

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)
Citation: Chin. Phys. B, 2025, 34 (**12**): 120702. DOI: 10.1088/1674-1056/ae172a
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

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

Inverse design of directional hybrid nanoantennas using neural networks

Ru-Lin Guan(管如林), Deng-Chao Huang(黄登朝), Ya-Qiong Li(李雅琼), Chen Wang(王晨), and Bin-Zi Xu(徐彬梓)
Chin. Phys. B, 2025, 34 (**9**): 090702. DOI: 10.1088/1674-1056/adeb5c

Graph neural networks unveil universal dynamics in directed percolation

Ji-Hui Han(韩继辉), Cheng-Yi Zhang(张程义), Gao-Gao Dong(董高高), Yue-Feng Shi(石月凤), Long-Feng Zhao(赵龙峰), and Yi-Jiang Zou(邹以江)
Chin. Phys. B, 2025, 34 (**8**): 080702. DOI: 10.1088/1674-1056/add505

Analysis and design of multivalued many-to-one associative memory driven by external inputs

Qiang Fang(方强) and Hao Zhang(张浩)
Chin. Phys. B, 2025, 34 (**8**): 080701. DOI: 10.1088/1674-1056/add24b

Deep learning-enabled inverse design of polarization-selective structural color based on coding metasurface

Haolin Yang(杨昊霖), Bo Ni(倪波), Junhong Guo(郭俊宏), Hua Zhou(周华), and Jianhua Chang(常建华)
Chin. Phys. B, 2025, 34 (**5**): 050702. DOI: 10.1088/1674-1056/adb94a

Text-guided diverse-expression diffusion model for molecule generation

Wenchao Weng(翁文超), Hanyu Jiang(蒋涵羽), Xiangjie Kong(孔祥杰), and Giovanni Pau
Chin. Phys. B, 2025, 34 (**5**): 050701. DOI: 10.1088/1674-1056/adbedd

---

COMPUTATIONAL PROGRAMS FOR PHYSICS

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

Tiannian Zhu(朱天念)[1,2], Zhong Fang(方忠)[1,2], Quansheng Wu(吴泉生)[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] *University of Chinese Academy of Sciences, Beijing* 100049, *China*

(Received 16 September 2025; revised manuscript received 19 October 2025; accepted manuscript online 24 October 2025)

Modern materials science generates vast and diverse datasets from both experiments and computations, yet these multi-source, heterogeneous data often remain disconnected in isolated "silos". Here, we introduce MaterialsGalaxy, a comprehensive platform that deeply fuses experimental and theoretical data in condensed matter physics. Its core innovation is a structure similarity-driven data fusion mechanism that quantitatively links cross-modal records — spanning diffraction, crystal growth, computations, and literature — based on their underlying atomic structures. The platform integrates artificial intelligence (AI) tools, including large language models (LLMs) for knowledge extraction, generative models for crystal structure prediction, and machine learning property predictors, to enhance data interpretation and accelerate materials discovery. We demonstrate that MaterialsGalaxy effectively integrates these disparate data sources, uncovering hidden correlations and guiding the design of novel materials. By bridging the long-standing gap between experiment and theory, MaterialsGalaxy provides a new paradigm for data-driven materials research and accelerates the discovery of advanced materials.

Keywords: MaterialsGalaxy, data fusion, materials gene, materials database

PACS: 07.05.Mh, 61.50.Ah, 71.15.–m, 61.05.cc

DOI: 10.1088/1674-1056/ae172a **CSTR:** 32038.14.CPB.ae172a

1. Introduction

The fields of condensed matter physics and materials science are undergoing a profound, data-intensive transformation. Decades of experimental exploration and theoretical computation have amassed invaluable data, from experimentally determined crystal structures (e.g., ICSD,[1] COD,[2,3] CSD,[4] and the Pauling File[5]) to properties derived from first-principles calculations (e.g., in databases like Materials Project,[6] AFLOW,[7] OQMD,[8] MatCloud,[9] NOMAD,[10] Materials Cloud,[11] Atomly[12]). This data wealth, coupled with the rise of the data-driven "fourth paradigm",[13] offers unprecedented opportunities for materials discovery.[14–20] It enables the systematic analysis of massive datasets to uncover and interpret hidden structure–property relationships,[21,22] accelerate the rational design of novel materials,[23–25] and predict their performance with increasing accuracy.[26–28] Indeed, the synergistic integration of artificial intelligence, high-performance computing, and automated experimentation is emerging as a powerful strategy to enrich and accelerate every stage of the discovery cycle.[29–31]

Despite this abundance of data from structured databases, the full potential of these data resources remains largely untapped due to the pervasive "data silo" phenomenon. An even larger reservoir of knowledge, including crucial synthesis details, resides within the unstructured text of scientific literature. Experimental data, theoretical calculations, and literature-extracted knowledge are fundamentally disparate, differing in formats, naming conventions, precision standards, and acquisition methods. This inherent heterogeneity, compounded by a lack of standardized interoperability protocols, severely hinders cross-source integration and analysis.[32,33] Consequently, researchers face significant challenges in comparing and integrating data from these distinct origins, a bottleneck that diminishes research efficiency.

To address this, the materials science community has initiated crucial data integration efforts. The OPTIMADE consortium,[34] for example, has made significant strides in providing a unified application programming interface (API) for computational materials databases, enhancing interoperability among them. Concurrently, specialized NLP models have been developed to extract information from scientific literature.[35–38] However, such efforts predominantly focus on homogeneous data sources (e.g., linking computational databases). The deep, cross-modal fusion of experimental data with theoretical computations — a far more complex challenge due to fundamental differences in data generation, semantics, and precision — remains a critical and largely un-

†Corresponding author. E-mail: quansheng.wu@iphy.ac.cn
‡Corresponding author. E-mail: hmweng@iphy.ac.cn

© 2025 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

solved frontier.

Bridging the divide between experimental and theoretical data holds immense scientific merit. High-quality experimental data provide the ground truth for validating and refining computational models.[39,40] Crucially, even experimental failures or "negative results" are invaluable, as they provide critical constraints that help define the boundaries of successful synthesis or desired properties, thereby further refining predictive models.[41] Conversely, theoretical calculations offer predictive guidance for exploring materials yet to be synthesized or characterized.[42,43] An effective fusion of these two data modalities would establish a powerful closed loop, where experiment and theory mutually validate and accelerate one another, fostering more accurate models and hastening the discovery of new materials.

To address these challenges, we developed the MaterialsGalaxy platform, designed for the deep fusion of heterogeneous experimental and theoretical databases in condensed matter physics. Our core innovation is a data-linking methodology centered on crystal structure similarity. We transform crystal structures into fixed-length numerical vectors, or "fingerprints", that encode key chemical and structural features. While advanced, end-to-end representation learning methods like graph neural networks (GNNs) and continuous-filter convolutional networks offer state-of-the-art performance,[44–46] for this foundational work, we opted for a robust and interpretable feature engineering approach from the matminer library.[47] These descriptor-based fingerprints can be generated orders of magnitude faster than deep learning embeddings and are deterministic, and their components (e.g., mean atomic radius, packing efficiency) have clear physical and chemical meaning. These vectors are then efficiently indexed in a vector database. Leveraging this index, we perform high-speed similarity searches to dynamically link disparate data records — from experimental synthesis to theoretical properties — that correspond to the same or similar materials, effectively dismantling data silos. Beyond this central fusion engine, MaterialsGalaxy integrates a synergistic suite of AI tools, including a domain-specific large language model TopoChat,[48] a generative model for crystal structure prediction (Con-CDVAE),[49] and machine learning models for property prediction. These tools, coupled with the fused data, create a powerful ecosystem for intelligent data analysis and accelerated materials discovery.

This paper elucidates the architecture and data fusion methodology of the MaterialsGalaxy platform. We demonstrate its capabilities through application examples that connect experimental and theoretical data to facilitate materials discovery. Finally, we discuss the broader implications of this approach for the research paradigm in condensed matter physics and materials science, aiming to provide a unified data infrastructure that fosters deeper synergy between experiment and theory.

2. Results

2.1. Platform architecture

The overall architecture of the MaterialsGalaxy platform, illustrated in Fig. 1, is engineered to systematically address the challenge of heterogeneous data integration in materials science. At its core, the architecture follows a multi-stage workflow designed for robust data processing and intelligent analysis. It begins with a Data Acquisition and Standardization Layer that ingests and harmonizes multimodal data from disparate sources, including public databases, electronic lab notebooks, and the scientific literature.

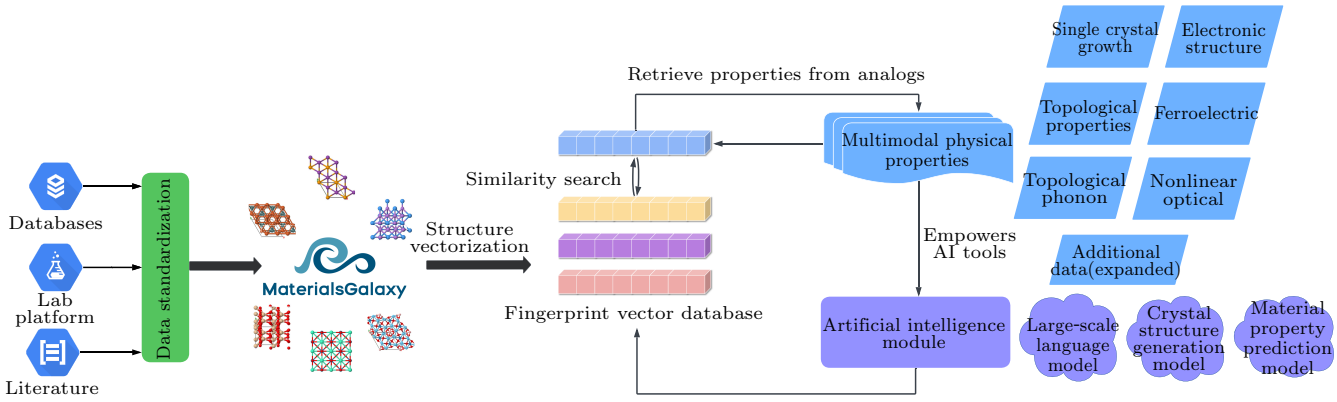


Fig. 1. Architecture of the MaterialsGalaxy platform. The platform employs a systematic workflow to fuse heterogeneous data from three primary channels: (1) existing public databases, (2) electronic laboratory notebooks, and (3) automated literature extraction. Raw data first undergo a rigorous standardization process. The core innovation is the structure vectorization module, which uses representation learning to generate a unique fingerprint for each crystal structure. These fingerprints are indexed in a high-performance vector database, enabling a similarity matching engine to dynamically link disparate records. The resulting fused data backbone supports a rich application layer featuring interactive querying, visualization tools, a RESTful API, and a suite of integrated AI tools (e.g., LLM-based assistants, generative models, and property predictors). Crucially, this architecture not only connects siloed experimental and theoretical data but also enriches them, creating a comprehensive, multi-modal profile for each material based on shared structural features.

The central component is a Structure-driven Fusion Engine. This engine leverages representation learning to vectorize crystal structures and employs a vector database for high-speed similarity matching. This mechanism is the key to linking otherwise disconnected records. Critically, this fusion process creates a holistic, multi-modal profile for each material. For instance, an experimental record, which might originally contain only synthesis conditions and a diffraction pattern, can be dynamically linked to its theoretical counterparts in the fused database, instantly augmenting it with computed properties like band structure, formation energy, and topological invariants. This cross-modal enrichment allows researchers to rapidly gain a comprehensive understanding of a material's properties, seamlessly bridging the gap between its experimental realization and theoretical characteristics.

This fused data backbone supports a versatile Application and Analysis Layer, which provides user-facing functionalities such as interactive querying, data visualization, API access, and a suite of integrated AI tools for property prediction and materials discovery. This cohesive design establishes a systematic solution to not only bridge data silos but also to create a synergistically enriched data ecosystem that empowers advanced, data-driven research.

2.2. Data sources and integration strategy

The MaterialsGalaxy platform is built upon a diverse and growing collection of multi-source, heterogeneous data from the field of condensed matter physics, hosted at Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences. Our data acquisition strategy follows a three-pronged approach: (1) aggregation of established public databases, (2) automated ingestion from our in-house electronic laboratory platform (MatElab),[50] which captures curated experimental records with full provenance, and (3) automated extraction of structural and property data from the scientific literature. This multi-channel approach ensures the construction of a comprehensive data ecosystem spanning material synthesis, characterization, and theoretical properties.

The platform currently integrates a wide spectrum of data modalities, including crystal structures, electronic band structures, topological classifications, phonon dispersions, ferroelectric properties, non-linear optical responses, and single-crystal growth recipes. The scale and diversity of these data sources are quantitatively summarized in Fig. 2(a). The experimental cornerstone of our collection is a large-scale crystal structure database[51] comprising nearly 500000 entries derived from the Crystallography Open Database (COD)[2,3] and supplemented by literature mining. This is complemented by a unique single-crystal growth database sourced from our MatElab platform,[50] which contains 2000 detailed experimental records documenting synthesis parameters and characterization results.

On the theoretical front, MaterialsGalaxy incorporates several high-throughput computation databases. Key among these is a comprehensive topological materials database,[52] containing over 8000 unique materials identified as topologically non-trivial (e.g., topological insulators, semimetals) from a screening of more than 28000 candidates.[53] Similarly, a topological phonon database provides phononic band structures and classifications for over 5000 materials.[54,55] Additional computational datasets cover properties such as 2D ferroelectricity[56] and non-linear optical coefficients.[57]

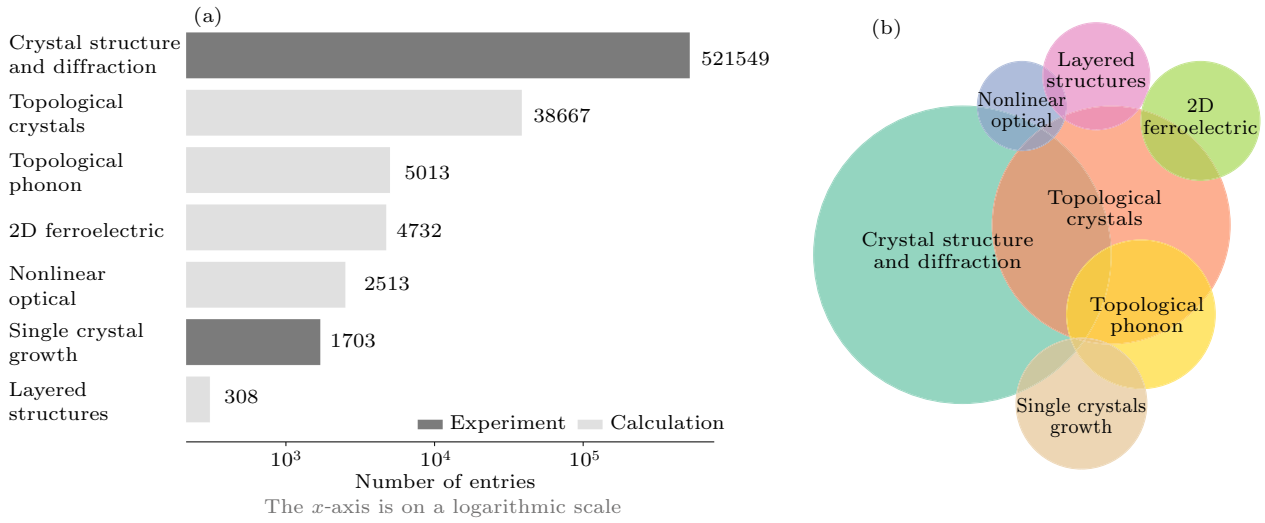


Fig. 2. Overview of integrated data sources and their heterogeneity. (a) Distribution of entries across the primary integrated databases, with experimental sources shown in blue and theoretical/computational sources in orange. The y-axis is on a logarithmic scale to accommodate the wide range of data volumes. The collection includes a large experimental crystal structure database (COD-derived), various computational property databases (e.g., topological materials, topological phonons), and a unique database of single-crystal growth experiments. (b) A conceptual Venn diagram illustrating the complex relationships of overlap and uniqueness among different data modalities. This highlights the core challenge of data heterogeneity, where, for instance, the materials space of experimental synthesis records, theoretically predicted topological materials, and the general crystal structure database are partially intersecting yet distinct, necessitating a robust data fusion strategy.

The primary challenge addressed by our platform stems from the inherent heterogeneity and fragmentation of these data sources. As conceptually illustrated by the Venn diagram in Fig. 2(b), these datasets exhibit complex overlaps and complementarities. For example, the broad chemical space of experimental crystal structures (including organics) contrasts with the inorganic focus of most computational databases. Different property databases may share materials but describe orthogonal physical phenomena. The experimental growth database provides unique synthesis context that is often decoupled from theoretical entries. This intricate landscape of data types, formats, precision levels, and semantic contexts necessitates the systematic standardization and fusion methodologies detailed in the following sections.

2.3. Data standardization

A rigorous and automated data standardization pipeline is the bedrock of the MaterialsGalaxy platform, transforming raw, heterogeneous inputs into a consistent, analysis-ready format. This initial step is critical for constructing a truly "AI-ready" dataset. By this, we mean that the data is not only clean and structured but is also semantically rich and primed for effective use by machine learning algorithms, thus preventing issues like data leakage or biased model training that arise from inconsistent inputs. This disciplined approach is essential, as disparities in crystallographic conventions, data formats, and physical units across sources would otherwise prevent reliable data fusion and undermine model performance.

The standardization process is centered on establishing a canonical representation for crystal structures, the universal anchors for data linking. All incoming structural data are processed using the pymatgen[58] and spglib[59] libraries to parse, validate, and resolve symmetries according to IUCr conventions, ensuring identical materials map to a single, unique representation. For associated property data, we developed a formal data schema that enforces standardized nomenclature and units. Crucially, this schema was not developed in isolation; it is the product of close collaboration with domain experts — both the original data producers and active researchers — ensuring that our standards reflect the nuanced requirements and best practices of the condensed matter physics community. This schema is programmatically enforced using data validation libraries like Pydantic, guaranteeing that every data point adheres to a predefined type and structure before ingestion. This is particularly crucial for parsing the output of diverse first-principles calculation packages, such as the widely used Vienna *Ab initio* Simulation Package (VASP),[60] ensuring that computational parameters are captured consistently. This systematic process captures not only the data values but also essential metadata (e.g., computational parameters or experimental conditions), providing the robust and reliable foundation required for the structure-based data fusion mechanism described next.

2.4. Core data fusion: Structure-based similarity linking

Having established a foundation of standardized, canonical data, we implement our core innovation: a dynamic data fusion mechanism driven by crystal structure similarity. This approach directly tackles the long-standing challenge of linking records across heterogeneous databases. Traditional methods for identifying similar crystal structures, such as the structure matching algorithms found in libraries like Pymatgen,[58] rely on direct, pairwise comparisons of atomic coordinates and cell parameters. While precise for near-identical structures, these methods suffer from two critical drawbacks for large-scale data fusion: they are computationally expensive, often scaling poorly to millions of comparisons, and they are brittle, struggling to identify structurally related but not identical phases (e.g., those with minor distortions or different elemental decorations). Our approach overcomes these limitations by moving from direct comparison to a highly efficient, vector-space similarity search.

Our fusion workflow comprises two key stages. First, in an offline pre-processing step, every standardized crystal structure is transformed into a fixed-length numerical vector, or structural fingerprint, using a representation learning algorithm. For this, we employ the SiteStatsFingerprint featurizer from the matminer library,[47] which encodes rich information about local atomic environments into a high-dimensional vector. This process effectively maps the complex, variable-sized crystal graph into a unified, machine-readable vector space where geometric and chemical similarity is represented by proximity.

Second, and most critically, data fusion occurs dynamically at query time through a high-performance vector search index. All structural fingerprints are indexed using approximate nearest neighbor (ANN) algorithms (e.g., HNSW-based graphs[61]), enabling sub-second similarity searches. When a user views a specific material entry, the platform performs multiple, context-aware searches to retrieve two classes of information for each property module: (1) direct data, which is any information directly linked to the queried material's exact structure, and (2) analog data, which comprises the properties of structurally analogous materials. These analogs are identified in real-time by launching a similarity search within the relevant data subset.

This "just-in-time" data augmentation is exceptionally powerful. If a queried material lacks data in a certain dimension, for instance, no experimental synthesis record exists, the platform can still provide crucial insights by displaying the synthesis conditions of its closest structural analogs. This mechanism effectively uses the collective knowledge of the

entire database to enrich the profile of a single material, offering researchers valuable predictive hints and experimental starting points. It is this dynamic, similarity-driven approach that robustly bridges data gaps and transforms a collection of siloed datasets into an interconnected, intelligent knowledge base.

2.5. Application example: Data fusion for $CrGeTe_3$

To illustrate the practical power of our structure-driven fusion mechanism, we present a case study on $CrGeTe_3$, a widely studied two-dimensional magnetic semiconductor.[62–64] Data for this material are typically fragmented across our platform, making it an ideal example to demonstrate how our platform enables materials analysis along two orthogonal axes: horizontal integration for a single material and vertical comparison across similar materials. This dual-faceted workflow is visually encapsulated in Fig. 3.

First, the platform facilitates horizontal integration, creating a comprehensive, multi-modal profile for the target material itself. When a user queries $CrGeTe_3$, the system aggregates all its direct data by linking records from disparate sources. This process connects, for example, its experimental crystal structure from a diffraction database with its calculated electronic band structure from a topological database. This unified view enables powerful cross-modal analyses, such as correlating experimental conditions with theoretical properties. Even if a direct record is missing, this horizontal integration can fill a gap; for instance, if an experimental structure lacks a corresponding theoretical calculation, our similarity mechanism links it to the closest available computational entry, providing a robust theoretical proxy.

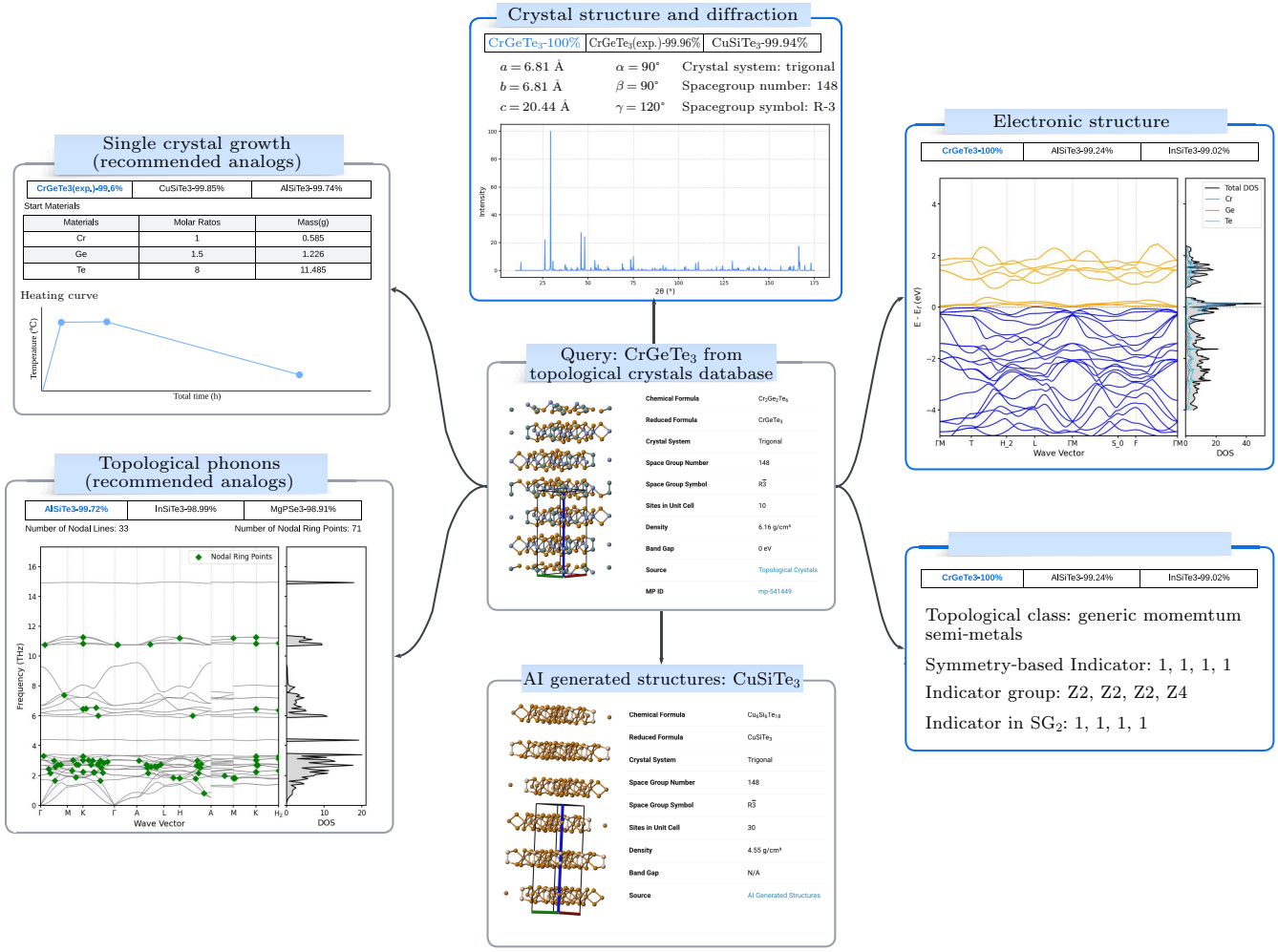


Fig. 3. Data fusion and discovery workflow for $CrGeTe_3$. The platform's dual-axis analysis is triggered by a query for a target material. Horizontal integration: Direct data for $CrGeTe_3$ are aggregated across multiple modules (e.g., "crystal structure", "electronic structure") to build a deep, cross-modal profile linking experiment and theory. Vertical comparison: The material profile is enriched with data from structural analogs. For modules where direct data is missing (e.g., "single crystal growth"), the platform provides actionable references from known similar materials (e.g., $AlSiTe_3$). This comparison is further extended to novel, AI-generated structures (e.g., $CuSiTe_3$), enabling the exploration of uncharted chemical space for accelerated materials discovery.

Building on this complete single-material profile, the platform enables vertical comparison, a powerful data-driven workflow for hypothesis generation and knowledge discovery. This process situates the target material within its broader "structural family" by identifying a cohort of its closest structural analogs via vector similarity. This vertical analysis pioneers a dual-pronged exploration strategy. It begins with knowledge extraction from known materials. By comparing

$CrGeTe_3$ against existing compounds in the database (e.g., $FePSe_3$, $AlSiTe_3$), researchers can systematically mine for structure–property relationships and identify trends, providing data-driven guidance to optimize experimental growth and reduce trial-and-error cycles.

More profoundly, this workflow accelerates the in silico discovery-synthesis loop by integrating AI-generated candidate structures. The vertical comparison extends beyond known materials into the vast, uncharted chemical space of theoretically plausible structures generated by our integrated deep generative models. By comparing the target material against these novel, yet-undiscovered candidates, researchers can identify promising new compositions that may exhibit superior functionalities. This process of generating theoretical candidates and comparing them against established materials provides a direct, actionable pathway for guiding future synthesis efforts towards the most promising frontiers, thereby embodying a key tenet of the data-driven materials discovery paradigm.

In summary, the $CrGeTe_3$ case study demonstrates a paradigm shift from static data retrieval to a dynamic, multi-faceted research process. The horizontal integration offers unprecedented data depth for a single material, while the vertical comparison — spanning both known data and AI-generated possibilities — provides the data breadth required for true discovery. This powerful combination establishes a virtuous cycle: integrated data fuels AI models, which in turn generate new knowledge and propose novel materials, setting the stage for an accelerated, inverse design approach to materials science.[65]

Additional application examples, including a topological phonon material (CoSi) and a nonlinear optical material ($LiNbO_3$), are provided in the supplementary materials (Figs. S4 and S5).

2.6. Platform features and functionality

Building on its core data fusion engine, the MaterialsGalaxy platform provides a multi-layered suite of features designed to maximize data accessibility, usability, and analytical power. The primary user entry point is a web-based portal offering a rich, interactive data exploration experience. Through this interface, users can perform complex queries using a combination of filters, such as chemical composition, space group symmetry, and calculated property ranges, and assess material properties through a suite of integrated visualization tools, including a 3D crystal structure viewer and interactive plots for electronic and phononic band structures.

To support the growing need for data-driven research and ensure interoperability, the platform is designed with the FAIR Guiding Principles as a cornerstone.[66] Data are programmatically accessible through a well-documented RESTful API. This API, which adheres to the OpenAPI specification and is implemented using the FastAPI framework, allows for automated, large-scale data retrieval in a structured JSON format suitable for direct integration into machine learning workflows and custom research pipelines. Full documentation, including endpoint specifications and detailed usage examples for the API, is provided in the supplementary materials. For bulk analysis, key datasets are also available for direct download.

A key distinguishing feature of MaterialsGalaxy is its seamless integration of state-of-the-art AI tools that operate directly on the fused data to accelerate the research cycle. These include our conversational agent, TopoChat, which is a specialized large language model for condensed matter physics.[48] For generative inverse design, the platform offers a modular framework supporting multiple distinct generative strategies, including the conditional variational autoencoder Con-CDVAE[49] and DiffCSP++,[67] a diffusion model that rigorously incorporates space group constraints. To bridge the experimental–computational gap, the platform also integrates PXRDGen,[68] an end-to-end model for de novo crystal structure determination directly from powder x-ray diffraction (PXRD) patterns. Finally, to enable high-throughput virtual screening, a suite of machine learning models provides on-the-fly predictions for key properties like formation energy on any user-provided or AI-generated structure. These powerful, integrated AI capabilities collectively transform the platform from a simple data repository into a dynamic and interactive discovery environment.

3. Discussion and conclusion

The implementation of the MaterialsGalaxy platform demonstrates that systematic data standardization, when coupled with dynamic, structure-driven data association, can effectively dismantle the long-standing data silos in condensed matter physics. Our platform provides a powerful infrastructure to accelerate data-driven materials discovery, yet key challenges and opportunities for future enhancement remain. The foremost challenge is the scarcity of high-quality experimental data, which remains a bottleneck for the field and directly affects the reliability of both data fusion and downstream AI models.[69,70] Another critical limitation arises from the dependence of our fusion scheme on structural similarity, which makes it sensitive to imperfections in experimental data, such as missing atoms or inaccurate atomic positions.[71] To enhance robustness, our future work will focus on developing more invariant structure representations under uncertainty,[72–74] for instance by leveraging graph neural network embeddings[44,45] and exploring multi-modal fusion strategies that integrate complementary data modalities such as electronic band structures, x-ray diffraction (XRD) patterns, and other spectroscopic features.[75] Given the limited sample

size within individual modalities, we further plan to investigate cross-modal alignment approaches, such as CLIP-based frameworks,[76] to better align heterogeneous data and enable scalable multi-modal representation learning.

4. Methods

4.1. Data acquisition and standardization

The MaterialsGalaxy platform integrates data from multiple sources as detailed in Section 2, including public databases, internal experimental data, and data extracted from scientific literature. Prior to fusion, all data undergo a rigorous standardization pipeline. Crystal structures, primarily from CIF files,[77] are parsed and validated using the pymatgen library.[58] Standardized representations (including primitive and conventional cells) are generated according to IUCr conventions, with symmetry analysis handled by the underlying spglib library,[59] ensuring consistent structural representation across all sources. Chemical formulas and calculated properties are also standardized, with metadata schemas (e.g., computational parameters) enforced in collaboration with domain experts to ensure data comparability. Automated scripts facilitate this efficient data ingestion and standardization process.

4.2. Structure vectorization

To enable structure-based similarity search, each standardized crystal structure is converted into a fixed-length numerical vector (structural fingerprint). We employed the SiteStatsFingerprint featurizer from the matminer library.[47] This method computes local atomic environment fingerprints and then calculates their statistics (mean and maximum) across all sites to generate a final, 122-dimensional structure-level vector. This representation effectively encodes key structural and chemical information into a format suitable for high-speed similarity comparison.

4.3. Vector database and similarity search implementation

The generated structural fingerprint vectors are indexed for efficient retrieval using an implementation of the approximate nearest neighbor (ANN) search algorithm. Specifically, we utilize an index based on the hierarchical navigable small world (HNSW) graph method,[61] a state-of-the-art technique for high-dimensional vector search, often implemented in libraries such as Faiss.[78] The index is constructed to balance memory usage and search accuracy. Similarity is quantified using the cosine similarity metric. Critically, data linking occurs dynamically at query time. A query structure's vector is used to retrieve its k-nearest neighbors (typically $k = 10$–50) from the index within the relevant data subset. This ANN approach enables sub-second query times on datasets of millions of vectors, providing a scalable and responsive fusion experience.

Data availability

The data supporting the findings of this study are publicly accessible through the MaterialsGalaxy platform and its associated services. The main web portal, available at https://materialsgalaxy.iphy.ac.cn, provides interactive browsing and visualization of all integrated data.

For programmatic access, a comprehensive, OpenAPI-compliant RESTful API is provided. The API offers tiered access: endpoints for searching and retrieving summary-level information for all materials are open to the public without authentication. Access to detailed data records for individual materials, such as full crystal structures and specific calculated properties, requires a free, user-registered API key. Full interactive documentation for the API is available at https://materialsgalaxy.iphy.ac.cn/docs, with usage guides at https://materialsgalaxy.iphy.ac.cn/guides/api.

To facilitate reproducibility and large-scale analysis, a static snapshot of the core data summary, containing essential information such as chemical formulas and crystal structures (in CIF format) for the materials presented in this study, is available for download on https://materialsgalaxy.iphy.ac.cn/downloads.

The underlying datasets integrated into MaterialsGalaxy are hosted and maintained by the Condensed Matter Physics Data Center (CMPDC) (https://cmpdc.iphy.ac.cn). Raw data originating from external public databases (e.g., COD) are subject to their original licenses and remain available from their respective repositories.

Progarm availability

The MaterialsGalaxy platform relies exclusively on publicly available, open-source libraries for its core scientific methodologies, as cited throughout the text. The key procedures for data standardization, structure vectorization, and similarity search can be reproduced using libraries such as Pymatgen,[58] Matminer,[47] and an appropriate ANN library implementing HNSW.[61] Further details are provided in Section 4.

Acknowledgements

We wish to express our sincere gratitude to the researchers and their teams who contributed the foundational datasets that constitute MaterialsGalaxy. Specifically, we thank Shifeng Jin for the Crystal Structure and Diffraction data;[51] Youguo Shi for the Single Crystal Growth experimental data;[79] Jinbo Pan for the Layered Structures Database;[80] Qingbo Yan for the 2D Ferroelectric

Materials Database;[56] Xingqiu Chen for the Topological Phonon Database;[55] and Yong Xu for the Nonlinear Optical Database.[57] This work was supported by the Science Center of the National Natural Science Foundation of China (Grant No. 12188101), the National Natural Science Foundation of China (Grant Nos. 12274436 and 11921004), the National Key R&D Program of China (Grant Nos. 2023YFA1607400 and 2022YFA1403800), H.W. acknowledges support from the New Cornerstone Science Foundation through the XPLORER PRIZE. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of the Chinese Academy of Sciences.

References

[1] Zagorac D, Müller H, Ruehl S, Zagorac J and Rehme S 2019 *Journal of Applied Crystallography* **52** 918
[2] Gražulis S, Chateigner D, Downs R T, Yokochi A F T, Quirós M, Lutterotti L, Manakova E, Butkus J, Moeck P and Le Bail A 2009 *Journal of Applied Crystallography* **42** 726
[3] Gražulis S, Daškevič A, Merkys A, Chateigner D, Lutterotti L, Quirós M, Serebryanaya N R, Moeck P, Downs R T and Le Bail A 2012 *Nucleic Acids Research* **40** D420
[4] Groom C R, Bruno I J, Lightfoot M P and Ward S C 2016 *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **72** 171
[5] Villars P, Berndt M, Brandenburg K, Cenzual K, Daams J, Hulliger F, Massalski T, Okamoto H, Osaki K, Prince A, Putz H and Iwata S 2004 *J. Alloys Compd.* **367** 293
[6] 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 Materials* **1** 011002
[7] Curtarolo S, Setyawan W, Hart G L W, Jahnatek M, Chepulskii R V, Taylor R H, Wang S, Xue J, Yang K, Levy O, Mehl M J, Stokes H T, Demchenko D O and Morgan D 2012 *Computational Materials Science* **58** 218
[8] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Computational Materials* **1**15010
[9] Yang X, Wang Z, Zhao X, Song J, Zhang M and Liu H 2018 *Computational Materials Science* **146** 319
[10] Draxl C and Scheffler M 2019 *J. Phys.: Mater.* **2** 036001
[11] 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** 299
[12] Miao L and Sheng M 2020 Atomly https://atomly.net/
[13] Hey T, Tansley S, Tolle K and Gray J 2009 *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research)
[14] Agrawal A and Choudhary A 2016 *APL Materials* **4** 053208
[15] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 *npj Computational Materials* **3** 54
[16] Lookman T, Balachandran P V, Xue D and Yuan R 2019 *npj Computational Materials* **5** 21
[17] Schleder G R, Padilha A C M, Acosta C M, Costa M and Fazzio A 2019 *J. Phys.: Mater.* **2** 032001
[18] Merchant A, Batzner S, Schoenholz S S, Aykol M, Cheon G and Cubuk E D 2023 *Nature* **624** 80
[19] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 *Nature* **559** 547
[20] Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, Park C W, Choudhary A, Agrawal A, Billinge S J L, Holm E, Ong S P and Wolverton C 2022 *npj Computational Materials* **8** 59
[21] Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A and Han T Y J 2022 *npj Computational Materials* **8** 204
[22] Vu T S, Ha M Q, Nguyen D N, Nguyen V C, Abe Y, Tran T, Tran H, Kino H, Miyake T, Tsuda K and Dam H C 2023 *npj Computational Materials* **9** 215
[23] Sanchez-Lengeling B and Aspuru-Guzik A 2018 *Science* **361** 360
[24] Gubernatis J E and Lookman T 2018 *Phys. Rev. Materials* **2** 120301
[25] Ma J, Cao B, Dong S, Tian Y, Wang M, Xiong J and Sun S 2024 *npj Computational Materials* **10** 59
[26] Oganov A R, Pickard C J, Zhu Q and Needs R J 2019 *Nature Reviews Materials* **4** 331
[27] Chan C H, Sun M and Huang B 2022 *EcoMat* **4** e12194
[28] Griesemer S D, Xia Y and Wolverton C 2023 *Nature Computational Science* **3** 934
[29] Stein H S and Gregoire J M 2019 *Chemical Science* **10** 9640
[30] Pyzer-Knapp E O, Pitera J W, Staar P W J, Takeda S, Laino T, Sanders D P, Sexton J, Smith J R and Curioni A 2022 *npj Computational Materials* **8** 84
[31] Szymanski N J, Rendy B, Fei Y, Kumar R E, He T, Milsted D, McDermott M J, Gallant M, Cubuk E D, Merchant A, Kim H, Jain A, Bartel C J, Persson K, Zeng Y and Ceder G 2023 *Nature* **624** 86
[32] Kalidindi S R and Graef M D 2015 *Annual Review of Materials Research* **45** 171
[33] Himanen L, Geurts A, Foster A S and Rinke P 2019 *Advanced Science* **6** 1900808
[34] Andersen C W, Armiento R, Blokhin E, *et al.* 2021 *Scientific Data* **8** 217
[35] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 *Nature* **571** 95
[36] Gupta T, Zaki M, Krishnan N M A and Mausam 2022 *npj Computational Materials* **8** 102
[37] Pyzer-Knapp E O, Manica M, Staar P, Morin L, Ruch P, Laino T, Smith J R and Curioni A 2025 *npj Computational Materials* **11** 61
[38] Jiang X, Wang W, Tian S, Wang H, Lookman T and Su Y 2025 *npj Computational Materials* **11** 79
[39] Green M L, Choi C L, Hattrick-Simpers J R, Joshi A M, Takeuchi I, Barron S C, Campo E, Chiang T, Empedocles S, Gregoire J M, Kusne A G, Martin J, Mehta A, Persson K, Trautt Z, Van Duren J and Zakutayev A 2017 *Applied Physics Reviews* **4** 011105
[40] Wu Y, Wang C F, Ju M G, Jia Q, Zhou Q, Lu S, Gao X, Zhang Y and Wang J 2024 *Nat. Commun.* **15** 138
[41] Raccuglia P, Elbert K C, Adler P D F, Falk C, Wenny M B, Mollo A, Zeller M, Friedler S A, Schrier J and Norquist A J 2016 *Nature* **533** 73
[42] Zhang H, Liu C X, Qi X L, Dai X, Fang Z and Zhang S C 2009 *Nat. Phys.* **5** 438
[43] Weng H, Fang C, Fang Z, Bernevig B A and Dai X 2015 *Phys. Rev. X* **5** 011029
[44] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[45] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 *Chemistry of Materials* **31** 3564
[46] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A and Müller K R 2018 *The Journal of Chemical Physics* **148** 241722
[47] Ward L, Dunn A, Faghaninia A, Zimmermann N E R, 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 *Computational Materials Science* **152** 60
[48] Xu H, Zhang B, Jin Z, Zhu T, Wu Q and Weng H 2024 Enhancing Large Language Models with Domain-Specific Knowledge: The Case in Topological Materials (*Preprint* 2409.13732)
[49] Ye C Y, Weng H M and Wu Q S 2024 *Computational Materials Today* **1** 100003
[50] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences MatElab: Electronic Laboratory for Material Science https://matelab.iphy.ac.cn
[51] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences Crystal Structure and Diffraction Database https://cmpdc.iphy.ac.cn/diff/
[52] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences Materiae: Topological Materials Database https://cmpdc.iphy.ac.cn/materiae/
[53] Zhang T, Jiang Y, Song Z, Huang H, He Y, Fang Z, Weng H and Fang C 2019 *Nature* **566** 475
[54] Li J, Liu J, Baronett S A, Liu M, Wang L, Li R, Chen Y, Li D, Zhu Q and Chen X Q 2021 *Nat. Commun.* **12** 1204
[55] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences and Institute of Metal Research, Chinese Academy of Sciences Topological Phonon Database http://www.phonon.synl.ac.cn/

[56] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences and Univeristy of Chinese Academy of Sciences 2D Ferroelectric Materials Database https://cmpdc.iphy.ac.cn/fedb/

[57] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences and Department of Physics, Tsinghua University Nonlinear Optical Database https://cmpdc.iphy.ac.cn/nlo/

[58] 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 *Computational Materials Science* **68** 314

[59] Togo A, Shinohara K and Tanaka I 2024 *Science and Technology of Advanced Materials: Methods* **4** 2384822

[60] Kresse G and Furthmüller J 1996 *Phys. Rev. B* **54** 11169

[61] Malkov Y A and Yashunin D A 2020 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42** 824

[62] Sivadas N, Daniels M W, Swendsen R H, Okamoto S and Xiao D 2015 *Phys. Rev. B* **91** 235425

[63] Xu C, Feng J, Xiang H and Bellaiche L 2018 *npj Computational Materials* **4** 57

[64] Lin G T, Zhuang H L, Luo X, Liu B J, Chen F C, Yan J, Sun Y, Zhou J, Lu W J, Tong P, Sheng Z G, Qu Z, Song W H, Zhu X B and Sun Y P 2017 *Phys. Rev. B* **95** 245212

[65] Wang J, Wang Y and Chen Y 2022 *Materials* **15** 1811

[66] Wilkinson M D, Dumontier M, Aalbersberg I J, *et al.* 2016 *Scientific Data* **3** 160018

[67] Jiao R, Huang W, Liu Y, Zhao D and Liu Y 2024 Space Group Constrained Crystal Generation (*Preprint* 2402.03992)

[68] Li Q, Jiao R, Wu L, Zhu T, Huang W, Jin S, Liu Y, Weng H and Chen X 2025 *Nat. Commun.* **16** 7428

[69] Kulik H J 2025 *Journal of Materials Research* **40** 833

[70] Dunn A, Wang Q, Ganose A, Dopp D and Jain A 2020 *npj Computational Materials* **6** 138

[71] Spek A L 2020 *Acta Crystallographica Section E: Crystallographic Communications* **76** 1

[72] Goodall R E A and Lee A A 2020 *Nat. Commun.* **6280** 2041

[73] Antunes L M, Butler K T and Grau-Crespo R 2024 *Nat. Commun.* **15** 10570

[74] Zhu R, Nong W, Yamazaki S and Hippalgaonkar K 2024 *Matter* **7** 3469

[75] Moro V, Loh C, Dangovski R, Ghorashi A, Ma A, Chen Z, Kim S, Lu P Y, Christensen T and Soljačić M 2025 *Newton* **1** 100016

[76] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I 2021 Learning Transferable Visual Models From Natural Language Supervision (*Preprint* 2103.00020)

[77] Hall S R, Allen F H and Brown I D 1991 *Acta Crystallographica Section A* **47** 655

[78] Johnson J, Douze M and Jégou H 2021 *IEEE Transactions on Big Data* **7** 535

[79] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences Single Crystal Growth Database https://cmpdc.iphy.ac.cn/mlab/

[80] Condensed Matter Physics Data Center, Institute of Physics, Chinese Academy of Sciences Layered Materials Database https://cmpdc.iphy.ac.cn/layered/