

Curation and featurization of multiple topological materials databases

Yuqing He(贺雨晴), Matteo Giantomassi, Gian-Marco Rignanese, and Hongming Weng(翁红明)
Citation: Chin. Phys. B, 2026, 35 (**5**): 050701. DOI: 10.1088/1674-1056/ae2bf2
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

MaterialsGalaxy: A platform fusing experimental and theoretical data in condensed matter physics

Tiannian Zhu(朱天念), Zhong Fang(方忠), Quansheng Wu(吴泉生), and Hongming Weng(翁红明)
Chin. Phys. B, 2025, 34 (**12**): 120702. DOI: 10.1088/1674-1056/ae172a

Optimized PID neural network closed-loop control for basal ganglia network in Parkinson's disease

Hengxi Zhang(张恒熙), Honghui Zhang(张红慧), Shuang Liu(柳爽), and Lin Du(都琳)
Chin. Phys. B, 2025, 34 (**12**): 120701. DOI: 10.1088/1674-1056/ae0d55

Improved physics-informed neural networks incorporating lattice Boltzmann method optimized by tanh robust weight initialization

Chenghui Yang(杨程晖), Minglei Shan(单鸣雷), Mengyu Feng(冯梦宇), Ling Kuai(蒯玲), Yu Yang(杨雨), Cheng Yin(殷澄), and Qingbang Han(韩庆邦)
Chin. Phys. B, 2025, 34 (**11**): 110701. DOI: 10.1088/1674-1056/adfc43

M2ANet: Multi-branch and multi-scale attention network for medical image segmentation

Wei Xue(薛伟), Chuanghui Chen(陈创慧), Xuan Qi(戚轩), Jian Qin(秦健), Zhen Tang(唐振), and Yongsheng He(何永胜)
Chin. Phys. B, 2025, 34 (**8**): 080703. DOI: 10.1088/1674-1056/adce96

Testing algorithm for the computation of the transverse emittance of the ion beams generated by the ECR mVINIS ion source based on a pepper-pot method

Viktor Jocić, Igor Telečki, and Ivan Trajić
Chin. Phys. B, 2025, 34 (**6**): 060701. DOI: 10.1088/1674-1056/adc674

DATA PAPER

Curation and featurization of multiple topological materials databases

Yuqing He(贺雨晴)[1], Matteo Giantomassi[2], Gian-Marco Rignanese[2,3,4,†], and Hongming Weng(翁红明)[1,‡]

[1] *Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*
[2] *Institute of Condensed Matter and Nanosciences, UCLouvain, Chemin des Étoiles* 8, 1348 *Louvain-la-Neuve, Belgium*
[3] *WEL Research Institute, Avenue Pasteur* 6, 1300 *Wavre, Belgium*
[4] *School of Materials Science and Engineering, Northwestern Polytechnical University, Xi'an* 710072, *China*

(Received 27 October 2025; revised manuscript received 11 December 2025; accepted manuscript online 12 December 2025)

The discovery of topological materials has advanced rapidly due to high-throughput computation and machine learning, but research progress is hampered by inconsistent classification standards and fragmented data resources. Existing databases differ in computational methods, material coverage, and labeling criteria, making it difficult to compare findings across studies. To overcome these challenges, we present a unified topological materials dataset that systematically combines and reconciles two major databases: Materiae and the Topological Materials Database. This dataset provides consistent topological classifications for 35608 materials, accessible through the Materials Galaxy platform for interactive exploration and available for bulk download via MatElab. We describe the featurization methodology that converts crystal structures into 4710 machine-learning-ready descriptors and present a comprehensive analysis of topological material distributions. This work serves as a complete guide for accessing, utilizing, and interpreting this unified resource, designed to enable reproducible machine learning applications and accelerate the discovery of topological materials.

Keywords: data, topological materials, machine learning

PACS: 07.05.Kf

DOI: 10.1088/1674-1056/ae2bf2 **CSTR:** 32038.14.CPB.ae2bf2

1. Introduction

Over the past two decades, topological materials have emerged as one of the most fascinating discoveries in condensed matter physics. They are characterized by the presence of topological invariants,[1–4] which are mathematical quantities that remain unchanged even under continuous deformations of the electronic structure. Such materials host unique topological configurations in their electronic band structures, making them robust against certain types of defects and disorders.[5–7] They exhibit remarkable phenomena, including quantum anomalous Hall effects,[3] topological superconductivity[6] and dissipationless edge conduction.[5,8]

Topologically non-trivial materials (NTMs) are primarily classified into two categories: topological semimetals (TSMs) and topological insulators (TIs).[9–14] TSMs possess protected electronic states, such as Weyl or Dirac fermions,[15–17] which emerge at specific points or lines in their electronic band structure. Weyl semimetals contain Weyl nodes — discrete points in the Brillouin zone where two non-degenerate bands cross. Dirac semimetals, on the other hand, feature Dirac nodes, which are four-fold degeneracies formed by the crossing of two doubly degenerate bands. Nodal-line semimetals exhibit one-dimensional band-touching lines in the Brillouin zone. Unlike conventional semimetals, where band crossings are unstable and easily gapped out, TSMs maintain stable band degeneracies protected by symmetries or topology, making them robust against small perturbations. TIs represent another important class characterized by insulating bulk properties with a finite band gap induced by strong spin–orbit coupling, while the time-reversal symmetry protects the topological phase.[5,6] Their most distinctive feature is the presence of conducting surface/edge states — either topological surface states or Dirac surface states. These states are protected by global topological invariants, making them resistant to backscattering and robust against impurities, defects, and disorder. This unique property enables efficient charge and spin transport, distinguishing TIs from ordinary insulators and offering promising potential for various technological applications.

Research on topological materials is advancing rapidly, motivated by both fundamental scientific interest and technological potential. Many advances have been made through the combined efforts of theoretical predictions and experimental realizations. Dirac semimetals such as cadmium arsenide (Cd_3As_2)[18,19] and Weyl semimetal tantalum arsenide (TaAs)[20] have been identified. Bismuth selenide (Bi_2Se_3) and bismuth telluride (Bi_2Te_3)[11–13] represent prototypical

†Corresponding author. E-mail: gian-marco.rignanese@uclouvain.be
‡Corresponding author. E-mail: hmweng@iphy.ac.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

topological insulators. In recent years, the discovery of NTMs has been greatly accelerated by the development of high-throughput computational screening methods, combining density functional theory (DFT)[21,22] with topological quantum chemistry theory (TQC).[23] In 2019, three independent research groups employed similar frameworks to systematically analyze the nonmagnetic crystalline compounds from the Inorganic Crystal Structure Database (ICSD)[24] or the Material Project (MP).[25] Despite applying slightly different screening criteria, these parallel efforts yielded generally consistent results, collectively identifying tens of thousands of NTMs by analyzing the symmetry of their wavefunctions.[26–28] This resulted in the prediction of thousands of topological phases, their catalogs are presented in corresponding online databases, such as Materiae,[26] Topological Materials Database (TMD),[27,29] and Topological Materials Arsenal.[28] The theoretical framework has since been extended to magnetic systems,[30–32] further expanding the scope of topological materials research.

Sufficient data and advances in machine learning (ML)[33] have introduced powerful new approaches for topological materials discovery; various ML techniques have been successfully applied. By transforming Hamiltonians into interpretable inputs, studies have employed neural networks to identify topological phases like Chern insulators[34] and Z_2 topological insulators.[35] Zhang *et al.*[36] trained a neural network to predict winding numbers in 1D insulators, demonstrating its ability to learn the discrete winding number formula. Later, Scheurer *et al.*[37] developed an unsupervised k-means clustering method to identify adiabatic deformation paths between Hamiltonians, enabling phase classification by comparison with reference trivial insulators (TrIs). Meanwhile, additional studies have made comparable advances in applying ML to *ab-initio* derived datasets. Compressed sensing[38] was utilized[39] to derive a descriptor for the Z_2 invariant, enabling the classification of metals, TrIs, and 2D TIs through analysis of 220 functionalized 2D honeycomb-lattice materials isoelectronic to graphene. Subsequent studies utilized SISSO (sure independence screening and sparsifying operator),[40] an algorithm derived from compressed sensing principles, for TI/TrI classification: Cao *et al.*[41] identified a minimal 2D descriptor (atomic number and electronegativity) considering 243 alloyed tetradymites, while Liu *et al.*[42] developed an expanded descriptor (adding the number of valence electrons) considering 116 ternary half-Heusler compounds. Several other studies employed larger datasets with general materials. Claussen *et al.*[43] achieved 87.0% accuracy using gradient boosted trees (GBTs) on 35009 symmetry-based entries from the TMD,[27,29] successfully distinguishing five quantum phases: TrIs, two TSM classes: enforced semimetal (ES) or enforced semimetal with Fermi degeneracy (ESFD), and two TI classes: split extended Brillouin zone representation (SEBR) or not equal to a linear combination (NLC). Using a dataset of TIs and TrIs from 2D materials databases,[44–48] a study[49] developed a method that integrates SISSO for feature space construction and GBTs for predicting topological classifications of insulating materials. Alternative approaches explored different physical signatures, Andrejevic *et al.*[50] developed a neural network classifier leveraging the x-ray absorption near-edge structure (XANES) spectra of 16458 inorganic compounds from ICSD and their topological classes from TMD. Their model achieved an F_1 score of 89% in distinguishing NTMs and TrIs from their XANES spectral signatures. The 'topogivity' approach developed by Ma *et al.*[51] in 2023 attained 82.7% average accuracy in 11×10-fold NCV for NTM/TrI classification. Based on 9026 materials from the work of Tang *et al.*,[28] this method uses support vector machines to derive element-specific topogivities that map chemical formulas to topological class indicators.

Advances in first-principles calculations have spurred the creation of multiple online databases[26–29,52] cataloging predicted topological phases. However, these resources often diverge in computational methodologies, classification criteria, and material coverage, complicating efforts to establish a comprehensive reference. Meanwhile, studies employing ML to classify topological materials have yielded promising yet inconsistent results. For instance, while Claussen *et al.*[43] demonstrated that topology may depend weakly on atomic positions within crystal lattices, other approaches such as those leveraging XANES spectra[50] or topogivity[51] reveal sensitivity to local chemical environments, suggesting a more nuanced role for atomic arrangements. These discrepancies are further exacerbated by the diversity of material systems studied (spanning 2D to bulk 3D crystals), variations in input databases, and differences in feature selection for ML models, making direct comparisons across studies difficult.

To address these challenges, we have systematically integrated two major topological materials databases, Materiae[26] and TMD[27,29] into a unified, curated dataset $M \cup T$. While our companion paper[53] provides detailed benchmark results and performance comparisons of classification methods applied to this dataset, the present work focuses on documenting the dataset itself. The following sections introduce the unified dataset and describe our use of the Matminer tool for systematic feature engineering, which transforms crystal structures into informative feature vectors — including electron localization function descriptors and other physically interpretable features compatible with machine learning models. This process ensures both algorithmic readiness and physical interpretability, supporting direct application in visualization and model training. Subsequent sections provide a statistical overview of the compiled materials, offer guidelines for using

the resource, and detail the methodology for dataset construction and integration.

2. Data curation

We present a curated dataset, $M \cup T$, created by integrating and reconciling materials from the Materiae and TMD databases. This final unified set is the union of two refined constituent datasets: M, which was extracted from Materiae, and T, which was constructed from TMD. Based on their origin, all entries can be systematically categorized into three distinct subsets: $M \backslash T$ for materials exclusive to the dataset M, $M \cap T$ for materials common to both, and $T \backslash M$ for materials found solely in the dataset T. A comprehensive account of the data acquisition and integration methodology from the original sources is available in Section 5. The present section focuses on describing the content of the datasets and the process employed to generate machine-learning-ready features.

The dataset $M \cup T$ comprises 35608 rigorously validated material entries. Each entry contains a pymatgen[54] structure object and a unified topological classification label derived from DFT and TQC analysis that includes spin–orbit coupling effects. Each entry carries a unique ID following Materiae's indexing conventions and is associated with at least one additional identifier, either an MP-ID from the MP or an ICSD-ID from the ICSD to facilitate cross-referencing and data validation. The topological classifications are labeled into five types: TrI and four types of NTMs: high-symmetry-line semimetal (HSLSM), high-symmetry-point semimetal (HSPSM), topological crystalline insulator (TCI) and topological insulator (TI). For materials in subset T, the original topological label from the TMD is also provided. For illustrative purposes, Table 1 presents five sample entries, where the chemical formula is displayed in the structure column instead of the full structural data for simplicity. The complete dataset with full structures is available and referenced in Section 4.

Table 1. 5 Example entries of dataset $M \cup T$. The structure column contains the chemical formula instead of the full structure for a simplified visualization. The MAT_TYPE is the topological type given by Materiae. The subset column shows the subset the entry belongs to. The MPID and ICSD_ID are the indexes of the material in the Material Project and the Inorganic Crystal Structure Database, respectively. The TM_TYPE is the original topological label in the Topological Materials Database.

ID	Structure	MAT_TYPE	Subset	MPID	ICSD_ID	TM_TYPE
MAT00000190	K2 Sc2 B2 P4 H2 O18	TrI	M/T	mp-23808		
MAT00025751	Zr1 Ru3 C1	HSLSM	$M \cap T$	mp-10706	77282.0	ES
MAT00002966	Y2 Fe2 Si2 C1	TI	$M \cap T$	mp-1025339	617825.0	NLC
MAT00029130	Ca2 Fe10 As6	HSPSM	T/M	mp-1105435	427443.0	ESFD
MAT00038040	Zr1 Al8 Fe4	TCI	T/M		607756.0	SEBR

This dataset is now suitable for application to specialized machine learning methods developed for materials science. However, for the application of general machine learning algorithms, such as random forests[55] or neural networks,[56] additional feature engineering[57] of the structural information is necessary. This process is a key component of the machine learning pipeline, involving the extraction, construction, and selection of informative, machine-understandable numerical representations from raw data by integrating domain-specific knowledge. These resulting features serve as the foundation for model training. In cases where certain features are missing from the data, feature imputation[58] is applied to systematically estimate or replace absent values. Standard approaches include single imputation (e.g., mean, median, or zero filling) and multiple imputation, where predictive models are used to infer missing values based on available data. Categorical variables, including target labels in classification tasks, must also be encoded–commonly through ordinal encoding, which assigns a distinct integer to each category, or one-hot encoding, which represents each category as a binary vector. Together, these procedures improve the expressiveness, completeness, and usability of the data, leading to enhanced model performance and generalization.

To make the dataset immediately useful for general machine learning methods that require numerical descriptors, we employ feature extraction functions from the MatMiner[59] library to convert all crystal structures into a comprehensive set of 4710 physically relevant features. This featurization process transforms abstract structural information into physically relevant quantities that can be consumed by standard ML algorithms. For instance, for a face-centered cubic Pd structure, the GlobalSymmetryFeatures class yields symmetry features such as [225, 'cubic', True], representing the space group number, crystal system, and centrosymmetry. We adopt the DeBreuck2020Featurizer[60] with accelerated oxidation state parameters as our primary feature set. Missing features are imputed with the default values (most are zero) of the corresponding featurizer. Table 2 shows the full list of featurizers, which are broadly categorized into three types: composition, structure, and atomic sites.

Composition-based featurizers generate descriptors based on the material's chemical formula, primarily producing element-related attributes. For example, the ElementProperty featurizer calculates statistics (e.g., mean, maximum) of elemental properties such as the atomic number or the electron orbital count. The ValenceOrbital featurizer performs statistical analyses on the valence electrons of constituent elements, such as the maximum number of p-electrons. This class of

features has proven effective for predicting formation energies in inorganic crystals.[59] Structure-based features incorporate spatial crystal structure information. The GlobalSymmetryFeatures class outputs the space group number, crystal system, and centrosymmetry. The DensityFeatures featurizer calculates properties such as the material's density and volume per atom. Site-based features describe the local environment of individual atomic sites within a structure. As an example, the CoordinationNumber quantifies the number of nearest neighbors for a specific site, often determined using Voronoi tessellation. More complex descriptors, such as those from CrystalNNFingerprint, evaluate how well a site's environment matches ideal structural motifs (e.g., body-centered cubic) to form a detailed feature vector. Collectively, these featurizers transform complex material attributes into a fixed-length vector of physically relevant numerical quantities for machine learning applications. A complete introduction to the different featurizers can be found in Ref. [61].

Table 2. The set of Matminer featurizers to generate the numerical descriptors.

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

3. Data distributional analysis

This section presents an analysis of the data distribution pertaining to the topological types of materials within dataset $M \cup T$ and its three subsets $M \backslash T$, $M \cap T$ and $T \backslash M$, thereby providing a comprehensive overview of the composition of the datasets.

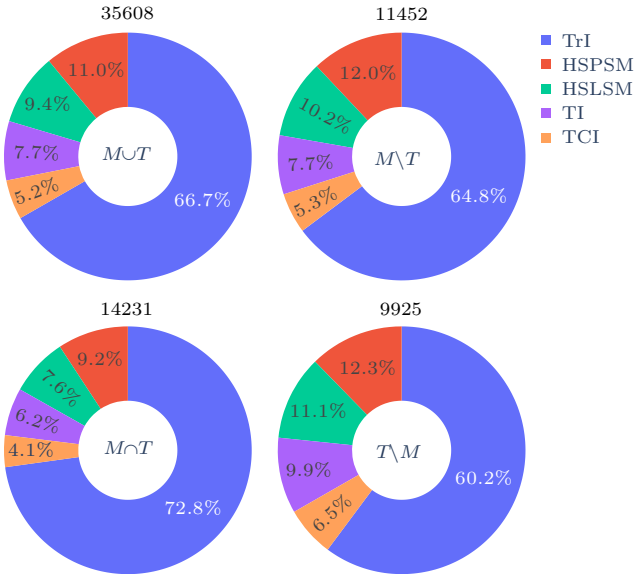


Fig. 1. Dataset composition by topological type. The analysis examines the combined dataset $M \cup T$ (where M originates from Materiae and T from the Topological Materials Database) along with its three subsets: $M \backslash T$ (materials unique to M), $M \cap T$ (materials shared by both), and $T \backslash M$ (materials unique to T).

Figure 1 displays the topological class distribution as pie charts, showing both the absolute counts and proportional representation for the combined dataset $M \cup T$ and its three subsets. It allows for immediate comparison of type frequencies between the complete dataset and its derived subsets, providing clear quantitative understanding of their respective topological classifications. TrI dominates across all configurations, comprising 60.2%–72.8% of materials, with the highest proportion observed in $M \cap T$ (72.8%) and the lowest in the intersection subset $T \backslash M$ (60.2%). NTMs consistently represent a small proportion across all datasets, with stable relative proportions among their four subtypes. TSMs appear at moderate levels, showing minor variations across different subsets. Meanwhile, both TIs and TCIs consistently appear as minority phases, with TCIs being especially scarce in $T \backslash M$ (4.1%).

The relative distribution across the seven different crystal systems is shown in Fig. 2. It presents a comprehensive analysis of the distribution of topological classes within each system, with the horizontal axis listing the crystal systems: monoclinic, orthorhombic, trigonal, tetragonal, triclinic, cubic, and hexagonal, while the vertical axis represents the count of materials. The data is color-coded to distinguish the topological classes. We observe that the monoclinic and orthorhombic systems have the highest total count, indicating that they are the most common crystal structures in the dataset. Following them in frequency are the tetragonal, cubic, and hexagonal systems, whereas the trigonal and triclinic systems appear relatively scarce. Notably, NTMs in orthorhombic systems display a distribution similar to that of the overall dataset, though with a greater proportion of TIs and TCIs. In the tetrago-

nal, cubic, and hexagonal systems, NTMs are observed more frequently, with HSPSM being especially prevalent in cubic systems and HSLSM more dominant in hexagonal systems. On the other hand, triclinic systems show only minimal occurrences of NTMs and a complete absence of TSMs. This analysis suggests a potential correlation between crystal systems and topological classification. For instance, the relatively more frequent presence of NTMs in the cubic system can be attributed to its high symmetry, which helps protect band degeneracy (e.g., PbTe in space group $Fm\bar{3}m$, a TCI belonging to the cubic system). It is also noteworthy that this phenomenon is related to the method that is used for determining topological classification in our dataset, as it relies on crystal symmetry.

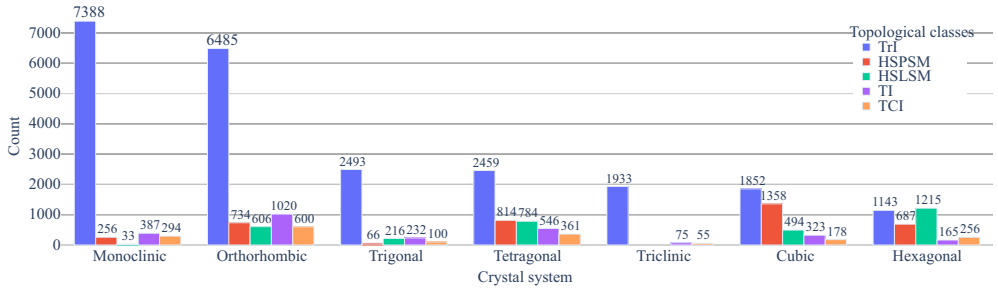


Fig. 2. Relative distribution of dataset $M \cup T$ under different crystal systems.

The distribution of materials according to the presence or absence of a center of symmetry is illustrated in Fig. 3. Centrosymmetric crystals are far more numerous ($\sim$ 27k) than non-centrosymmetric ones ($\sim$ 8k) and also show a higher occurrence of symmetry-protected topological states. In contrast, the non-centrosymmetric subset demonstrates significantly less diversity in NTMs, with their counts being an order of magnitude lower than those in centrosymmetric systems. The asymmetry becomes even more evident when analyzing specific topological phases, the TrIs decrease by 3-fold, while TSMs (HSLSM and HSPSM) reduce by roughly 6-fold. Meanwhile, TIs and TCIs experience a sharp 10-fold reduction, indicating that TI/TCI phases are particularly sensitive to the loss of inversion symmetry. Overall, the data suggests that although symmetry-protected topological states exist in both crystal types, centrosymmetric systems not only outnumber non-centrosymmetric ones but also offer more conducive conditions for the formation and stability of diverse topological phases–especially those like TIs and TCIs, which rely heavily on stringent symmetry protection.

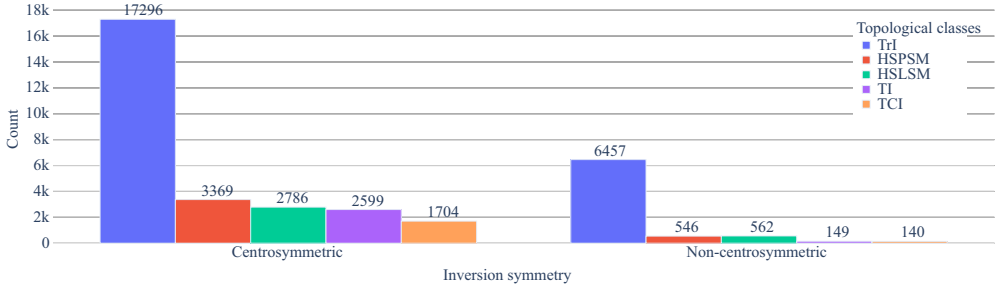


Fig. 3. Relative distribution of dataset $M \cup T$ with different inversion symmetries.

Figure 4 presents the chemical composition statistics of compounds in the dataset $M \cup T$, where each element is represented by its chemical symbol accompanied by two data rows. The lower row provides the total number of compounds containing a certain element and the upper row indicates the occurrence frequency (in percentage) of NTMs among these compounds. A color scale visually represents the NTM frequency values, with gray indicating elements absent from the dataset. For instance, hydrogen (H)-containing compounds amount to 4544 with a 7% NTM frequency. Additional plots for subsets $M \backslash T$, $M \cap T$ and $T \backslash M$ are provided in the appendix. From this comprehensive visualization, we identify both the universal behaviors and dataset-specific peculiarities that emerge.

The NTM frequencies across datasets remain generally consistent, with noticeable variations only occurring when the total number of compounds containing certain elements becomes relatively small. Transition metals, particularly Mn, Fe, Co, Ni, Rh, and Ir, exhibit both high NTM frequencies (50%–85%) and large compound counts, underscoring their significance in NTMs. In contrast, noble gas elements occur in fewer compounds with minimal NTM frequencies. Alkali metals, alkaline earth metals, and reactive nonmetals appear in numerous compounds, but demonstrate moderate NTM frequencies (7%–30%).

Differences can be identified among subsets. The exclu-

sion of certain rare-earth elements from the dataset T means that materials containing these elements appear exclusively in dataset $M\backslash T$. Furthermore, since the TMD database incorporated more comprehensive ICSD structures during its construction, the dataset T/M includes some materials containing Po, Rn, Ra, or Am that are absent in other subsets. For materials containing noble gas element, the dataset T includes a few NTMs that Materiae did not identify. Although both datasets are of similar size, the dataset T contains substantially more materials incorporating specific elements, such as Mn (444 vs. 121) or Fe (740 vs. 222). This discrepancy may relate to potential residual magnetic materials in the dataset $T\backslash M$ (where 5122 materials lack a successful link to the MP and thus have unknown magnetic moments), given that magnetic materials typically contain Fe.

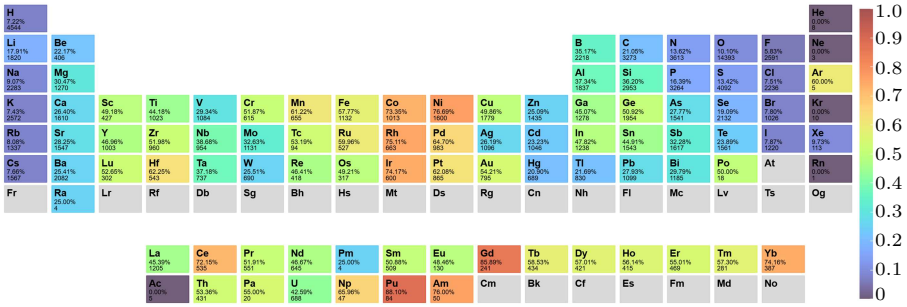


Fig. 4. Chemical composition of compounds in dataset $M \cup T$. Beneath each atomic symbol, the bottom row displays the total count of compounds containing that element, while the top row presents the corresponding NTM occurrence frequency (visualized through the color scale).

4. Data availability and application

The $M \cup T$ dataset is publicly available through multiple channels designed for different use cases. For interactive exploration, users can interactively query and explore individual material entries through the Materials Galaxy web platform (https://materialsgalaxy.iphy.ac.cn/materials). For programmatic access, the entire dataset can be accessed via API requests to Materials Galaxy, enabling integration into automated workflows and scripts. Furthermore, the complete dataset is available for download as a JSON-serialized DataFrame from MatElab (https://in.iphy.ac.cn/eln/link.html#/119/P6g5) and Materials Cloud Archive[62] (https://doi.org/10.24435/materialscloud:xx-xb). This is ideal for offline analysis and large-scale machine learning training. Additionally, the dataset can be interactively visualized in a reduced feature space using the chemiscope tool,[63] allowing users to explore correlations and clusters. And finally, all resources are centrally accessible at https://topoclass.modl-uclouvain.org.

The unified dataset serves as a foundational resource for computational materials science, particularly in the accelerated discovery of topological materials. As demonstrated in our existing studies,[53,64] this dataset is not merely a static repository, but a dynamic tool that powers both predictive analytics and generative discovery. For instance, its large scale and consistent labeling make it an ideal benchmark for developing and evaluating machine learning models, such as the top-performing XGBoost classifier used in our first paper to predict a material's topological class at various levels of granularity. Beyond predictive modeling, the dataset enables sophisticated data mining, allowing researchers to probe the correlations between chemical composition, crystal symmetry, and the emergence of topological states. It is also crucial for the inverse design of new materials. As showcased in another study involving some of us, predictive models trained on this dataset can be integrated into generative AI frameworks like PODGen,[64] where the model acts as a guide to steer the generation of novel crystal structures with a high probability of possessing desired topological properties. This approach significantly increases the efficiency of discovering new materials, successfully yielding thousands of candidates — including hard-to-find gapped topological insulators — and enabling a hypothesis-driven research paradigm where users can define specific criteria to receive curated lists of candidate materials for targeted applications.

Furthermore, we have developed a comprehensive benchmarking framework based on this dataset that provides not only standardized data but also establishes unified evaluation protocols for the quantitative comparison of model performance. These efforts facilitated the creation of the broader MatSciBench platform, which embodies a systematic "data integration–algorithm optimization–benchmark validation" framework for multiple materials informatics tasks. This platform offers a unified standard for fair and reproducible algorithm comparisons across the research community.

5. Data acquisition methodology

This section presents the comprehensive data acquisition process to assemble the unified dataset $M \cup T$. Initially, we construct a preliminary dataset MAT from the Materiae database[26] and a preliminary dataset TM from the TMD database.[27,29] They are then related and subjected to

a conflict-resolution procedure, resulting in two cleaned and harmonized datasets M and T. The final dataset $M \cup T$ is the union of the two refined datasets M and T. Our methodology involves systematic data collection, rigorous cleaning procedures, and standardized formatting to ensure consistency. It is important to note that while both databases were constructed using identical theoretical methodologies, their final classification criteria for NTMs exhibit discrepancies. To address this inconsistency, we have established comprehensive cross-references between the four NTM categories across the two databases. These mappings are described in detail in Subsection 5.2.

5.1. Dataset *M*

Dataset M is derived from the Materiae database. This database comprises topological properties for 26120 materials, which were computed employing VASP[65] and SymTopo[66] in accordance with the workflow depicted in Fig. 5. We will provide a brief introduction to this workflow before proceeding to detail the construction of dataset M. Although it is not a requirement for dataset M specifically, this workflow will act as a reference to perform data cleaning during the construction of dataset T from the TMD database, as will be discussed in Subsection 5.2.

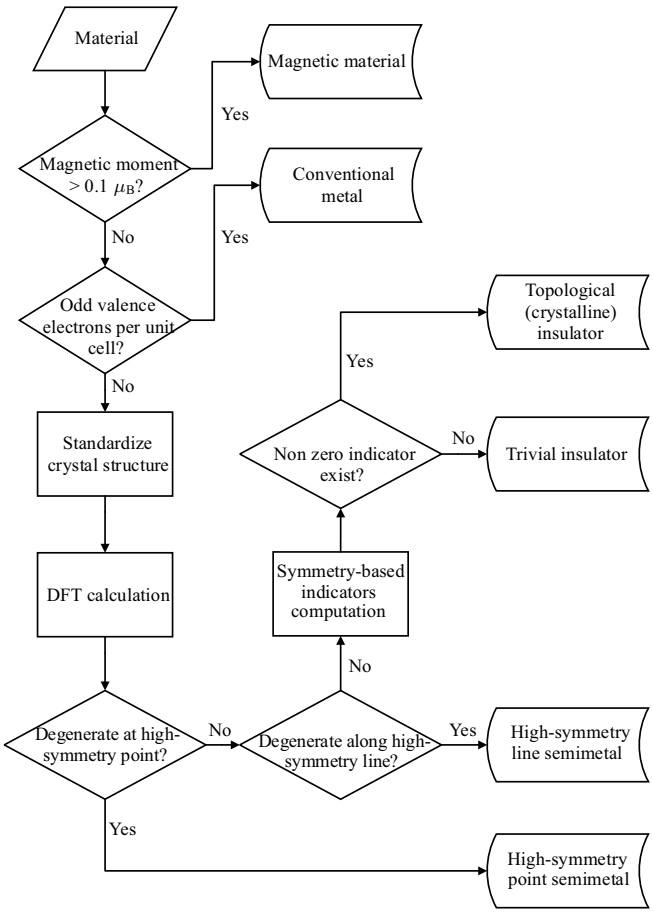


Fig. 5. Workflow of high-throughput calculations for building dataset MAT. Image adjusted from Ref. [66].

The selection process of the Materiae database began with 39519 materials co-registered in both MP and ICSD databases. As a preliminary step before DFT computations, they implemented an exclusion protocol to filter the dataset. First, they identified and removed magnetic materials, defined as those with magnetic moments exceeding 0.1 μ_B per unit cell according to the MP records. Additionally, they flagged conventional metals, determined by odd electron counts per unit cell based on the employed pseudopotentials (PBE-GGA functional, the generalized gradient approximation of the Perdew–Burke–Ernzerhof type exchange–correlation potential[67]) from the VASP software package. These exclusions were necessary because SymTopo is unsuitable for analyzing magnetic systems and conventional metals are not considered as potential NTMs candidates.

Structural standardization is a critical preprocessing step. Since the initial structures acquired from the MP database were not in a unified convention — with arbitrary primitive cell orientations and varying origin points, they went through a symmetry verification via PHONOPY. This standardization process, which included symmetrization of atomic positions following lattice parameter and atomic coordinates loading, enabled reliable high-throughput calculations and ensured that VASP could correctly identify their symmetries. During this phase, 166 materials were excluded due to discrepancies between the space groups identified by PHONOPY and those listed in MP, with a tolerance of up to 0.1 Å in atomic position differences. All materials in space group 1 were also excluded from the final database, as according to TQC analysis, these materials are invariably classified as trivial insulators. The remaining materials then progressed to DFT characterization. For TSMs, certain bands below the Fermi level maintain degeneracy and resist separation from other bands due to unsatisfied compatibility relations. Their topological nodes persist against symmetry-preserving perturbations, with their specific degeneracy patterns and dimensionalities determining classification as either HSPSM or HSLSM.[26] For TIs, the valence bands satisfy compatibility relations but cannot be expressed as a sum of elementary band representations. This fundamental characteristic leads to non-zero topological indicators, where the parity of the symmetry indicator serves to distinguish between TIs and TCIs.

In this work, we incorporated all Materiae's published results along with their PHONOPY-standardized VASP input files (POSCAR format). Additionally, we expanded the dataset by reintroducing the materials in space group 1, structures sourced directly from MP which underwent identical PHONOPY processing, and labeling all of them as TrIs.

At this stage, we established an initial dataset (named MAT) comprising 25895 materials, each with topological classifications derived from calculations that included spin–orbit

coupling. This collection includes 17813 TrI and 8082 NTMs, distributed as follows: 2292 HSLSMs, 2713 HSPSMs, 1238 TCIs, and 1839 TIs. Subsequently, the dataset M is constructed by removing those materials exhibiting conflicting labels between the database TMD and Materiae (as detailed in the following section). All material entries preserve their original MP-IDs from Materiae to ensure consistent cross-referencing capabilities.

5.2. Dataset T

To construct dataset T, we performed cleaning and augmentation using data from the database TMD, which was constructed by screening 96196 ICSD entries, yielding 73234 convergent entries linked to ICSD-IDs. These entries were derived from DFT calculations performed with VASP and were post-processed using the VASP2Trace[68] and Check Topological Mat[69] programs. The materials were categorized into five subclasses: TrIs, two types of TSMs (ES or ESFD) and two types of TIs (SEBR or NLC). Applying specific criteria, where compounds are considered identical if they share the same chemical formula, space group, and stable topology at E_{F}, they were grouped into 38298 groups, each group shares one unique-ID.

The original database TMD provides 73234 entries with details such as chemical compositions, ICSD-IDs, spacegroup numbers, MP-IDs, unique-IDs, and topological states (including topological type and indices, where applicable). The corresponding structures were retrieved based on ICSD-IDs, which were originally sourced from ICSD and post-processed with PHONOPY. We performed a comprehensive cleaning process on these data, which consisted of two main phases: first, establishing reliable MP-ID assignments, and second, ensuring compatibility and integration with the dataset MAT. It is important to note that the topological type classifications in these two databases differ in naming conventions. As a result, we established a mapping between them during the data curation process to ensure consistency.

During the initial phase, upon discovering inaccuracies in some of the pre-assigned MP-IDs, a critical issue for cross-referencing with the MAT dataset through MP-IDs, we implemented a systematic validation procedure using the *structure_matcher* of pymatgen[54] (with default tolerances) to verify and establish reliable MP-ID assignments. Our approach began by selecting at least one representative structure for each unique-ID that had a confirmed MP-ID, we compiled a comprehensive dataset indexed by 39166 unique ICSD-IDs. To enhance understanding of our cleaning process, Appendix A provides illustrative examples organized by unique-IDs. For compounds belonging to the same unique material group, we considered three scenarios to select the appropriate entry. First, for materials without any assigned MP-ID in 5351 unique-ID groups, we randomly retained one ICSD entry. Second, among the materials with a single associated MP-ID in 32392 unique-ID groups, we randomly preserved one corresponding ICSD entry and verified the MP-ID match, notable examples being materials with unique-IDs 2 or 32. Third, for materials associated with multiple MP-IDs in 555 unique-ID groups (such as those with unique-IDs 58 or 174), we conducted structural similarity assessments. Our structural analysis examined materials in 20 unique-ID groups (spanning 231 ICSD items corresponding to 44 MP-IDs), observing that structures tied to different MP-IDs were generally distinct. Based on these findings, our procedure for materials with the same unique-ID, but multiple MP-IDs involved: initially retaining one random ICSD entry per MP-ID; then ordering the related MP-IDs by ascending energy above hull (using MP data); and finally confirming each ICSD entry's association to one possible MP-ID with the lowest energy above hull. The process concluded by removing duplicate entries sharing identical MP-IDs. For MP-ID verification in both the single MP-ID and multiple MP-ID cases, we performed structural comparisons of ICSD-ID-associated structures against the structure from the MP with the *structure_matcher* with default settings. The comparison parameters included: a fractional length tolerance of 0.2, a site tolerance of 0.3 (calculated as $(V/N_{\mathrm{sites}})^{1/3}$) and an angle tolerance of 5°. For entries without a matching MP-ID, the value is marked as 'not available'. After successful confirmation of the new MP-IDs, we deduplicated the whole dataset to 38916 entries. All subsequent cleaning steps relied on these verified MP-IDs unless otherwise specified.

In the second phase, the data underwent cleaning according to Materiae's methodology and was linked to the MAT dataset using MP-IDs. After the entire cleaning process, 24156 entries remained from the database TMD, forming dataset T. Notably, TMD exhibits differences from Materiae in several key aspects: (1) TMD incorporates materials with magnetic moments exceeding 0.1 μ_{B} per unit cell (as documented in MP), even though calculations were conducted in a paramagnetic state. Given that the classification is done in paramagnetic state, we excluded potentially magnetic materials. (2) TMD categorizes NTMs into four distinct types: ES, ESFD, SEBR, and NLC, with conventional metals also labeled as ESFD. Meanwhile, although both TMD and Materiae employ VASP with its PBE-GGA functional, the parity of pseudopotential valency for certain elements varies, a critical discrepancy since this factor is used to identify conventional metals.

Based on these observations, we further refined the data to ensure consistency using the following approach. First, we removed 4275 materials containing rare-earth elements (Pr, Nd, Pm, Sm, Tb, Dy, Ho, Er, Tm, Yb, Lu, or Sc) due to pseudopotential discrepancies. We further detected and excluded

4729 conventional metals and 4597 magnetic materials, leaving 25315 entries. For the remaining materials, we aligned their classifications with Materiae's framework. We built the mapping between the two databases as follows: according to Ref. [27], ESFD features Bloch states degenerate at a high-symmetry point, corresponding to HSPSM. ES involves high-symmetry-point eigenstates violating compatibility relations, indicating connections to other states along high-symmetry lines or planes, corresponding to HSLSM. SEBR's stable topological bands can be expressed as integer-valued linear combinations of elementary band representations, whereas NLC does not. For them, no direct mapping existed. We labeled SEBR and NLC as either TI or TCI as follows: materials in space groups 174, 187, 189, 188, or 190 were all labeled as TCI, while others were labeled based on the parity of their last topological indices (odd for TI, even for TCI). During relabeling, we encountered conflicting topological types for entries sharing an MP-ID, due to minor structural variations in their ICSD records. (Recalculations using DFT and symmetry indicators on ICSD structures confirmed consistency with TMD's results.) After removing 672 such conflicting entries, we obtained dataset TM of 24643 materials, consisting of 16595 TrIs, 2229 HSLSMs, 2560 HSPSMs, 1282 TCIs, and 1977 TIs. While all TM records maintain ICSD-ID indexing, only a subset possesses corresponding MP-IDs.

The dataset TM was subsequently linked with dataset MAT using MP-IDs as the connection key. Within the overlapping portion of these two datasets, we identified 223 items (comprising 223 ICSD-IDs associated with 212 MP-IDs) that exhibited conflicting type labels, a discrepancy arising from subtle structural differences, and these were consequently removed. For dataset TM, a deduplication process was performed on 522 items (522 ICSD-IDs corresponding to 258 MP-IDs), resulting in their consolidation into 258 entries. This deduplication is the last step resulting in dataset T.

Datasets M and T have been successfully constructed from their respective source databases, Materiae and TMD. This comprehensive process culminates in the creation of our final combined dataset $M \cup T$, which incorporates a total of 35608 entries.

6. Conclusion

The final curated dataset $M \cup T$ comprises 35608 rigorously validated material samples, each annotated with at least one unique material identifier (MP-ID or ICSD-ID). Every entry includes a pymatgen structure object and is accompanied by its corresponding topological classification derived from DFT calculations and TQC analysis incorporating spin–orbit coupling effects. In addition, each entry is enriched with 4710 feature columns specifically engineered to support general machine learning algorithms. This comprehensive dataset integrates the strengths of existing databases while resolving inconsistencies through systematic curation, offering cross-validated topological classifications, uniform feature representations, and comprehensive annotations. It exhibits key characteristics that ensure scientific reusability and reliability, including a standardized data architecture for multi-source integration, a traceable dual-identifier system enabling cross-platform validation, and numerically encoded structure–property relationships suitable for algorithmic input. By providing a consistent, richly annotated and machine-learning-ready resource, this dataset establishes a robust foundation for accelerating the discovery of topological materials through reproducible and comparable machine learning studies.

The current dataset focuses on non-magnetic solid materials, with topological classifications based on TQC theory. This approach relies on crystal symmetry and band representations at high-symmetry points, providing a systematic and efficient method suitable for large-scale automated screening. However, it becomes inadequate when crystal symmetry is reduced, leading to the misdiagnosis of certain topological materials. As a complementary pathway, Wilson loop analysis directly traces the global evolution of electronic wavefunctions across the Brillouin zone, thereby circumventing the symmetry constraints inherent to TQC. For instance, the TopoMat[52] database includes materials identified through this method. On the other hand, recent extensions of symmetry-based classification theories have incorporated magnetic materials through wavefunction symmetry representations.[30–32] Although some data resources for these materials exist, they remain fragmented and disconnected from established datasets. Therefore, constructing a more extensive and integrated repository that incorporates Wilson loop results and the topological classifications of magnetic materials is essential for driving further progress in the field.

Acknowledgements

We extend our gratitude to N. Regnault for providing the data from the Topological Materials Database, which was instrumental in the data curation process. Y. H. and H. W. acknowledge financial support from the National Key Research and Development Program of China (Grant No. 2022YFA1403800), the National Natural Science Foundation of China (Grant Nos. 12188101 and 11925408), and the Chinese Academy of Sciences (Grant No. XDB33000000). H. W. also acknowledges support from the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

[1] Thouless D J, Kohmoto M, Nightingale M P and den Nijs M 1982 *Phys. Rev. Lett.* **49** 405
[2] Tsui D C, Stormer H L and Gossard A C 1982 *Phys. Rev. Lett.* **48** 1559

[3] Laughlin R B 1983 *Phys. Rev. Lett.* **50** 1395
[4] Kane C L and Mele E J 2005 *Phys. Rev. Lett.* **95** 146802
[5] Hasan M Z and Kane C L 2010 *Rev. Mod. Phys.* **82** 3045
[6] Qi X L and Zhang S C 2011 *Rev. Mod. Phys.* **83** 1057
[7] Bansil A, Lin H and Das T 2016 *Rev. Mod. Phys.* **88** 021004
[8] Kitaev A 2009 *Periodic table for topological insulators and superconductors, AIP conference proceedings* (American Institute of Physics) Vol. 1134 pp. 22–30
[9] Konig M, Wiedmann S, Brune C, Roth A, Buhmann H, Molenkamp L W, Qi X L and Zhang S C 2007 *Science* **318** 766
[10] Fu L, Kane C L and Mele E J 2007 *Phys. Rev. Lett.* **98** 106803
[11] Zhang H, Liu C X, Qi X L, Dai X, Fang Z and Zhang S C 2009 *Nat. Phys.* **5** 438
[12] Xia Y, Qian D, Hsieh D, Wray L, Pal A, Lin H, Bansil A, Grauer D, Hor Y S, Cava R J, *et al*. 2009 *Nat. Phys.* **5** 398
[13] Chen Y, Analytis J G, Chu J H, Liu Z, Mo S K, Qi X L, Zhang H, Lu D, Dai X, Fang Z, *et al*. 2009 *Science* **325** 178
[14] Benalcazar W A, Bernevig B A and Hughes T L 2017 *Science* **357** 61
[15] Wan X, Turner A M, Vishwanath A and Savrasov S Y 2011 *Phys. Rev. B* **83** 205101
[16] Wang Z, Sun Y, Chen X Q, Franchini C, Xu G, Weng H, Dai X and Fang Z 2012 *Phys. Rev. B* **85** 195320
[17] Young S M, Zaheer S, Teo J C, Kane C L, Mele E J and Rappe A M 2012 *Phys. Rev. Lett.* **108** 140405
[18] Wang Z, Weng H, Wu Q, Dai X and Fang Z 2013 *Phys. Rev. B* **88** 125427
[19] Liu Z, Jiang J, Zhou B, Wang Z, Zhang Y, Weng H, Prabhakaran D, Mo S K, Peng H, Dudin P, *et al*. 2014 *Nat. Mater.* **13** 677
[20] Lv B, Muff S, Qian T, Song Z, Nie S, Xu N, Richard P, Matt C E, Plumb N C, Zhao L, *et al*. 2015 *Phys. Rev. Lett.* **115** 217601
[21] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[22] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[23] Bradlyn B, Elcoro L, Cano J, Vergniory M G, Wang Z, Felser C, Aroyo M I and Bernevig B A 2017 *Nature* **547** 298
[24] Hellenbrandt M 2004 *Crystallogr. Rev.* **10** 17
[25] Jain A, Ong S, Hautier G, Chen W, Richards W, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K 2013 *APL Mater.* **1** 011002
[26] Zhang T, Jiang Y, Song Z, Huang H, He Y, Fang Z, Weng H and Fang C 2019 *Nature* **566** 475
[27] Vergniory M G, Elcoro L, Felser C, Regnault N, Bernevig B A and Wang Z 2019 *Nature* **566** 480
[28] Tang F, Po H C, Vishwanath A and Wan X 2019 *Nature* **566** 486
[29] Vergniory M G, Wieder B J, Elcoro L, Parkin S S, Felser C, Bernevig B A and Regnault N 2022 *Science* **376** eabg9094
[30] Watanabe H, Po H C and Vishwanath A 2018 *Science Advances* **4** aat8685
[31] Elcoro L, Wieder B J, Song Z, Xu Y, Bradlyn B and Bernevig B A 2021 *Nat. Commun.* **12** 5965
[32] Peng B, Jiang Y, Fang Z, Weng H and Fang C 2022 *Phys. Rev. B* **105** 235138
[33] Samuel A L 1959 *IBM J. Res. Dev.* **3** 210
[34] Zhang Y and Kim E A 2017 *Phys. Rev. Lett.* **118** 216401
[35] Zhang Y, Ginsparg P and Kim E A 2020 *Phys. Rev. Res.* **2** 023283
[36] Zhang P, Shen H and Zhai H 2018 *Phys. Rev. Lett.* **120** 066401
[37] Scheurer M S and Slager R J 2020 *Phys. Rev. Lett.* **124** 226401
[38] Donoho D 2006 *IEEE Transactions on Information Theory* **52** 1289
[39] Acosta C M, Ouyang R, Fazzio A, Scheffler M, Ghiringhelli L M and Carbogno C 2018 arXiv 1805.10950
[40] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 *Phys. Rev. Mater.* **2** 083802
[41] Cao G, Ouyang R, Ghiringhelli L M, Scheffler M, Liu H, Carbogno C and Zhang Z 2020 *Phys. Rev. Mater.* **4** 034204
[42] Liu J, Cao G, Zhou Z and Liu H 2021 *J. Phys. Condens. Matter* **33** 325501
[43] Claussen N, Bernevig B A and Regnault N 2020 *Phys. Rev. B* **101** 245117
[44] Marrazzo A, Gibertini M, Campi D, Mounet N and Marzari N 2019 *Nano Lett.* **19** 8431
[45] Haastrup S, Strange M, Pandey M, Deilmann T, Schmidt P S, Hinsche N F, Gjerding M N, Torelli D, Larsen P M, Riis-Jensen A C, *et al*. 2018 *2D Mater.* **5** 042002
[46] Zhou J, Shen L, Costa M D, Persson K A, Ong S P, Huck P, Lu Y, Ma X, Chen Y, Tang H, *et al*. 2019 *Scientific Data* **6** 86
[47] Olsen T, Andersen E, Okugawa T, Torelli D, Deilmann T and Thygesen K S 2019 *Phys. Rev. Mater.* **3** 024005
[48] Wang D, Tang F, Ji J, Zhang W, Vishwanath A, Po H C and Wan X 2019 *Phys. Rev. B* **100** 195108
[49] Schleder G R, Focassio B and Fazzio A 2021 *Appl. Phys. Rev.* **8** 031409
[50] Andrejevic N, Andrejevic J, Bernevig B A, Regnault N, Han F, Fabbris G, Nguyen T, Drucker N C, Rycroft C H and Li M 2022 *Adv. Mater.* **34** 2204113
[51] Ma A, Zhang Y, Christensen T, Po H C, Jing L, Fu L and Soljacic M 2023 *Nano Lett.* **23** 772
[52] Wu Q S, Autes G, Mounet N and Yazyev O V 2019 *TopoMat: A Database of High-Throughput First-Principles Calculations of Topological Materials*, Materials Cloud Archive 2019.0019/v1
[53] He Y, De Breuck P P, Weng H, Giantomassi M and Rignanese G M 2025 *npj Computational Materials* **11** 181
[54] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 *Comput. Mater. Sci.* **68** 314
[55] Ho T K 1995 *Random decision forests, Proceedings of 3rd International Conference on Document Analysis and Recognition* (Montreal, Canada, 14–16 August 1995) (IEEE) pp. 278–282
[56] McCulloch W S and Pitts W 1943 *The bulletin of mathematical biophysics* **5** 115
[57] Knobbe A J, Siebes A, van der Wallen D 1999 *Multi-relational Decision Tree Induction* In: Żytkow J M, Rauch J (eds) *Principles of Data Mining and Knowledge Discovery*. PKDD 1999 Lecture Notes in Computer Science, Vol. 1704 (Berlin: Springer)
[58] Barnard J and Meng X L 1999 *Statistical methods in medical research* **8** 17
[59] Ward L, Dunn A, Faghaninia A, Zimmermann N E, Bajaj S, Wang Q, Montoya J, Chen J, Bystrom K, Dylla M, Chard K, Asta M, Persson K A, Snyder G J, Foster I and Jain A 2018 *Comput. Mater. Sci.* **152** 60
[60] De Breuck P P, Hautier G and Rignanese G M 2021 *npj Comput. Mater.* **7** 83
[61] De Breuck P P 2025 *Vibrational properties of solids: a machine learning approach* (UCLouvain)
[62] Talirz L, Kumbhar S, Passaro E, Yakutovich A V, Granata V, Gargiulo F, Borelli M, Uhrin M, Huber S P, Zoupanos S, Adorf C S, Andersen C W, Schütt O, Pignedoli C A, Passerone D, VandeVondele J, Schulthess T C, Smit B, Pizzi G and Marzari N 2020 *Scientific Data* **7**
[63] Fraux G, Cersonsky R and Ceriotti M 2020 *Journal of Open Source Software* **5** 2117
[64] Ye C, Wang Y, Xie X, Zhu T, Liu J, He Y, Zhang L, Zhang J, Fang Z, Wang L, *et al.* 2025 *npj Computational Materials* **12** 63
[65] Kresse G and Furthmüller J 1996 *Computational Materials Science* **6** 15
[66] He Y, Jiang Y, Zhang T, Huang H, Fang C and Jin Z 2019 *Chin. Phys. B* **28** 087102
[67] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[68] Wang Z 2021 Vasp2trace https://github.com/zjwang11/irvsp
[69] Vergniory M Check topological mat https://www.cryst.ehu.es/cgi-bin/cryst/programs/magnetictopo.pl?tipog=gesp