

MatSciBench: A flexible benchmarking platform for machine learning in condensed matter science

Yuqing He(贺雨晴), Juncheng Xiao(肖俊成), Caiyuan Ye(叶财渊), Zhiyuan Gao(高致远), Mengshu Ge(葛梦舒), Ruijuan Xiao(肖睿娟), Weida Wang(王蔚达), Mark Meulemans, Shifeng Jin(金士峰), Quansheng Wu(吴泉生), Shufei Zhang(张舒飞), Yunqi Cai(蔡云麒), Hongming Weng(翁红明)
Citation: Chin. Phys. B. 35, 090702 (2026). DOI: 10.1088/1674-1056/ae88e3 | CSTR: 32038.14.CPB.ae88e3
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

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

Reviews of algorithm-driven terahertz metamaterials: From intelligent design to multidisciplinary applications

Wenyue Cao(曹文钺), Yuying Jiang(蒋玉英), Hongyi Ge(葛宏义), Juncheng Cao(曹俊诚)
Chin. Phys. B. 35, 080702 (2026). DOI: 10.1088/1674-1056/ae6b40 | CSTR: 32038.14.CPB.ae6b40

Multi-task deep-learning optimization of trade-off properties for superior-performance Fe-based soft magnetic alloys

Kang-Yuan Li(李康源), Mao-Zhi Li(李茂枝), Wei-Hua Wang(汪卫华)
Chin. Phys. B. 35, 070705 (2026). DOI: 10.1088/1674-1056/ae5b5f | CSTR: 32038.14.CPB.ae5b5f

Physics-informed neural network for material identification via distortion-robust polychromatic x-ray attenuation correction in photon-counting detectors

Xin Yan(闫欣), Jie Zhang(张杰), Kai He(何凯), Yiheng Liu(刘毅恒), Yuetong Zhao(赵悦彤), Gang Wang(王刚), Xinlong Chang(常新龙), Youwei Zhang(张有为)
Chin. Phys. B. 35, 050702 (2026). DOI: 10.1088/1674-1056/ae5a13 | CSTR: 32038.14.CPB.ae5a13

Curation and featurization of multiple topological materials databases

Yuqing He(贺雨晴), Matteo Giantomassi, Gian-Marco Rignanese, Hongming Weng(翁红明)
Chin. Phys. B. 35, 050701 (2026). DOI: 10.1088/1674-1056/ae2bf2 | CSTR: 32038.14.CPB.ae2bf2

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), Hongming Weng(翁红明)
Chin. Phys. B. 34, 120702 (2025). DOI: 10.1088/1674-1056/ae172a | CSTR: 32038.14.CPB.ae172a

---

COMPUTATIONAL PROGRAMS FOR PHYSICS

MatSciBench: A flexible benchmarking platform for machine learning in condensed matter science

Yuqing He(贺雨晴)[1,2], Juncheng Xiao(肖俊成)[1,2,3], Caiyuan Ye(叶财渊)[1,2,3], Zhiyuan Gao(高致远)[1,2,5], Mengshu Ge(葛梦舒)[1,2], Ruijuan Xiao(肖睿娟)[1,2], Weida Wang(王蔚达)[4,6], Mark Meulemans[1,2], Shifeng Jin(金士峰)[1,2], Quansheng Wu(吴泉生)[1,2], Shufei Zhang(张舒飞)[4], Yunqi Cai(蔡云麒)[1,2,†], and Hongming Weng(翁红明)[1,2,‡]

[1] *Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*
[2] *Condensed Matter Physics Data Center, Chinese Academy of Sciences, Beijing* 100190, *China*
[3] *University of Chinese Academy of Sciences, Beijing* 100190, *China*
[4] *Shanghai Artificial Intelligence Laboratory, Shanghai* 200232, *China*
[5] *University of Science and Technology of China, Hefei* 230026, *China*
[6] *Fudan University, Shanghai* 200433, *China*

(Received 14 April 2026; revised manuscript received 16 June 2026; accepted manuscript online 10 July 2026)

With the increasing application of deep learning and big data technologies in materials science, researchers face challenges such as fragmented private datasets, inconsistent evaluation standards, and difficulties in reproducing results. To address these issues, this paper presents MatSciBench: an open and extensible machine learning benchmark platform for materials science. The platform integrates multi-source data from materials databases such as the Materials Project and Materiae, as well as open experimental datasets, to provide standardized datasets. It also establishes a multi-dimensional evaluation system (assessing accuracy, validity, RMSE, etc.) covering multiple task types: property prediction, structure generation, object detection, and scientific reasoning. Through online model submission and leaderboard functionalities, MatSciBench enables an integrated workflow for data acquisition, model evaluation, and result sharing. Case studies on topological materials classification, crystal structure generation, and XRD structure resolution demonstrate the platform's effectiveness in model performance comparison, data distribution visualization, and result reproducibility. The open-access and iterative nature of MatSciBench provides a fair, transparent, and efficient evaluation environment for machine learning research in condensed matter physics and materials science.

Keywords: machine learning, benchmarking platform, computational materials science

PACS: 07.05.Kf, 29.85.Ca

DOI: 10.1088/1674-1056/ae88e3 **CSTR:** 32038.14.CPB.ae88e3

1. Introduction

Materials science research is entering a new data-driven era. The integration of machine learning (ML)[1] has emerged as a transformative force, driving the emergence of a new data-driven paradigm for materials research.[2] This paradigm is propelled by the vast amounts of structural and property data generated from diverse sources, including experimental measurements and high-throughput computations. It complements traditional theoretical and experimental approaches, as well as the third paradigm of high-throughput computational simulation exemplified by density functional theory (DFT),[3,4] by leveraging patterns within vast datasets to accelerate the discovery and design of novel materials. ML models have made significant progress in various tasks including predicting fundamental properties such as formation energy[5,6] and band gap,[7] determining crystal structures[8,9] and planning synthetic routes.[10] Pioneering studies have demonstrated this potential: by extracting knowledge from the enormous repositories of both computational and experimental data, ML promises to shorten the decades-long materials development cycle, offering not only rapid property predictions but also new physical insights derived from the underlying data patterns.

Despite this considerable promise, the rapid proliferation of ML applications in materials science has exposed a critical benchmarking gap, posing significant challenges to reproducibility, reliable model comparison, and the overall rigor of the field. This gap stems from several interconnected factors intrinsic to the discipline. The data underpinning these models are profoundly heterogeneous, originating from diverse sources such as DFT computations, experimental measurements, and scientific literature, which leads to inherent inconsistencies in format, scale, and quality that complicate aggregation and standardized processing. Compounding this issue is the vast diversity of ML tasks themselves, from regressing formation energies and classifying material types to generating novel crystal structures and image segmentation for

[†]Corresponding author. E-mail: caiyq@iphy.ac.cn
[‡]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.

microstructural analysis. Each task presents unique challenges and demands distinct and tailored evaluation protocols. Furthermore, the landscape of algorithms is continuously expanding, encompassing both general-purpose models and specialized architectures designed for materials data, whose performance can vary dramatically across different data regimes and tasks. In the absence of community-wide standards, this complexity culminates in inconsistent evaluation practices, where studies employ divergent data splits, preprocessing steps, and performance metrics, rendering direct comparisons between proposed algorithms unreliable and often misleading. Consequently, the field faces a growing reproducibility crisis, where published results are difficult to verify or build upon, ultimately impeding scientific progress and eroding trust in data-driven findings.

Recognizing these challenges, several initiatives have been established to develop standardized benchmarks. Foundational data repositories such as the Materials Project (MP)[11] have played a pivotal role by providing extensive datasets derived from DFT calculations. These resources have served as a foundation for numerous custom benchmarks, including the prediction of formation energies and band gaps across their databases. Building upon such infrastructures, more formalized benchmarking platforms have emerged. Among these, Matbench[12] stands out as one of the most comprehensive efforts focused on crystalline materials. It provides a rigorous k-fold cross-validation framework encompassing 13 supervised learning tasks, such as predicting dielectric constants, band gaps, formation energies, and steel yield strength, using publicly available DFT and experimental datasets to emulate diverse materials discovery scenarios. A key strength of Matbench lies in its standardized evaluation protocol, which mitigates data leakage through predefined train-test splits. However, this approach also constitutes a primary limitation: the retrospective evaluation framework results in training and test sets being drawn from identical underlying distributions, thereby failing to assess model generalization under out-of-distribution conditions. In the molecular domain, a parallel and influential effort is MoleculeNet,[13] which offers not only a curated collection of datasets with standardized data splits but also incorporates challenging scaffold-based splits to evaluate generalization to novel molecular structures. For more specialized applications, benchmarks such as the Open Catalyst 2020 (OC20)[14] have been introduced, focusing explicitly on catalyst modeling and advancing the application of machine learning for predicting energies and atomic forces in catalyst-adsorbate systems. This benchmark has been particularly instrumental in driving progress in equivariant graph neural networks. Another notable effort is JARVIS-Leaderboard,[15] which aims to provide a comprehensive, user-friendly, and community-extensible framework covering diverse modalities including artificial intelligence, electronic structure, force-field potentials, quantum computation, and experiments.

While each platform has made valuable contributions, they are collectively constrained by a common and pivotal limitation: their reliance on static and fixed dataset structures. While JARVIS-Leaderboard broadens the scope of tasks and integrates multi-modal data, each benchmark remains constrained by the same fundamental shortcoming: a design inherently centered on algorithm comparison. This prioritizes the evaluation of algorithmic performance while overlooking the dynamic, data-centric workflows that are increasingly essential in materials science. As a result, these frameworks fail to accommodate processes such as training set optimization, data augmentation, and iterative data curation. They do not support the simultaneous optimization of both algorithms and training data composition for a given task, nor do they facilitate the study of data-centric strategies like active learning within a benchmarking context. Consequently, they cannot adequately capture real-world research workflows in which the continuous curation of high-quality training data is a critical component of iterative model improvement.

To address the need for a more adaptive and task-oriented benchmark that accurately reflects the complete machine learning pipeline, where the curation of high-quality training data is as critical as algorithmic innovation, we introduce MatSciBench (https://cmpdc.iphy.ac.cn/benchmarks). This platform is a flexible and comprehensive benchmarking solution for machine learning in materials science. It offers a diverse collection of tasks spanning a wide range of material properties and data types, thereby facilitating a task-oriented analysis of model performance. By integrating multi-source data, standardizing evaluation protocols, and employing task-specific metrics, MatSciBench establishes a unified, fair, and reproducible evaluation environment aimed at fostering standardized development in materials science. The core contribution of our platform lies in its dynamic and holistic evaluation framework, which represents a significant departure from traditional static benchmarks. This design empowers users to actively augment, refine, and curate their training data while evaluating against a fixed and held-out test set. Consequently, it directly enables and assesses data-centric machine learning strategies within a rigorous benchmarking context.

2. Platform design and architecture

The design of MatSciBench is guided by a core philosophy: to move beyond static, algorithm-centric benchmarks towards a dynamic, task-oriented evaluation ecosystem that reflects the complete machine learning pipeline in materials science. To this end, MatSciBench is structured around a modular architecture that standardizes evaluation while providing

the flexibility to explore both model architecture and training data composition. This section details the system components of the platform.

2.1. Task modules

MatSciBench is organized into a suite of distinct task modules, each targeting an application scenario within condensed matter science. The initial release encompasses six core tasks, as summarized in Table 1, covering topological material classification, structure determination from powder x-ray diffraction (PXRD) data, crystal structure generation, automated selected area electron diffraction (SAED) center identification, bond-valence migration barrier prediction, and graduate-level condensed matter physics large language model evaluation, called CMPhysBench. This diversity ensures that the platform assesses a model's ability to handle various data modalities, including crystal structures, spectroscopic data and images, and different learning paradigms, from regression and classification to generative modeling. The modular design of these tasks is a foundational feature of MatSciBench. Each task is independent, with standardized data and evaluation scripts. To ensure reliability across heterogeneous data sources, each task module employs a tailored data curation pipeline; all data within a given task are processed to a unified format and evaluated against a consistent prediction target, with ground-truth labels drawn from an internally coherent source. This design allows for the straightforward integration of new tasks and datasets from the community, and ensures the benchmark remains at the forefront of the field's evolving challenges.

Table 1. Summary of benchmark tasks in MatSciBench, covering diverse data modalities, prediction targets and evaluation metrics.

Task name	Data modality	Primary data source (s)	Dataset size	Prediction target	Evaluation metric
Topological materials classification	Crystal structure	Materiae,[16] topological materials database,[17,18]	$\sim$ 35k	Topological type of the material	Accuracy, F1-score, Precision, Recall, ROC AUC
XRD structure parsing	XRD pattern, crystal structure	MP[11]	$\sim$ 152k	Lattice parameters, atomic positions	Match rate, RMSE
Crystal structure generation	Crystal structure	MP,[8,11] OQMD[7]	$\sim$ 178k	Stable crystal structure	Validity, diversity, energy above hull
SAED center identification	SAED pattern	Experimental data	$\sim$ 0.2k	Center coordinates (x, y)	RMSE, center error failure rate
Bond-valence barrier prediction	Crystal structure	BVpath ionic transport database[19]	$\sim$ 72k	Migration barrier (eV)	Mean absolute error, RMSE
Large language model evaluation	Text	17 graduate textbooks[20]	520	Answers to the problem	Scalable expression edit distance, accuracy

2.2. Standardized benchmark layout: Leaderboard and dataset explorer

As illustrated in the architecture diagram in Fig. 1, every task module in MatSciBench follows a consistent and standardized structure composed of three primary components: the task introduction, the leaderboard module, and the dataset explorer module. This uniform structure lowers the barrier to entry, allowing researchers to quickly engage with any benchmark.

Upon selecting a task, users are first presented with a concise introduction that outlines the task's scientific motivation, defines the input and output formats, and specifies the primary evaluation metrics. This section is designed to enable researchers to rapidly ascertain the task's relevance to their own work and to understand the precise problem formulation.

The centerpiece of each task is the leaderboard module, which provides a comparative and dynamic ranking of submitted models to facilitate assessment and selection. The leaderboard is built upon a fixed, held-out test set that is carefully curated to be representative of real-world application scenarios and to present a meaningful challenge, thereby ensuring all model performances are directly comparable. In contrast, the training set is flexible, as users are encouraged to augment, curate, or optimize it. This key innovation enables the study of data-centric ML strategies, such as active learning and data augmentation, which are critical to real research workflows.

Beyond this foundational framework for evaluation, the platform is further designed to deliver deep and actionable insights. It goes beyond a single performance score by providing multi-dimensional comparative visualizations, such as bar charts and line graphs, that break down model performance across different data subgroups like crystal systems or element compositions. This allows users to identify model strengths and weaknesses and select the most suitable model for their specific sub-task. The platform enhances accessibility and practical utility by providing pre-trained models through its web interface for rapid material predictions, enabling researchers to quickly test models on new materials of interest without requiring local high-performance computing resources. For models available for download, the platform also provides details about the software environment needed to run them locally. Concurrently, to ensure transparency and reproducibility, MatSciBench offers open-source evaluation scripts for local model validation. Additionally, users may

submit their verified predictions to our leaderboard module, where they will be displayed publicly after passing our verification process; the full submission workflow, verification protocol, and data-centric training policies are detailed in Subsection 2.3. Each entry on the leaderboard is enriched with metadata, including a link to the exact training data used, model architecture details, and submission date, while an integrated commenting functionality promotes the formation of a community where researchers can discuss results, share insights, and collaborate to improve model performance.

Beneath the leaderboard lies the Dataset Explorer module, a powerful tool for understanding the data that underpins the models. This module provides a suite of interactive visualizations, such as pie charts, bar graphs, and periodic table heatmaps; they characterize the composition and properties of the test set and training sets. Users can visualize distributions of the dataset under different key attributes, for example, crystal systems, space groups, and elemental diversities. This functionality is crucial for diagnosing dataset biases, understanding the scope of a model's capabilities, and assessing its potential for out-of-distribution (OOD) generalization.

This structured workflow enables researchers to benchmark new models against existing ones under consistent conditions, with the diagnostic visuals aiding in the interpretation of performance gaps.

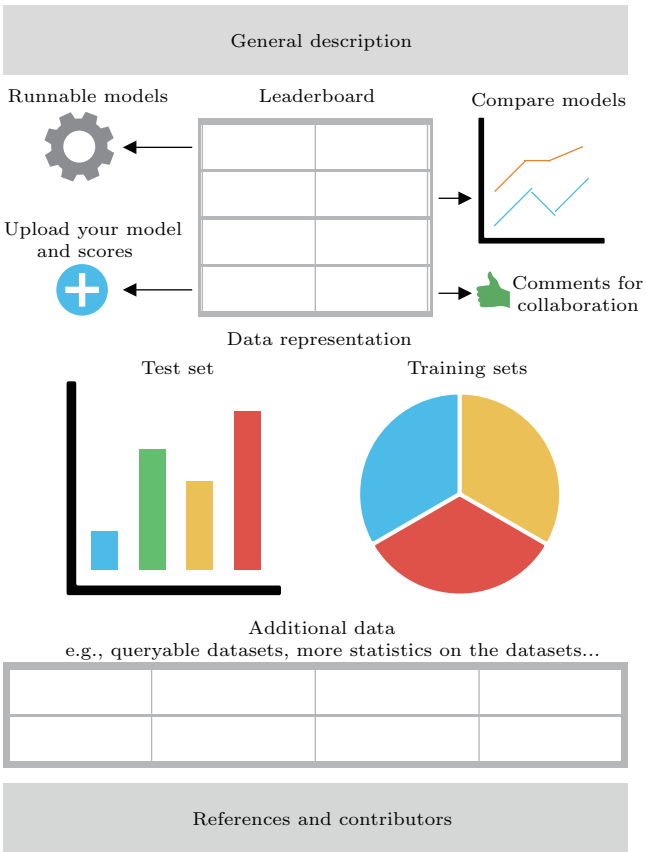


Fig. 1. The architectural schematic diagram of the benchmark page.

2.3. Task contribution and model submission

MatSciBench is designed to be inherently extensible, enabling the community to contribute new task modules and submit model results through standardized protocols. This section describes the model submission workflow, the data-centric paradigm that governs it, and the mechanisms that ensure reproducibility and trustworthiness of results.

Each task page provides a submission interface located directly beneath the leaderboard. Submission requirements are task-specific and documented on the respective page, specifying the required prediction file format and how model scores are extracted. A standard submission package comprises prediction results on the sequestered test set, model metadata (architecture and hyperparameters), and documentation of the training data and any curation steps applied. Upon upload, users affirm the truthfulness of their submission. The platform then performs random verification checks, which may include re-running the submitted model to verify reported scores, retraining with the user-provided training parameters, and screening for data leakage between the claimed training data and the test set. These measures ensure that leaderboard rankings remain trustworthy and results reproducible.

MatSciBench embeds a data-centric philosophy directly into its submission protocol. In conventional benchmarks such as Matbench and JARVIS-Leaderboard, all data splits are fixed at release. While this supports clean algorithmic comparisons, it can conflate model limitations with data quality issues and restrict the search for better performance to architectural innovations alone. MatSciBench addresses this by fixing only the test set, thereby treating training data curation as an integral part of the submission workflow rather than a predetermined constraint.

Under this protocol, users may start from any training dataset of their choice, whether the baseline provided on the platform, their own data, or a hybrid, provided that it contains no samples from the test set. They may then apply data-centric strategies locally (data augmentation, label correction, class rebalancing, etc.), train a model on the curated data, and generate predictions on the fixed test set. The curated training set and resulting predictions are submitted together through the verification pipeline described above. This design is demonstrated in the topological materials classification case study (Subsection 3.1), where the same XGBoost model trained on a baseline and a refined dataset exhibited distinctly different performance profiles against the same test set, directly revealing the impact of data curation on model behavior.

To ensure that historical results remain reproducible as the platform evolves, MatSciBench retains all datasets, models, and evaluation scripts together with their associated metadata. Every dataset and leaderboard entry carries a timestamp, enabling clear distinction between versions. When datasets

or evaluation scripts are updated, prior versions remain accessible, and all historical leaderboard entries retain references to the exact data and script versions against which they were evaluated. In addition to timestamps, version distinction can be based on references to scientific publications, ensuring traceability to the original data sources and methodologies. This versioning scheme preserves the integrity of past benchmarking results while allowing the platform to incorporate improved datasets and refined evaluation protocols over time.

Researchers interested in adding a new benchmark task are invited to contact the platform team to discuss requirements and coordinate integration. A new task submission requires the following components: a curated dataset with a clearly defined test set; evaluation scripts capable of scoring model predictions according to the task's metrics; at least one baseline leaderboard entry to initialize the benchmark; a task introduction including title, scientific motivation, input/output format specifications, and evaluation protocol descriptions; and relevant explanatory graphics such as data distribution visualizations. This modular structure mirrors the layout of the six existing tasks (Table 1), ensuring that all contributed tasks follow a consistent and immediately usable format.

3. Case studies: Platform workflow in practice

To demonstrate how MatSciBench functions in practical research scenarios, we present three case studies covering distinct computational tasks in materials science. These examples illustrate the platform's standardized workflow from data exploration to model evaluation, highlighting how the platform facilitates direct comparison and reproducible analysis. Each case study follows a consistent pathway: researchers begin by exploring the dataset characteristics, proceed to evaluate their models using provided scripts, and finally analyze their results through the platform's comparative visualization tools.

3.1. Topological materials classification

The identification of topological materials[21–23] is a crucial frontier in quantum materials research.[16,17,24–27] To this end, MatSciBench has established a benchmark system for predicting the topological classification of crystalline materials. We utilize a curated dataset[27] that integrates data from the Materiae[16] database and the Topological Materials Database.[17,18] This dataset includes 35608 materials, with all classifications consistently derived from DFT[3,4] calculations and topological quantum chemistry (TQC) theory.[28] The construction of this dataset involved a rigorous multi-stage curation pipeline — including symmetry standardization, removal of magnetic materials and conventional metals, MP-ID cross-validation, and resolution of inter-database labeling conflicts — as detailed in Ref. [27]. All entries have thus been processed to ensure consistent labeling and cross-platform comparability. The dataset encompasses trivial insulators (TrIs) and several major topological phases, such as topological insulators (TIs), topological crystalline insulators (TCIs), high-symmetry-point semimetals (HSPSMs) and high-symmetry-line semimetals (HSLSMs). The proportion of TrIs to non-trivial topological materials (NTMs) in this dataset is similar to that in most high-throughput screening studies,[16,17,24] a distribution we consider to be a reasonable approximation of their natural occurrence. Accordingly, we employed stratified sampling at a 1:10 ratio to construct a test set that maintains a distribution consistent with the original dataset across key dimensions, including crystal system, number of atoms and material type.

The leaderboard for this task is shown in Fig. 2(a), presenting results from algorithms including Random Forest[29] and XGBoost[30] trained on two different datasets, T_20250501 and T_20250520. The dataset T_20250501 constitutes the remainder of the original dataset after removing the test set and thus follows the same distribution. The XGBoost model trained on this dataset achieved the highest overall accuracy (84.02%) and F1-score (83.17%) among all submitted models. The refined dataset T_20250520 was created by removing potentially mislabeled trivial insulators of specific space groups from T_20250501. This modification was motivated by limitations in the theoretical framework of TQC, where symmetry-based diagnostic methods can fail to correctly identify certain TIs and TCIs, leading to potential mislabeling in the original database. Figure 3 shows the data explorer module, which compares the relative distribution of the test set and the training set T_20250520. When the XGBoost model was retrained on T_20250520, the overall accuracy decreased to 83.04% and the F1-score saw a slight reduction to 82.40%, suggesting that the reduction in training volume had a measurable impact on aggregate performance. However, a detailed analysis of the confusion matrices in Fig. 2(b) reveals that the refined dataset improved the model's discriminative capability for certain NTMs. Specifically, while the accuracy for predicting TrIs and TCIs declined, the recognition rates for the other three categories HSLSMs, HSPSMs and TIs improved. This trade-off arises from the nature of the refinement process. In T_20250501, the potentially mislabeled trivial insulators acted as "easy negatives" for the classifier, allowing the model to achieve high overall accuracy. However, the decision boundaries learned by the model could become overly sensitive to symmetry-related features rather than learning more meaningful physical features. After removing these samples in T_20250520, the training set becomes smaller and the relative proportion of NTMs increases, forcing the model to learn more discriminative features. Fewer TrI examples remain which slightly hurts the model's ability to identify true

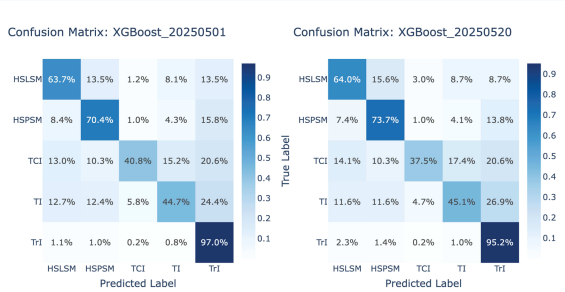
TrIs, but favors the detection of genuine topological phases. This explains why accuracy for TrIs and TCIs declined while recognition for other NTMs improved.

The crystal system-wise accuracy trends in Fig. 2(c) corroborate this interpretation: the model trained on T_20250520 performs better in cubic and hexagonal crystal systems, precisely where NTMs are enriched (as shown in Fig. 3). This demonstrates that the dataset refinement successfully mitigated labeling noise and enhanced the model's capability to identify genuine NTMs, despite the slight drop in aggregate metrics. Importantly, the fact that refinement improved NTM discrimination despite a smaller training set suggests that the original labels in T_20250501 contain systematic, non-random noise — concentrated specifically in TrIs of certain space groups where symmetry-based methods are theoretically known to be unclear. Therefore, dataset optimization based on domain knowledge can effectively enhance model performance on target categories, despite potential trade-offs in aggregate metrics. These analytical comparisons visually highlight performance disparities across different topological phases and crystal systems. Based on these insights, users can select the most appropriate algorithm according to their actual task requirements. Specifically, users targeting new HSLSMs or HSPSMs may prefer models trained on T_20250520. In addition, users can download the standardized dataset to reproduce and verify the listed results, as well as to make further improvements. Through the data explorer module, users can quickly understand data structures and distributions through interactive visualizations like stacked bar charts and heatmaps, enabling them to assess data representativeness and identify potential biases.

(a)

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)	ROC AUC (%)	Training set	Comments	Date
Random Forest [1]	82.34	81.04	80.72	82.34	75.98	T_20250501	0	2025-05-06
XGBoost_20250501 [2]	84.02	83.17	83.01	84.02	78.76	T_20250501	0	2025-05-06
XGBoost_20250520 [2]	83.04	82.40	82.37	83.04	78.63	T_20250520	0	2025-05-20

(b)



(c)

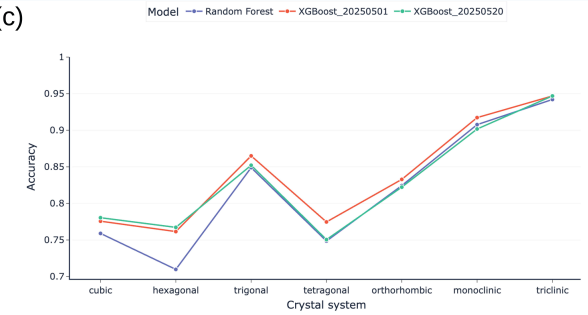


Fig. 2. The benchmark results for topological materials classification models. (a) Overall comparison of three models across accuracy, F1 score, precision, recall and ROC AUC score. (b) Comparison of the confusion matrix for two XGBoost models trained on dataset T_20250501 and T_20250520. (c) Accuracy as a function of the crystal system.



Fig. 3. A screenshot of the data explorer module for topological materials classification task on the MatSciBench website. It compares the relative distribution of the test set and the training set T_20250520.

3.2. XRD structure resolution

The crystal structure forms the fundamental basis for understanding the intrinsic properties and functional behavior of materials and directly determines their physical, chemical, and mechanical performance. Consequently, accurate structure determination has become a core task across multiple disciplines, including materials science, chemistry, physics, and geology. In practical research and industrial applications, many materials are only available in polycrystalline or powdered form, which makes powder x-ray diffraction (PXRD) an indispensable tool for structural characterization.[31,32] To advance machine learning methods for resolving crystal structures from PXRD patterns, we established a dedicated benchmark on the MatSciBench platform and constructed a curated dataset comprising 152587 high-quality "crystal structure-XRD pattern" pairs through rigorous data processing and quality control.

The structural data used in this task are mainly drawn from the MP[11] and several open experimental diffraction databases.[31] All crystal structures are processed into both conventional and primitive cell representations and further reduced to Niggli cells, so that symmetry information is standardized and ambiguities arising from duplicated or symmetry-equivalent representations are removed. We then compute PXRD patterns under a unified configuration of radiation source, 2θ scan range and angular step size, ensuring cross-comparability among samples originating from different databases and boundary conditions. Finally, all necessary crystallographic and computational parameters are completed, yielding a fully reproducible dataset that can be directly used for visualization, sampling and model training.

To gain a comprehensive understanding of the dataset composition and distribution, we perform systematic statistical analysis from multiple perspectives. As a macroscopic manifestation of crystal symmetry, the crystal system strongly influences material properties at the macroscopic scale by constraining anisotropic physical behavior. As shown in Fig. 4, the distribution of the seven crystal systems in our dataset indicates that monoclinic structures are the most abundant, with 35646 samples accounting for 23.4% of the total, whereas cubic structures are intrinsically less frequent due to their high symmetry, with 20285 samples (13.3%). This distribution is broadly consistent with the natural abundance of crystal systems in inorganic materials, providing a well-balanced and realistic sample basis for structure-related prediction tasks.

To ensure comprehensive evaluation and a fair train-test split, we adopt a partitioning strategy that combines crystal-system balancing with complexity-based filtering. Specifically, we first perform equal sampling from each of the seven crystal systems to construct an initial test set containing approximately one-tenth of all samples. We then move all samples with more than seven distinct atomic species to the test set in order to strengthen its coverage of materials with complex structures. In the final partition, the training set contains 137347 samples and the test set 15240 samples. This strategy preserves balance across crystal systems while substantially increasing the fraction of high-complexity structures in the test set, enabling a systematic assessment of model generalization across different structural complexities and providing a more challenging benchmark for future ML models.

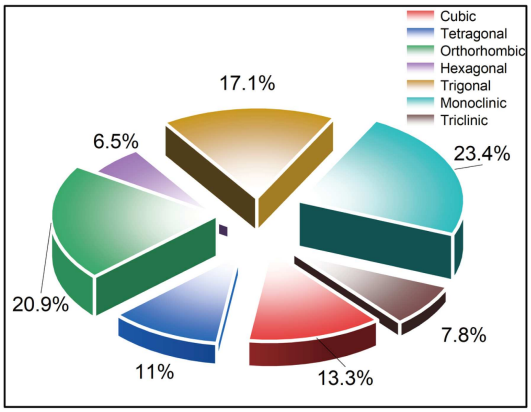


Fig. 4. Distribution of the seven crystal systems in the PXRD dataset. Monoclinic structures are the most frequent (23.4%), followed by orthorhombic (20.9%), reflecting a realistic distribution of inorganic materials.

To support intuitive data exploration, the interactive interface provides five complementary visualization views: a pie chart of the crystal-system distribution, a pie chart of the number of distinct atomic species per structure, a binned pie chart of the total atom count per structure, a periodic-table heatmap showing the occurrence frequency of each element, and a statistical summary table of the space-group distribution.

To objectively compare different models on the PXRD structure resolution task, MatSciBench provides a unified evaluation framework that integrates representative structural reconstruction models such as XtalNet[33] and PXRDGen.[34] We adopt the match rate and root mean squared error (RMSE) as the primary evaluation metrics. The leaderboard results on the retained test set (summarized in Fig. 5(a)) reveal a clear performance gap among different models. Overall, the PXRDGen models achieve the strongest performance: both the diffusion-based (PXRDGen-Diffusion) and flow-based (PXRDGen-Flow) variants reach a match rate of 0.74. PXRDGen-Diffusion further achieves the best RMSE of 0.08, while PXRDGen-Flow obtains an RMSE of 0.11. In comparison, XtalNet reaches a match rate of 0.56 with an RMSE of 0.15. Other baseline models perform substantially worse, including Uni-3DAR (match rate 0.34, RMSE 0.05), Uni-3DAR finetuned (0.46, 0.09),[35] Crystalyze (0.35, 0.06)[36] and PXRDNet (0.28, 0.18).[37] These results indicate that current diffusion- and flow-based generative frameworks remain the most effective solutions for the highly nonlinear inverse problem posed by PXRD-based structure determination.

This comparison not only provides a clear ranking of model performance, but also highlights several systematic trends from the finer-grained analysis shown in Fig. 5(b). First, model performance consistently degrades as the number of atoms in the structure increases, with all methods showing significantly lower match rates for structures containing more than 52 atoms. Second, crystal symmetry has a strong influence on prediction quality: most models perform substantially better on highly symmetric cubic and tetragonal systems, while triclinic and monoclinic structures remain much more challenging. Third, increasing chemical complexity also makes the task harder, as evidenced by the rapid decline in match rate with increasing element count. These observations suggest that large-scale, low-symmetry and chemically complex structures remain the major bottlenecks for current PXRD structure prediction models. Through systematic data construction, multidimensional statistics and unified evaluation, the PXRD structure resolution task in MatSciBench offers not only a fair and transparent platform for algorithm comparison, but also fine-grained diagnostics of model strengths and weaknesses across different regions of materials science.

(a)

Model	Match rate	RMSE	Training set	Comments	Date	Publisher
pxrdgen_flow	0.74	0.11	XRDGen	0	2025-03-29	肖俊成, CMPDC; IOP CAS
pxrdgen_diffusion	0.74	0.08	XRDGen	0	2025-07-01	肖俊成, CMPDC; IOP CAS
xtalnet	0.56	0.15	XRDGen	0	2025-07-08	肖俊成, CMPDC; IOP CAS
PXRDNet (cdvae_xrd)	0.28	0.18	XRDGen	0	2026-03-09	高致远, CMPDC; IOP CAS
Crystalyze	0.35	0.06	XRDGen	0	2026-03-09	高致远, CMPDC; IOP CAS
Uni-3DAR	0.34	0.05	XRDGen	0	2026-03-09	高致远, CMPDC; IOP CAS
Uni-3DAR (finetuned)	0.46	0.09	XRDGen	0	2026-03-09	高致远, CMPDC; IOP CAS

Add records of testing on this dataset for a new model

(b)

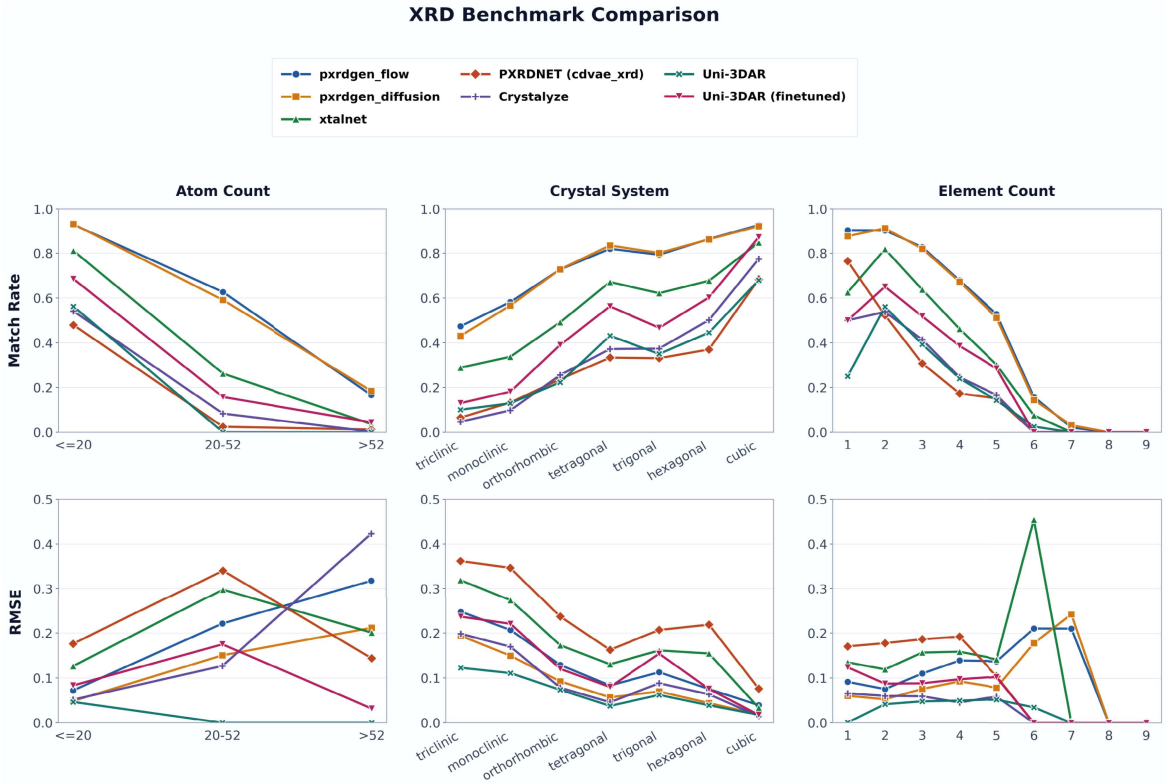


Fig. 5. Benchmark performance of PXRD crystal structure identification models on the retained test set. (a) Overall leaderboard summarizing match rate and RMSE for all evaluated models. (b) Match rate and RMSE as functions of structural complexity, analyzed by atom count, crystal system and element count.

3.3. Crystal structure generation

At present, a wide variety of generative models have been applied to crystal structure generation,[8,38–46] and AI-based approaches are regarded as a new opportunity to enhance the efficiency of discovering novel materials. In MatSciBench, we design benchmarks focusing primarily on two types of generative tasks. The first category is general crystal structure generation, where no additional information is provided to the model and structures are generated purely at random. The second category is conditional generation, in which target physical properties are supplied as inputs to the model, with the goal of guiding the generator to produce crystal structures that satisfy the desired requirements.

For general (unconditional) generation tasks, we typically expect a model to produce crystal structures that are valid, stable, and diverse. To comprehensively assess model performance, we employ several evaluation metrics, including structure validity, composition validity, match rate, normalized RMSE, uniqueness, and novelty.

A generated structure is considered structurally valid if the interatomic distance between any pair of atoms exceeds 0.5 Å. Composition validity is determined using Semiconducting Materials from Analogy and Chemical Theory (SMACT),[47] which evaluates whether the elemental valences and stoichiometric ratios are chemically reasonable. The match rate checks whether a generated structure remains similar to its relaxed counterpart. The normalized RMSE measures the magnitude of structural changes before and after relaxation. Uniqueness quantifies whether each generated crystal appears only once within the model's own generated dataset, while novelty measures whether the generated crystals do not duplicate any structures in the reference dataset. In this workflow, we replace first-principles relaxation with the OpenLAM[48] universal machine-learning potential. Although OpenLAM produces relaxed structures that closely resemble those obtained from ab initio calculations, it reduces the computational cost by approximately three orders of magnitude.

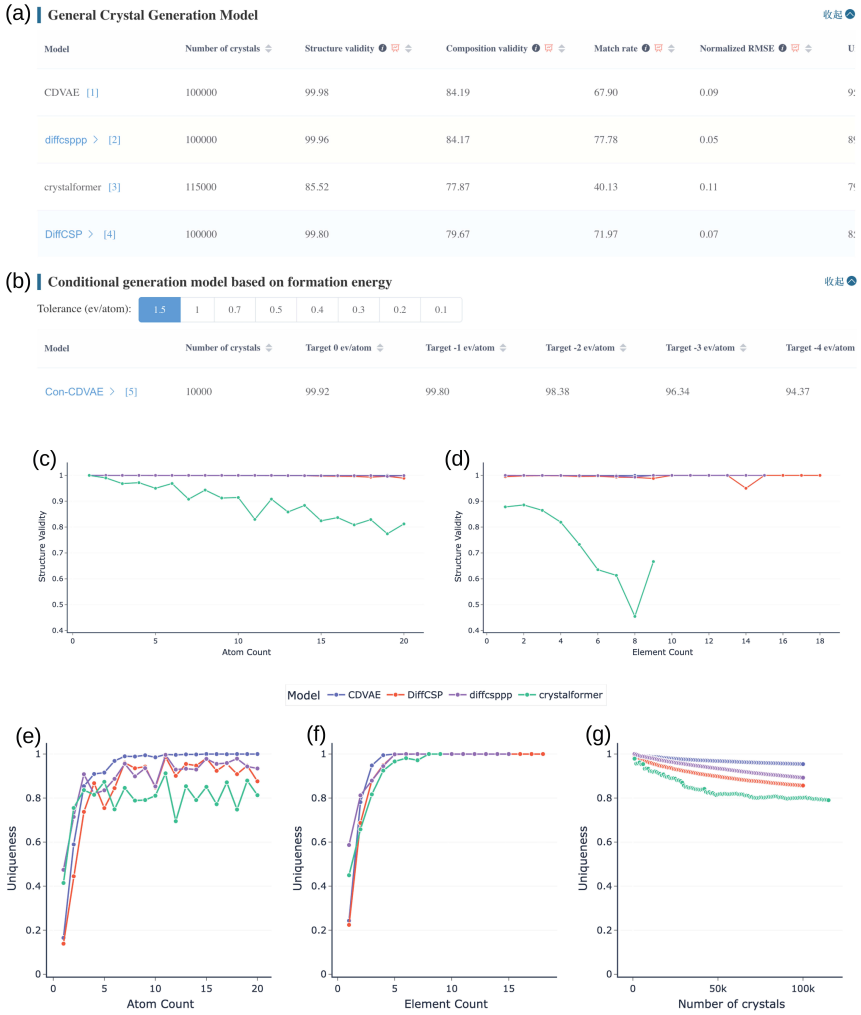


Fig. 6. Benchmarking results for general and conditional crystal structure generation. (a) Overall comparison of four generative models across structure validity, composition validity, match rate, normalized RMSE, uniqueness and novelty. (b) Success rate of the Con-CDVAE model under conditional generation with target formation energies. (c)–(d) Structural validity as a function of the number of atoms in the primitive cell and the number of distinct elements. (e)–(g) Uniqueness of generated crystals with respect to the number of atoms, number of elements and the total number of generated samples.

Table 2. Performance comparison of four generative models on the general crystal structure generation task across multiple evaluation metrics.

Model	Number of crystals	Structure validity	Composition validity	Match rate	Normalized RMSE	Uniqueness	Novelty in MP_0220	Novelty in training set
CDVAE	100000	99.98%	84.19%	67.90%	0.09	95.43%	95.59%	97.94%
DiffCSP++	100000	99.96%	84.17%	77.78%	0.05	89.25%	94.78%	96.99%
CrystalFormer	115000	85.52%	77.87%	40.13%	0.11	79.04%	86.44%	90.96%
DiffCSP	100000	99.80%	79.67%	71.97%	0.07	85.69%	85.10%	91.09%

In addition, we employ the pymatgen[49] StructureMatcher to assess structural similarity and quantify differences between crystal structures. For uniqueness and novelty evaluation, we adopt the default StructureMatcher settings (ltol = 0.2, stol = 0.3, angle_tol = 5). When comparing pre- and post-relaxation structures, we use more permissive thresholds (ltol = 0.3, stol = 0.5, angle_tol = 10) to increase the number of matched pairs and enable broader statistical analysis.

We first evaluate several representative generative models, including CDVAE,[8] DiffCSP++,[41] CrystalFormer[42] and DiffCSP.[40] Among them, CrystalFormer is an autoregressive model, whereas the other three are diffusion-based models. For each model, we generate 100000 samples, except for CrystalFormer, which produces 115000 samples. This is because CrystalFormer requires specifying a space group before generation; therefore, we generate 500 samples for each of the 230 space groups. DiffCSP++ also requires predefined space groups and Wyckoff positions. To satisfy these constraints during sampling, we randomly draw valid combinations of space groups and Wyckoff positions from the training set. Similarly, DiffCSP requires specifying the number of atoms in the primitive cell, which we also determine through random sampling from the training data. In contrast, CDVAE does not require any predefined structural information for generation. The evaluation results are summarized in Table 2 and Fig. 6(a). Figures 6(c) and 6(d) show how the structural validity of the four models varies with respect to the number of atoms in the primitive cell and the number of distinct elements. Figures 6(e)–6(g) illustrate how the uniqueness of the generated structures changes as a function of the number of atoms per primitive cell, the number of elements and the total number of generated samples.

For the conditional generation task, we use formation energy as the target property. The formation energies of the newly generated crystals are also evaluated using the OpenLAM machine-learning potential. Figure 6(b) shows the success rate of the Con-CDVAE[43] model when conditioned on formation energies of 0, 1, 2, 3, and 4 eV/atom, under a tolerance of 1.5 eV/atom. On the interactive webpage, users can adjust the tolerance and examine how the success rate varies accordingly.

4. Conclusion

In this work, we have introduced MatSciBench, an integrated benchmarking platform designed to establish standardized, reproducible, and transparent evaluation protocols for machine learning in materials science. By integrating multi-source data from several authoritative databases and incorporating representative tasks such as property prediction, structure generation, object detection, and scientific reasoning, the platform enables consistent and fair performance comparisons across a wide range of algorithms.

Through case studies in topological phase classification, XRD-based structure determination, and crystal structure generation, we have demonstrated MatSciBench's capacity to deliver nuanced insights into model behavior, facilitate dataset refinement, and support model selection under realistic scenarios. Key design features — including scripted evaluation interfaces, fixed test sets with flexible training data, and interactive visualization tools — collectively foster reproducibility and help users diagnose model strengths and limitations across different data regimes.

Looking forward, MatSciBench is developed as an extensible and community-oriented platform. Its modular design supports the seamless incorporation of new tasks and datasets, while its open-source evaluation framework encourages collaborative improvement and methodological transparency. To strengthen the platform's connection with real-world applications, we are initiating collaborations with experimental research groups to co-develop benchmark tasks that directly mirror laboratory workflows, such as predicting the synthesizability of theoretically proposed materials. To foster community engagement, we are organizing a MatSciBench Challenge, inviting researchers to benchmark state-of-the-art models, contribute new tasks, and collectively advance the frontier of ML for materials science. By bridging data curation, model evaluation, and performance analysis within a unified environment, MatSciBench aims to accelerate the development of robust, data-driven ML approaches in condensed matter and materials science. We envision the platform serving as a foundational tool for fostering reproducibility, enabling benchmarking at scale, and ultimately contributing to the maturation of ML-based research workflows in the field.

Acknowledgements

This project acknowledges financial support from the Beijing Natural Science Foundation (Grant No. Z250005), the National Key Research and Development Program of China (Grant No. 2022YFA1403800), the National Natural Science Foundation of China (Grants 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. This work was additionally supported by the Shanghai Municipal Science and Technology Major Project (Grant No. 2025SHZDZX025D06). Computational resources were provided by the Condensed Matter Physics Data Center, Chinese Academy of Sciences.

References

[1] Samuel A L 1959 *IBM J. Res. Dev.* **3** 210
[2] Agrawal A and Choudhary A 2016 *APL Mater.* **4** 053208
[3] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[4] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[5] Jha D, Ward L, Paul A, Liao W K, Choudhary A, Wolverton C and Agrawal A 2018 *Sci. Rep.* **8** 17593
[6] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[7] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Comput. Mater.* 1
[8] Xie T, Fu X, Ganea O E, Barzilay R and Jaakkola T 2021 arXiv:2110.06197
[9] Dong R, Zhao Y, Song Y, Fu N, Omee S S, Dey S, Li Q, Wei L and Hu J 2022 *ACS Appl. Mater. Interfaces* **14** 40102
[10] Segler M H, Preuss M and Waller M P 2018 *Nature* **555** 604
[11] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 *APL Mater.* **1** 011002
[12] Dunn A, Wang Q, Ganose A, Dopp D and Jain A 2020 *npj Comput. Mater.* **6** 159
[13] Wu Z, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K and Pande V 2018 *Chem. Sci.* **9** 513
[14] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, Tran K, Heras-Domingo J, Ho C, Hu W, *et al.* 2021 *ACS Catal.* **11** 6059
[15] Choudhary K, Wines D, Li K, Garrity K F, Gupta V, Romero A H, Krogel J T, Saritas K, Fuhr A, Ganesh P, *et al.* 2024 *npj Comput. Mater.* **10** 93
[16] Zhang T, Jiang Y, Song Z, Huang H, He Y, Fang Z, Weng H and Fang C 2019 *Nature* **566** 475
[17] Vergniory M G, Elcoro L, Felser C, Regnault N, Bernevig B A and Wang Z 2019 *Nature* **566** 480
[18] Vergniory M G, Wieder B J, Elcoro L, Parkin S S P, Felser C, Bernevig B A and Regnault N 2022 *Science* **376** eabg9094
[19] Xiao R, Li H and Chen L 2015 *Sci. Rep.* **5** 14227
[20] Wang W, Huang D, Li J, Yang T, Zheng Z, Zhang D, Han D, Chen B, Luo B, Liu Z, *et al.* 2025 arXiv:2508.18124
[21] Kitaev A 2009 *AIP Conf. Proc.* **1134** 22
[22] Hasan M Z and Kane C L 2010 *Rev. Mod. Phys.* **82** 3045
[23] Qi X L and Zhang S C 2011 *Rev. Mod. Phys.* **83** 1057
[24] Tang F, Po H C, Vishwanath A and Wan X 2019 *Nature* **566** 486
[25] Claussen N, Bernevig B A and Regnault N 2020 *Phys. Rev. B* **101** 245117
[26] Ma A, Zhang Y, Christensen T, Po H C, Jing L, Fu L and Soljacic M 2023 *Nano Lett.* **23** 772
[27] He Y, De Breuck P P, Weng H, Giantomassi M and Rignanese G M 2025 *npj Comput. Mater.* **11** 181
[28] Bradlyn B, Elcoro L, Cano J, Vergniory M G, Wang Z, Felser C, Aroyo M I and Bernevig B A 2017 *Nature* **547** 298
[29] Ho T K 1995 *Proc. Int. Conf. Doc. Anal. Recognit. (ICDAR)* **1** 278
[30] Chen T and Guestrin C 2016 *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)* **785**
[31] Hollarek D, Schopmans H, Östreicher J, Teufel J, Cao B, Alwen A, Schweidler S, Singh M, Kodalle T, Hu H, Heymans G, Abdelsamie M, Hardiagon A, Wieczorek A, Zhuk S, Schwaiger R, Siol S, Coudert F X, Wolf M, Sutter-Fella C M, Breitung B, Hodge A M, Zhang T Y and Friederich P 2025 *Adv. Intell. Discovery* 202500044
[32] Dinnebier R E, Leineweber A and Evans J S O 2019 *Rietveld Refinement: Practical Powder Diffraction Pattern Analysis Using TOPAS* (De Gruyter)
[33] Lai Q, Xu F, Yao L, Gao Z, Liu S, Wang H, Lu S, He D, Wang L, Wang C and Ke G 2025 *Adv. Sci.* **12** 2410722
[34] Li Q, Jiao R, Wu L, Zhu T, Huang W, Jin S, Liu Y, Weng H and Chen X 2025 *Nat. Commun.* **16** 7428
[35] Lu S, Lin H, Yao L, Gao Z, Ji X, Liang Y, Zhang L, Ke G, *et al.* 2025 arXiv:2503.16278
[36] Riesel E A, Mackey T, Nilforoshan H, Xu M, Badding C K, Altman A B, Leskovec J and Freedman D E 2024 *J. Am. Chem. Soc.* **146** 30340
[37] Guo G, Saidi T L, Terban M W, Valsecchi M, Billinge S J and Lipson H 2025 *Nat. Mater.* **24** 1726
[38] Chen Y, Wang X, Deng X, Liu Y, Chen X, Zhang Y, Wang L and Xiao H 2024 arXiv:2408.07608 [cond-mat.mtrl-sci]
[39] Zeni C, Pinsler R, Zügner D, Fowler A, Horton M, Fu X, Wang Z, Shysheya A, Crabbé J, Ueda S, *et al.* 2025 *Nature* 1
[40] Jiao R, Huang W, Lin P, Han J, Chen P, Lu Y and Liu Y 2023 *Adv. Neural Inf. Process. Syst.* **36** 17464
[41] Jiao R, Huang W, Liu Y, Zhao D and Liu Y 2024 arXiv:2402.03992 [cs.LG]
[42] Cao Z, Luo X, Lv J and Wang L 2024 arXiv:2403.15734 [cond-mat.mtrl-sci]
[43] Ye C Y, Weng H M and Wu Q S 2024 *Comput. Mater. Today* **1** 100003
[44] Luo X, Wang Z, Gao P, Lv J, Wang Y, Chen C and Ma Y 2024 *npj Comput. Mater.* **10** 254
[45] Ye C, Wang Y, Xie X, Zhu T, Liu J, He Y, Zhang L, Zhang J, Fang Z, Wang L, *et al.* 2025 arXiv:2505.00076 [cond-mat.mtrl-sci]
[46] Cao Z and Wang L 2025 arXiv:2504.02367 [cond-mat.mtrl-sci]
[47] Davies D W, Butler K T, Jackson A J, Skelton J M, Morita K and Walsh A 2019 *J. Open Source Softw.* **4** 1361
[48] Peng A, Liu X, Guo M Y, Zhang L and Wang H 2025 arXiv:2501.16358 [cs.LG]
[49] 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