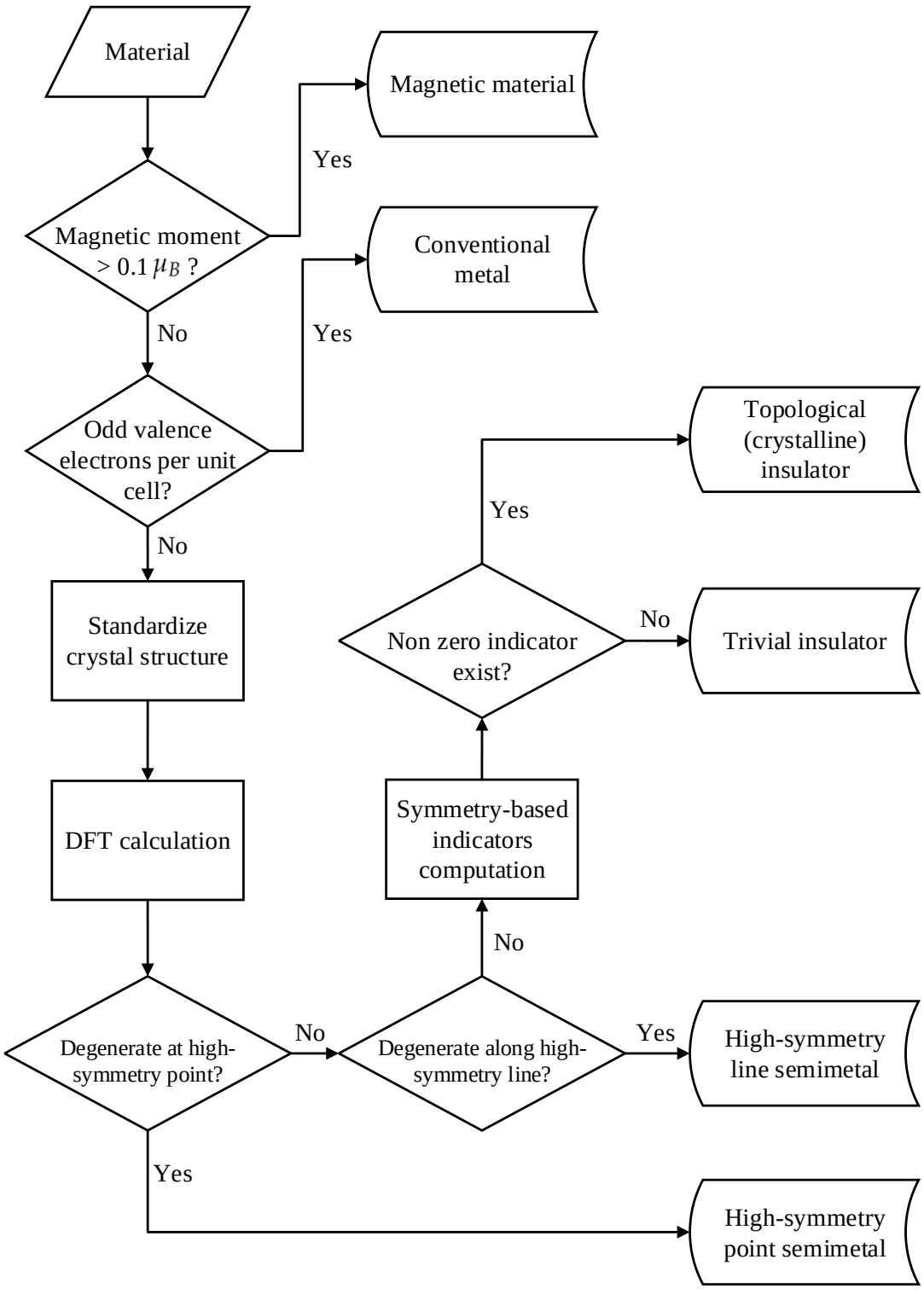


Figure 2.1: Workflow of high-throughput calculations for building dataset *MAT*. Image adjusted from Ref. [140]

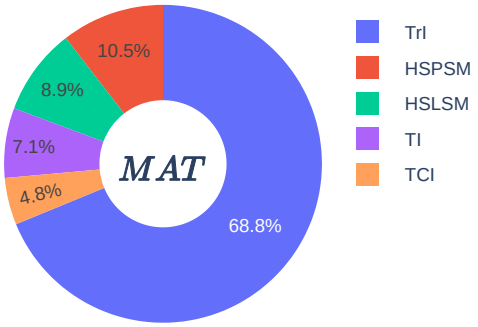


Figure 2.2: Composition of dataset *MAT*

a HSPSM or HSLSM [28]. TIs have a set of valence bands that satisfy the compatibility relations but cannot be decomposed as a sum of elementary band representations, thus owning non-zero topological indicators. Distinguished by the parity of the symmetry indicator a material is categorized into a TI or TCI.

In this work, we collect the results displayed on Materiae and take all the structures they used in VASP (POSCAR files provided by the builder of Materiae), which are after the standardization of PHONOPY. In addition, we supplement the materials in spacegroup 1 to our dataset, by taking structures from MP followed by post-processing with PHONOPY and labeling all of them as trivial insulators.

After data cleaning, we collect a dataset MAT, which includes 25,895 samples, 17,813 of them are TrI, and 8,082 of them are topologically nontrivial, sorted into 2,292 HSLSMs, 2,713 HSPSMs, 1,238 TCIs, and 1,839 TIs. Their topological classes distribution is shown in Figure 2.2, their relative distribution across different crystal systems and inversion symmetry are shown in Figure 2.3, and Figure 2.5. Figure 2.4 plots the chemical composition of compounds in this dataset. It displays the total amount of compounds containing a certain element and the frequency of NTMs among compounds.

Furthermore, dataset M is constructed from dataset MAT by removing conflicting entries between database TMD and Materiae (see next section). All the records from Materiae are indexed by a unique MP-ID.

2.2 DATASET T

Dataset T is constructed from the TMD. TMD was compiled by sweeping 96,196 ICSD entries, resulting in 73,234 convergent entries (associated with ICSD-IDs). They are results from DFT calculations with VASP and post-processed with VASP2Trace [143] and Check Topological Mat [144] programs. They are grouped into 38,298 unique materials with 38,298 unique-IDs following the criteria that two compounds are the same if they share the following attributes: chemical formula, spacegroup, and stable topology at E_F. From an original dataset provided by the builder of TMD, including information for 73,234 entries with their chemical compositions, ICSD-IDs, spacegroup numbers, MP-IDs, unique-IDs, and topological classes of 3 types (TSMs, TIs, and TrIs), we proceeded with a complete cleaning process. We first take at least one representative of each unique-ID with a confirmation of MP-ID, then we clean these data and connect them with dataset MAT by MP-IDs. During this process, we check TMD again to get the detailed topological state (topological type and topological indices if exist) of these 38,298 unique materials, and all the structures related to

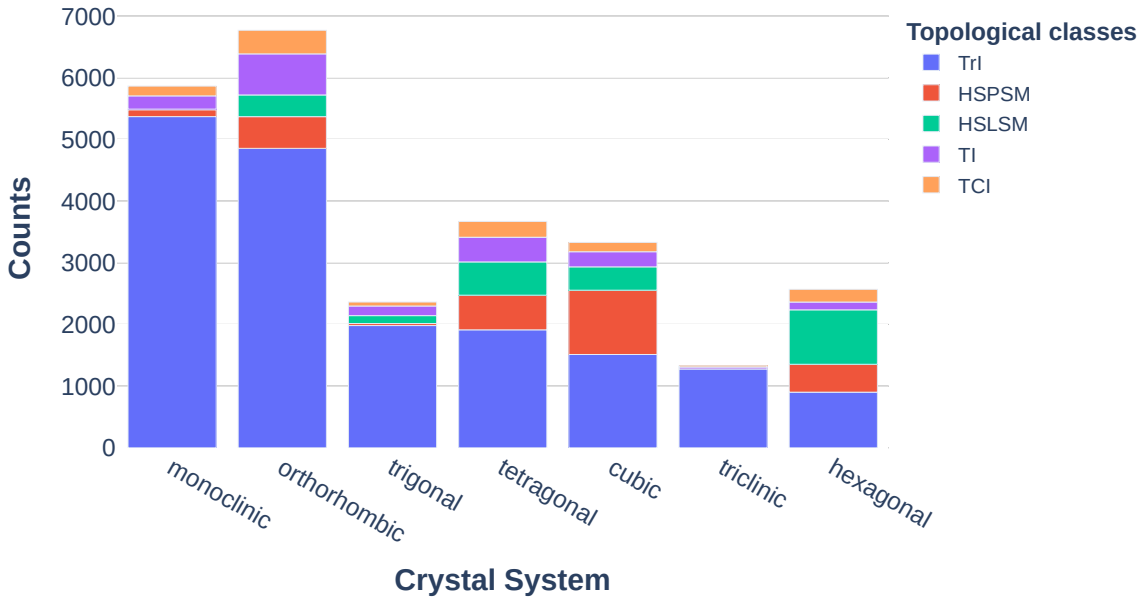


Figure 2.3: Relative distribution of dataset *MAT* under different crystal systems

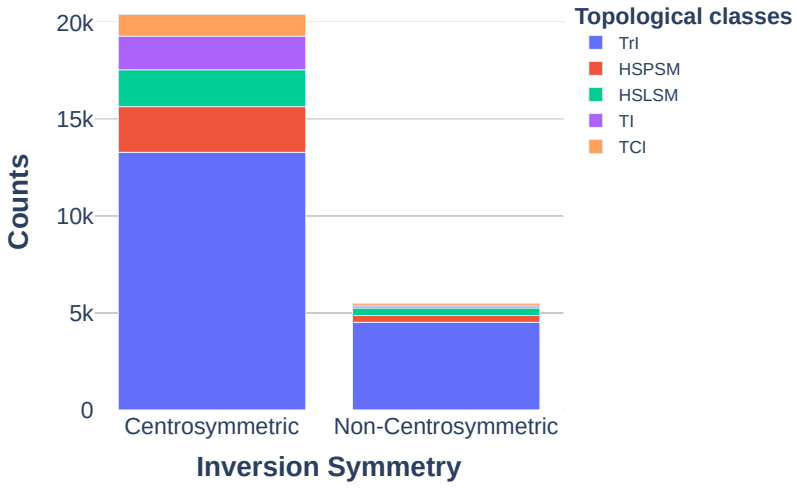


Figure 2.4: Relative distribution of dataset *MAT* with different inversion symmetry.

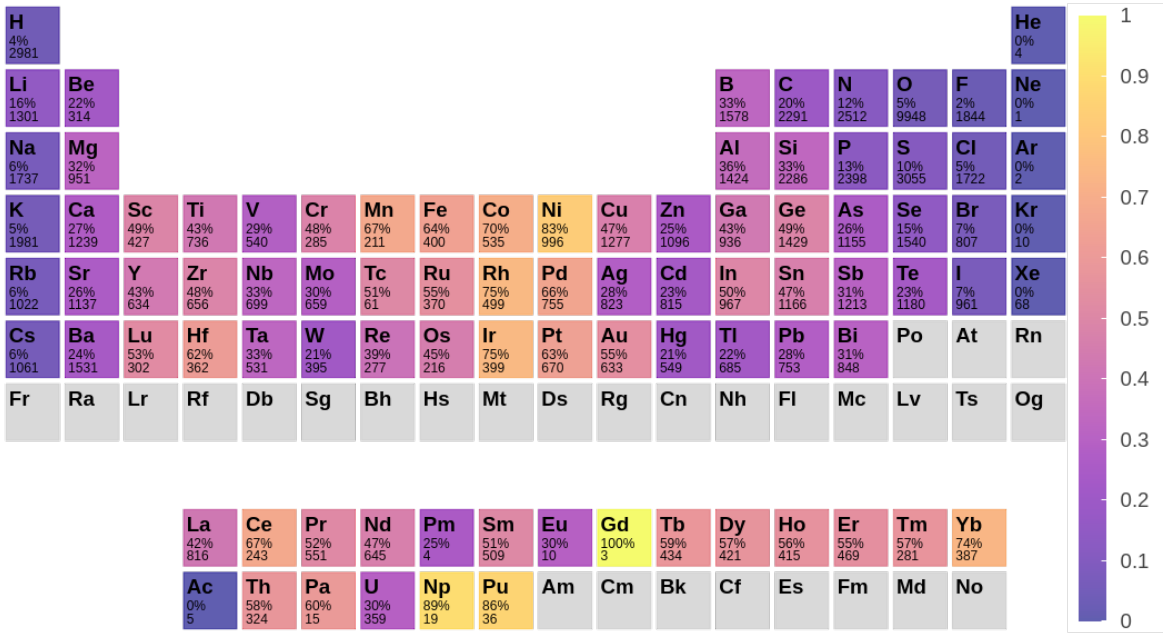


Figure 2.5: Chemical composition of compounds in dataset *MAT*. Under the atomic symbol, the first row illustrates the frequency of NTMs in compounds containing this element (the color scale is defined by this value). The second row illustrates the total amount of compounds containing this element.

this dataset are fetched based on the ICSD-ID, which are the structures from ICSD followed by post-processing with PHONOPY.

In the first stage, we collect a dataset that is indexed with 39,166 unique ICSD-IDs. The complete information of these entries are provided in the supplementary. To illustrate the cleaning process, we list some examples in Appendix B in order of unique-IDs. Given a set of compounds in the same group of unique materials, we distinguish three cases, in order to retrieve the corresponding structure. First, for 5,351 unique materials, no single structure has an assigned MP-ID, we then keep one ICSD entry at random. Second, 32,392 unique materials have one associated MP-ID, one ICSD entry with that MP-ID is kept at random, followed by a confirmation of the MP-ID, for example, the material whose unique-ID is 2 or 32. Finally, as a third case, multiple MP-IDs might be associated, for example, the material whose unique-ID is 58 or 174. In this case, there are 555 unique materials, we did a test comparing the similarity of the structures: for 20 unique materials, there are 231 ICSD items that related to 44 MP-IDs, and we found generally structures associated with different MP-IDs are different. So for each unique material, we first keep one ICSD entry associated with each MP-ID; then list all the related MP-IDs in order of increasing energy above hull given at MP, and confirm the association of every ICSD entry to one possible MP-ID with the lowest energy above hull; at the end, we remove the duplicated entries with same MP-ID. For the confirmation of the MP-ID in the second and the third cases, it is done by comparing the structure associated with ICSD-ID to the structure from MP by default setting structure_matcher in Pymatgen [125]. The fractional length tolerance is 0.2, the site tolerance, which represents the proportion of the average free length per atom (calculated as $(V/Nsites)^{1/3}$), is at 0.3. Additionally, the angle tolerance is 5 degrees. In case no match MP-ID is found, NaN is assigned. Based on the newly confirmed MP-IDs,

we deduplicate the dataset into 38,916 entries. All further cleanings are also done based on the new confirmed MP-IDs without explicit mention.

Next, we clean these data followed the same way as materiae and connect them with dataset MAT based on MP-IDs. After the whole cleaning a total of 24,156 items from TMD remain, named dataset T, details are provided below.

In TMD, there are several essential differences from Materiae: they also computed and classified the material whose magnetic moment records on MP is higher than 0.1 μB per unit cell; NTMs are grouped into four other types: ES, ESFD, SEBR, and NLC; pseudo-potentials of different valency are used in their DFT calculation. Since their computation and classification are based on nonmagnetic band topology, in our dataset we excluded the potential magnetic materials according to the record magnetic moment from MP. Defined in Ref. [30], ESFD has a set of Bloch states that are degenerate at a high-symmetry point, the conventional metals are also categorized as this type. ES has a set of high-symmetry-point Bloch eigenstates that violate the compatibility relations, indicating the connection to other states along a high-symmetry line or plane. The stable topological bands of SEBR could be written as integer-valued linear combinations of the elementary band representation, while NLC does not adhere to this. Although they also use VASP with the generalized gradient approximation of the Perdew–Burke–Ernzerhof type exchange-correlation potential, the valency for certain elements in their pseudo-potentials varies from Materiae. This is a critical difference because conventional metals are defined based on this factor.

According to these facts, we remove 4,275 materials that include at least one of the following rare-earth elements: Pr, Nd, Pm, Sm, Tb, Dy, Ho, Er, Tm, Yb, Lu, or Sc due to the difference of pseudo-potentials. Then we detect and exclude 4,729 conventional metals and 4,597 magnetic materials. After this step, the dataset is cleaned into 25,315 items. For the remaining items, we map them into Materiae's definition. For TSMs, we map ESFD into HSPSM and ES into HSLSM. But for TIs, no simple mapping exists. We label SEBR and NLC as TI or TCI. For materials in spacegroup 174, 187, 189, 188, or 190, they are all labeled as TCI. For the rest, they are labeled according to the parity of the last topological indices, odd as TI and even as TCI. After the relabeling, we found some entries could have conflicting topological types though related to the same MP-ID. This is due to the slight difference between ICSD structures since when we redo the DFT calculation and symmetry indicator analysis with the ICSD structures, our results remain consistent with the result from TMD. After removing 672 entries of this case, dataset TM of 24,643 items is constructed, and sorted into 16,595 TrI, 2,229 HSLSMs, 2,560 HSPSMs, 1,282 TCIs, and 1,977 TIs. All the records from TMD are indexed by ICSD-ID.

Subsequently, we connect this dataset with dataset MAT based on MP-ID. In the intersection of the two databases, there are 223 items (223 ICSD-IDs corresponding to 212 MP-IDs) labeled with different types (again due to the slight difference between structures), we remove these conflicting items. Then for dataset TM, we deduplicate 522 items (522 ICSD-IDs corresponding to 258 MP-IDs) into 258 items. Until now, datasets M and T are constructed from datasets MAT and TM, we group all these entries into three sub-dataset, $M\backslash T$, the materials only in M; $M \cap T$, the materials that are common to both; and $T\backslash M$, the materials only in T. The topological class distribution of these five datasets is shown in Figure 2.6. Finally, our full dataset $M \cup T$ with 35,608 entries is constructed.

For dataset T, we display the relative distribution between different crystal systems and inversion symmetry in Figure 2.7 and Figure 2.9. Additionally, we present the chemical composition in Figure 2.8. Comparing them to dataset MAT, their relative distribution of

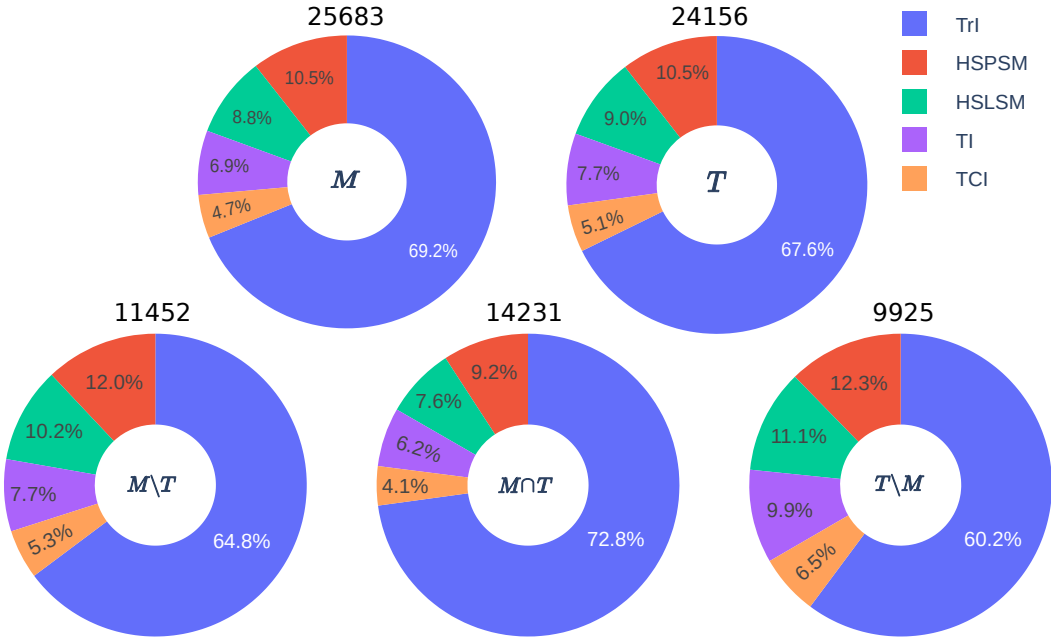


Figure 2.6: Composition of datasets. M is constructed from the Materiae, T is constructed from the Topological Materials Database, $M \setminus T$ is the difference from M to T, $M \cap T$ is the intersection of M and T, and $T \setminus M$ is the difference from T to M.

topological types appears remarkably similar, but the chemical composition of compounds exhibits variations. Besides the removal of materials containing certain rare-earth elements from dataset T attributed to differences in pseudo-potentials, several more notable distinctions arise. Dataset T has a few materials including Po, Rn, Ra, or Am that are absent in dataset MAT. Although the size of the two datasets is similar, dataset T encompasses more materials including certain elements, such as Mn (534 to 211) or Fe (910 to 400). For materials containing noble gas elements, dataset T identified some NTMs while MAT did not.

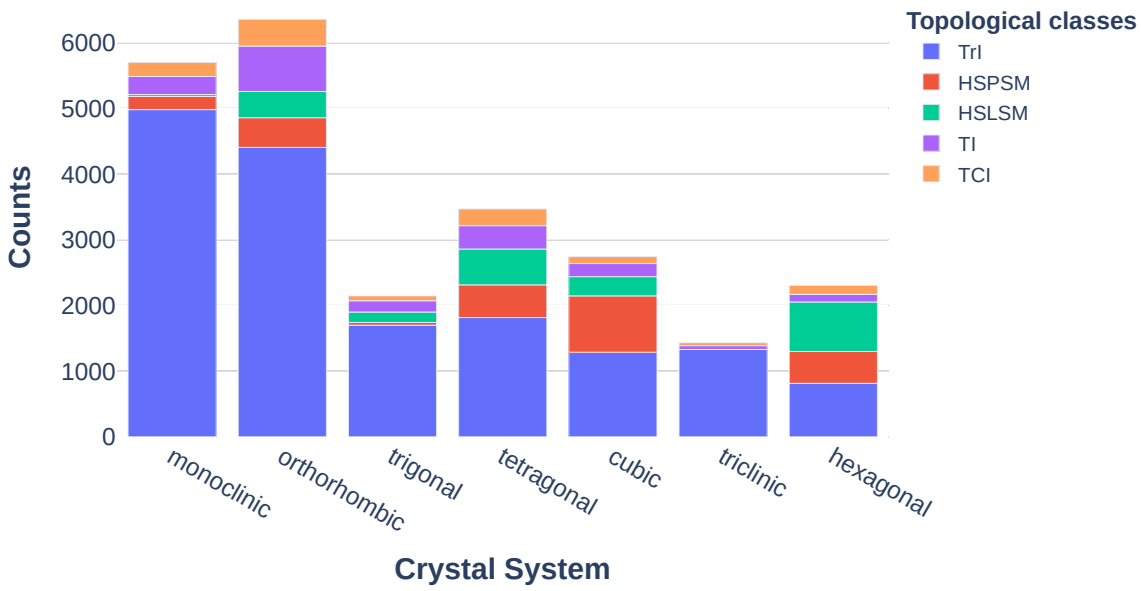


Figure 2.7: Relative distribution of dataset T under different crystal systems.

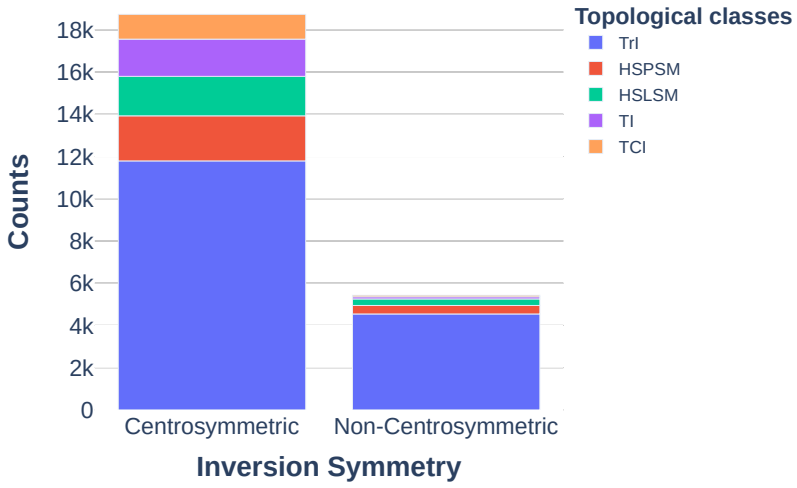


Figure 2.8: Relative distribution of dataset T with different inversion symmetry.

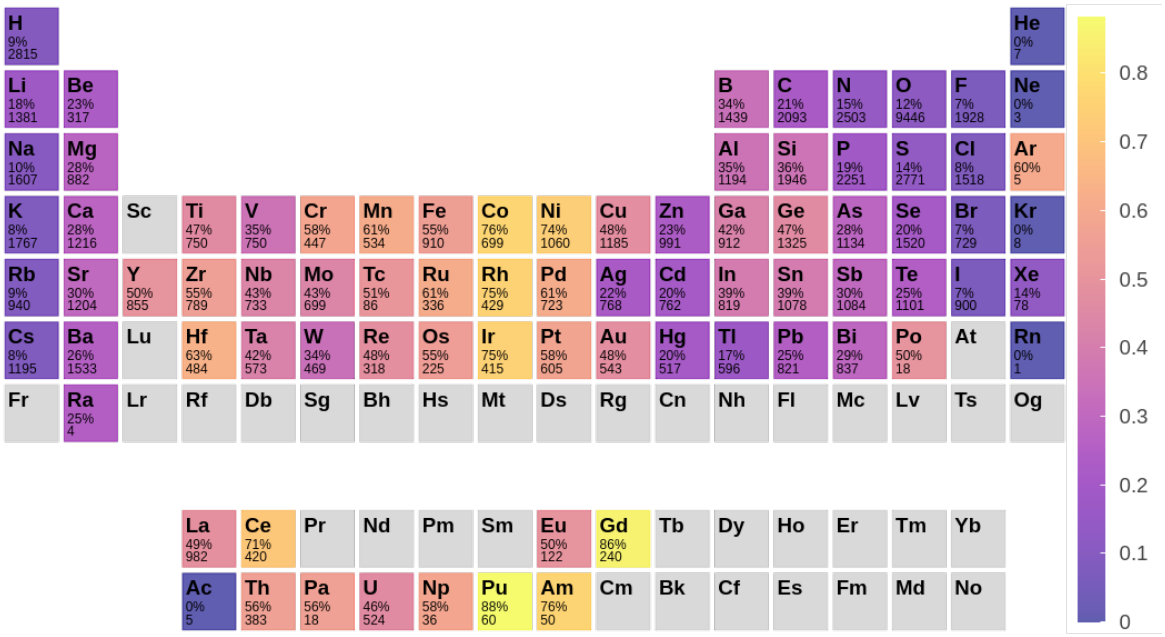


Figure 2.9: Chemical composition of compounds in dataset T. Under the atomic symbol, the first row illustrates the frequency of NTMs in compounds containing this element (the color scale is defined by this value). The second row illustrates the total amount of compounds containing this element.

3 METHODOLOGY

In this work, we complete two tasks, first a 5 types classification to categorize the materials into TrI, HSLSM, HSPSM, TCI, and TI; and second a binary task of distinguishing TrI and NonTr to analyze the important features that affect the topology of materials. In this chapter, we will discuss the procedure and the algorithms used in our work.

3.1 5 TYPES CLASSIFICATION

For 5 types classification, a benchmark is done on the dataset MAT to find the best method, then a generalization test is done on the whole dataset $M \cup T$.

3.1.1 *Benchmark with dataset MAT*

As shown in Figure 3.1, we comparatively analyze two procedures to categorize materials into 5 types. The first is a hierarchical approach named tree 1, to categorize the material with four binary classifications (nodes 1-4). These steps are similar to the workflow of first-principle calculations thereby facilitating interpretability. Furthermore, from a modeling perspective, this hierarchical framework relieved concerns regarding data imbalance at each node. Conversely, in the second approach one directly submits the entirety of materials across 5 types to the model and classifies them in one go, despite the imbalance problem, the model can incorporate larger amount of data simultaneously.

For this classification task, 6 algorithms are considered: MEGNet, Automatminer, MODNet, SISSO, RF and GBTs. Consistent NCV testing procedure following the MatBench [50] recommendations is used to evaluate their performance. As the example outer loop shown in Figure 1.7, an identical train/test split, a five-fold outer loop with scikit-learn [123] stratified KFold (random seed 18012019) is applied on dataset MAT where data are labeled as 5 types. Within each of the five splits, 20% are kept for testing, and 80% are given to the algorithm for applying the internal validation for features and hyperparameters optimization. For the procedure of tree 2, these split data are directly fed. For the procedure of tree 1, the split of the materials is kept at each node by mapping their label into the corresponding parent class. The internal validation and model selection process are dependent on each algorithm, the detailed process and their setting are given below.

MEGNet

MEGNet v1.2.9 from their official git repository is used here. In the NCV procedure, without specifying a validation set, all valid structures (discard the structure containing isolated atoms) in the training set are fed into the MEGNet model. Following their instruction, a model defined with CrystalGraph whose cutoff radius equals 4 Å is used for extracting the atomic number of elements and spatial distance. The training is constrained to be done within 500 epochs. The scores of MEGNet are calculated only based on the number of valid structures. Specifically for the task of tree-2, we adjust the algorithm to change the

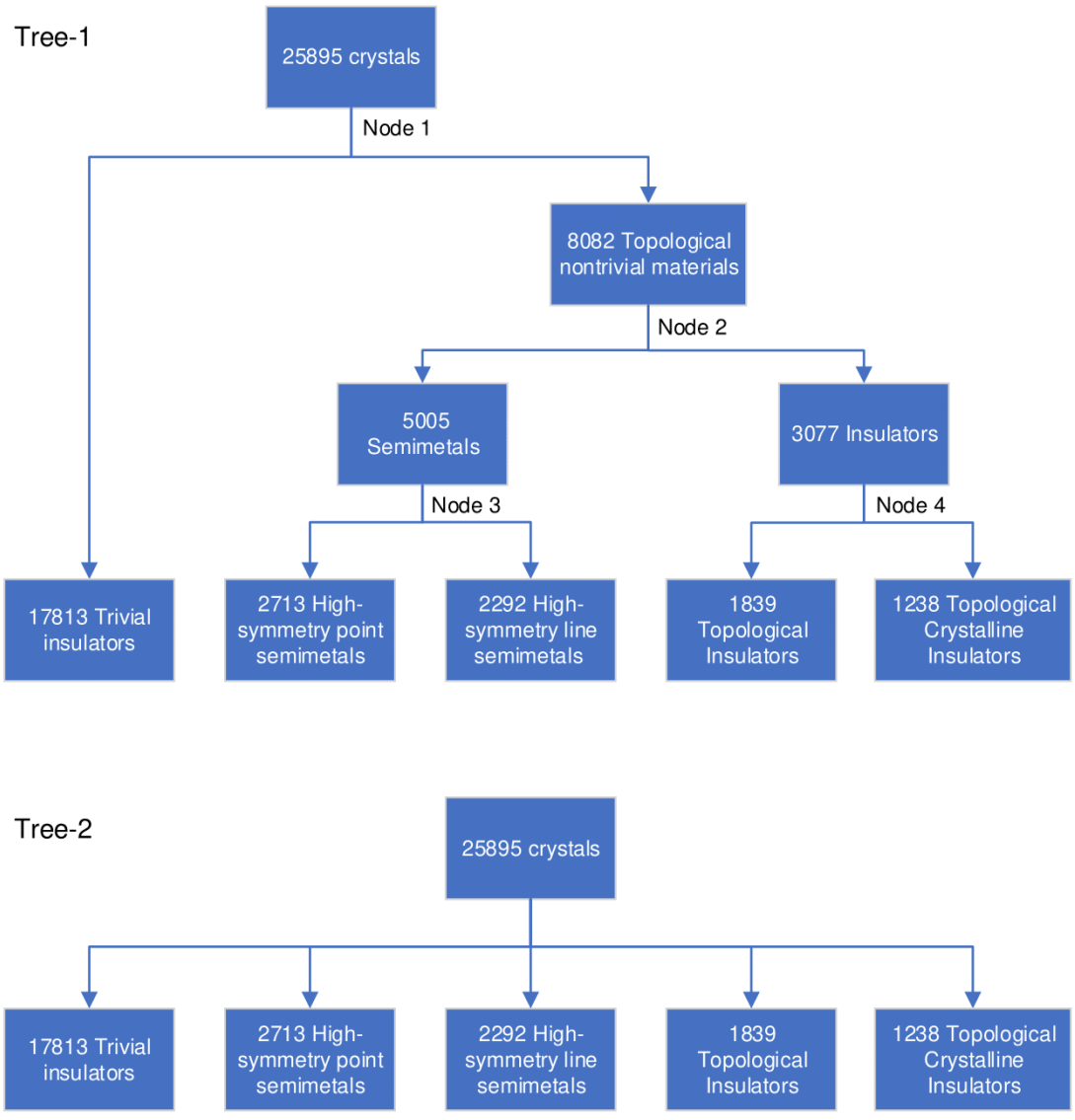


Figure 3.1: Categorize the materials with two different procedures.

Table 3.1: Hyperparameters used in SISSO.

Hyperparameter	Value	Description
desc_dim	4	Dimension of the descriptor
rung	2	Rung of the feature space to be constructed
subs_sis	50	Size for SIS-selected subspace

last layer to softmax and loss function to "categorical_crossentropy" for suiting multiclass classification.

Automatminer

A customized 'express' preset pipeline of Automatminer v1.0.3.20191111 is used on our training set ('EwaldEnergy' is excluded from the default AutoFeaturizer since it blocked the pipeline). In this pipeline, features are automatically generated from structures, and missing features are imputed with the mean of the known data. Following the feature extraction, the number of features is reduced based on Pearson correlation coefficients [145] and a RF method. Finally, TPOT v.10.1 is used to train and internally validate (five-fold CV within the training data) competing ML pipelines, and pick the winning model to make test predictions. For the inner validation, 'balanced_accuracy' is used as the scoring method.

MODNet

For MODNet, we first use a predefined set of featurizers (DeBreuck2020Featurizer with accelerated oxidation state parameters) to generate features from structures and fill the missing features with the default value (most are zero) of the corresponding featurizer. Then all training data in the outer 5-fold loop of the NCV procedure are given to the algorithm to perform feature selection and train an optimized model with a genetic algorithm.

SISSO

To apply SISSO, we pick 10 features with the largest NMI score used by MODNet to construct the initial dataset of numerical descriptors. Given that the classification method of SISSO is quite sensitive to noisy data, we map categories into integers to use the regression approach for training. The testing results of 20% data from each fold are normalized and interpreted from a probabilistic perspective. Since the accuracy of a SISSO regression model is positively related to its complexity, we opt for the most complex model feasible within the constraints of computing resources. However, the performance of SISSO is not impressive. Therefore, during the benchmark, it has been evaluated only for one single task: node 1 in tree -1, the complete list of hyperparameters for that model is detailed in Table 3.1.

Random Forest

We use RF as a baseline representing commonly used methods. First, we do feature extraction to transform structure data into numerical descriptors. This involves utilizing featurizers of "magpie" elemental statistics [128], Sine Coulomb Matrix [136], density features and global symmetry features from Matminer to do feature extraction, and imputing missing values by the mean of the known data. Different from previous algorithms, for

Table 3.2: Hyperparameter grid for GBTs

Parameter	Value	Parameter	Value
Learning rate η	0.23	L^2 regularization λ	1.33
Maximal tree depth	[9, 10, 11]	Minimal child weight	[0.1, 0.3, 0.5]
Column subsampling by tree	[0.75, 0.78]	Column subsampling by node	[0.75, 0.78]

RF a stratified three-fold inner cross-validation (with 'balanced_accuracy' as the scoring method) is explicitly used to determine an optimal model before final training and testing. Within this framework, a set of RF classifiers from scikit-learn defined on hyperparameter grid: {'max_features' : ['auto', 'sqrt', 'log2'], 'criterion' : ['gini', 'entropy']} are tested.

Gradient boost trees

An algorithm similar to Ref. [113] is adopted here. On each fold of the training data, we use the implementations of XGBOOST [146] defined by hyperparameters chosen from Table 3.2, it is selected by a stratified five-fold inner cross-validation (with F_1 score as the assessing function). The input of the model is a full set of features same as the input of MODNet before feature selection.

3.1.2 *Generalization test with dataset $M \cup T$*

In further research on all the data, we adopt the multiclass approach with GBTs which gain the highest accuracy according to the benchmarking test on dataset MAT. To appraise the relation between dataset M and T and assess how well this model generalizes to new, unseen data in practical applications (where the new dataset might have a different distribution from the training dataset), a series of generalization tests are conducted on different datasets, i.e. M, T and their mathematical operations. As shown in Figure 3.2), the dataset is represented by a distinct shape, and the conduct on it is indicated by color. In each of the seven generalization tests, we initially perform a 5-fold NCV on the green segment, followed by a generalization test on the yellow segment using the model trained on the entire green portion.

Heterogeneity metric

After the generalization test, we use a heterogeneity metric defined by the k-nearest-neighbor (k-NN) distance in the feature space for analyzing the results. A heterogeneity metric is a quantitative measure used to assess the degree of variation, dissimilarity, or diversity within a dataset or between multiple datasets. It provides valuable insights into the distribution of the dataset to understand the underlying patterns and characteristics of the data. For example, in clustering analysis, heterogeneity metrics measure the dissimilarity or similarity between data points based on their feature values to evaluate the quality of cluster assignments, a lower heterogeneity score indicates that data points within the same cluster are more similar to each other, while a higher score suggests greater dissimilarity. The choice of a specific heterogeneity metric depends on the nature of the data and the research objectives. It could be a simple summary statistic, such as the mean

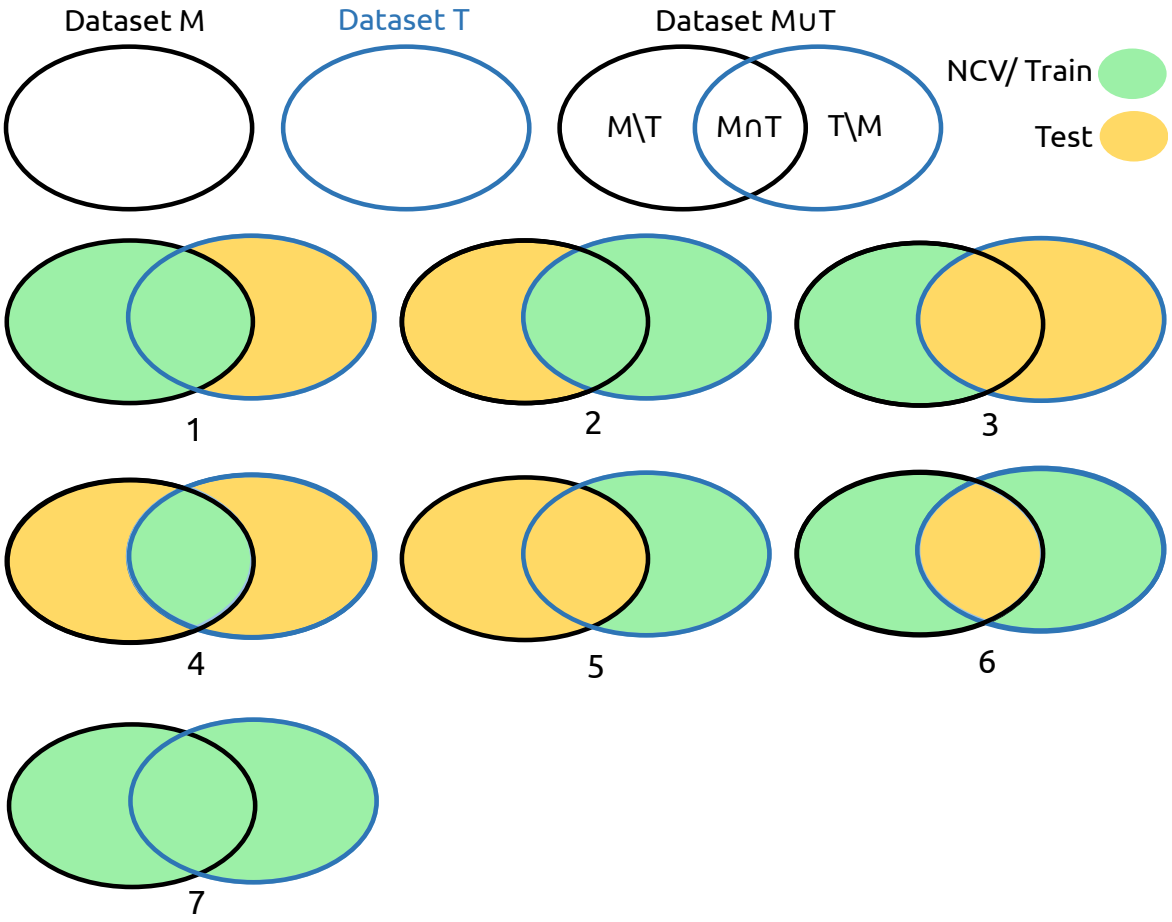


Figure 3.2: Diagram of generalization tests. The black and blue circle represent dataset M and T, the union dataset $M \cup T$ is split into three part: $M \backslash T$, $M \cap T$, and $T \backslash M$. We did 7 groups of tests, in each of them, the set for NCV and training is colored in green, the set for testing the training model is colored in yellow.

Table 3.3: Selected features for measuring the heterogeneity metric between datasets. The first column is the name of the featurizer, the second column indicates which statistics state of this feature is taken, and the third column refers to the type of this quantity

Featurizer Name	Description
Miedema	Formation enthalpies of intermetallic compounds, from Miedema et al. [133].
ElementProperty ("magpie" preset)	Weighted elemental statistics from Ward et al. [128].
OxidationStates	Statistics about the oxidation states for each specie. Features are concentration-weighted statistics of the oxidation states.
AtomicOrbitals	The highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) estimated from atomic orbital energies of the composition [129].
GlobalSymmetryFeatures	Determines symmetry features

or median distance, or a more sophisticated measure that takes into account the entire distribution. Common heterogeneity metrics include measures of dispersion (e.g., variance, standard deviation), distance-based metrics (e.g., Euclidean distance), information-theoretic metrics (e.g., entropy), and statistical tests (e.g., analysis of variance , Chi-squared test). An appropriate heterogeneity metric provides a quantitative basis for making informed decisions and drawing meaningful conclusions from the data.

The k-NN distance is a fundamental concept in ML and data analysis. It calculates the similarity between a data point (or observation) and its k-nearest neighbors within a dataset based on a chosen distance metric. It is widely used in tasks like classification, clustering, and outlier detection. In classification, it helps determine a data point's class based on the majority class among its nearest neighbors. In clustering, it assists in grouping similar data points together, while in outlier detection, it identifies anomalies by assessing how different a point is from its neighbors. The choice of distance metric and parameter k is crucial in k-NN distance calculations depending on the nature of the data and the research objectives. A smaller k makes the calculation more sensitive to local patterns, while a larger k provides a more robust estimate but may smooth out local variations.

For our dataset, we first construct a feature space with 47 important features extracted by Matminer. They are the union of top 20 important features with highest gain importance during training GBTs model on sub-datasets $M\backslash T$, $M \cap T$, and $T\backslash M$, a description of these features is listed in Table 3.3

In this 47-dimensional feature space of dataset $M \cup T$, we scale all values to the range of 0 to 1 using MinMaxScaler from scikit-learn. Subsequently, for each point in one sub-dataset, we calculate the distance from it to the nearest 5 neighbors in the other dataset, where the distance metrix $D_e(x, y)$ between point x and y is a Euclidean-like distance defined as:

$$D_e(x,y) = \sqrt{\sum_{i=1}^{n_c}(x_i - y_i)^2 + \sum_{j=n_c+1}^{n_c+n_d}(1/2(x_j == y_j))^2}\,, \tag{3.1}$$

where n_c is the number of continuous features, and n_d is the number of discrete features.

The heterogeneity metric $H(A,B)$ from datasets A to dataset B (can be the same dataset A as well) is defined based on the mean of the 5-NN distances as

$$H(A,B) = \sum_{i=1}^{N_A}(1/5\sum_{j=1}^{5} D_e(p_i,q_j))/N_A, p_i \in A, q_j \in B \ , \tag{3.2}$$

where q_j goes through the 5 nearest neighbor of p_i.

3.2 BINARY CLASSIFICATION

Furthermore, we tackle the binary classification problem of distinguishing the TrI from NTM on the whole dataset $M \cup T$ for detecting the essential factors that matter to the topology of a material. The performance of three algorithm: GBTs, Topogivity, and an unsupervised algorithm t-SNE are compared through a NCV approach of identical 5-Fold outer loop (with scikit-learn stratified kFold function, random seed 18012019). In the inner loop, CV are applied on GBTs and Topogivity to find the best hyperparameter sets, while for t-SNE, we simply input all the training data without the target value.

Same setting as applying GBTs for 5-types classification (explained at 3.1.1) is used here, the best model of XGBOOST [146] is selected on the hyperparameter grid defined in Table 3.2, with the assessing function F_1 score for validation. The input of is the same set of features, and the target is mapped into two types of materials: TrI and NTM.

For Topogivity, we follow the method defined in Ref. [113]. First, generate a subset from dataset $M \cup T$ which excluded elements occurring less than 25 times, leaving 83 elements that occur among the remaining 35522 materials. The scores of Topogivity are calculated based on the number of materials in the subset. Then construct a soft-margin linear SVM [57] model with the scikit-learn library (specifically, the sklearn.svm.SVC class), the model is tuned by F_1 score on a hyperparameter grid with 15 values of the hyperparameter C that are evenly spaced on a log scale ranging from 10^4 to 10^6.

The t-SNE [66] is an unsupervised algorithm. We construct a set of important features employed in GBTs, and project them into 2-dimension to visualize the feature space. The implementation of t-SNE in scikit-learn defined with PCA initialization is applied.

4 RESULTS OF 5 TYPES CLASSIFICATION

This chapter introduces the results of 5 types classification, to categorize the materials into TrI, HSLSM, HSPSM, TCI, and TI.

4.1 BENCHMARK

The benchmark is done on the dataset MAT for finding the best method, two procedures in Figure 3.1 are tested with 6 algorithms following consistent NCV procedure.

4.1.1 *Hierarchical approach - tree 1*

For the hierarchical approach of tree 1, we process with four steps of binary classifications (nodes 1-4). At each node, algorithms are tested with a corresponding subset. Ultimately, we combine all the results and map their prediction into one of the five types.

Node 1

At node 1 of tree 1, the task is to distinguish 17,813 TrIs and 8,082 NTMs. We compared the performance of 6 algorithms: MEGNet, Automatminer, MODNet, SISSO, RF, and GBTs. The normalized confusion matrices based on their boolean prediction are shown in Figure 4.1. The values are corresponding to Table 4.1. They all perform well on this task by taking advantage of enough data, especially the large amount of TrIs. Thus their true negative rate (all more than 90%) are always higher than their recall (all less than 90%, for SISSO, just 79.7%). The scores of Automatminer, MODNet, and RF in Figure 4.1 are similar, while MEGNet is a bit worse and GBTs is a bit better.

According to the probability prediction of each algorithm, Figure 4.2 displays their ROC curve and precision-recall curve. The differences between algorithms are consistent between probability prediction as shown in Figure 4.2 and boolean prediction as shown in the normalized confusion matrix, indicating the appropriate thresholds are selected for all the algorithms. Table 4.2 provides an overview of each algorithm, the accuracy, precision, recall, and F_1 score are computed based on the boolean predictions, and the ROC-AUC is based on the probability prediction.

For distinguishing TrIs and NTMs, a high accuracy, 94% is reached by GBTs, indicating the ability of ML to reproduce the DFT results relatively precisely. Other methods also get consistent and good results. This is partly due to sufficient data, but also because this

Table 4.1: The normalized confusion matrix over the actual conditions.

	Positive-Predicted	Negative-Predicted
Positive-Actual	True Positive Rate (recall)	False Negative Rate
Negative-Actual	False Positive Rate	True Negative Rate

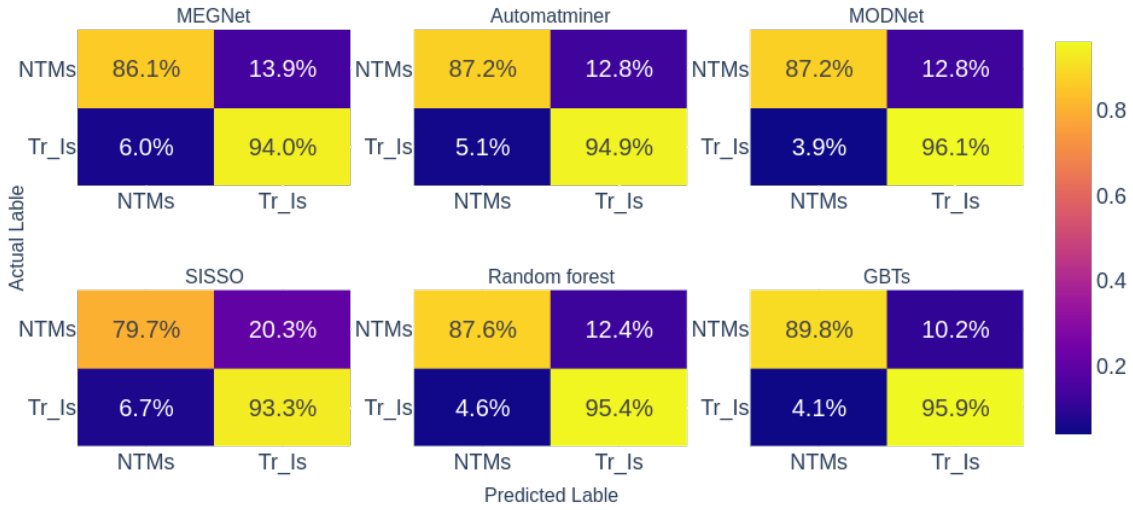


Figure 4.1: Normalized confusion matrices for distinguishing topologically nontrivial material (NTM) from trivial insulator (TrI) in dataset MAT. The values are normalized over the actual conditions (in rows).

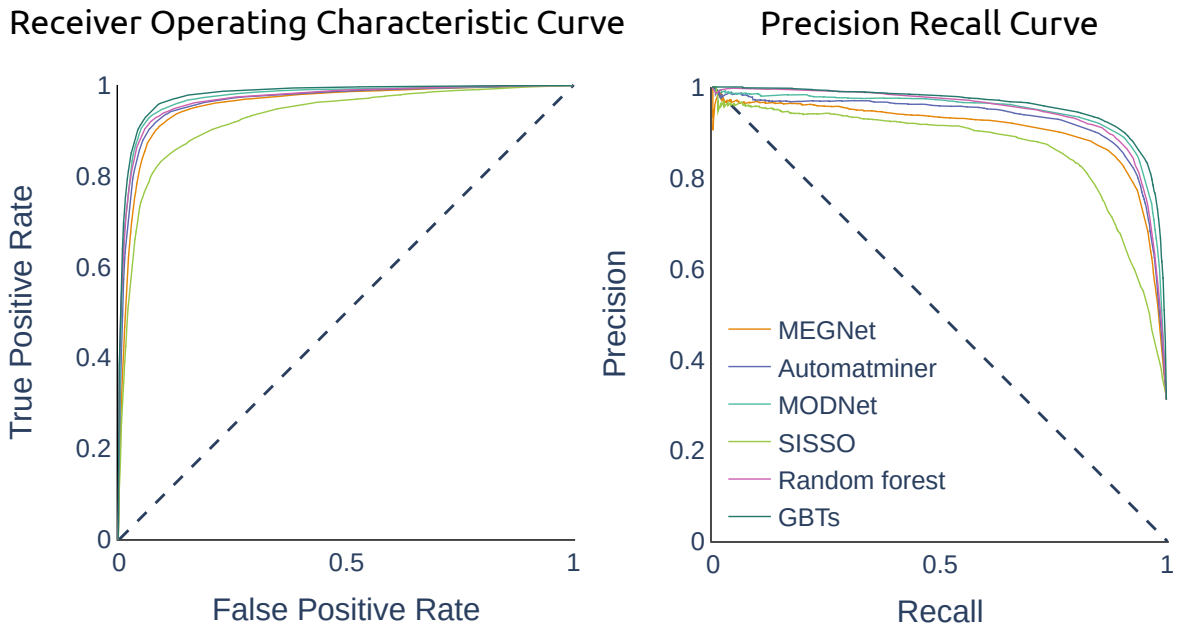


Figure 4.2: Receiver operating characteristic (ROC) curve and precision-recall curve for distinguishing topologically nontrivial material (NTM) from trivial insulator (TrI) in dataset MAT. They demonstrated the model performance across different probability thresholds. A higher ROC curve means the NTM predictions made by that model are more accurate. A higher precision-recall curve means the model could both make correct NTM predictions, and find out more NTMs across all actual NTMs.

Table 4.2: Comparison of the score for distinguishing NTM from TrI on dataset MAT. (in %)

	MEGNet	Automatminer	MODNet	SISSO	Random forest	GBTs
Accuracy	91.6	92.5	93.3	89.1	93.0	94.0
F_1 score	86.4	87.9	89.1	82.0	88.6	90.3
Precision	86.7	88.7	91.0	84.4	89.7	90.8
Recall	86.1	87.2	87.2	79.7	87.6	89.8
ROC_AUC	95.7	96.5	97.3	92.9	97.0	98.0

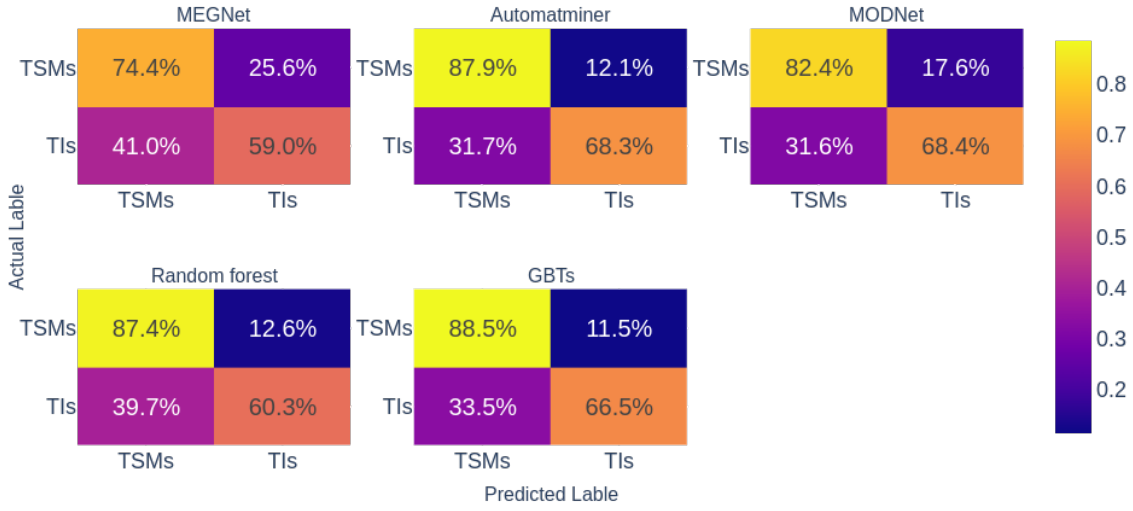


Figure 4.3: Normalized (over each actual conditions) confusion matrices for distinguishing topological semimetals (TSMs) and topological insulators (TIs) in dataset MAT.

task is relatively simple. However, possibly limited by insufficient computing resources (further resulting in a limited feature space), the performance of SISSO is not impressive. For further tasks of benchmark, SISSO will not compared.

Node 2

At node 2 of tree 1, the task is to distinguish 5,005 TSMs and 3,077 TIs. Figure 4.3 shows the normalized confusion matrices based on the boolean prediction of each algorithm. ROC curve and precision-recall curve are plotted in Figure 4.4 according to the probability prediction. An overview of the accuracy, precision, recall, F_1 score, and ROC-AUC is illustrated in Table 4.3.

The performance of each algorithm decreases, mainly due to the reduction of the dataset size and the increased task difficulty. Specifically, identifying TIs has proven to be exceedingly challenging, as the algorithms tend to classify materials as TSMs. Even the best-performing algorithm (highest true negative rate), Automatminer, can only correctly identify 68.3% of TIs, with a substantial 31.7% of TIs being misclassified as TSMs. Various indicators consistently show that the performance of GBTs and AM are relatively good,

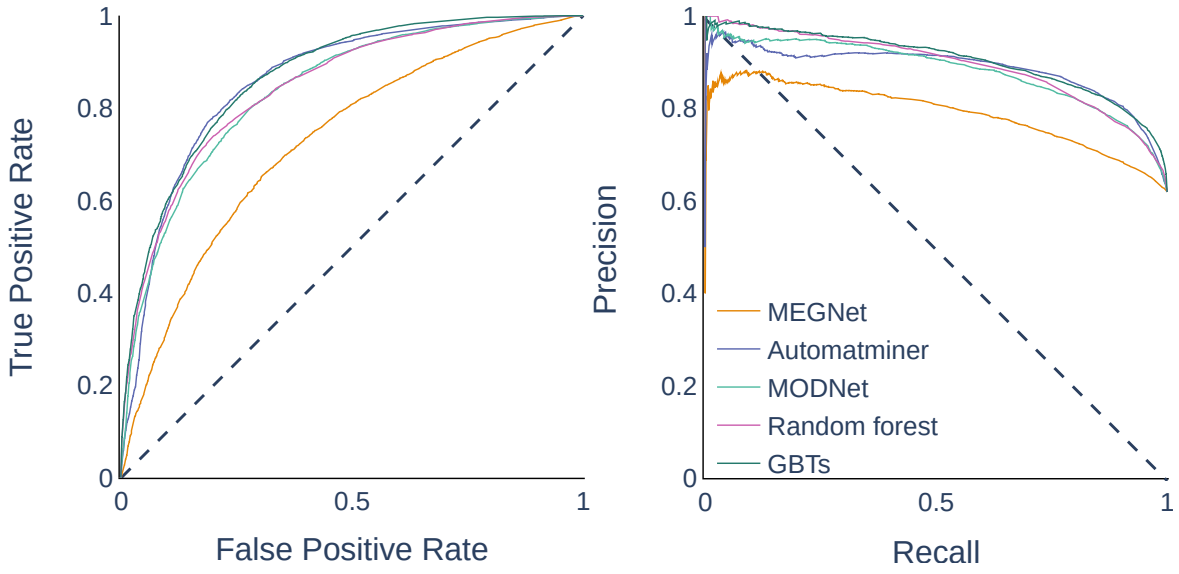


Figure 4.4: Receiver operating characteristic (ROC) curve and precision-recall curve for distinguishing TSMs and TIs in dataset MAT. They demonstrated the model performance across different probability thresholds. A higher ROC curve means the NTM predictions made by that model are more accurate. A higher precision-recall curve means the model could both make correct NTM predictions, and find out more NTMs across all actural NTMs.

Table 4.3: Comparison of the score for distinguishing TSMs and TIs in dataset MAT. (in %)

	MEGNet	Automatminer	MODNet	Random forest	GBTs
Accuracy	68.5	80.5	77.0	77.1	80.1
F_1 score	74.6	84.8	81.6	82.5	84.7
Precision	74.7	81.9	80.9	78.2	81.1
Recall	74.4	87.9	82.4	87.4	88.5
ROC_AUC	72.5	85.8	83.9	84.6	86.8

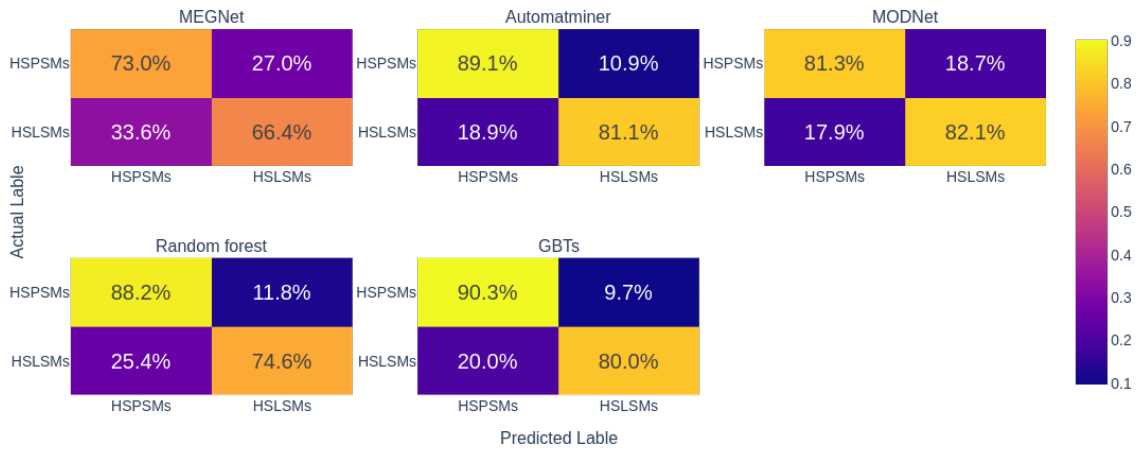


Figure 4.5: Normalized (over each actual conditions) confusion matrices for distinguishing high-symmetry-point semimetal (HSPSMs) and high-symmetry-line semimetal (HSLSMs) in dataset MAT.

Table 4.4: Comparison of the score for distinguishing HSPSMs and HSLSMs in dataset MAT. (in %)

	MEGNet	Automatminer	MODNet	Random forest	GBTs
Accuracy	70.0	85.5	81.7	81.9	85.6
F_1 score	72.5	86.9	82.8	84.1	87.1
Precision	72.0	84.8	84.3	80.4	84.2
Recall	73.0	89.1	81.3	88.2	90.3
ROC_AUC	75.5	92.0	89.5	89.8	92.9

modnet and RF are slightly worse, while MEGNet significantly drops due to its reliance on dataset size.

Node 3

At node 3 of tree 1, the task is to distinguish 2,713 HSPSMs and 2,292 HSLSMs. Based on the bool prediction of each algorithm, their normalized confusion matrices are compared in Figure 4.5, and corresponding accuracy, precision, recall, and F_1 score are listed in Table 4.4. According to their probability prediction, the ROC curve and precision-recall curve are plotted in Figure 4.6, with ROC-AUC listed in Table 4.4.

This task is relatively simpler than classifying TSMs and TIs. Despite further reductions in the dataset size, the performances of models have generally improved. While the differences in performance between algorithms are roughly consistent with the task of node 2.

Node 4

At node 4 of tree 1, the task is to distinguish 1,839 TIs and 1,238 TCIs. Figure 4.7 shows the normalized confusion matrices. Figure 4.8 plots their ROC curve and precision-recall curve. Table A.1 provides their score.

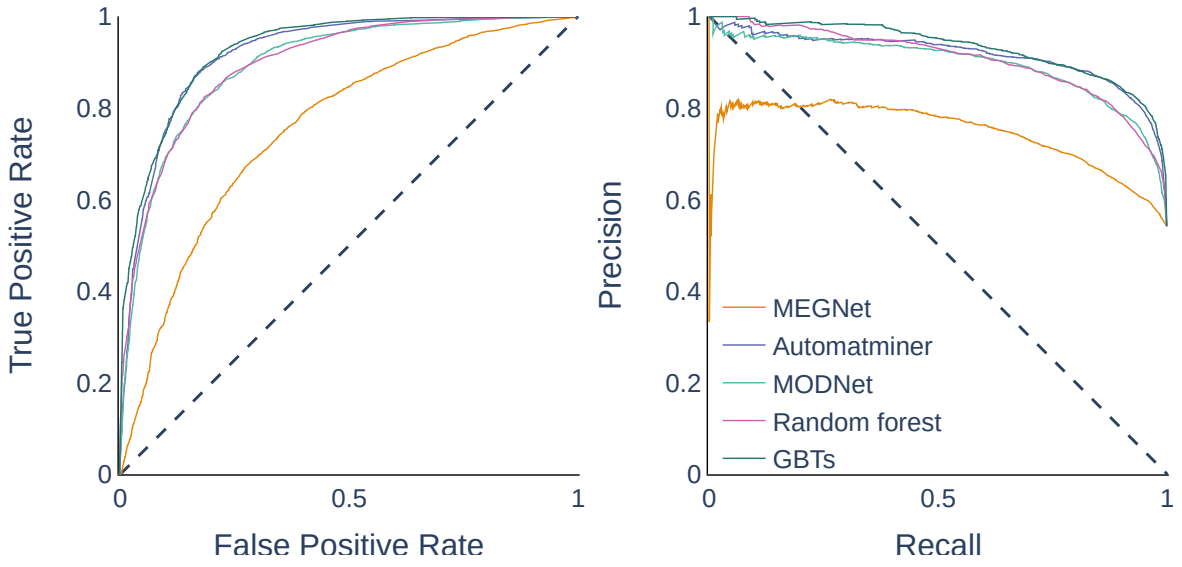


Figure 4.6: Receiver operating characteristic (ROC) curve and precision-recall curve for distinguishing HSPSMs and HSLSMs in dataset MAT.

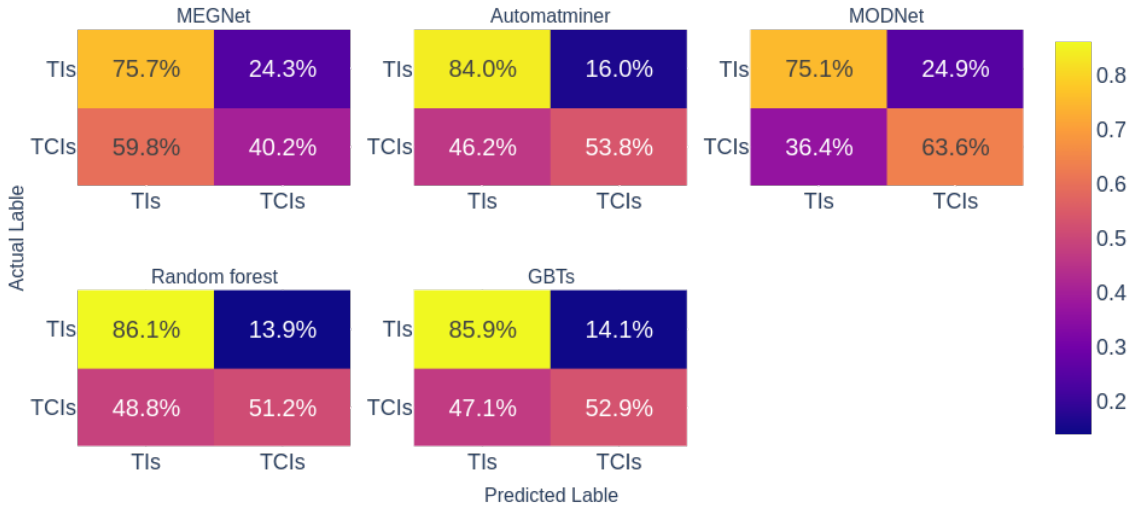


Figure 4.7: Normalized (over each actual conditions) confusion matrices for distinguishing topological insulator (TI) from topological crystalline insulator (TCI) in dataset MAT.

Table 4.5: Comparison of the score for distinguishing TI from TCI in dataset MAT. (in %)

	MEGNet	Automatminer	MODNet	Random forest	GBTs
Accuracy	61.4	71.8	70.5	72.1	72.6
F_1 score	70.1	78.1	75.2	78.6	78.9
Precision	65.3	73.0	75.4	72.4	73.0
Recall	75.7	84.0	75.1	86.1	85.9
ROC_AUC	61.4	77.1	76.1	76.8	78.4

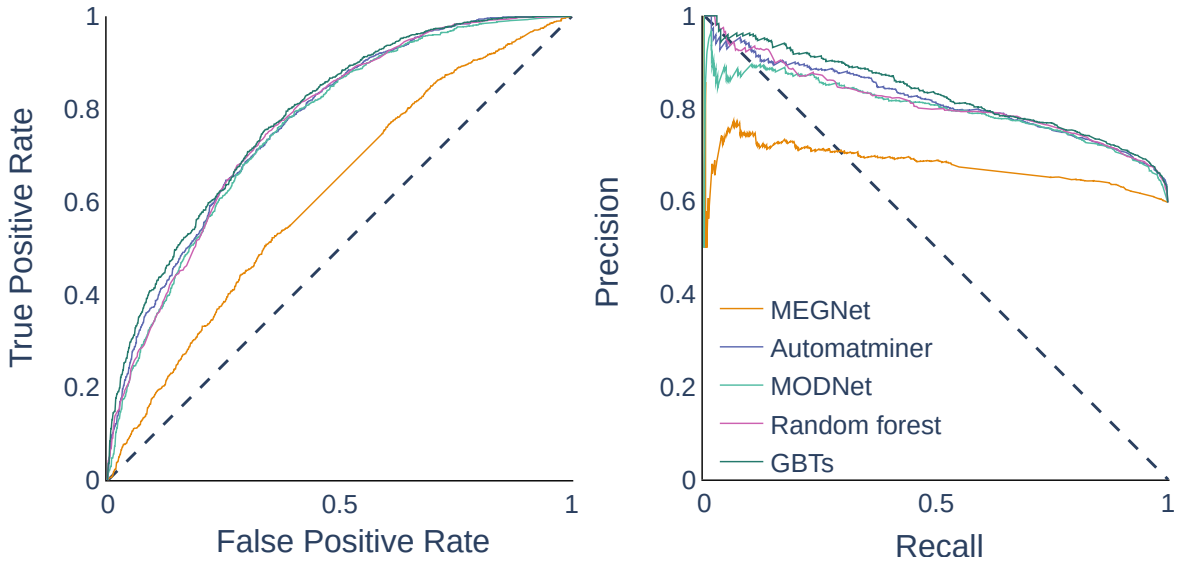


Figure 4.8: Receiver operating characteristic (ROC) curve and precision-recall curve for distinguishing TI from TCI in dataset MAT.

The difference between TI and TCI is difficult to find, even in the method of calculating symmetry indicator based on electron bands, the two are only distinguished based on the parity of the indicator. Therefore, the performance of each model further declined, and MEGNet even misclassified 59.8% of TCIs as TIs. MODNet correctly classified them two relatively fairly (recall equal to 75.1% and true negative rate equal to 63.6%), with a cost that 24.9% of TIs are classified as TCIs. While GBTs achieved the best accuracy at 72.6% and the highest F_1 score at 78.9%.

Figure 4.9 summarises the accuracy of predictions in 5 algorithms: MEGNet, Automatminer, MODNet, RF, and GBTs.

We combine all the results on nodes 1-4 and map their predictions into a final prediction, one of the five types. For each entry, the final prediction will be counted as correct only if all its related predictions (both itself at the end-node and classifications at its parent node) are correct. The accuracy of this approach is compared with the accuracy of the multiclass approach (tree 2) in the following section.

4.1.2 *Multiclass approach - tree 2*

5 algorithms (MEGNet, Automatminer, MODNet, RF, and GBTs) are tested on the full dataset MAT labeled with 5 types of materials, including 17,813 TrIs, 2,292 HSLSMs, 2,713 HSPSMs, 1,238 TCIs, and 1,839 TIs. In Figure 4.10, we display the accuracy of predictions made from two approaches, tree 1, the hierarchical approach, and tree 2, the multiclass approach. Table 4.6 provides a detailed comparison of each algorithm.

The normalized confusion matrices of each algorithm are shown in Figure 4.11 to Figure 4.15. They exhibit same trends as hierarchical approach: 1. Most TrIs are correctly labeled. 2. Distingushing 4 types of NTMs are difficult, algorthms tend to classify materials as TSMs more than TIs. 3. Among the five algorithms considered, GBTs gave the best performance. (Only for diagnosing TCIs and TIs, it is worse than MODNet.)

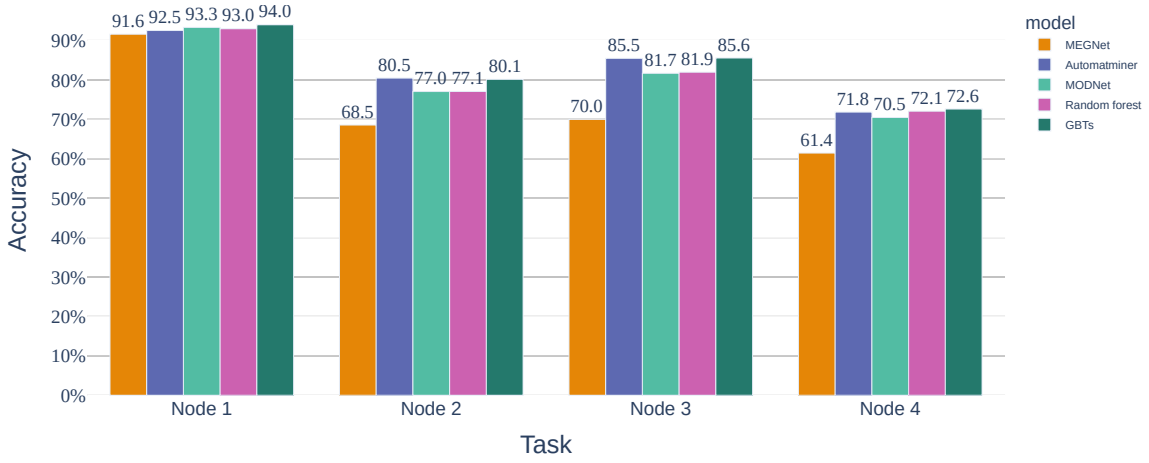


Figure 4.9: Comparison of the accuracy of each node. (in %)

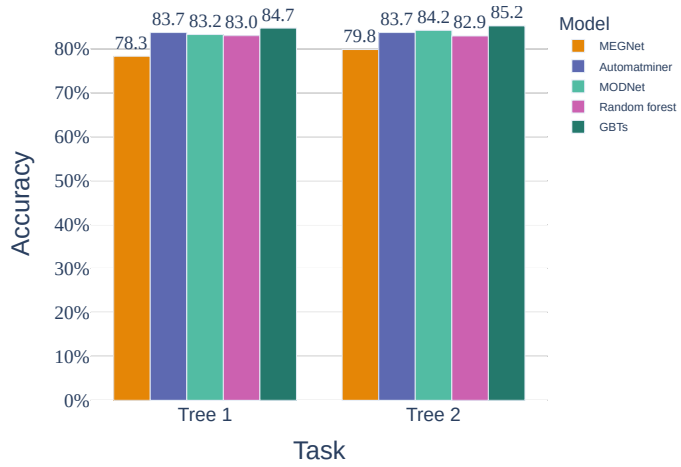


Figure 4.10: Comparison of the accuracy of tree-1 and tree-2. (in %)

Table 4.6: Comparison of the score for category materials in dataset MAT into 5 types. (in %). Accuracy is for 5 types of classification, other metrics are for classifying NTMs and TrIs, by mapping the 5 types of predictions into their parent class.

	MEGNet	Automatminer	MODNet	Random forest	GBTs
Accuracy	79.8	83.7	84.2	82.9	85.2
F_1 score	85.6	86.8	88.7	85.9	88.7
Precision	87.7	92.5	92.0	92.8	93.6
Recall	83.6	81.7	85.7	79.9	84.3

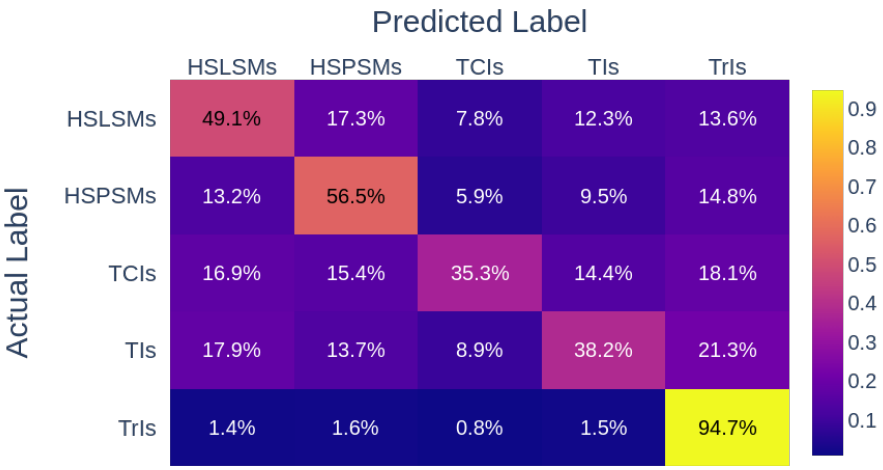


Figure 4.11: Normalized (over actual) confusion matrices of MEGNet.

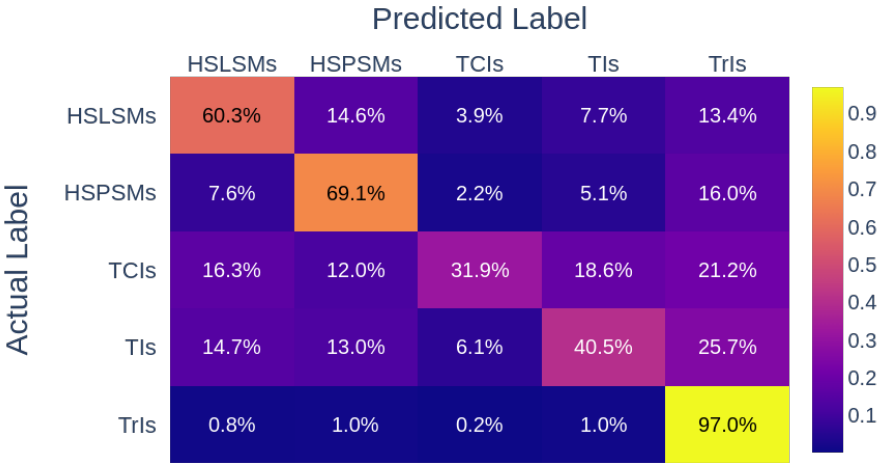


Figure 4.12: Normalized (over actual) confusion matrices of Automatminer.

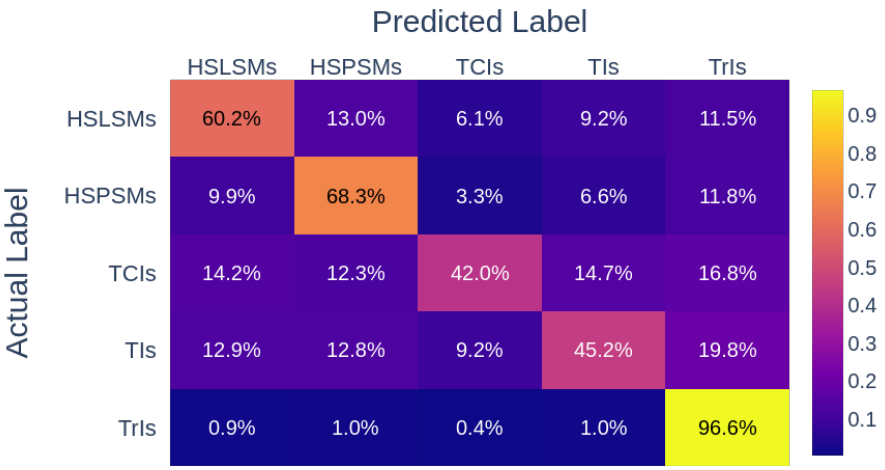


Figure 4.13: Normalized (over actual) confusion matrices of MODNet.

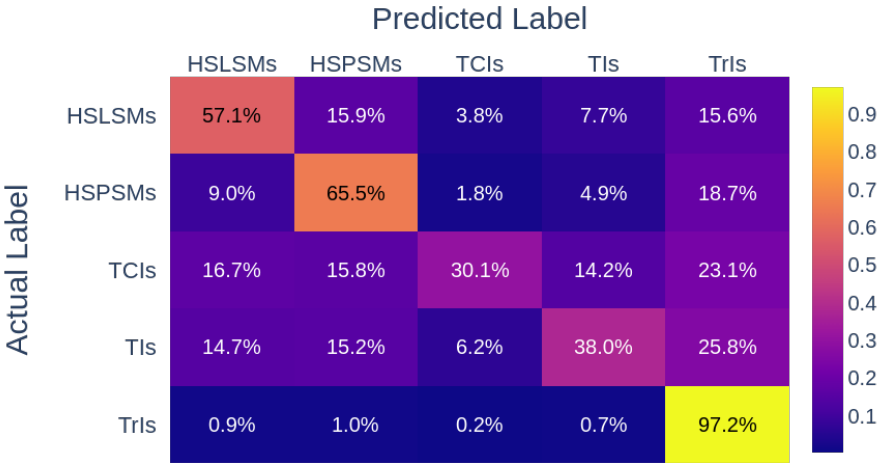


Figure 4.14: Normalized (over actual) confusion matrices of Random forest.

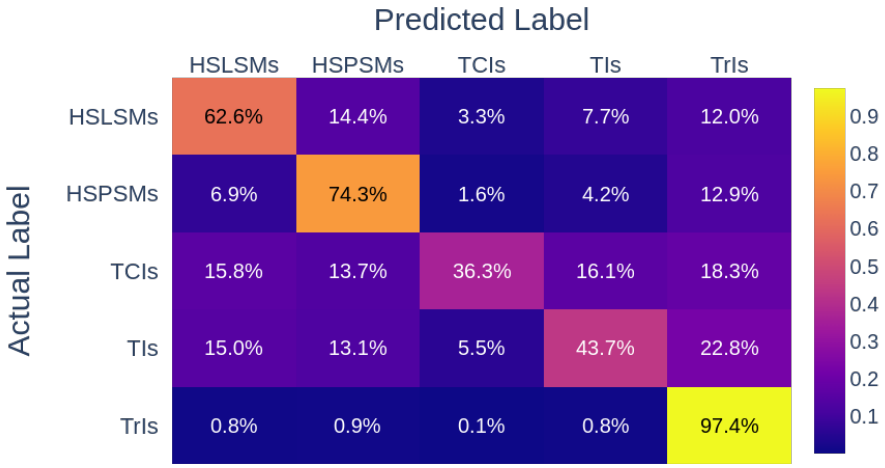


Figure 4.15: Normalized (over actual) confusion matrices of Gradient boost trees.

Table 4.7: Nested cross-validation and train-test results (accuracy in %)

Test\Train	M	T	M\T	$M \cap T$	T\M	(M\T)∪(T\M)	$M \cup T$
M\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\M	71.8	73.2*	69.8	70.2	72.1*	73.3*	74.3*
NCV	85.7	80.7	84.1	85.1	72.1	79.9	82.9
$M \cup T$	81.8	80.6	79.3	79.0	77.5	81.4	82.9

Overall, thanks to the fact that the model can learn on a larger amount of data at the same time, the multiclass approach performs better than the hierarchical approach, MEGNet, MODNet, and GBTs reach higher accuracy, Automatminer keeps the same and random forest is 1% worse. Among different algorithms, GBTs is adopted for further study, which achieves the highest accuracy and F_1 score.

4.2 GENERALIZATION TEST

We adopt the multiclass approach with GBTs to do 7 groups of generalization tests on our full dataset $M \cup T$, as shown in Figure 3.2. Dataset $M \cup T$ is counted as three sub-datasets: $M \backslash T$, $M \cap T$, and $T \backslash M$ for illustrating the results. Table 4.7 contains the classification accuracy for these different train-test combinations. Complementary metrics (i.e. F_1 score, precision, and recall) are computed based on the classification between NTMs and TrIs, by mapping the 5 types of predictions into their parent class, shown in Table 4.8. Results on the training set evaluated by NCV are indicated by a "*".

In principle, the NCV score provides a comprehensive assessment of how the training model will perform on new data. However, in our tests, the benchmark model trained on M, whose NCV accuracy is 85.7% (84.8% on dataset $M \backslash T$, 86.1% on dataset $M \cap T$) does not generalize well on dataset $T \backslash M$, resulting in an accuracy of 71.8%, corresponding to a decrease of 13.9%. Although this drop appears in most of the tests: for 4 groups of

Table 4.8: Nested cross-validation and train-test results(in %).

Test \ Train		M	T	M\T	$M \cap T$	T\M	(M\T)∪(T\M)	$M \cup T$
F_1 score	M\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\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\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\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\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\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

generalization tests (along column M, T, $M\backslash T$, and $M \cap T$), the NCV accuracy is higher than test accuracy on non-overlapping sets; And in all sub-datasets (along every row), the accuracy of the NCV (indicated by a "*") is always higher than its test accuracy. It is worth to notice that for 2 generalization tests (along column T\M and (M\T)∪(T\M)), the test accuracy on non-overlapping sets is higher than the NCV accuracy. Moreover, along each column, the accuracy on $T\backslash M$ is always the lowest, which indicates that predicting the topological class on materials provided in dataset $T\backslash M$ seems to be more difficult than on materials provided from Materiae. Other scoring metrics represented in Table 4.7 also show the same trend. We advance four main hypotheses to explore this bias.

The first noticeable difference relies in the label distribution. As can be seen from Figure 2.6, the proportion of TrIs is higher in $M\backslash T$ than in $T\backslash M$. According to the benchmark test of the hierarchical approach, task of node 1, TrIs are easier to distinguish from NTMs, and the refinement classifications of NTMs are relatively difficult. Therefore the proportion of TrIs affects the global accuracies.

A second difference can be found in the elemental distribution of the datasets. It is found that the following elements are more present in $T\backslash M$ than any other dataset: Ne, Mn, Fe, Eu, Gd, Po, Rn, Ra, Am. (The element is selected according to its amount in dataset $M\backslash T$, $M \cap T$, and $T\backslash M$: $N_{M\backslash T}$, $N_{M\cap T}$, and $N_{T\backslash M}$. when they satisfy equation: $\log_{10}[N_{T\backslash M}/(N_{M\backslash T} + N_{M\cap T})] > 0.2$.) Therefore we recompute the model's performance without these elements. Table 4.9 contains the corresponding accuracies, F_1 score, precision, recall, and the proportion of these materials. In general, an increase in performance is found when excluding the above elements. Predictions for materials containing these atoms are therefore more error-prone. As expected, this drop in accuracy is more important for dataset $T\backslash M$ (3% compared to 0.5%), as it contains more of the above-mentioned elements. Driven

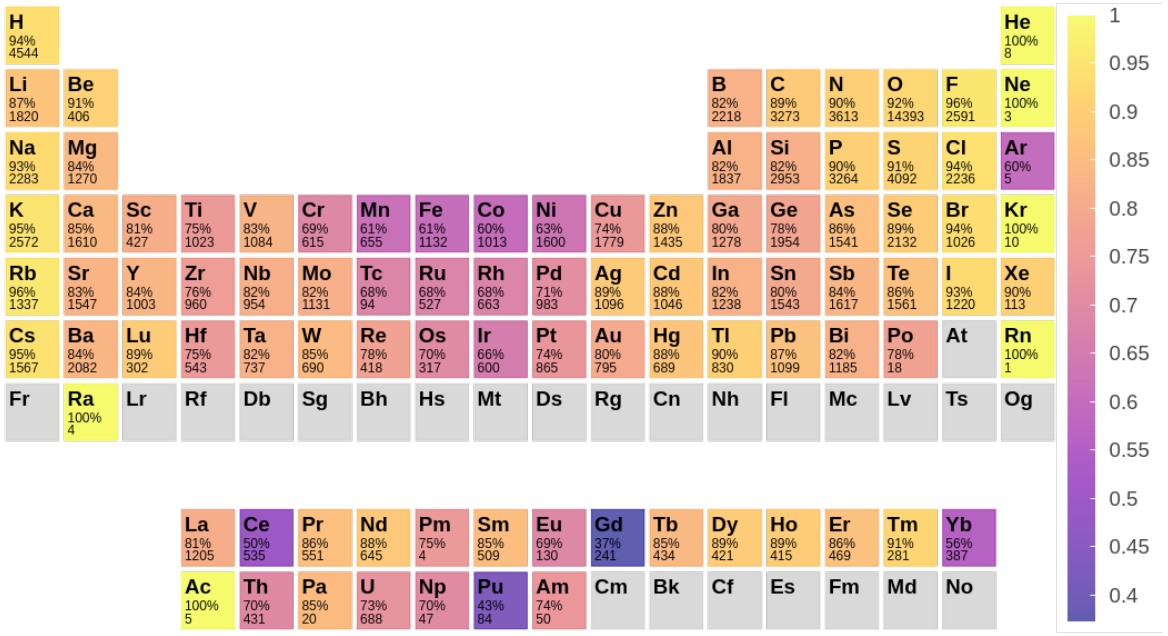


Figure 4.16: The accuracy of materials in nested cross validation on dataset $M \cup T$. Displayed is the accuracy of materials containing a given element

by this fact, we further analyze error-prone elements on dataset $M \cup T$. A filter criteria for an error-prone element is set as follows. First, the amount of materials containing the error-prone element should be larger than 30. Second, the accuracy for materials containing the error-prone element should be lower than 75%. Finally, the F_1 score for materials containing the error-prone element should be lower than the F_1 score for materials not containing the error-prone element. The following elements are identified as such: Cr, Mn, Fe, Cu, Tc, Eu, Os, and Np. Table 4.10 contains the accuracies, F_1 score, precision, and recall based on the presence of the previous elements.

A third factor that might be playing a role is that around half of the compounds in dataset $T \backslash M$ have an unknown magnetic attribute, as they have not been linked back successfully to the Materials Project. Table 4.11 investigates both the impact of elements and the presence of magnetic information. As can be seen, excluding certain elements (the element either occurs less than 30 times in dataset $M \backslash T$, $M \cap T$ or $T \backslash M$, or it is an error-prone element) 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 all three above factors, we define datasets $\widetilde{M \backslash T}$, $\widetilde{M \cap T}$ and $\widetilde{T \backslash M}$. They are formed by sampling 3372 materials with the same criteria (as shown in Table 4.11) from each corresponding dataset, such that the amount of compounds per class is the same across the three datasets (i.e. 2339 TrI, 315 NPSM, 279 NLSM, 271 TI, and 168 TCI). Table 4.12 summarizes the NCV results, the accuracy becomes 78.9%, 78.9%, and 77.3%, with a considerable decrease in the observed differences (1.6% instead of 13.9%).

Finally, a fourth reason involves a heterogeneous distribution of the datasets in feature space. We use the Euclidean-like distance defined in Eq. 3.1 and heterogeneity metric defined in Eq. 3.2 to estimate the presence of 'similar' points in the dataset. Close neighbors should roughly result in a better estimate. Figure 4.17 depicts the heterogeneity metric

Table 4.9: Comparisom of the score (in %) on datasets $M\backslash T$, $M \cap T$ and $T\backslash M$ when excluding certain materials. The material is excluded if it contains at least one of the following elements: Ne, Mn, Fe, Eu, Gd, Po, Rn, Ra, Am. The first three columns are from NCV (nested cross-validation) on each sub-dataset, and the last three columns are from NCV on dataset $M \cup T$.

		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
proportion		3.1	1.9	15.9	3.1	1.9	15.9

Table 4.10: Comparing the score (in %) of nested cross-validation result on the dataset $M \cup T$ for excluding materials contains from Cr, Mn, Fe, Cu, Eu, Os and Np elements.

		Cr	Mn	Fe	Cu	Eu	Os	Np
Accuracy	inc	67.0	59.8	60.8	73.5	65.4	67.2	72.3
	exc	83.2	83.4	83.7	83.4	83.0	83.1	82.9
F_1 score	inc	83.8	78.9	80.3	81.2	75.2	83.3	87.3
	exc	86.9	87.1	87.2	87.2	86.9	86.8	86.8
Precision	inc	91.8	90.4	87.4	94.2	89.1	89.1	100.0
	exc	92.9	92.9	93.2	92.8	92.9	92.9	92.9
Recall	inc	77.1	70.1	74.3	71.3	65.1	78.2	77.4
	exc	81.6	81.9	81.9	82.3	81.6	81.5	81.5

Table 4.11: Comparing the score of NCV result on the dataset $T\backslash M$ for excluding materials containing certain elements and further excluding materials without MP-ID. (the element either occurs less than 30 times in dataset $M\backslash T$, $M \cap T$ or $T\backslash M$, or it is Cr, Mn, Fe, Cu, Eu, Os or Np)

	Count	Accuracy	F_1 score	Precision	Recall
$T\backslash M$	9925	72.0%	78.2%	89.1%	69.7%
Exc_Eles	7364	77.3%	79.9%	90.1%	71.8%
Exc_Eles_MP	3550	78.3%	82.9%	90.4%	76.6%

Table 4.12: The NCV result on the dataset $\widetilde{M\backslash T}$, $\widetilde{M\cap T}$ and $\widetilde{T\backslash M}$. (in %)

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

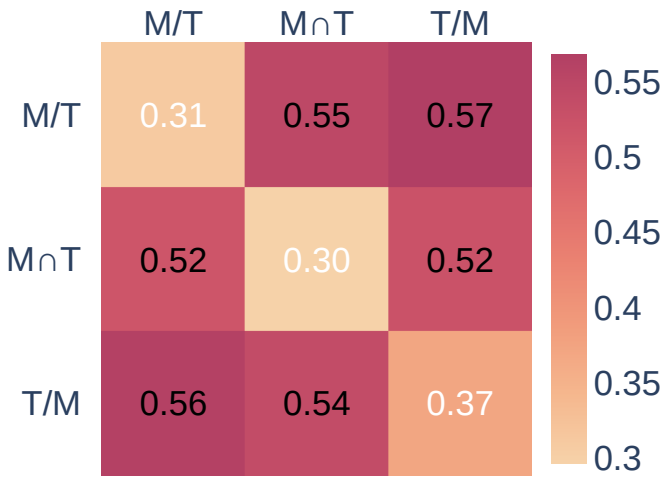


Figure 4.17: Heterogeneity metric between datasets: $M\backslash T$, $M\cap T$ and $T\backslash M$

between dataset $M\backslash T$, $M\cap T$ and $T\backslash M$. As can be seen, lower values lie on the diagonal, showing that points are more closely placed inside a dataset than between datasets. This may explain why the models generally do not generalize well to the other datasets. Moreover, the trend of the mean distance inside each dataset (on the diagonal) may explain the trend of the NCV accuracy score on datasets $\widetilde{T\backslash M}$, $\widetilde{M\backslash T}$ and $\widetilde{M\cap T}$, namely, the dataset which is more closely placed has a higher NCV score.

4.3 PREDICTIONS

Although the model constructed based on GBTs is not flawless, it already demonstrates the ability to make reasonably accurate predictions. Furthermore, given a relatively comprehensive data space is established from dataset $M\cup T$, a combination of two databases Materiae and TMD, it is probable that a model trained on this combined dataset will have better predictive capabilities. Specifically, new data points are likely to fall within this space. We conducted two tests to understand the capabilities of the model: 1. On existing materials that have been analyzed with DFT calculations. 2. On a few newly created structures inspired by existing materials.

During the cleaning process for dataset T (introduced in sect. 2.2), we removed 672 entries in database TMD due to conflicting topological types while being related to the same MP-ID. These were grouped into set A. 223 entries were removed from the intersection of database Materiae and TMD since they were labeled with different types in the two databases. Those

were grouped into set B. Here, we predict the topological type of these 895 materials with a model trained on dataset $M \cup T$. The complete results are shown in Table C.1 and Table C.2 in order of ascending MP-IDs. For the materials in set A, their topological type is only provided on TMD. 63 entries (related to 29 MP-IDs) have a prediction different from any of their labeled type, for instance, materials whose MP-ID is "mp-1079901". For the other 609 entries (related to 290 MP-IDs), the prediction agrees at least with one of their labeled type of materials with the same MP-ID, for instance, materials whose MP-ID is "mp-1002133" or "mp-1018094". For the materials in set B, their topological type is provided both on Materiae and TMD. 62 entries have a prediction that agrees with their labeled type on Materiae, for instance, the material whose ICSD-ID is "168895". 100 entries have a prediction that agrees with their labeled type on TMD, for instance, the material whose ICSD-ID is "167854". 61 entries have a prediction different from any of their labeled type, for instance, materials whose ICSD-ID is "612769" or "290935".

Besides these two sets, we pick one material to illustrate the possible prediction for newly created materials. The material $CeNiGa_2$ (ICSD-ID:188329) is a TI from space group 65 (listed in Table C.1). We respectively replaced one element in its structure, for instance, changing "Ce" to "Ti" to form $TiNiGa_2$. Following a structural relaxation with Abiml from Abipy, we predicted the topological type of the new structure. The results are detailed in Table C.3. All the new structures maintain the same space group, with the majority predicted as TI and a few as TCI. These predictions are relatively reasonable, especially considering a same material, whose ICSD-ID as 621131 is predicted as TCI by DFT calculations. The verfication of these predictions through DFT calculations is still in progress.

5 RESULTS OF DISTINGUISHING TOPOLOGICAL MATERIALS FROM TRIVIAL INSULATORS

5.1 COMPARISON

In this binary classification task, we consider NTMs as positive and TIs as negative, thus precision measures the reliability of NTM predictions, and recall measures the ability to detect all NTMs. The comparison between results from GBTs, topogivity, and t-SNE is presented in Figure 5.1, Table 5.1 and Figure 5.2.

Figure 5.1 and Table 5.1 show results with the boolean predictions of each algorithm. GBTs shows the best performance with the highest accuracy, F_1 score, and precision. Thanks to its usage of high-dimensional feature space to represent materials that fully describe the properties of materials. Figure 5.2 shows the tradeoff of scores under different thresholds. GBTs always has better score, at the intersection point of Topogivity and t-SNE, the two methods achieve consistent scores. (Topogivity is done on a subset of 35522 materials.)

While the scores of Topogivity and t-SNE may be lower than those of GBTs, their advantage lies in their simplicity, making them easy to interpret. Topogivity making prediction by a chemical rule where the topogivity of element E approximately represents its inclination to form an NTM. Figure 5.3 represents a table of our trained topogivities. They are the result from a model selected with 5-fold CV procedure, specifically, fitted with C = 3.73×10^4 on a subset of 35522 materials. It obtained 87.3% accuracy, 79.5% F_1 score, 85.7% precision, and 74.1% recall when evaluated on the whole training set.

To compare the topogivities for 54 elements between our results (T_m) and results from Ref. [121] (T_n), we use Eq. 5.1 to quantify their relative difference, shown in Figure 5.4. They are mostly consistent, except for three elements: Li, Si, and Sb, which exhibit different signs. We test them on a subset of 18637 materials (within 54 elements), and the scores are compared in Table 5.2.

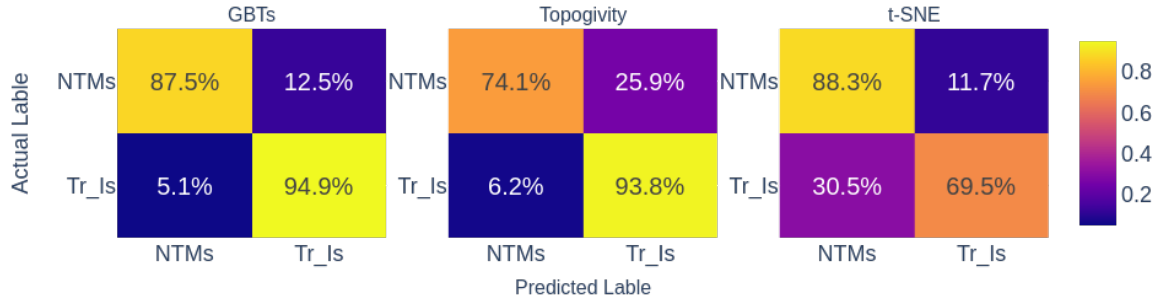
$$\frac{2sign(T_m * T_n) * |T_m - T_n|}{2 + |T_m| + |T_n|} \tag{5.1}$$

We consider the unsupervised algorithm t-SNE to investigate features that are important for topological materials. The results above are based on two features: "max packing efficiency" (MPE) and "fraction p valence electrons" (FPV). We represent the whole dataset into the feature space and apply t-SNE to project them into 2 other variables aiming for a natural separation of NTMs and TrIs. The visualization of the result is shown in Figure 5.5,

Table 5.1: Comparing the score of GBTs, Topogivity, and t-SNE (in %).

	Accuracy	F_1 score	Precision	Recall	ROC_AUC
GBTs	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

Figure 5.1: Normalized confusion matrices for distinguishing topologically nontrivial material (NTM) from trivial insulator (TrI) on dataset $M \cup T$.



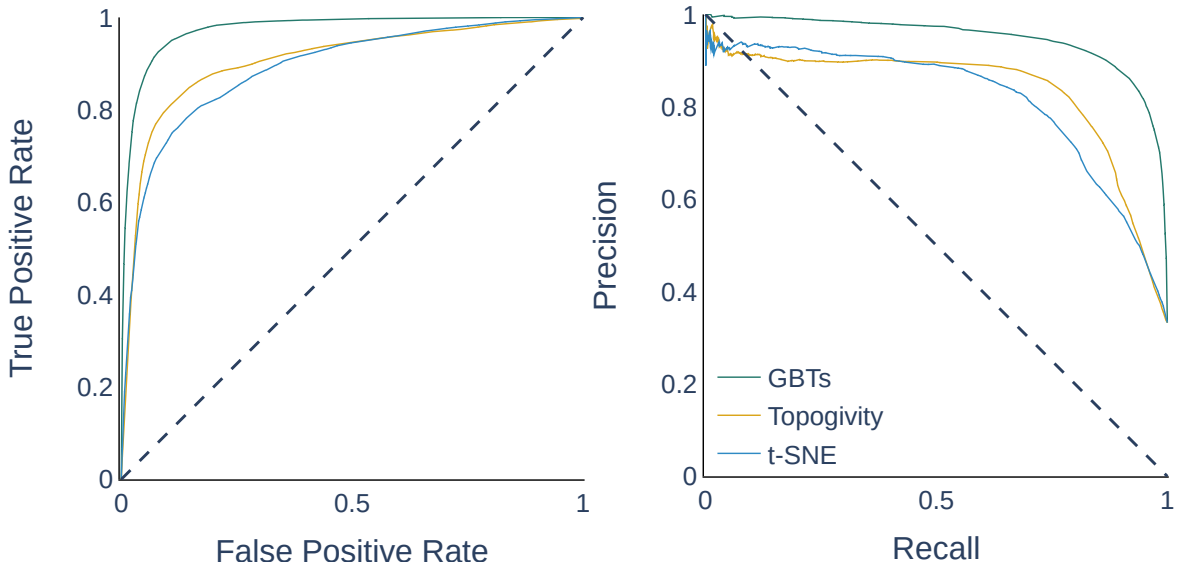


Figure 5.2: Receiver operating characteristic (ROC) curve and precision-recall curve for distinguishing topologically nontrivial material (NTM) from trivial insulator (TrI) on dataset $M \cup T$.

Table 5.2: Comparing the test score of topogivities from Ma [121] (T_m) and our dataset (T_n) on 18637 materials.

	Accuracy	F_1 score	Precision	Recall
T_n	90.2%	76.8%	83.5%	71.2%
T_m	89.6%	77.3%	77.1%	77.5%

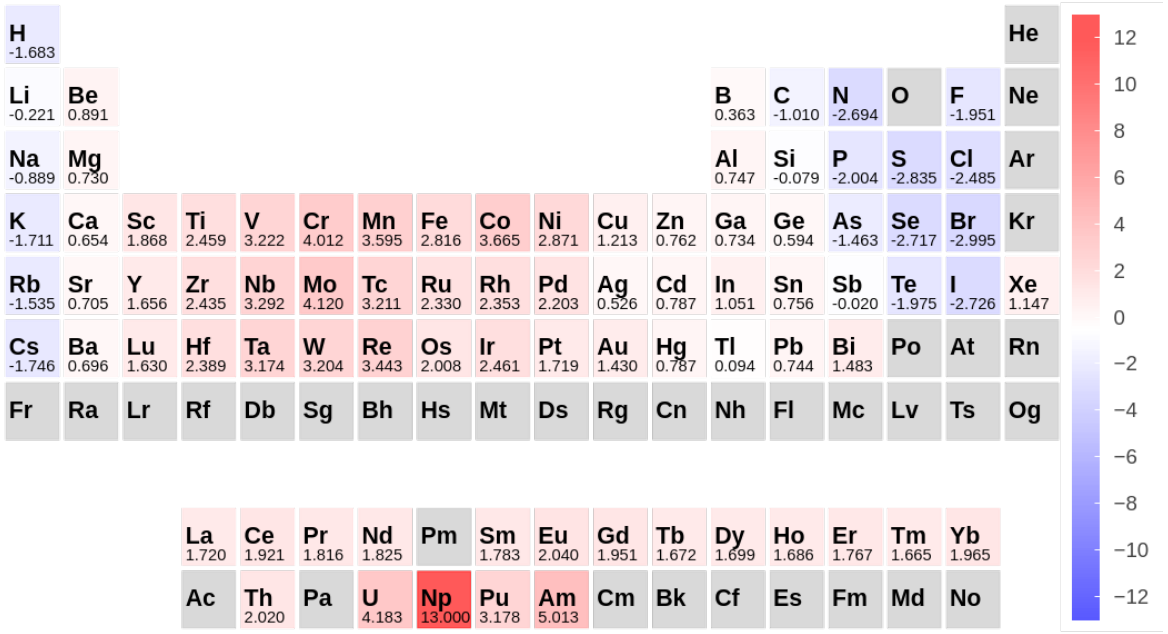


Figure 5.3: Periodic table of topogivities trained from dataset $M \cup T$. Existing topogivities are represented through numerical values with color coding, others are displayed in gray.

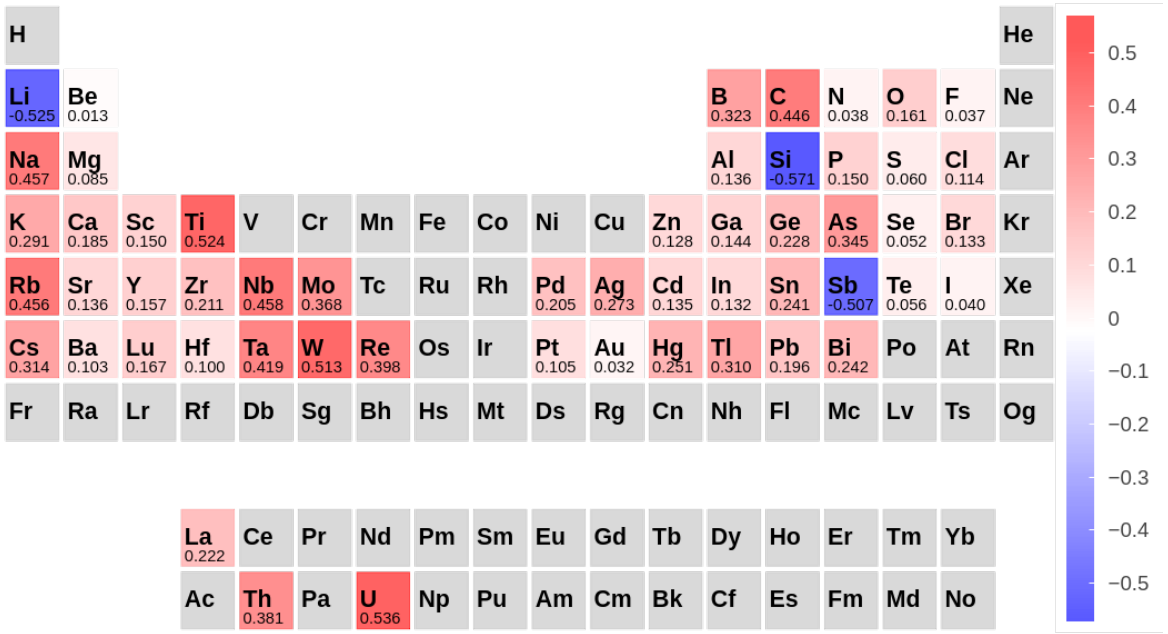


Figure 5.4: Comparison of topogivities for 54 elements between our results and the results from Ref. [121]. The color coding quantifies their relative difference. When the signs are different, the value is negative, highlighted in blue.

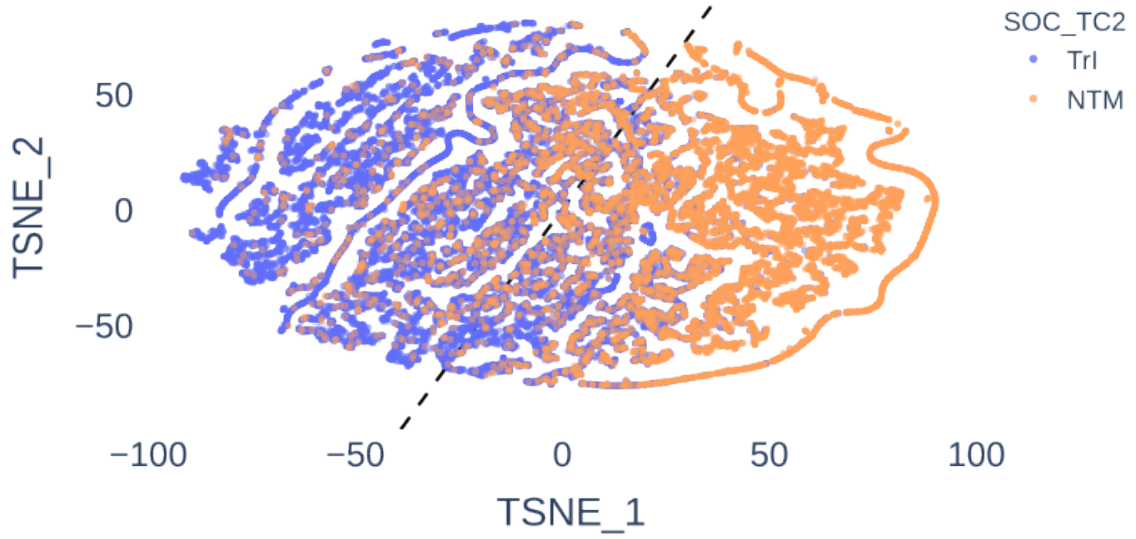


Figure 5.5: Visualization of t-SNE result on dataset $M \cup T$.

the points are located with the projected variable from t-SNE and colored based on their actual type. It shows that NTMs and TrIs are rather clearly located on two sides of the space. If simply take the vertical line where TSNE_1 is equal to zero as the criterion to split NTMs and TrIS, we are already able to predict 75.7% materials correctly and detect 88.3% NTMs (recall). Furthermore, by the dashed dividing line (found with a soft-margin linear SVM in which C is equal to 1.0), the accuracy could reach 84.7%. This is still a bit lower than GBTs and Topogivity, but it shows that without using the target value, the model could find the underlying relations between features and the topology of materials. The method to select the features is explained in the following section.

5.2 FEATURES INTERPRETATIONS

For each feature, GBTs provides gain importance that measures the improvement in the loss function of the model when the feature is used. We first rank the features based on their importance for 5 types classification on three datasets: $M \backslash T$, $M \cap T$ and $T \backslash M$ in descending order, then take the intersection of the top 15 features, results to 3 features: MPE, FPV, and "Miedema deltaH_inter" (deltaH). MPE is a structure-based feature that represents the maximum possible packing efficiency within that crystal lattice. FPV is a composition-based feature that calculates the proportion of valence electrons in p orbitals among all valence electrons. The distributions of their values in dataset $M \cup T$ are displayed in Figures 5.6 and 5.7.

Figure 5.6 shows 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 arising of the topological phase with crystal-symmetry protection.

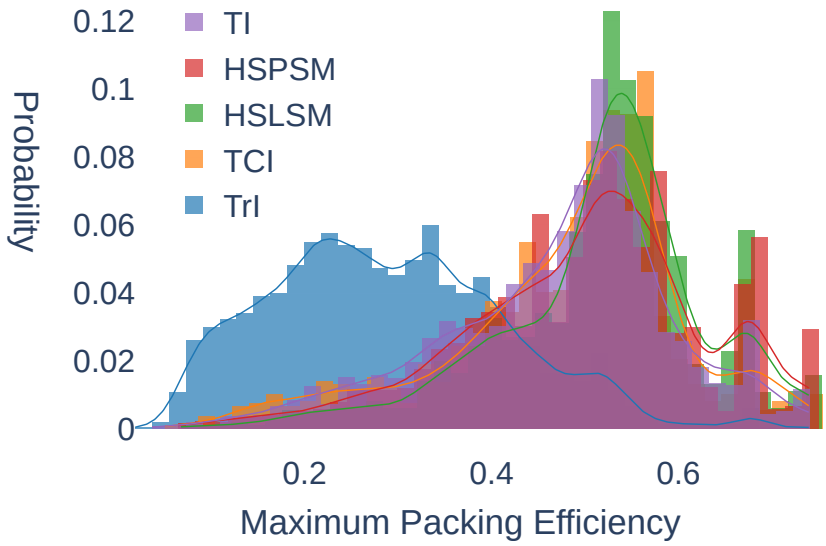


Figure 5.6: Distribution of 'max packing efficiency' in dataset $M \cup T$ under 5 topological types.

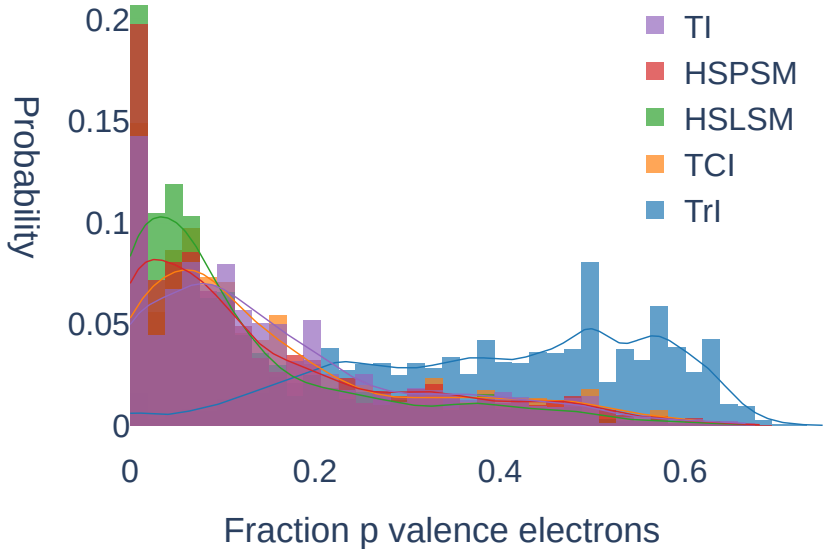


Figure 5.7: Distribution of 'fraction p valence electrons' in dataset $M \cup T$ under 5 topological types.

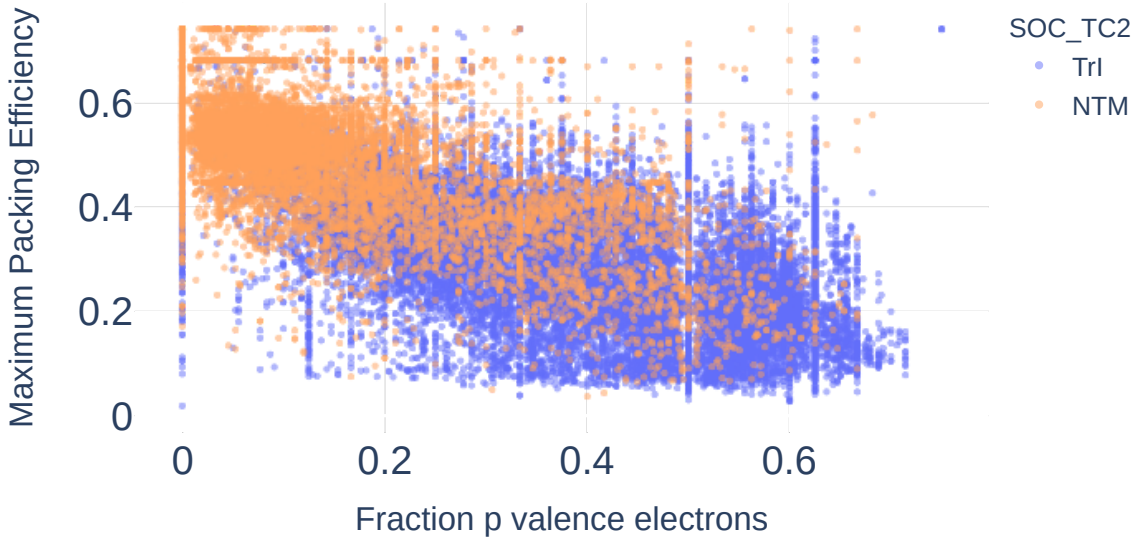


Figure 5.8: Visualization of dataset $M \cup T$ with two features: 'fraction p valence electrons' and 'max packing efficiency'.

According to Figure 5.7, NTMs tend to be formed more with atoms that have a lower fraction of p-orbitals valence electrons. This aligns with the trends in element topogivity, as depicted in Figure 5.3. Elements that have a higher fraction of valence electrons in p-orbitals, located in the top-right parts of the periodic table, display negative topogivities, indicating their inclination to form TrIs.

Figure 5.8 provides the visualization of dataset $M \cup T$ with the above two features. It demonstrates a simple distinction between TrIs and NTMs based on these two features.

The deltaH is the formation enthalpies (eV/atom) of the intermetallic compound based on the semi-empirical Miedema model [133]. The distribution of its values in dataset $M \cup T$ is displayed in Figure 5.10. A large number of TrIs have a zero value, this is due to the fact that some materials are not supported by this feature extractor in Matminer, and these missing values were all filled with zero. Although this does not carry to any physical meaning, in ML, it is one of the common methods for handling missing values. For the above t-SNE method, we discard this feature to obtain a well-interpreted model. As for the GBTs method, we explored the possible impact of this filling and illustrate the results in Table C.4. After removing the entries without the value of deltaH, 12518 entries are left, sorted into 3622 TrIs, 2843 HSPSMs, 2734 HSLSMs, 1339 TCIs, and 1980 TIs. Figure 5.10 displays the composition of topological classes within datasets $M \cup T$ containing known or unknown values of deltaH. It reveals that the distribution across each category is relatively even in cases where the value of deltaH is available. However, in instances where the value of deltaH is unavailable, a significant majority of materials fall into the TrIs. For the subset with known deltaH, the distribution of its value across the five classes is shown in Figure 5.11. They show the possible reasons for the utilization of this feature in the GBTs model. When this value is filled with 0, the material is more likely to be a TrI. For cases where deltaH is known, the values for TrIs are mostly not less than zero and the values for

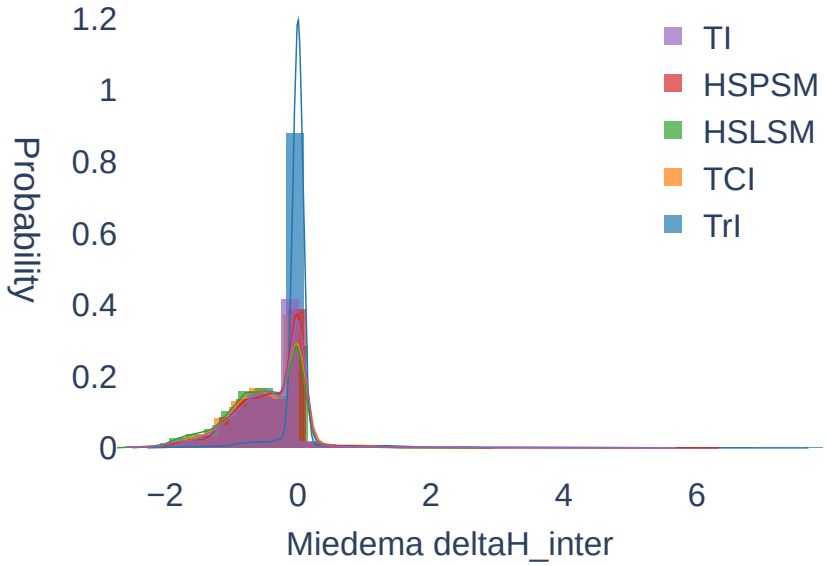


Figure 5.9: Distribution of 'Miedema_deltaH_inter' in dataset $M \cup T$ under 5 topological types.

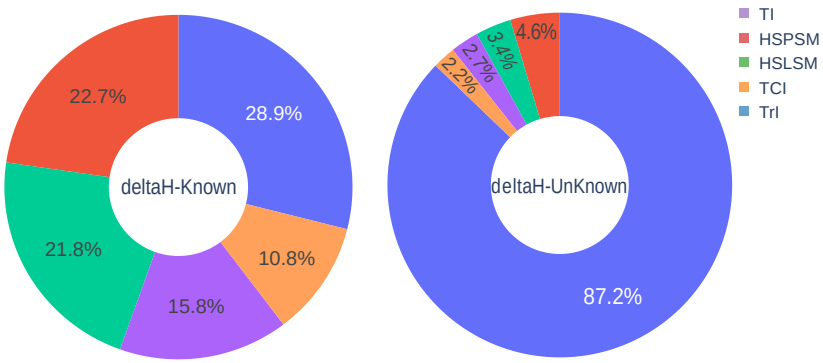


Figure 5.10: Composition of dataset $M \cup T$ when the value of 'Miedema_deltaH_inter' is available or unavailable.

NTMs are mostly less than zero. Echoing the classic notion in ML that "correlation does not imply causation", we currently lack a clear understanding of the reasons behind this phenomenon. One potentially enlightening direction could be to gain a comprehensive understanding of the implementation of this featurizer to uncover the underlying reasons.

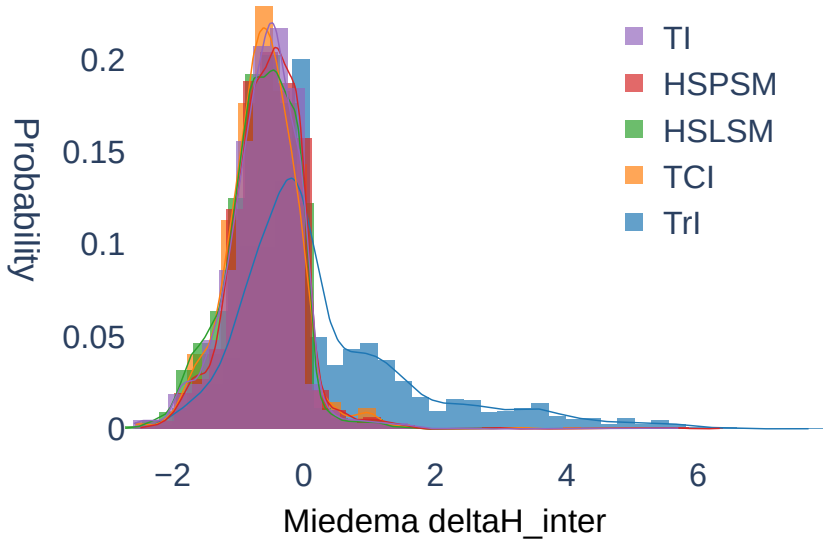


Figure 5.11: Distribution of existing 'Miedema_deltaH_inter' in dataset $M \cup T$ under 5 topological types.

CONCLUSION

Remarkable progress has been made in the field of topological materials, revealing both valuable new materials and fascinating physical phenomena. Through high-throughput calculations grounded in the analysis of symmetry-based indicators, thousands of topologically nontrivial materials have been predicted from the first principle, marking a transformative milestone in materials science. In this context, ML emerges as a powerful tool to advance current research by harnessing the wealth of existing data.

Through an exhaustive data collection and cleaning process, we have compiled an extensive dataset comprising 35,608 entries from two symmetry-indicator-based databases, namely Materiae and Topological Materials Database. This is the first integration of materials from distinct data sources, a development that paves the way for more comprehensive and profound ML research. Our research objectives were twofold. Firstly, we aimed to categorize materials into five distinct types, opening doors to novel material discoveries. We conducted a thorough benchmarking process to compare various ML methods, and GBTs was identified as the most promising model. We further delved into the influence of dataset differences on model performance with a series of generalization tests and discovered several differences between these datasets. Our robustly trained model achieved an impressive 82.9% accuracy for the classification into 5 topological types across a wide range of materials. We illustrated the capability of the model by predicting the topological type for both existing structures and newly created structures. This opens up the possibility of large material space scanning. For instance, combining with other ML models that predict the band gap, we can selectively choose some TIs with a large band gap to do further calculations for screening the ideal topological materials. Secondly, we embarked on a quest to identify the key factors influencing material topology. In the task of distinguishing NTMs from TrIs, we compared GBTs with two simple methods: Topogivity and t-SNE. GBTs can predict the topological type of a given material with an accuracy of 92.4%. Furthermore, we demonstrated that utilizing only two key features employed in GBTs, combined with the unsupervised technique t-SNE, a model with an accuracy of 84.7% was produced. Upon analyzing the distribution of these features, we found notable disparities between NTMs and TrIs. Together with Topogivity, they offer element-based intuitions for designing topological materials. This breakthrough highlights the potential of discovering critical features through ML approaches.

The application of ML has not only enabled us to discover new topological materials but has also allowed us to gain physical insights. However, this study represents just a starting point, with promising avenues for further exploration. One such direction involves conducting more comprehensive searches for new topological materials using ML models. Additionally, there is potential for gaining a deeper understanding of a fundamental question: what factors affect whether certain materials exhibit topological properties or not? Our study has revealed the critical role of a comprehensive database in these endeavors. The acquisition of a more extensive dataset, encompassing not only symmetry-indicator-based topological materials but also simulation outcomes derived from Wilson loops, along with experimental data, holds immense importance for advancing the further development of this research.

In the field of ML, data curation is crucial for obtaining robust models. We invested a significant amount of time in this task as data was not always clear and easily accessible. The inconsistencies between the two databases made it even more challenging. It required a detailed understanding of each database to complete the entire data-cleaning process. Therefore, in various domains, the establishment of open and uniform databases would greatly facilitate the development of the corresponding ML models.

APPENDIX A

This Appendix illustrates a group of example materials featurized with some attributes mentioned in sect. 1.2.3

Table A.1: A sample dataset with 9 features.

ICSD-ID	68091	192086	185442	39621
Chem. formula	Li Pt3 B	Mg Au Ga	Nb N	Rb4 (Cd Br6)
Space group number	189	189	186	167
Maximum Packing Efficiency	0.4479	0.6020	0.3331	0.4169
Miedema_deltaH_inter	-0.3455	-0.6115	-1.1143	0.0
Frac p valence electrons	0.0132	0.0250	0.3	0.2542
HOMO_character	1.0	1.0	1.0	2.0
HOMO_element	78	79	41	35
MagpieData maximum MendeleevNumber	72.0	74.0	82.0	95.0
Transition metal fraction	0.6	0.0	0.5	0.0
Maximum Electronegativity difference	1.2200	0.0	1.44	2.14

APPENDIX B

B

This Appendix provides a sample dataset from TMD in table B.1. Materials are listed in order of unique IDs (U-ID). Original info from TMD is listed, except the confirmed MP-ID (Conf. MP-ID). It is checked based on the similarity between the structure linked to ICSD and MP-ID. Details are explained in 2.2.

Table B.1: A sample dataset from TMD. Each unique ID (U-ID) is associated with at least one ICSD-ID, they share identical chemical formula (Chem. F.), space group number (SG), and topological type. In addition to these attributes, the MP-ID is retrieved from TMD. Notably, the confirmed MP-ID (Conf. MP-ID) is a new piece of information we added by verifying the structural similarity.

U-ID	Chem. F.	ICSD-ID	SG	MP-ID	Conf. MP-ID	Type_TMD
1	Mn2 (Ge Te4)	165681	62	NaN	NaN	trivial (LCEBR)
2	La Sb Te	601624	129	mp-1018747	mp-1018747	TI (SEBR) TCI
30	Cs (Sb Cl6)	155691	9	NaN	NaN	trivial (LCEBR)
31	Tl Fe (Mo O4)2	250337	62	mp-645796	mp-645796	SM (ES)
32	Mg Ge H12 (OF)6	401096	14	mp-541747	NaN	trivial (LCEBR)
58	K Mn Cl3	173944	62	mp-567530	mp-567530	SM (ES)
58	K Mn Cl3	32710	62	mp-23049	mp-23049	SM (ES)
61	Tb Ag As2	420615	62	mp-1186868	NaN	TI (NLC) TCI
174	V2 O5	43132	59	mp-624689	NaN	trivial (LCEBR)
174	V2 O5	94904	59	mp-25279	mp-25279	trivial (LCEBR)
178	K O2	38246	15	mp-1071539	NaN	SM (ESFD)
178	K O2	87184	15	mp-21325	NaN	SM (ESFD)
3675	Zn S	107149	156	mp-556280	mp-556280	trivial (LCEBR)
3675	Zn S	107140	156	mp-556716	mp-556716	trivial (LCEBR)
3675	Zn S	42848	156	mp-554713	mp-554713	trivial (LCEBR)
3675	Zn S	42844	156	mp-555583	mp-555583	trivial (LCEBR)
3675	Zn S	42192	156	mp-555773	mp-555773	trivial (LCEBR)
3675	Zn S	42995	156	mp-554829	mp-554829	trivial (LCEBR)
3675	Zn S	42991	156	mp-555543	mp-555543	trivial (LCEBR)
3675	Zn S	107137	156	mp-556485	NaN	trivial (LCEBR)
3675	Zn S	107177	156	mp-553880	mp-553880	trivial (LCEBR)
3675	Zn S	107158	156	mp-554630	mp-554630	trivial (LCEBR)
3675	Zn S	107157	156	mp-554999	mp-554999	trivial (LCEBR)
3675	Zn S	107141	156	mp-556815	mp-556815	trivial (LCEBR)

Continued on next page

Table B.1 – continued from previous page

U-ID	Chem. F.	ICSD-ID	SG	MP-ID	Conf. MP-ID	Type_TMD
3675	Zn S	42785	156	mp-556784	mp-556784	trivial (LCEBR)
3675	Zn S	107148	156	mp-556732	mp-556732	trivial (LCEBR)
3675	Zn S	42780	156	mp-557418	mp-557418	trivial (LCEBR)
3675	Zn S	15741	156	mp-556392	mp-556392	trivial (LCEBR)
3675	Zn S	42792	156	mp-554253	mp-554253	trivial (LCEBR)
3675	Zn S	42807	156	mp-555782	mp-555782	trivial (LCEBR)
3675	Zn S	42191	156	mp-556152	mp-556152	trivial (LCEBR)
3675	Zn S	42193	156	mp-557346	mp-557346	trivial (LCEBR)
3675	Zn S	42194	156	mp-557175	mp-557175	trivial (LCEBR)
3675	Zn S	42816	156	mp-554820	mp-554820	trivial (LCEBR)
3675	Zn S	42817	156	mp-554681	mp-554681	trivial (LCEBR)
3675	Zn S	42836	156	mp-556161	mp-556161	trivial (LCEBR)
3675	Zn S	42799	156	mp-557058	NaN	trivial (LCEBR)
3675	Zn S	42190	156	mp-556363	mp-556363	trivial (LCEBR)
3675	Zn S	42990	156	mp-555594	mp-555594	trivial (LCEBR)
3675	Zn S	42989	156	mp-555664	mp-555664	trivial (LCEBR)
3675	Zn S	42843	156	mp-554115	mp-554115	trivial (LCEBR)
3675	Zn S	15738	156	mp-553916	mp-553916	trivial (LCEBR)
3675	Zn S	15740	156	mp-557308	mp-557308	trivial (LCEBR)
3675	Zn S	42189	156	mp-556448	mp-556448	trivial (LCEBR)
3675	Zn S	42786	156	mp-556543	mp-556543	trivial (LCEBR)
3675	Zn S	37400	156	mp-557009	mp-557009	trivial (LCEBR)
3675	Zn S	15744	156	mp-555151	mp-555151	trivial (LCEBR)
3675	Zn S	42787	156	mp-556000	mp-556000	trivial (LCEBR)
3675	Zn S	42195	156	mp-557026	mp-557026	trivial (LCEBR)
3675	Zn S	42196	156	mp-556989	mp-556989	trivial (LCEBR)
3675	Zn S	42779	156	mp-556155	mp-556155	trivial (LCEBR)
3675	Zn S	15743	156	mp-554608	mp-554608	trivial (LCEBR)
3675	Zn S	107139	156	mp-557054	mp-557054	trivial (LCEBR)
3675	Zn S	107134	156	mp-556005	mp-556005	trivial (LCEBR)
3675	Zn S	107184	156	mp-554503	mp-554503	trivial (LCEBR)
3675	Zn S	107133	156	mp-556395	mp-556395	trivial (LCEBR)
3675	Zn S	37397	156	mp-557062	mp-557062	trivial (LCEBR)
3675	Zn S	107185	156	mp-554004	mp-554004	trivial (LCEBR)
3675	Zn S	107182	156	mp-555079	mp-555079	trivial (LCEBR)
3675	Zn S	107150	156	mp-556207	mp-556207	trivial (LCEBR)

Continued on next page

Table B.1 – continued from previous page

U-ID	Chem. F.	ICSD-ID	SG	MP-ID	Conf. MP-ID	Type_TMD
3675	Zn S	107178	156	mp-555311	mp-555311	trivial (LCEBR)
3675	Zn S	107138	156	NaN	NaN	trivial (LCEBR)
3675	Zn S	107183	156	mp-554961	mp-554961	trivial (LCEBR)
3675	Zn S	42993	156	mp-554889	mp-554889	trivial (LCEBR)
3675	Zn S	42994	156	mp-555628	mp-555628	trivial (LCEBR)
3675	Zn S	107179	156	mp-555214	mp-555214	trivial (LCEBR)
3675	Zn S	43068	156	mp-555779	mp-555779	trivial (LCEBR)
3675	Zn S	43069	156	mp-556950	mp-556950	trivial (LCEBR)
3675	Zn S	43070	156	mp-556775	mp-556775	trivial (LCEBR)
3675	Zn S	42992	156	mp-555381	mp-555381	trivial (LCEBR)
14667	Ba Ga Sn D	173565	156	mp-1018094	mp-1018094	trivial (LCEBR)
15627	La Ga	634476	63	mp-1002133	mp-1002133	TI (SEBR) TCI
18478	Ga La	423623	63	mp-1002133	mp-1002133	TI (NLC) TI
35261	Ba2 Cr Mo O6	191680	225	mp-1079901	mp-1079901	SM (ES)
37226	Ba2 Cr Mo O6	184907	225	mp-1079901	mp-1079901	trivial (LCEBR)
37495	Ba Ga Sn H	173574	156	mp-1018094	mp-1018094	SM (ES)

APPENDIX C

This appendix contains a complete list of the predictions made by the model trained on Dataset $M \cup T$. (Details are given in sect. 4.3)

Table C.1: Predictions of materials in set A. Materials are list in order of MP-IDs, related to one same MP-ID, there are more than one ICSD-ID entries labeled as different type on TMD. The chemical formula (Chem. F.) and space group number (SG) are provided on TMD (related to the ICSD-ID).

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
423623	Ga La	63	TI	TCI	mp-1002133
634476	La Ga	63	TCI	TCI	mp-1002133
44004	Mn Ru As	189	HSLSM	HSLSM	mp-10049
610918	Mn Ru As	189	TCI	HSLSM	mp-10049
190430	In Sb	186	TrI	HSLSM	mp-1007661
640432	In Sb	186	HSLSM	HSLSM	mp-1007661
44661	Ir P2	14	TrI	TrI	mp-10155
174231	Ir P2	14	TI	TrI	mp-10155
186436	La N	129	TrI	TrI	mp-1018049
423892	La N	129	TI	TrI	mp-1018049
173565	Ba Ga Sn D	156	TrI	TrI	mp-1018094
173574	Ba Ga Sn H	156	HSLSM	HSLSM	mp-1018094
646388	Ni Ti2 S4	12	TCI	TI	mp-1025263
646394	Ni Ti2 S4	12	TI	TI	mp-1025263
621131	Ce Ga2 Ni	65	TCI	TrI	mp-1025446
188329	Ce Ni Ga2	65	TI	TI	mp-1025446
638987	Hf V2	74	TI	TI	mp-1043
104283	Hf V2	227	HSPSM	HSPSM	mp-1043
157607	Ni Ti	11	TCI	TI	mp-1048
105414	Ni Ti	11	TI	TI	mp-1048
105416	Ni Ti	11	TrI	TI	mp-1048
181724	Ni Ti	11	TI	TI	mp-1048
56503	Sr	74	TrI	HSLSM	mp-10617
109024	Sr	141	HSLSM	HSLSM	mp-10617
99452	Mo N	186	HSPSM	HSLSM	mp-1065394
185560	Mo N	194	HSPSM	HSPSM	mp-1065394

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
151774	Mo N	186	TrI	HSLSM	mp-1065394
187469	Ti Ni	11	TI	TI	mp-1067248
252716	Ca C2	12	TI	TI	mp-1071152
237874	Ni2 In Sb O6	146	HSLSM	TrI	mp-1078367
237876	Ni2 In Sb O6	146	TrI	TrI	mp-1078367
635134	Pt3 Ga	127	TI	TI	mp-1078666
187685	Pb (Ni O3)	161	HSPSM	TrI	mp-1078668
187684	Pb (Ni O3)	161	TrI	TrI	mp-1078668
649557	Pt3 Sb	139	HSLSM	HSLSM	mp-1078755
649555	Pt3 Sb	139	TI	HSLSM	mp-1078755
237385	Cr La As O	129	TrI	TrI	mp-1079055
237386	Cr La As O	129	TI	TrI	mp-1079055
191680	Ba2 Cr Mo O6	225	HSLSM	HSPSM	mp-1079901
184907	Ba2 Cr Mo O6	225	TrI	HSPSM	mp-1079901
187020	Cu2 Hg Ge S4	121	TI	TrI	mp-10952
94988	Cu2 Hg Ge S4	121	TrI	TrI	mp-10952
617977	Ti3 Ge C2	194	HSLSM	TI	mp-1095505
180424	Ti3 Ge C2	194	TI	TI	mp-1095505
619750	Cu2 Cd (Ge Se4)	121	TrI	TrI	mp-10967
181295	Cu2 Cd (Ge Se4)	121	TI	TrI	mp-10967
246809	La Si5	12	TCI	TrI	mp-1101787
246807	La Si5	12	TI	TrI	mp-1101787
636761	Ir Y Ge	62	TI	TI	mp-1102373
427251	Y Ir Ge	62	TrI	TI	mp-1102373
261456	Bi2 Pd3 Se2	12	TI	TI	mp-1103697
261458	Bi2 Pd3 Se2	12	TI	TI	mp-1103697
261457	Bi2 Pd3 Se2	12	TCI	TrI	mp-1103697
238148	Ca2 Fe Os O6	14	TI	TrI	mp-1105304
238149	Ca2 Fe Os O6	14	TrI	TrI	mp-1105304
194435	Nb Ru B	51	TCI	TI	mp-1105360
426468	Nb Ru B	51	TI	TI	mp-1105360
150736	Mg17 Sr2	194	HSLSM	HSLSM	mp-1116
104875	Mg17 Sr2	194	TI	HSLSM	mp-1116
58643	Ba Cd2	74	TI	TI	mp-11266
260668	Ba Cd2	74	TCI	TI	mp-11266
75194	Ti3 O5	12	TI	TI	mp-1147

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
26492	Ti3 O5	12	TrI	TI	mp-1147
44328	Ga Sb	216	HSPSM	HSPSM	mp-1156
41675	Ga Sb	216	TrI	HSPSM	mp-1156
634145	Fe Zr2	140	HSLSM	HSLSM	mp-1159
103712	Fe Zr2	140	TI	HSLSM	mp-1159
618149	La2 C3	220	TrI	TI	mp-1184
153809	La2 C3	220	TI	TI	mp-1184
636396	Gd2 Se3	62	TI	TrI	mp-1188175
99997	Gd2 Se3	62	TrI	TrI	mp-1188175
625574	Y2 Co7	166	HSLSM	HSLSM	mp-1188333
625563	Y2 Co7	166	TI	HSLSM	mp-1188333
427410	Pb7 Bi4 Se13	12	TI	TrI	mp-1190436
427411	Pb7 Bi4 Se13	12	TrI	TrI	mp-1190436
247112	Ba Re H9	11	TI	TrI	mp-1190554
247111	Ba Re H9	11	TrI	TrI	mp-1190554
165457	Co3 W3 C	227	HSLSM	HSPSM	mp-1192636
166747	Co3 W3 C	227	TI	HSLSM	mp-1192636
168948	Ti11 Ni9 Pt4	15	TCI	TI	mp-1196249
168947	Ti11 Ni9 Pt4	15	TI	TI	mp-1196249
427249	Y2 Ru B6	55	TrI	TI	mp-1196778
603493	Y2 Ru B6	55	TI	TI	mp-1196778
185952	Sb2 Te3	166	TI	TI	mp-1201
187495	Sb2 Te3	166	TCI	TI	mp-1201
413690	Mo Br3	59	TI	TrI	mp-1205355
9422	Mo Br3	59	TrI	TrI	mp-1205355
44717	Ir Sb3	204	HSPSM	HSPSM	mp-1239
15458	Ir Sb3	204	TrI	HSPSM	mp-1239
42522	Zr2 Si	140	TI	HSLSM	mp-1278
652615	Zr2 Si	140	HSLSM	HSLSM	mp-1278
248490	Pt2 Si	139	TI	HSLSM	mp-1299
77973	Pt2 Si	139	TCI	HSLSM	mp-1299
53301	P	166	TCI	TCI	mp-130
600019	P	166	TI	TI	mp-130
160935	Al Ni Y	189	TCI	TCI	mp-13095
608947	Al Ni Y	189	HSLSM	HSLSM	mp-13095
16719	Cr5 S6	163	TCI	TCI	mp-1311

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
43045	Cr5 S6	163	HSLSM	TCI	mp-1311
44715	Co Sb3	204	HSPSM	HSPSM	mp-1317
34048	Co Sb3	204	TrI	TrI	mp-1317
43724	Rh P3	204	TrI	HSPSM	mp-1357
23712	Rh P3	204	HSPSM	HSPSM	mp-1357
23073	Se	152	HSLSM	HSLSM	mp-14
23072	Se	152	HSPSM	HSLSM	mp-14
22251	Se	152	TrI	TrI	mp-14
420576	Ca Cd2	74	TCI	TCI	mp-1444
58875	Ca Cd2	74	TI	TCI	mp-1444
40799	Ce Ru2 B2	70	TI	TI	mp-15588
40800	Ce Ru2 B2	22	TrI	TrI	mp-15588
27847	P	64	TrI	TI	mp-157
187464	P	64	TI	TI	mp-157
70100	As	64	TI	TI	mp-158
609828	As	64	TrI	TI	mp-158
62486	Mo2 S3	11	TI	TI	mp-1627
602115	Mo2 S3	11	TCI	TI	mp-1627
76371	Mo2 S3	11	TI	TI	mp-1627
95119	Cu2 Hg Sn Se4	121	TI	TI	mp-16566
262976	Cu2 Hg Sn Se4	121	TrI	TI	mp-16566
26813	Tl2 O3	206	HSPSM	TrI	mp-1658
74090	Tl2 O3	206	TrI	TrI	mp-1658
418614	Ca Zn5	191	TI	TI	mp-1734
58947	Ca Zn5	191	HSLSM	HSLSM	mp-1734
636541	Hf5 Ge3	193	HSLSM	TI	mp-1739
44361	Hf5 Ge3	193	TI	TI	mp-1739
69881	Tl5 Te3	140	HSPSM	HSPSM	mp-174
69028	Tl5 Te3	87	TrI	HSPSM	mp-174
57534	Al8 Ca Co2	55	TI	TCI	mp-17871
57533	Al8 Ca Co2	55	TCI	TrI	mp-17871
105420	Ni Ti2	227	HSPSM	HSLSM	mp-1808
15808	Ni Ti2	227	TI	TI	mp-1808
31842	Hg Te	225	HSPSM	HSPSM	mp-1811
60204	Hg Te	225	HSLSM	HSLSM	mp-1811
58660	Ba2 Mg17	166	TI	HSLSM	mp-1813

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
150735	Ba2 Mg17	166	HSLSM	HSLSM	mp-1813
42139	Ca5 Ge3	140	HSLSM	HSLSM	mp-1884
181076	Ca5 Ge3	108	HSLSM	HSLSM	mp-1884
181075	Ca5 Ge3	130	TI	HSLSM	mp-1884
93749	Ca (Mn2 O4)	57	TI	TrI	mp-18844
93750	Ca (Mn2 O4)	57	TrI	TrI	mp-18844
1473	V2 O3	167	TCI	HSLSM	mp-18937
201109	V2 O3	167	TI	HSLSM	mp-18937
157024	Sr2 (Ni Mo O6)	87	TCI	TrI	mp-18940
155274	Sr2 Ni (Mo O6)	87	TrI	TrI	mp-18940
922	Ba Co O3	194	HSPSM	TrI	mp-18965
88670	Ba (Co O3)	187	HSLSM	TrI	mp-18965
426980	Te	152	TrI	HSPSM	mp-19
23058	Te	152	HSPSM	HSPSM	mp-19
23067	Te	152	HSLSM	TrI	mp-19
202244	Ca Ni Si2 O6	15	TI	TrI	mp-19220
159545	Ni Ca (Si2 O6)	15	TrI	TrI	mp-19220
24496	Zn Fe2 O4	227	HSLSM	TrI	mp-19313
91943	Zn Fe2 O4	227	TI	TrI	mp-19313
166205	Zn (Fe2 O4)	227	HSPSM	TrI	mp-19313
201105	Cr2 O3	167	TrI	HSLSM	mp-19399
173470	Cr2 O3	167	TI	HSLSM	mp-19399
290245	Cr2 O3	167	HSLSM	HSLSM	mp-19399
171889	Zn Cr2 O4	227	HSLSM	HSPSM	mp-19410
290592	Zn (Cr2 O4)	141	TI	TCI	mp-19410
290590	Zn (Cr2 O4)	141	HSLSM	TCI	mp-19410
50047	Zn Cr2 O4	227	TI	TCI	mp-19410
290591	Zn (Cr2 O4)	70	TI	TCI	mp-19410
648615	Pb Te	225	TCI	TCI	mp-19717
38295	Pb Te	225	TrI	TCI	mp-19717
32562	Y Rh Si	62	TrI	TrI	mp-19728
650347	Rh Y Si	62	TI	HSLSM	mp-19728
96072	Fe2 O3	167	HSLSM	TI	mp-19770
15840	Fe2 O3	167	TI	TI	mp-19770
96076	Fe2 O3	167	TrI	TrI	mp-19770
76799	Cr7 C3	62	TrI	TrI	mp-19855

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
181713	Cr7 C3	62	TCI	TrI	mp-19855
41728	Ni Sb2	58	TI	TrI	mp-19895
646409	Ni Sb2	34	TrI	TrI	mp-19895
107372	In Ni2 Zr2	127	HSPSM	HSPSM	mp-19906
405563	Zr2 Ni2 In	136	HSLSM	HSLSM	mp-19906
24519	In Sb	216	HSPSM	HSPSM	mp-20012
190418	In Sb	216	TrI	HSPSM	mp-20012
42327	Fe Li P	129	HSLSM	HSLSM	mp-20049
166456	Li Fe P	129	TCI	TrI	mp-20049
55537	U Fe Si	62	TI	TI	mp-20121
95009	U Fe Si	62	TCI	TI	mp-20121
156996	In Sb	63	TI	TCI	mp-20253
157947	In Sb	63	TCI	TCI	mp-20253
24518	In As	216	HSPSM	HSPSM	mp-20305
190415	In As	216	TrI	TrI	mp-20305
165957	Fe Se	67	TI	TI	mp-20311
26889	Fe Se	129	HSLSM	TI	mp-20311
169283	Fe Se	67	TrI	TrI	mp-20311
163555	Fe Se	67	TI	TI	mp-20311
169267	Fe Se	129	TCI	HSPSM	mp-20311
166441	Fe Se	67	TCI	HSPSM	mp-20311
2575	W (V2 O6)	136	HSLSM	TrI	mp-20312
28489	W V2 O6	136	TI	TrI	mp-20312
612865	Co3 B	62	TI	HSLSM	mp-20373
603543	Co3 B	62	TCI	TCI	mp-20373
640190	In P	225	HSLSM	HSPSM	mp-20457
53104	In P	225	TrI	HSPSM	mp-20457
185172	In Se	194	TrI	TrI	mp-20485
30377	In Se	194	TI	TI	mp-20485
605453	Ag In Se2	122	TrI	TrI	mp-20554
28751	Ag In Se2	122	TI	TrI	mp-20554
76720	Ni2 U Si2	139	HSLSM	HSLSM	mp-20596
25684	U Ni2 Si2	139	TI	HSLSM	mp-20596
102803	Pd Pb2	140	HSLSM	TI	mp-20599
105590	Pb2 Pd	140	TI	TI	mp-20599
9982	V2 C	60	TrI	TrI	mp-20648

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
9965	V2 C	60	TI	TrI	mp-20648
103987	Ga2 Th	141	TI	HSPSM	mp-2067
150533	Th Ga2	141	HSLSM	HSLSM	mp-2067
66602	Ba2 Y Cu3 O7	47	TI	TI	mp-20674
41646	Y Ba2 Cu3 O7	47	TrI	TI	mp-20674
65224	Ba2 Y Cu3 O7	47	TI	TI	mp-20674
63335	Ba2 Y Cu3 O7	47	TCI	TI	mp-20674
66601	Ba2 Y Cu3 O7	47	TCI	TI	mp-20674
66623	Y Ba2 Cu3 O7	47	TCI	TI	mp-20674
77737	Ba2 Y Cu3 O7	47	TI	TI	mp-20674
66616	Y Ba2 Cu3 O7	47	TI	TI	mp-20674
5258	Fe Si2	123	TI	TCI	mp-20738
24360	Fe Si2	123	TCI	TCI	mp-20738
106471	Co Ge La	129	TI	TI	mp-20761
85860	La Co Ge	129	TrI	TI	mp-20761
640195	In (P S4)	82	TrI	TrI	mp-20790
23612	In (P S4)	82	TI	TrI	mp-20790
51004	Zr Te3	11	TrI	TrI	mp-2089
653232	Zr Te3	11	TI	TI	mp-2089
645088	Nb Ni P	62	TCI	TCI	mp-20932
49728	Nb Ni P	62	TI	TCI	mp-20932
106847	Ce Co2 Ge2	139	HSLSM	HSLSM	mp-20933
107209	Ce Co2 Ge2	139	TrI	HSLSM	mp-20933
181712	Cr3 C2	62	TrI	TCI	mp-20937
15086	Cr3 C2	62	TCI	TCI	mp-20937
77882	Pb Te	221	TrI	TrI	mp-20943
290708	Pb Te	221	HSLSM	HSLSM	mp-20943
35008	Fe S	62	TI	HSLSM	mp-2099
633265	Fe S	194	HSLSM	HSLSM	mp-2099
68847	Fe S	186	HSLSM	HSLSM	mp-2099
76024	Ta O2	136	HSPSM	HSPSM	mp-20994
60628	Ta O2	88	TrI	TrI	mp-20994
182427	Ti3 Sn	194	TI	HSLSM	mp-21030
150657	Ti3 Sn	194	HSLSM	HSLSM	mp-21030
77865	Pb S	221	TI	TI	mp-21039
183242	Pb S	221	TrI	TrI	mp-21039

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
652420	Ti5 Si3	193	HSLSM	HSLSM	mp-2108
62591	Ti5 Si3	193	TI	HSLSM	mp-2108
638947	Hf5 Sn3	193	HSLSM	HSPSM	mp-21104
638939	Hf5 Sn3	193	TI	TI	mp-21104
162108	Rh N3	204	HSPSM	HSPSM	mp-2118
162107	Rh N3	204	TrI	TrI	mp-2118
44753	Fe Te	129	HSLSM	HSPSM	mp-21273
180602	Fe Te	129	TCI	HSPSM	mp-21273
62190	Pb S	225	TCI	TrI	mp-21276
38293	Pb S	225	TrI	TrI	mp-21276
42894	Mo Co B	62	TI	TI	mp-21347
20461	Mo Co B	62	TCI	TI	mp-21347
621436	Ce Ir2 Sn2	129	TCI	HSLSM	mp-21358
106407	Ce Ir2 Sn2	129	TI	TI	mp-21358
165247	V Co Si	62	TCI	TCI	mp-21371
625120	V Co Si	62	TrI	HSLSM	mp-21371
24189	Os Te2	205	TI	TrI	mp-2142
647831	Os Te2	205	TrI	TrI	mp-2142
25580	Mg6 Si7 Cu16	225	HSLSM	HSLSM	mp-21612
16624	Mg6 Si7 Cu16	225	HSPSM	HSPSM	mp-21612
412752	Hf Co Ga2	107	TrI	TrI	mp-21660
623079	Co Ga2 Hf	107	HSLSM	TrI	mp-21660
107268	Cu5 In Th	62	TrI	TrI	mp-21670
152504	Th Cu5 In	62	TI	TI	mp-21670
31018	Fe7 C3	62	TrI	TrI	mp-21717
167346	Fe7 C3	62	TI	TrI	mp-21717
174591	Pb2 Mn O4	114	TI	TrI	mp-21823
33990	Pb2 (Mn O4)	114	TrI	TrI	mp-21823
159208	Pb2 (Mn O4)	114	HSLSM	TrI	mp-21823
43237	Mo4 P3	62	TCI	TI	mp-21833
65652	Mo4 P3	62	TI	TI	mp-21833
238502	Pb4 Se4	225	TCI	TCI	mp-2201
183008	Pb Se	225	TrI	TrI	mp-2201
38294	Pb Se	225	TCI	TCI	mp-2201
280597	Pb Cu6 O8	225	TCI	HSLSM	mp-22037
27585	Cu6 Pb O8	225	HSLSM	TrI	mp-22037

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
16305	Nb Se2	194	HSLSM	HSPSM	mp-2207
16304	Nb Se2	194	HSPSM	HSPSM	mp-2207
42572	Rh As	62	TCI	TCI	mp-22079
657339	Rh As	62	TI	TCI	mp-22079
161515	Ca2 (Fe2 O5)	62	TrI	TrI	mp-22113
26474	(Fe2 O3) (Ca O)2	62	TCI	TrI	mp-22113
14296	Ca2 Fe2 O5	62	TI	TrI	mp-22113
25677	In N	186	HSLSM	TrI	mp-22205
190428	In N	186	TrI	TrI	mp-22205
74703	In Ta S2	187	TCI	TCI	mp-22332
640378	In Ta S2	187	HSLSM	TCI	mp-22332
51281	Ba Cu2 (Si2 O7)	62	TrI	TrI	mp-22407
289997	Ba Cu2 Si2 O7	62	TCI	TrI	mp-22407
626685	Cr Sb2	58	HSLSM	TrI	mp-22498
42601	Cr Sb2	58	TCI	TrI	mp-22498
84529	Ce Ni Sb2	129	TCI	HSPSM	mp-22539
658103	Ce Ni Sb2	129	TI	HSPSM	mp-22539
42988	Zr2 Se	58	TI	TI	mp-22642
250208	Zr2 Se	58	TrI	TI	mp-22642
42127	Cu In S2	122	TI	TI	mp-22736
628051	Cu In S2	122	TrI	TI	mp-22736
89097	U Ir2 Si2	129	HSLSM	HSLSM	mp-22758
62716	U Ir2 Si2	129	TI	HSLSM	mp-22758
162105	Co N3	204	TCI	HSPSM	mp-22762
162106	Co N3	204	HSPSM	HSPSM	mp-22762
84558	Cu (W O4)	2	TI	TrI	mp-22773
4189	Cu W O4	2	TCI	TrI	mp-22773
182750	Cu (W O4)	2	TI	TrI	mp-22773
68334	Cu In Se2	122	TrI	TI	mp-22811
183469	Cu In Se2	122	TI	TI	mp-22811
250247	Cu2 In2 Se4	1	TrI	TI	mp-22811
107367	Ni2 Sn U2	127	TI	TI	mp-22813
602807	Ni2 Sn U2	127	HSLSM	HSLSM	mp-22813
40328	Ni S2	205	HSPSM	TI	mp-2282
169569	Ni S2	205	TI	TI	mp-2282
37367	Bi2 O3	224	HSPSM	TrI	mp-22891

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
27150	Bi2 O3	224	HSLSM	HSLSM	mp-22891
67073	Ba Bi O3	12	TrI	TrI	mp-22942
61499	Ba Bi O3	12	TCI	TrI	mp-22942
1430	Ba2 Bi2 O6	12	TrI	TrI	mp-22942
181781	Cs Pb Cl3	221	TI	TrI	mp-23037
29072	Cs Pb Cl3	221	TrI	TrI	mp-23037
166533	Ca2 P I	166	TI	TrI	mp-23040
65217	Ca2 P I	166	TrI	TrI	mp-23040
240479	Cs2 (Pd Br4) I2	12	TI	TI	mp-23099
240483	Cs2 (Pd Br4) I2	12	TrI	TrI	mp-23099
422870	Ti I3	59	TrI	TrI	mp-23223
23172	Ti I3	193	HSPSM	HSPSM	mp-23223
422868	Ti I3	59	TrI	TrI	mp-23264
38236	Ti Cl3	148	TrI	TrI	mp-23275
39426	Ti Cl3	148	TI	TI	mp-23275
74648	Zr I3	59	TrI	TrI	mp-23282
35317	Zr I3	193	HSPSM	HSPSM	mp-23282
58849	Bi Rb3	225	HSPSM	HSPSM	mp-23304
616996	Bi Rb3	225	TrI	HSPSM	mp-23304
80252	Cu I	166	TI	TrI	mp-23306
78267	Cu I	166	HSLSM	TrI	mp-23306
191908	Mo2 B4	166	TCI	TI	mp-2331
167732	Mo B2	166	TI	TI	mp-2331
39554	Mo2 B4	166	TI	TI	mp-2331
40907	Mo B2	166	TI	TI	mp-2331
24593	Hg5 O4 Cl2	72	TI	TI	mp-23358
75935	Hg5 Cl2 O4	72	TrI	TrI	mp-23358
174405	Bi (Cr O3)	15	TI	TrI	mp-23456
160455	Bi (Cr O3)	15	TCI	TrI	mp-23456
417371	Cs2 Au Au Cl6	139	TrI	TrI	mp-23484
37453	Cs2 Au2 Cl6	139	TCI	TrI	mp-23484
151991	In Pb2 I5	140	TrI	TrI	mp-23520
152028	In Pb2 I5	140	TI	TrI	mp-23520
151984	K Sn2 I5	140	TrI	TrI	mp-23534
152012	K Sn2 I5	140	HSLSM	TrI	mp-23534
38605	(Hg Br2) (Hg O)4	72	TrI	TrI	mp-23543

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
28232	Hg (Hg O)4 Br2	72	TI	TI	mp-23543
88819	C	206	TrI	TrI	mp-24
5390	C	206	HSPSM	TrI	mp-24
74655	C8	206	HSPSM	TrI	mp-24
615493	Sr B6	221	TI	TrI	mp-242
659503	Sr B6	221	TrI	TrI	mp-242
189201	Zr Ni5	216	HSLSM	HSPSM	mp-2439
601018	Zr Ni5	216	TI	TI	mp-2439
181134	Cd3 P2	137	HSLSM	TrI	mp-2441
158864	Cd3 P2	137	TrI	TrI	mp-2441
55210	Ca Sn	63	TCI	TI	mp-2450
55214	Ca Sn	63	TI	TI	mp-2450
25787	Re3 As7	229	TI	HSPSM	mp-2473
611260	Re3 As7	229	HSPSM	HSPSM	mp-2473
611257	Re3 As7	229	HSLSM	HSPSM	mp-2473
186435	La N	225	HSLSM	TrI	mp-256
169212	La N	225	TrI	TrI	mp-256
616162	Ba Te	221	TrI	TrI	mp-2600
108097	Ba Te	221	TCI	TCI	mp-2600
78997	Sr Si	63	TCI	TCI	mp-2661
25534	Sr Si	63	TI	TCI	mp-2661
156443	Ni Ti	147	HSLSM	HSLSM	mp-2716
150943	Ti Ni	147	TI	HSLSM	mp-2716
166370	Ti Ni	147	TI	HSLSM	mp-2716
290600	Mg Mn2 O4	141	TrI	TrI	mp-27510
16858	Mg (Mn2 O4)	141	TCI	TrI	mp-27510
652385	Si Te2	164	HSLSM	TrI	mp-2755
80205	Si Te2	164	TI	TrI	mp-2755
24353	Cr B4	71	TI	TI	mp-27710
239596	Cr B4	71	TCI	TI	mp-27710
150527	Zr Al2	194	TCI	TI	mp-2772
55594	Zr Al2	194	TI	TI	mp-2772
35007	Fe S	190	HSLSM	HSLSM	mp-2779
51003	Fe S	190	TCI	HSLSM	mp-2779
15043	Fe3 Se4	12	TI	TCI	mp-2780
150568	Fe3 Se4	12	TCI	TCI	mp-2780

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
24560	Fe3 Se4	12	TCI	TI	mp-2780
26976	Cr3 Se4	12	TCI	TI	mp-27840
601348	Cr3 Se4	12	TI	TI	mp-27840
626716	Cr3 Se4	12	TCI	TI	mp-27840
27394	Ti N I	59	TI	TI	mp-27848
290037	Ti N I	59	TCI	TI	mp-27848
614639	La2 Re3 B7	54	TCI	TI	mp-27894
20878	La2 Re3 B7	54	TI	TI	mp-27894
22264	Zr C	225	TI	TI	mp-2795
181094	Zr C	225	HSLSM	HSLSM	mp-2795
166534	Ca2 As I	166	HSLSM	TrI	mp-28554
65218	Ca2 As I	166	TrI	TrI	mp-28554
39228	Th11 Ru12 C18	217	TrI	HSPSM	mp-2868
79240	Th11 Ru12 C18	217	HSPSM	HSPSM	mp-2868
422342	Al2 Fe3 Si3	2	TI	TrI	mp-29110
83664	Fe3 Al2 Si3	2	TrI	TrI	mp-29110
617113	Tl Bi Se2	166	TI	TrI	mp-29662
43314	Bi Tl Se2	166	TrI	TrI	mp-29662
68415	Th B2 C	166	HSLSM	HSLSM	mp-2997
68414	Th B2 C	166	TrI	HSLSM	mp-2997
604241	Tl4 Sn Te3	140	TCI	TrI	mp-3019
73087	Sn Tl4 Te3	140	TrI	TrI	mp-3019
184361	Ca Zn	63	TI	TCI	mp-30483
58944	Ca Zn	63	TCI	TCI	mp-30483
58948	Ca3 Zn	63	TI	TI	mp-30485
184359	Ca3 Zn	63	TCI	TI	mp-30485
633886	Fe3 Th	166	TCI	TI	mp-30638
103654	Fe3 Th	166	TI	TI	mp-30638
99175	Zr Rh Ge2	55	TCI	TCI	mp-31084
637736	Zr Rh Ge2	55	TI	TCI	mp-31084
202550	Mn4 (Nb2 O9)	165	TrI	TrI	mp-31901
29216	Mn4 (Nb2 O9)	165	HSLSM	TrI	mp-31901
43422	Ge	227	HSPSM	HSPSM	mp-32
54237	Ge	227	TrI	HSPSM	mp-32
15575	Zr Ge Te	129	HSLSM	TCI	mp-3208
25740	Zr Ge Te	129	TCI	TCI	mp-3208

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
109221	Ti6 Ni16 Si7	225	HSLSM	HSLSM	mp-3337
159263	Ti6 Ni16 Si7	225	HSPSM	HSPSM	mp-3337
55795	Ce Co2 Si2	139	HSLSM	HSLSM	mp-3437
106838	Ce Co2 Si2	139	TCI	HSLSM	mp-3437
154184	Ba (V S3)	8	TrI	TrI	mp-3451
52692	Ba V S3	36	HSPSM	TrI	mp-3451
63501	La2 Ni5 B4	12	TI	TI	mp-3471
170618	La2 Ni5 B4	12	TCI	TI	mp-3471
109220	Nb6 Ni16 Si7	225	HSLSM	HSLSM	mp-3640
159264	Nb6 Ni16 Si7	225	HSPSM	HSLSM	mp-3640
150471	Th O Te	129	TrI	TrI	mp-3718
33587	Th O Te	129	HSLSM	TrI	mp-3718
52892	Ce Pd2 Si2	139	TCI	TCI	mp-3826
40913	Ce Pd2 Si2	139	HSLSM	TCI	mp-3826
659748	Fe Tl S2	12	TCI	TrI	mp-3849
30011	Tl Fe S2	12	TrI	TrI	mp-3849
106259	Al3 Zr	139	TI	TI	mp-395
107130	Al3 Zr	139	HSLSM	TI	mp-395
188172	Ca Ni2 P2	139	TCI	TI	mp-4064
10461	Ca Ni2 P2	139	TI	TCI	mp-4064
52896	Ce Rh2 Si2	139	HSLSM	HSLSM	mp-4090
621948	Ce Rh2 Si2	139	TCI	HSLSM	mp-4090
165119	Ba Ru2 As2	139	TI	TI	mp-4109
602110	Ba Ru2 As2	139	HSLSM	HSLSM	mp-4109
96594	Ta	113	HSPSM	HSPSM	mp-42
96595	Ta	3	TrI	HSPSM	mp-42
54208	Ta30	113	HSPSM	HSPSM	mp-42
151208	Cu2 Ni Sn	225	HSPSM	HSLSM	mp-4367
103069	Cu2 Ni Sn	225	HSLSM	HSLSM	mp-4367
163869	Sr Fe2 As2	139	TI	TI	mp-4488
163209	Sr Fe2 As2	139	HSLSM	TI	mp-4488
38229	Co As3	204	HSPSM	HSPSM	mp-452
610045	Co As3	204	TI	TrI	mp-452
73105	Ba4 Nb14 O23	65	TI	TCI	mp-4564
73444	Ba4 Nb14 O23	65	TCI	TCI	mp-4564
183409	Ti	194	TrI	HSPSM	mp-46

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
43416	Ti	194	HSLSM	HSLSM	mp-46
57055	Zr2 Sn C	194	TI	TI	mp-4613
161068	Zr2 Sn C	194	HSLSM	HSLSM	mp-4613
188174	Ca Ni2 As2	139	TCI	TCI	mp-4744
23004	Ca Ni2 As2	139	HSLSM	HSLSM	mp-4744
188173	Ca Ni2 As2	139	TCI	HSLSM	mp-4744
648186	Ta3 P	86	HSLSM	HSLSM	mp-504502
10125	Ta3 P	86	TI	HSLSM	mp-504502
20067	Bi2 Fe4 O9	55	TCI	TrI	mp-504615
186442	Bi2 Fe4 O9	55	TI	TrI	mp-504615
202193	Mn Fe F5 (H2 O)2	44	TrI	TrI	mp-504945
62362	Mn Fe F5 (H2 O)2	74	TI	TrI	mp-504945
173073	Rb Os2 O6	227	HSPSM	HSPSM	mp-5050
161127	Rb (Os2 O6)	216	TI	HSPSM	mp-5050
81059	Ni (Mo O4)	12	TI	TrI	mp-505273
174488	Ni (Mo O4)	12	TrI	TrI	mp-505273
168417	Si4 Ti5	92	TrI	HSLSM	mp-505527
43080	Ti5 Si4	92	HSLSM	HSLSM	mp-505527
162272	Fe3 Mo3 N	227	HSPSM	HSPSM	mp-510619
88266	Fe3 Mo3 N	227	HSLSM	HSLSM	mp-510619
36372	Ca Pd2 As2	139	TCI	TCI	mp-5166
188178	Ca Pd2 As2	139	HSLSM	HSLSM	mp-5166
18203	Zn Sn As2	122	TI	TI	mp-5190
250389	Zn Sn As2	122	TrI	TI	mp-5190
49666	Cr S	194	TCI	TCI	mp-523
626593	Cr S	194	HSLSM	HSLSM	mp-523
94501	La Al Si2	164	TCI	TI	mp-5330
94502	La Al Si2	164	TI	TI	mp-5330
74980	Ba3 Nb16 O23	65	TCI	TCI	mp-5360
74979	Ba3 Nb16 O23	65	TI	TCI	mp-5360
631764	Fe3 Ga4	12	TI	TrI	mp-540538
1823	Ga4 Fe3	12	TrI	TrI	mp-540538
76969	Ag3 (As O4)	218	TrI	TrI	mp-5408
28836	Ag3 (As O4)	218	TI	TrI	mp-5408
196175	V O2	12	TrI	TrI	mp-541404
73855	V O2	12	TI	TrI	mp-541404

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
658197	Mo3 U2 Ge4	14	TCI	TI	mp-541491
80002	U2 Mo3 Ge4	14	TI	TI	mp-541491
629343	Cu2 Ti Te3	12	TI	TrI	mp-541754
402631	Cu2 Ti Te3	12	TrI	TrI	mp-541754
42909	Ir4 Ge5	116	TI	TrI	mp-541844
43053	Ir4 Ge5	116	TrI	TrI	mp-541844
42667	In2 Zn S4	160	HSLSM	HSLSM	mp-541937
640407	Zn In2 S4	160	TrI	TrI	mp-541937
102718	Co3 V	194	HSLSM	HSLSM	mp-542614
625533	V Co3	194	TCI	HSLSM	mp-542614
85953	V8 C7	212	TrI	HSPSM	mp-542730
22177	V8 C7	212	HSLSM	HSLSM	mp-542730
164189	Cu2 (V2 O7)	1	TrI	TrI	mp-542860
23479	Cu2 (V2 O7)	15	TCI	TrI	mp-542860
161961	Ba (Ir O3)	12	TrI	TrI	mp-542897
173848	Ba Ir O3	12	TCI	TrI	mp-542897
620930	Ce Cu2 Si2	139	TCI	HSLSM	mp-5452
620944	Ce Cu2 Si2	139	HSLSM	HSLSM	mp-5452
75660	Si O2	1	TrI	TrI	mp-556319
75667	Si O2	1	TrI	TrI	mp-556588
35122	Ti7 O13	2	TI	TI	mp-556724
9040	Ti7 O13	2	TrI	TrI	mp-556724
75663	Si O2	1	TrI	TrI	mp-556880
75656	Si O2	1	TrI	TrI	mp-556985
75661	Si O2	1	TrI	TrI	mp-556994
637429	Ge2 Ni2 U	139	TCI	TI	mp-5591
150835	U Ni2 Ge2	139	TI	TI	mp-5591
280596	Pb Cu6 O8	225	TCI	TCI	mp-559633
615773	Ba	194	TrI	TrI	mp-56
52680	Ba	194	TI	HSLSM	mp-56
35634	Fe V2 S4	12	TrI	TI	mp-561419
633361	Fe V2 S4	12	TI	TI	mp-561419
190562	Ti3 Si C2	194	HSLSM	TI	mp-5659
25762	Ti3 Si C2	194	TI	TI	mp-5659
99178	Ce P5	11	HSPSM	TrI	mp-566
621759	Ce P5	11	TrI	TrI	mp-566

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
52322	Ti Sb2	140	TCI	HSPSM	mp-568
409797	Ti Sb2	140	TI	TI	mp-568
106089	Sn5 Ti6	71	TrI	TCI	mp-568232
169010	Sn5 Ti6	71	TCI	TCI	mp-568232
643685	Mn2 U Si2	139	TCI	HSLSM	mp-5683
55933	U Mn2 Si2	139	TI	HSLSM	mp-5683
609749	Hf Al3	139	TI	TI	mp-568718
608102	Hf Al3	139	TCI	TI	mp-568718
182275	Ba Fe2 As2	139	TI	TI	mp-568961
166019	Ba Fe2 As2	139	HSLSM	TI	mp-568961
166018	Ba Fe2 As2	139	TI	TI	mp-568961
182272	Ba Fe2 As2	139	TI	TI	mp-568961
165120	Sr Rh2 As2	139	TI	TI	mp-569006
417002	Sr Rh2 As2	139	TCI	TI	mp-569006
616539	Bi Ca Li	62	TI	TI	mp-569501
58762	Bi Ca Li	62	TrI	TI	mp-569501
54205	Ta30	81	TrI	HSPSM	mp-569794
652898	Ta	136	HSPSM	HSPSM	mp-569794
64721	Cu3 (Sb S4)	121	TrI	TrI	mp-5702
42673	Cu3 Sb S4	121	TI	TI	mp-5702
108159	Cr2 Zr	194	HSLSM	HSLSM	mp-570608
626949	Zr Cr2	194	TI	HSLSM	mp-570608
160382	Bi Sn	64	TI	TI	mp-570686
160383	Bi Sn	64	TCI	TI	mp-570686
61556	K Br	221	TrI	TrI	mp-570891
290549	K Br	221	TCI	TrI	mp-570891
107252	Al4 U	74	TI	HSPSM	mp-574122
107243	Al4 U	74	TrI	HSPSM	mp-574122
91075	Ba (Ru O3)	166	TI	TI	mp-5773
172175	Ba Ru O3	166	TCI	TrI	mp-5773
41260	Al6 C3 N2	166	TI	TrI	mp-5798
654993	Al6 C3 N2	160	TrI	TrI	mp-5798
620765	Ce Co2 P2	139	HSLSM	TI	mp-5820
85893	Ce Co2 P2	139	TI	TI	mp-5820
73804	Ag (Ta S3)	63	TI	TrI	mp-5821
84910	Ag (Ta S3)	63	TrI	TrI	mp-5821

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
48027	Co As	33	TrI	TCI	mp-583
48030	Co As	62	TCI	TCI	mp-583
648313	Zr7 P4	12	TCI	TI	mp-583740
40281	Zr7 P4	12	TI	TI	mp-583740
56577	Sr3 (Ru2 O7)	139	TCI	TCI	mp-5868
83189	Sr3 (Ru2 O7)	139	TrI	TCI	mp-5868
1625	Al4 Cu9	215	TI	TrI	mp-593
151371	Al4 Cu9	215	TrI	TrI	mp-593
201285	Cs Pb Br3	221	TI	TrI	mp-600089
29073	Cs Pb Br3	221	TrI	TrI	mp-600089
166011	Ti Ni	51	TI	TCI	mp-603347
157606	Ni Ti	51	TCI	TCI	mp-603347
412834	(Cs2 (Pd Cl4)) (I2)	139	TrI	TrI	mp-607298
240484	Cs2 (Pd Cl4) I2	139	HSLSM	TrI	mp-607298
77938	La3 Ni2 B2 N3	139	TI	HSLSM	mp-6114
391192	La3 Ni2 B2 N3	139	HSLSM	HSLSM	mp-6114
624657	Co2 U P2	129	HSLSM	TCI	mp-611623
67932	U Co2 P2	129	TI	TCI	mp-611623
604219	Pu	11	TCI	TI	mp-613989
43335	Pu	11	TrI	TI	mp-613989
262925	Cs Sn I3	127	TI	TrI	mp-616378
69995	Cs Sn I3	127	TrI	TrI	mp-616378
656928	U B4	127	TI	TI	mp-619
24681	U B4	127	TI	TI	mp-619
615628	U B4	127	HSLSM	HSLSM	mp-619
402420	Ti2 Se	58	TI	TI	mp-620032
98687	Ti2 Se	58	TrI	TI	mp-620032
105403	Ni Th	62	TI	TI	mp-623836
105404	Ni Th	62	TCI	HSLSM	mp-623836
260983	Ni U6	140	TI	HSPSM	mp-636821
105435	Ni U6	140	HSLSM	HSPSM	mp-636821
108870	Ta3 Ge	86	TI	TI	mp-639805
637966	Ta3 Ge	86	TCI	TI	mp-639805
159540	Ti Na (Si2 O6)	15	HSPSM	TrI	mp-6465
166313	Na Ti (Si2 O6)	2	TI	TrI	mp-6465
181722	Fe Pt3	221	HSLSM	HSPSM	mp-649

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
56275	Fe Pt3	221	HSPSM	HSPSM	mp-649
182164	Na3 As	185	HSLSM	TrI	mp-655
79586	Na3 As	185	TrI	TrI	mp-655
174030	Au3 Zn	142	HSLSM	TI	mp-669566
58628	Au3 Zn	142	TI	TI	mp-669566
65614	Y2 Ba3 Cu2 Pt O10	12	TCI	TrI	mp-6829
62390	Ba3 Y2 Pt Cu2 O10	12	TI	TrI	mp-6829
106556	Cu9 In4	215	HSPSM	HSPSM	mp-683917
187898	Cu9 In4	215	TrI	HSPSM	mp-683917
188175	Ca Pd2 P2	139	HSLSM	TI	mp-6926
36371	Ca Pd2 P2	139	TI	TI	mp-6926
155247	Si O2	154	HSLSM	TrI	mp-6930
153455	Si O2	154	TrI	TrI	mp-6930
39830	Si O2	1	TrI	TrI	mp-6930
2623	Pt Si	62	TCI	TCI	mp-696
248491	Pt Si	62	TrI	TCI	mp-696
95762	V2 O3	15	TI	TI	mp-715514
6286	V2 O3	15	TCI	TI	mp-715514
42937	Hf3 Zn3 N	227	HSLSM	HSLSM	mp-722913
638671	Hf3 Zn3 N	227	TI	HSLSM	mp-722913
167668	Cr23 C6	225	HSLSM	HSPSM	mp-723
43377	Cr23 C6	225	HSPSM	HSPSM	mp-723
2837	Cr23 C6	225	TI	HSPSM	mp-723
31400	Ti10 O18	2	TI	TI	mp-748
9038	Ti5 O9	2	TrI	TrI	mp-748
642395	Sr3 Li2	136	HSLSM	HSLSM	mp-7507
15703	Sr3 Li2	136	TI	TI	mp-7507
42614	Co Sb2	14	TI	TrI	mp-755
161492	Co Sb2	14	TrI	TrI	mp-755
23865	Pd4 S	114	TI	TI	mp-7819
648748	Pd4 S	114	TrI	TrI	mp-7819
612693	Ca Ni4 B	191	HSLSM	HSLSM	mp-8307
36504	Ca Ni4 B	191	TI	TI	mp-8307
248526	Ca Si2	164	TCI	TI	mp-8372
41450	Ca Si2	164	TI	TI	mp-8372
187903	Cu5 Sn4	14	TI	TI	mp-845

Continued on next page

Table C.1 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	MP-ID
150125	Cu5 Sn4	14	HSLSM	TI	mp-845
151466	Hf2 Ni	140	HSLSM	TI	mp-861
102808	Ni Hf2	140	TI	TI	mp-861
643231	Mn P S3	12	TrI	TrI	mp-8613
61391	Mn P S3	12	TI	TrI	mp-8613
103923	Ga Pt3	127	TI	TI	mp-862621
635145	Pt3 Ga	140	HSPSM	HSPSM	mp-862621
102630	Co2 Pu	227	TrI	HSLSM	mp-863
624789	Co2 Pu	227	HSPSM	HSPSM	mp-863
26753	Ca B6	221	TrI	TrI	mp-865
612687	Ca B6	221	TI	TrI	mp-865
625712	Cr	139	HSLSM	HSPSM	mp-90
191754	Cr	229	HSPSM	HSPSM	mp-90
252734	Ca C2	12	TI	TrI	mp-917
54185	Ca C2	12	TrI	TrI	mp-917
76166	U	136	TCI	TCI	mp-93
43218	U	136	HSLSM	HSLSM	mp-93
426933	Cd	194	TrI	TI	mp-94
53770	Cd	194	TI	TI	mp-94
248454	Zn Te	225	TrI	TrI	mp-948
31840	Zn Te	225	HSPSM	HSPSM	mp-948
56210	Hg Se	225	TrI	TrI	mp-957
56211	Hg Se	225	HSPSM	HSPSM	mp-957
108270	Ce Zn	221	TI	TrI	mp-986
622313	Ce Zn	221	TrI	HSLSM	mp-986
196147	As2 Te3	166	TrI	TI	mp-9897
68110	As2 Te3	166	TI	TI	mp-9897
633328	Ti2 Fe S4	12	TrI	TCI	mp-9942
42563	Fe Ti2 S4	12	TCI	TCI	mp-9942
42921	Ti2 Ge C	194	TI	TI	mp-9958
180769	Ti2 Ge C	194	HSLSM	TI	mp-9958
168369	Mo N	186	HSPSM	HSPSM	mp-999503
56608	Si O2	1	TrI	TrI	
155252	Si O2	154	TrI	TrI	

Table C.2: Predictions of materials in set B. Materials are list in order of MP-IDs, their topological type on Materiae (Type_MAT) is different from the one on TMD (Type_TM). The chemical formula (Chem. F.) and space group number (SG) are provided on TMD (related to the ICSD-ID).

ICSD-ID	Chem. F.	SG	Type_TMD	Predicted Type	Type_MAT	MP-ID
167854	Tc N	225	HSPSM	HSPSM	TI	mp-1002182
168895	Tc B	187	TCI	HSLSM	HSLSM	mp-1002188
612769	Ce Fe4 B4	86	TCI	HSLSM	TrI	mp-1006322
184575	Tl N	216	HSPSM	TrI	TI	mp-1007778
167897	W2 C	164	TCI	TI	TI	mp-1008625
187430	Hf	191	HSLSM	HSLSM	TI	mp-1009460
180457	Ni C	225	TI	HSPSM	TCI	mp-1009582
167862	Re N	225	HSLSM	HSPSM	HSPSM	mp-1009694
166865	Ca2 As	139	TrI	TCI	TCI	mp-10106
186556	Tl N	186	HSLSM	TrI	TrI	mp-1017982
173564	Ba Ga Ge D	156	TrI	TrI	HSLSM	mp-1018095
173573	Ba Ga Ge H	156	TrI	TrI	HSLSM	mp-1018095
640056	In Nb Se2	187	TI	TCI	TrI	mp-1018116
638225	La H3	225	HSPSM	TrI	TrI	mp-1018144
649608	Pt2 Si	189	HSLSM	HSLSM	TCI	mp-1024071
627929	Cu2 Hg (Sn S4)	121	TrI	TrI	TI	mp-1025467
634721	Ti2 Ga N	194	HSLSM	TI	TI	mp-1025550
90795	Rb Ag3 Se2	12	TrI	TrI	TI	mp-10477
180453	Mg	225	TCI	TI	TI	mp-1056702
639119	Hg Ni	123	HSLSM	HSLSM	TI	mp-1057320
189288	Ce O2	139	HSLSM	HSLSM	TrI	mp-1062989
247972	Zn Si P2	225	HSPSM	HSPSM	HSLSM	mp-1064667
52697	Ba Si	63	TCI	TCI	TI	mp-1067235
291483	Al Sr H3	221	HSLSM	TrI	TrI	mp-1068011
615353	Zr Rh3 B	221	TCI	HSLSM	HSPSM	mp-1068661
161481	Cs (Pb I3)	221	TrI	TrI	TI	mp-1069538
616807	Mg2 Sr Bi2	164	TrI	HSLSM	HSLSM	mp-1069695
174263	Ba Cd2 Ge2	139	TI	TI	TCI	mp-1069712
190810	Hf2 Pt3	63	TI	HSPSM	HSLSM	mp-1070103
657608	Ga2 Te3	166	TI	TrI	TrI	mp-1070116
424597	Ca2 Pd3 Ge	166	TI	TI	HSLSM	mp-1071127
251834	La Os2 Al2 B	123	TI	TI	TrI	mp-1071208
429139	La Rh3 Ga2	74	TI	TCI	HSLSM	mp-1071252
290835	Ca C2	71	TrI	TrI	TI	mp-1072492

Continued on next page

Table C.2 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_MD	Predicted Type	Type_MAT	MP-ID
651864	Th Si Se	139	TrI	TCI	HSLSM	mp-1076936
423608	Mg In2	227	TI	HSPSM	TrI	mp-1077016
650891	Th Si S	139	TrI	HSLSM	TI	mp-1077374
648843	Pd3 Tl2 Se2	166	HSLSM	HSLSM	TCI	mp-1077903
648762	Pd3 Tl2 S2	166	TCI	HSLSM	HSLSM	mp-1077950
609280	Zn Al2 S4	160	TrI	TrI	HSLSM	mp-1078069
605631	Ag Pd5 P	123	TI	HSLSM	TCI	mp-1078079
430934	Y Pd2 Al	194	HSLSM	HSPSM	TI	mp-1078242
193226	Ca2 Sn Ge2	127	TI	TI	HSLSM	mp-1078274
238202	Na2 Mg Pb	194	HSLSM	HSLSM	TCI	mp-1078372
251629	Mo N	194	TrI	HSLSM	HSLSM	mp-1078389
196334	Ba2 Zn Te O6	225	TrI	TrI	HSPSM	mp-1078410
646193	Zr Ni P2	129	TI	HSPSM	TCI	mp-1078514
604338	Pd2 Sr Bi2	129	TCI	TCI	HSLSM	mp-1078567
262036	Na Zn4 As3	166	HSLSM	TrI	TrI	mp-1078703
262070	Li Tl	194	HSLSM	HSLSM	TI	mp-1078753
637658	Pt2 Th Ge2	129	TI	TCI	HSLSM	mp-1078793
165094	Cu2 Cd Ge Te4	121	TI	TrI	TrI	mp-1078909
628554	Cu2 Ni (Si S4)	10	TI	TrI	TCI	mp-1078922
260563	Sr In2 P2	194	TrI	TrI	TI	mp-1078973
262029	Cs Cd4 As3	123	TrI	TrI	TCI	mp-1079080
634496	La Ni Ga2	65	TI	TI	TCI	mp-1079217
649613	Pt3 Si	127	TI	TI	HSLSM	mp-1079269
262066	Na Tl	194	HSLSM	HSLSM	TCI	mp-1079298
656582	Ce Ni Zn	189	HSLSM	HSLSM	TrI	mp-1079342
130041	Zr2 Au2 In	136	HSLSM	HSLSM	HSPSM	mp-1079489
610094	Co Ni As	189	HSLSM	HSLSM	TCI	mp-1079542
262071	K Tl	194	TrI	HSPSM	HSLSM	mp-1079544
169390	Ir V	38	TrI	TrI	TI	mp-1079582
169392	Ir V	8	TrI	TrI	TI	mp-1079582
260562	Ca In2 P2	194	TrI	TrI	TI	mp-1079772
167644	Ti3 Pd	223	HSLSM	HSPSM	HSPSM	mp-1079796
601346	Th Os2 B2	22	TrI	TrI	TI	mp-1079803
130035	Hf2 Pd2 In	136	HSLSM	HSLSM	HSPSM	mp-1079822
601371	Gd Os2 B2	22	TrI	TrI	TI	mp-1079837
174041	Cs Fe Br3	194	TrI	TrI	HSLSM	mp-1079847

Continued on next page

Table C.2 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_MD	Predicted Type	Type_MAT	MP-ID
189460	Na	62	TrI	TI	TI	mp-1079952
604350	Ba Pd2 Sb2	129	HSLSM	TCI	TCI	mp-1079967
186645	Tl3 Ir	223	HSPSM	HSLSM	TI	mp-1080016
165995	K	62	TrI	HSLSM	TI	mp-1080043
168807	Bi2 O3	132	HSLSM	TrI	TrI	mp-1080122
656679	Al Pt3	127	TI	TI	HSPSM	mp-1080448
262033	K Cd4 P3	166	HSLSM	TrI	TrI	mp-1080467
53666	La Pt2 Ge2	4	HSLSM	TrI	HSPSM	mp-1080508
192086	Mg Au Ga	189	TI	HSLSM	HSLSM	mp-1080605
601365	Th Ru2 B2	22	TrI	TrI	TI	mp-1080672
658703	Cs2 Pd Bi2	63	TrI	TrI	TI	mp-1080693
236308	Ba Cu2 Sb2	129	TI	TCI	TCI	mp-1080702
601367	Ce Os2 B2	22	TrI	TrI	TI	mp-1080734
190537	Na Cl7	200	TI	HSPSM	TCI	mp-1080771
150974	Cs Hg	2	TCI	TrI	TI	mp-1080797
62000	Cs Hg	2	TCI	TrI	TI	mp-1080797
262030	Cs Zn4 As3	123	TCI	TrI	TI	mp-1084778
641754	La Rh Sn2	63	TI	HSPSM	TCI	mp-1086668
642380	Li2 Zn Si	164	HSLSM	TrI	TI	mp-1087492
610012	Ce Rh2 As2	129	TCI	TI	HSLSM	mp-1087498
261206	Ba Tl4	12	TI	TCI	TCI	mp-1091391
167204	Ge	194	TrI	TrI	HSLSM	mp-1091415
190543	Na3 Cl2	83	TCI	TI	TrI	mp-1095060
623462	Co La Ge2	63	TI	HSPSM	TCI	mp-1095134
601366	La Os2 B2	22	TrI	TrI	HSPSM	mp-1095136
251586	Ba Ge5	74	TCI	TI	TI	mp-1095284
426005	Fe Bi4 S7	12	TrI	TrI	TI	mp-1095303
290900	Y Zn2	194	HSLSM	HSLSM	TI	mp-1095495
190717	C	139	TrI	TrI	HSLSM	mp-1095534
104252	Hf Os	221	TCI	TI	TI	mp-11452
290935	Os Hf	221	TCI	HSPSM	TI	mp-11452
41016	Tc2 P3	12	TI	TI	TCI	mp-11690
41017	Tc2 P3	2	TI	TI	TCI	mp-11690
651600	Ta Sb2	12	TCI	TCI	TI	mp-11697
99787	Fe3 Pt	139	TI	HSLSM	HSLSM	mp-11798
56694	Ti5 O5	12	TCI	TCI	TI	mp-1203

Continued on next page

Table C.2 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_MD	Predicted Type	Type_MAT	MP-ID
10148	Ti4 O7	2	TrI	TI	TI	mp-12205
162605	Hg Te	221	HSLSM	TCI	HSPSM	mp-13141
61353	Ta2 Pt Se7	12	TI	TrI	TrI	mp-14474
44041	Pa As	225	TI	TI	HSPSM	mp-15694
53287	Cu2 Hf P2	164	HSLSM	TrI	TCI	mp-15986
58656	Ba Hg11	221	TI	HSLSM	HSLSM	mp-1600
26898	Pd6 P	14	TrI	TI	HSLSM	mp-1618
623418	Co2 Ge	194	TCI	HSLSM	HSLSM	mp-1667
100286	Sr Ni12 B6	166	HSLSM	HSLSM	TCI	mp-16830
189417	Ba4 (Ru3 O10)	64	TrI	TrI	TI	mp-17304
150693	Au3 Zn	64	TI	TI	TrI	mp-1755
1203	Sr5 Sn3	140	TI	HSPSM	HSLSM	mp-17720
107444	Au K3 Sn4	59	TrI	TrI	TI	mp-18500
41547	In N	216	HSPSM	TrI	TI	mp-20411
412061	Co4 Ga12	136	HSLSM	HSLSM	TI	mp-20559
102425	Co Ga3	136	HSLSM	HSLSM	TI	mp-20559
107004	Cu Pb Y	186	HSLSM	HSLSM	TrI	mp-20865
647750	Os S2	205	TrI	HSPSM	TI	mp-20905
102242	Ce Os2	227	TCI	HSPSM	HSLSM	mp-2098
105625	Pb3 Sr	123	TI	TCI	HSPSM	mp-21162
411678	Ce Ir2 Si2	129	TCI	HSLSM	HSLSM	mp-21900
621422	Ce Ir2 Si2	129	TCI	TI	HSLSM	mp-21900
58166	Al2 Sr	74	TrI	HSPSM	TI	mp-22318
79364	Bi Te I	156	TrI	TrI	HSLSM	mp-22965
69181	Ca3 Cu2 O4 Cl2	139	TrI	TI	HSLSM	mp-23095
87811	Bi4 Ti3 O12	139	TrI	TrI	HSLSM	mp-23335
239291	Pd5 As	12	TI	TCI	TCI	mp-2512
105567	Os3 Sn7	229	HSLSM	HSPSM	HSPSM	mp-2729
27747	Ce Re4 Si2	65	TI	TI	HSPSM	mp-27861
104268	Hf Ru	221	TCI	TI	TI	mp-2802
39621	Rb4 (Cd Br6)	167	TrI	TrI	TI	mp-28315
68091	Li Pt3 B	189	HSLSM	HSLSM	TCI	mp-28613
71440	Ca4 Ni3 C (C2)2	12	TrI	TrI	TI	mp-28742
75936	Hf3 Te2	139	TCI	TCI	TrI	mp-28919
84825	Zr2 Co P	11	TrI	HSPSM	TI	mp-29152
261988	Bi2 Ca Mg2	164	TrI	TI	TI	mp-29208

Continued on next page

Table C.2 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_MD	Predicted Type	Type_MAT	MP-ID
408938	Ba2 Mg3 Si4	12	TI	TrI	TrI	mp-29564
41478	Si H	164	TI	TrI	TrI	mp-29803
35121	Ti6 O11	2	TI	TI	TrI	mp-30524
99133	Ce Rh2 Si	63	TCI	HSPSM	TrI	mp-31078
106290	Au7 Rb3	65	TCI	TrI	TI	mp-31144
107555	Au8 K4 Ga	12	TCI	TrI	TI	mp-31484
262673	Sr Br2	85	TrI	TrI	TI	mp-32711
603713	Hf Te2	164	HSLSM	TrI	TrI	mp-32887
173153	Tc O2	14	TrI	TI	TI	mp-33137
161063	Ti2 Sn C	194	TI	TI	HSLSM	mp-3871
25315	Ca Cu2 Ge2	139	HSLSM	TCI	TCI	mp-4441
25629	Sr (Pb F6)	131	HSLSM	TrI	TI	mp-504677
42590	Fe Ni2 S4	227	TrI	HSPSM	TI	mp-505522
71217	Ta4 Pd3 Te16	12	TrI	TrI	TI	mp-510426
157925	La Ni Sn H	33	TrI	TI	TI	mp-510577
165169	Ag2 (Sb2 O6)	227	TrI	TrI	HSPSM	mp-540872
250794	Ag Sb O3	227	TrI	TrI	HSPSM	mp-540872
63666	Ce2 Sn5	65	TI	TCI	TCI	mp-541131
61352	Ni (Ta2 Se7)	12	TI	TrI	TrI	mp-541183
78851	(O2) (Pt F6)	206	TrI	TrI	HSPSM	mp-541474
412410	Ca3 Ni3 Si2	62	TrI	TI	TI	mp-542138
409435	K4 Sn4	142	TrI	TrI	TCI	mp-542374
104614	K Sn	142	TrI	TrI	TCI	mp-542374
42912	Pt2 Ge3	62	TI	HSLSM	TCI	mp-542587
43216	Sn	139	TrI	HSLSM	HSLSM	mp-55
70053	Ni3 Bi2 S2	12	TI	TrI	TCI	mp-553994
5434	Ni3 Bi2 S2	12	TI	TI	TCI	mp-553994
616874	Ni3 Bi2 S2	12	TI	TI	TCI	mp-553994
172572	Tl Bi S2	166	TrI	TrI	TI	mp-554310
79536	Na2 (Cu2 Zr S4)	12	TI	TrI	TrI	mp-556536
12104	Cu Bi2 O4	79	TrI	TrI	HSPSM	mp-560165
105422	Ni4 Ti3	148	TCI	HSLSM	TI	mp-567653
171472	Li Ag2 In	225	HSPSM	HSLSM	TI	mp-567787
633778	Fe2 Ta	194	HSLSM	HSLSM	TI	mp-568077
1559	Bi I	12	TCI	TrI	TI	mp-568388
104317	Hg Mg3	155	HSLSM	TrI	TrI	mp-569103

Continued on next page

Table C.2 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_MD	Predicted Type	Type_MAT	MP-ID
6056	Th Br4	88	TrI	TrI	HSLSM	mp-570229
150631	Mg5 Tl2	72	TI	TI	TCI	mp-570849
54102	K2 Ag4 Se3	12	TrI	TrI	TI	mp-573891
249562	Ba Au2 In2	62	TI	TI	TCI	mp-574176
103789	Ga Mg	88	HSLSM	HSPSM	TCI	mp-574266
171243	Ca6 Cu2 Sn7	12	TI	TCI	TCI	mp-579544
418388	Sr13 (Al6 Si8) O	12	TCI	TrI	TI	mp-581263
54553	Ce Pt2 In2	11	TI	TI	TCI	mp-602328
28419	C	36	TrI	TrI	TCI	mp-606949
41026	Co2 Nb4 Pd Se12	12	TI	TrI	TrI	mp-624253
88815	C	67	TrI	TrI	TCI	mp-624889
92981	Pb3 Bi2 S6	12	TI	TrI	TrI	mp-629690
240191	Ni10 Zr7	64	TCI	TI	TI	mp-636525
150276	Fe Ge2	64	TrI	TrI	TI	mp-640708
2853	Cl2 Te S7	62	TrI	TrI	TI	mp-672186
190914	Ge Te	221	TrI	TrI	HSLSM	mp-672698
610361	Cu3 (As Se4)	121	TI	TrI	TrI	mp-675626
78592	Tl2 Ba2 Ca Cu2 O8	139	TCI	TrI	TrI	mp-6885
50333	H2 S	142	TrI	TrI	HSLSM	mp-697135
42460	Ba Mg Ge	129	TI	HSLSM	HSLSM	mp-754515
24757	Ta S2	166	TI	TrI	TrI	mp-755523
262327	Mg2 N F	141	TrI	TrI	TI	mp-7589
35585	Zr Cu2 P2	164	TI	TrI	HSLSM	mp-8219
60632	Hf3 Al2	136	TI	HSLSM	TCI	mp-8462
280783	Cu H10 Se O9	2	TI	TrI	TCI	mp-863429
262063	K Tl	64	TrI	TrI	TI	mp-863730
41521	W	225	TI	HSPSM	TCI	mp-8641
415033	Ba8 (Bi4) O H2	139	TI	TrI	TrI	mp-865213
608909	Al16 Ni7 Ti6	225	TCI	HSPSM	HSLSM	mp-865235
167904	W	229	HSPSM	HSPSM	TCI	mp-91
611582	Y As	225	TrI	TI	TI	mp-933
420725	Rh3 Bi2 S2	12	TCI	TrI	TI	mp-977592
402643	K (Ag3 Se2)	12	TrI	TrI	TI	mp-9782
609161	Al16 Pt7 Ti6	225	HSPSM	HSPSM	HSLSM	mp-978976
165985	Sr Pd2 Ge2	139	HSLSM	TCI	TCI	mp-978986
168181	Sn H4	11	TCI	TI	TI	mp-978992

Continued on next page

Table C.2 – continued from previous page

ICSD-ID	Chem. F.	SG	Type_MD	Predicted Type	Type_MAT	MP-ID
167649	Ti Pd	11	TrI	TI	TCI	mp-979261
167648	Ti Pd	51	TCI	TI	TI	mp-980094
602652	Th Ni2 Sn2	129	TI	TI	TCI	mp-980110
76027	Nb P	141	TI	TI	HSLSM	mp-9830
41761	Y B2 C	135	TI	TI	HSLSM	mp-9880
191657	Co Fe Ti Al	216	TrI	HSPSM	HSPSM	mp-998980
247974	Hg Si P2	225	TI	HSPSM	HSPSM	mp-999187
185442	Nb N	186	HSLSM	TrI	HSPSM	mp-999355
76384	Nb N	194	HSLSM	HSLSM	TI	mp-999357

Table C.3: Predictions on new materials. They are created by replacing one element from material $CeNiGa_2$.

ID	Chem. F.	SG	Predicted Type
13	Ti Ni Ga2	65	TI
14	Zr Ni Ga2	65	TI
15	Hf Ni Ga2	65	TI
18	Th Ni Ga2	65	TI
24	Ce Pd Ga2	65	TI
25	Ce Pt Ga2	65	TCI
27	Ce Gd Ga2	65	TI
31	Ce Ni B2	65	TI
32	Ce Ni Al2	65	TI
35	Ce Ni Tl2	65	TCI
37	Ce Ni Ho2	65	TI

To explore the potential impact of filling the missing value of "Miedema deltaH_inter" with zero, we replaced the "Miedema deltaH_inter" values for these 895 materials with 0 and reran the predictions. Only 15 materials were affected which resulted in changed predictions. This indicates that zero filling is acceptable.

Table C.4: Changes of predictions for 895 materials in set A and set B. The chemical formula (Chem. F.) and space group number (SG) are provided on TMD (related to the ICSD-ID). The old predicted type (Old P.) is the result before replacing, new predicted type (New P.) is the result after replacing.

ICSD	Chem. F.	SG	Type_TMD	Type_MAT	MP-ID	Old P.	New P.
1203	Sr5 Sn3	140	TI	HSLSM	mp-17720	HSPSM	TI
58166	Al2 Sr	74	TrI	TI	mp-22318	HSPSM	TrI
602652	Th Ni2 Sn2	129	TI	TCI	mp-980110	TI	HSPSM
411678	Ce Ir2 Si2	129	TCI	HSLSM	mp-21900	HSLSM	TI
616807	Mg2 Sr Bi2	164	TrI	HSLSM	mp-1069695	HSLSM	TrI
615353	Zr Rh3 B	221	TCI	HSPSM	mp-1068661	HSLSM	TCI
167854	Tc N	225	HSPSM	TI	mp-1002182	HSPSM	HSLSM
638947	Hf5 Sn3	193	HSLSM	/	mp-21104	HSPSM	TI
102630	Co2 Pu	227	TrI	/	mp-863	HSLSM	HSPSM
57534	Al8 Ca Co2	55	TI	/	mp-17871	TCI	TI
248491	Pt Si	62	TrI	/	mp-696	TCI	TrI
156996	In Sb	63	TI	/	mp-20253	TCI	TrI
67932	U Co2 P2	129	TI	/	mp-611623	TCI	TI
25684	U Ni2 Si2	139	TI	/	mp-20596	HSLSM	TCI
1625	Al4 Cu9	215	TI	/	mp-593	TrI	HSPSM

BIBLIOGRAPHY

[1] R. P. Feynman, R. B. Leighton, and M. L. Sands, *The Feynman lectures on physics: mainly mechanics, radiation, and heat*, Vol. 1 (Addison-Wesley Reading, 1964).

[2] P. W. Anderson, Science **177**, 393 (1972).

[3] H. P. McKean Jr, Proceedings of the National Academy of Sciences **56**, 1907 (1966).

[4] B. J. Alder and T. E. Wainwright, The Journal of Chemical Physics **31**, 459 (1959).

[5] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[6] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[7] J. Carrasquilla and R. G. Melko, Nature Physics **13**, 431 (2017).

[8] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Nature **559**, 547 (2018).

[9] A. Kitaev, in *AIP conference proceedings*, Vol. 1134 (American Institute of Physics, 2009) pp. 22–30.

[10] M. Z. Hasan and C. L. Kane, Reviews of Modern physics **82**, 3045 (2010).

[11] X.-L. Qi and S.-C. Zhang, Reviews of Modern Physics **83**, 1057 (2011).

[12] A. Bansil, H. Lin, and T. Das, Reviews of Modern Physics **88**, 021004 (2016).

[13] N. Armitage, E. Mele, and A. Vishwanath, Reviews of Modern Physics **90**, 015001 (2018).

[14] C.-K. Chiu, J. C. Teo, A. P. Schnyder, and S. Ryu, Reviews of Modern Physics **88**, 035005 (2016).

[15] M. Zhang, X. Wang, F. Song, and R. Zhang, Advanced Quantum Technologies **2**, 1800039 (2019).

[16] H. Zhang, C.-X. Liu, X.-L. Qi, X. Dai, Z. Fang, and S.-C. Zhang, Nature physics **5**, 438 (2009).

[17] Y. Xia, D. Qian, D. Hsieh, L. Wray, A. Pal, H. Lin, A. Bansil, D. Grauer, Y. S. Hor, R. J. Cava, *et al.*, Nature physics **5**, 398 (2009).

[18] D. Hsieh, Y. Xia, D. Qian, L. Wray, J. Dil, F. Meier, J. Osterwalder, L. Patthey, J. Checkelsky, N. P. Ong, *et al.*, Nature **460**, 1101 (2009).

[19] Y. Chen, J. G. Analytis, J.-H. Chu, Z. Liu, S.-K. Mo, X.-L. Qi, H. Zhang, D. Lu, X. Dai, Z. Fang, *et al.*, science **325**, 178 (2009).

[20] B. Bradlyn, J. Cano, Z. Wang, M. Vergniory, C. Felser, R. J. Cava, and B. A. Bernevig, Science **353**, aaf5037 (2016).

[21] H. C. Po, A. Vishwanath, and H. Watanabe, Nature communications **8**, 50 (2017).

[22] J. Kruthoff, J. De Boer, J. Van Wezel, C. L. Kane, and R.-J. Slager, Physical Review X **7**, 041069 (2017).

[23] Z. Song, T. Zhang, Z. Fang, and C. Fang, Nature communications **9**, 3530 (2018).

[24] E. Khalaf, H. C. Po, A. Vishwanath, and H. Watanabe, Physical Review X **8**, 031070 (2018).

[25] Z. Song, T. Zhang, and C. Fang, Physical Review X **8**, 031069 (2018).

[26] M. Hellenbrandt, Crystallography Reviews **10**, 17 (2004).

[27] Mariette, Icsd, `http://www2.fiz-karlsruhe.de/icsd_home.html` (2015).

[28] T. Zhang, Y. Jiang, Z. Song, H. Huang, Y. He, Z. Fang, H. Weng, and C. Fang, Nature **566**, 475–479 (2019).

[29] F. Tang, H. C. Po, A. Vishwanath, and X. Wan, Nature **566**, 486–489 (2019).

[30] M. G. Vergniory, L. Elcoro, C. Felser, N. Regnault, B. A. Bernevig, and Z. Wang, Nature **566**, 480–485 (2019).

[31] M. G. Vergniory, B. J. Wieder, L. Elcoro, S. S. Parkin, C. Felser, B. A. Bernevig, and N. Regnault, Science **376**, eabg9094 (2022).

[32] S. Wu, Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, *et al.*, Npj Computational Materials **5**, 66 (2019).

[33] X. Qian, S. Peng, X. Li, Y. Wei, and R. Yang, Materials Today Physics **10**, 100140 (2019).

[34] L. Chen, H. Tran, R. Batra, C. Kim, and R. Ramprasad, Computational Materials Science **170**, 109155 (2019).

[35] F. Legrain, A. van Roekeghem, S. Curtarolo, J. Carrete, G. K. Madsen, and N. Mingo, Journal of chemical information and modeling **58**, 2460 (2018).

[36] P.-P. De Breuck, G. Hautier, and G.-M. Rignanese, npj Computational Materials **7**, 10.1038/s41524-021-00552-2 (2021).

[37] M. Nunez, Computational Materials Science **158**, 117 (2019).

[38] B. Olsthoorn, R. M. Geilhufe, S. S. Borysov, and A. V. Balatsky, Advanced Quantum Technologies **2**, 1900023 (2019).

[39] J. Behler and M. Parrinello, Physical review letters **98**, 146401 (2007).

[40] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, Physical review letters **120**, 143001 (2018).

[41] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Physical Review X **4**, 011019 (2014).

[42] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, *et al.*, Nature materials **15**, 1120 (2016).

[43] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, *et al.*, Nature **581**, 178 (2020).

[44] W. Sun, C. J. Bartel, E. Arca, S. R. Bauers, B. Matthews, B. Orvañanos, B.-R. Chen, M. F. Toney, L. T. Schelhas, W. Tumas, *et al.*, Nature materials **18**, 732 (2019).

[45] Y. Zhang, A. Mesaros, K. Fujita, S. Edkins, M. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. S. Davis, E. Khatami, *et al.*, Nature **570**, 484 (2019).

[46] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, Nature **589**, 59 (2021).

[47] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, Computational Materials Science **152**, 60 (2018).

[48] L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Computer Physics Communications **247**, 106949 (2020).

[49] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Chemistry of Materials **31**, 3564–3572 (2019).

[50] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, npj Computational Materials **6**, 10.1038/s41524-020-00406-3 (2020).

[51] T. Xie and J. C. Grossman, Physical review letters **120**, 145301 (2018).

[52] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, Physical Review Materials **2**, 10.1103/physrevmaterials.2.083802 (2018).

[53] M. Vervloet, H. A. Kiers, W. Van den Noordgate, and E. Ceulemans, Journal of Statistical Software **65**, 1 (2015).

[54] P.-P. De Breuck, M. L. Evans, and G.-M. Rignanese, Journal of Physics: Condensed Matter **33**, 404002 (2021).

[55] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).

[56] L. Breiman, J. Friedman, R. Olshen, and C. Stone, ISBN-13 , 978 (1984).

[57] C. Cortes and V. Vapnik, Machine learning **20**, 273 (1995).

[58] W. S. McCulloch and W. Pitts, The bulletin of mathematical biophysics **5**, 115 (1943).

[59] K. Pearson, The London, Edinburgh, and Dublin philosophical magazine and journal of science **2**, 559 (1901).

[60] I. T. Jolliffe, *Principal component analysis for special types of data* (Springer, 2002).

[61] T. K. Ho, in *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1 (IEEE, 1995) pp. 278–282.

[62] J. H. Friedman, Annals of statistics , 1189 (2001).

[63] B. E. Boser, I. M. Guyon, and V. N. Vapnik, in *Proceedings of the fifth annual workshop on Computational learning theory* (1992) pp. 144–152.

[64] JavaTpoint, Support vector machine algorithm (2021).

[65] D. Arden, Applied deep learning - part 1: Artificial neural networks (2017).

[66] L. Van der Maaten and G. Hinton, Journal of machine learning research **9** (2008).

[67] T. B. Guide, 6 machine learning steps explained for the business (2020).

[68] A. J. Knobbe, A. Siebes, and D. van der Wallen, in *Principles of Data Mining and Knowledge Discovery: Third European Conference, PKDD'99, Prague, Czech Republic, September 15-18, 1999. Proceedings 3* (Springer, 1999) pp. 378–383.

[69] J. Barnard and X.-L. Meng, Statistical methods in medical research **8**, 17 (1999).

[70] R. J. Carroll and S. Pederson, Journal of the Royal Statistical Society: Series B (Methodological) **55**, 693 (1993).

[71] A. C. Cameron and F. A. Windmeijer, Journal of econometrics **77**, 329 (1997).

[72] S. V. Stehman, Remote sensing of Environment **62**, 77 (1997).

[73] Y. Sasaki *et al.*, Teach tutor mater **1**, 1 (2007).

[74] A. P. Bradley, Pattern recognition **30**, 1145 (1997).

[75] M. Stone, Journal of the Royal Statistical Society: Series B (Methodological) **36**, 111 (1974).

[76] K. Ajitesh, Python – nested cross validation for algorithm selection (2020).

[77] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, Physical review letters **49**, 405 (1982).

[78] C. L. Kane and E. J. Mele, Physical review letters **95**, 146802 (2005).

[79] D. C. Tsui, H. L. Stormer, and A. C. Gossard, Physical Review Letters **48**, 1559 (1982).

[80] R. B. Laughlin, Physical Review Letters **50**, 1395 (1983).

[81] L. Fu and C. L. Kane, Physical Review B **74**, 195312 (2006).

[82] L. Fu, C. L. Kane, and E. J. Mele, Physical review letters **98**, 106803 (2007).

[83] T. H. Hsieh, H. Lin, J. Liu, W. Duan, A. Bansil, and L. Fu, Nature communications **3**, 982 (2012).

[84] Y. Tanaka, Z. Ren, T. Sato, K. Nakayama, S. Souma, T. Takahashi, K. Segawa, and Y. Ando, Nature Physics **8**, 800 (2012).

[85] X. Wan, A. M. Turner, A. Vishwanath, and S. Y. Savrasov, Physical Review B **83**, 205101 (2011).

[86] Z. Wang, Y. Sun, X.-Q. Chen, C. Franchini, G. Xu, H. Weng, X. Dai, and Z. Fang, Physical Review B **85**, 195320 (2012).

[87] S. M. Young, S. Zaheer, J. C. Teo, C. L. Kane, E. J. Mele, and A. M. Rappe, Physical review letters **108**, 140405 (2012).

[88] M. Konig, S. Wiedmann, C. Brune, A. Roth, H. Buhmann, L. W. Molenkamp, X.-L. Qi, and S.-C. Zhang, Science **318**, 766 (2007).

[89] Z. Wang, H. Weng, Q. Wu, X. Dai, and Z. Fang, Physical Review B **88**, 125427 (2013).

[90] Z. Liu, J. Jiang, B. Zhou, Z. Wang, Y. Zhang, H. Weng, D. Prabhakaran, S. K. Mo, H. Peng, P. Dudin, *et al.*, Nature materials **13**, 677 (2014).

[91] B. Lv, S. Muff, T. Qian, Z. Song, S. Nie, N. Xu, P. Richard, C. E. Matt, N. C. Plumb, L. Zhao, *et al.*, Physical review letters **115**, 217601 (2015).

[92] S.-Y. Yang, H. Yang, E. Derunova, S. S. Parkin, B. Yan, and M. N. Ali, Advances in Physics: X **3**, 1414631 (2018).

[93] W. A. Benalcazar, B. A. Bernevig, and T. L. Hughes, Science **357**, 61 (2017).

[94] L. Fu, Physical review letters **106**, 106802 (2011).

[95] S.-Y. Xu, I. Belopolski, N. Alidoust, M. Neupane, G. Bian, C. Zhang, R. Sankar, G. Chang, Z. Yuan, C.-C. Lee, *et al.*, Science **349**, 613 (2015).

[96] A. A. Soluyanov, D. Gresch, Z. Wang, Q. Wu, M. Troyer, X. Dai, and B. A. Bernevig, Nature **527**, 495 (2015).

[97] Z. Liu, B. Zhou, Y. Zhang, Z. Wang, H. Weng, D. Prabhakaran, S.-K. Mo, Z. Shen, Z. Fang, X. Dai, *et al.*, Science **343**, 864 (2014).

[98] A. Burkov, M. Hook, and L. Balents, Physical Review B **84**, 235126 (2011).

[99] B. Bradlyn, L. Elcoro, J. Cano, M. G. Vergniory, Z. Wang, C. Felser, M. I. Aroyo, and B. A. Bernevig, Nature **547**, 298–305 (2017).

[100] A. Jain, S. Ong, G. Hautier, W. Chen, W. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. Persson, APL Materials **1**, 011002 (2013).

[101] G. Kresse and J. Hafner, Physical review B **47**, 558 (1993).

[102] G. Kresse and J. Furthmüller, Physical review B **54**, 11169 (1996).

[103] P. Blaha, K. Schwarz, G. K. Madsen, D. Kvasnicka, J. Luitz, *et al.*, An augmented plane wave+ local orbitals program for calculating crystal properties **60** (2001).

[104] K. G. Wilson, Physical review D **10**, 2445 (1974).

[105] Y. Zhang and E.-A. Kim, Physical review letters **118**, 216401 (2017).

[106] Y. Zhang, P. Ginsparg, and E.-A. Kim, Physical Review Research **2**, 023283 (2020).

[107] P. Zhang, H. Shen, and H. Zhai, Physical review letters **120**, 066401 (2018).

[108] M. S. Scheurer and R.-J. Slager, Physical review letters **124**, 226401 (2020).

[109] D. Donoho, IEEE Transactions on Information Theory **52**, 1289 (2006).

[110] C. M. Acosta, R. Ouyang, A. Fazzio, M. Scheffler, L. M. Ghiringhelli, and C. Carbogno, arXiv preprint arXiv:1805.10950 (2018).

[111] G. Cao, R. Ouyang, L. M. Ghiringhelli, M. Scheffler, H. Liu, C. Carbogno, and Z. Zhang, Physical Review Materials **4**, 034204 (2020).

[112] J. Liu, G. Cao, Z. Zhou, and H. Liu, Journal of Physics: Condensed Matter **33**, 325501 (2021).

[113] N. Claussen, B. A. Bernevig, and N. Regnault, Physical Review B **101**, 245117 (2020).

[114] G. R. Schleder, B. Focassio, and A. Fazzio, Applied Physics Reviews **8**, 031409 (2021).

[115] A. Marrazzo, M. Gibertini, D. Campi, N. Mounet, and N. Marzari, Nano letters **19**, 8431 (2019).

[116] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, *et al.*, 2D Materials **5**, 042002 (2018).

[117] J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, *et al.*, Scientific data **6**, 86 (2019).

[118] T. Olsen, E. Andersen, T. Okugawa, D. Torelli, T. Deilmann, and K. S. Thygesen, Physical Review Materials **3**, 024005 (2019).

[119] D. Wang, F. Tang, J. Ji, W. Zhang, A. Vishwanath, H. C. Po, and X. Wan, Physical Review B **100**, 195108 (2019).

[120] N. Andrejevic, J. Andrejevic, B. A. Bernevig, N. Regnault, F. Han, G. Fabbris, T. Nguyen, N. C. Drucker, C. H. Rycroft, and M. Li, Advanced Materials **34**, 2204113 (2022).

[121] A. Ma, Y. Zhang, T. Christensen, H. C. Po, L. Jing, L. Fu, and M. Soljacic, Nano Letters **23**, 772 (2023).

[122] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, Relational inductive biases, deep learning, and graph networks (2018), arXiv:1806.01261 [cs.LG] .

[123] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[124] F. Chollet *et al.*, Keras, `https://keras.io` (2015).

[125] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).

[126] A. M. Deml, R. O'Hayre, C. Wolverton, and V. Stevanović, Phys. Rev. B **93**, 085142 (2016).

[127] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, Nature **571**, 95 (2019).

[128] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Computational Materials **2**, 10.1038/npjcompumats.2016.28 (2016).

[129] S. Kotochigova, Z. H. Levine, E. L. Shirley, M. D. Stiles, and C. W. Clark, Phys. Rev. A **55**, 191 (1997).

[130] K. Laws, D. Miracle, and M. Ferry, Nature communications **6**, 8123 (2015).

[131] M. A. Butler and D. S. Ginley, Journal of the Electrochemical Society **125**, 228 (1978).

[132] C. Kittel, *Introduction to solid state physics* (John Wiley & sons, inc, 2005).

[133] R. Zhang, S. Zhang, Z. He, J. Jing, and S. Sheng, Computer Physics Communications **209**, 58 (2016).

[134] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Physical Review B **96**, 024104 (2017).

[135] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).

[136] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, International Journal of Quantum Chemistry **115**, 1094 (2015).

[137] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore, Applications of evolutionary computation: 19th european conference, evoapplications 2016, porto, portugal, march 30 – april 1, 2016, proceedings, part i (Springer International Publishing, 2016) Chap. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pp. 123–137.

[138] J. Fan and J. Lv, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 849 (2008).

[139] G. Hooker, L. Mentch, and S. Zhou, Statistics and Computing **31**, 1 (2021).

[140] Y. He, Y. Jiang, T. Zhang, H. Huang, C. Fang, and Z. Jin, Chinese Physics B **28**, 087102 (2019).

[141] A. Togo and I. Tanaka, Scripta Materialia **108**, 1 (2015).

[142] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical review letters **77**, 3865 (1996).

[143] Z. Wang, Vasp2trace (2021).

[144] M. Vergniory, Check topological mat.

[145] D. Freedman, R. Pisani, and R. Purves, *Statistics (international student edition)* (2007).

[146] T. Chen and C. Guestrin, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016) pp. 785–794.