Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: 2766-0400 (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Starrydata: from published plots to shared materials data

Yukari Katsura, Masaya Kumagai, Tomoya Mato, Yu Takada, Yuki Ando, Erina Fujita, Fumikazu Hosono, Eiji Koyama, Farhan Mudasar, Ton Nu Thanh Phuong, Naoto Saito, Yoshihiro Sakamoto, Atsumi Tanaka, Dewi Yana, Kaoru Kimura, Koji Tsuda & Masahiko Demura

To cite this article: Yukari Katsura, Masaya Kumagai, Tomoya Mato, Yu Takada, Yuki Ando, Erina Fujita, Fumikazu Hosono, Eiji Koyama, Farhan Mudasar, Ton Nu Thanh Phuong, Naoto Saito, Yoshihiro Sakamoto, Atsumi Tanaka, Dewi Yana, Kaoru Kimura, Koji Tsuda & Masahiko Demura (2025) Starrydata: from published plots to shared materials data, Science and Technology of Advanced Materials: Methods, 5:1, 2506976, DOI: 10.1080/27660400.2025.2506976

To link to this article: https://doi.org/10.1080/27660400.2025.2506976

Published online: 12 Jun 2025.

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2025, VOL. 5, NO. 1, 2506976
https://doi.org/10.1080/27660400.2025.2506976

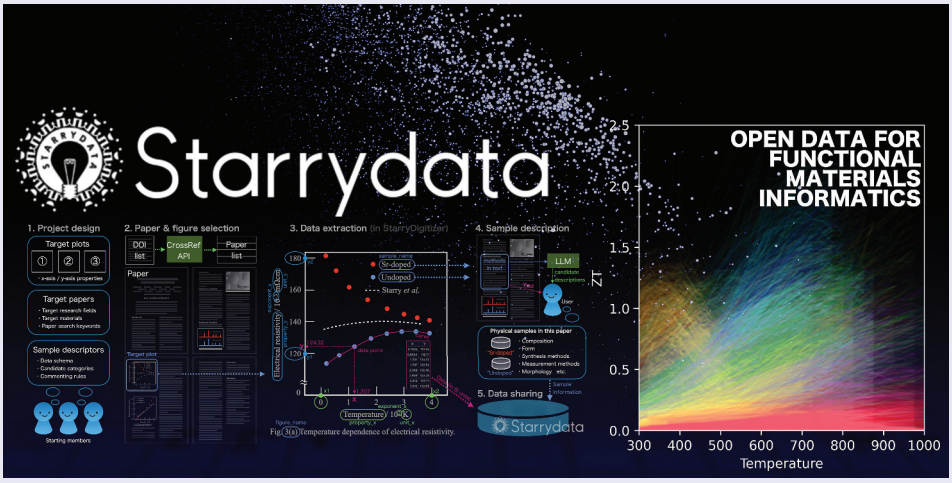
Starrydata: from published plots to shared materials data

Yukari Katsura[a,b,c], Masaya Kumagai[c,d,e], Tomoya Mato[a], Yu Takada[a], Yuki Ando[f], Erina Fujita[g], Fumikazu Hosono[a], Eiji Koyama[a], Farhan Mudasar[h,i,j], Ton Nu Thanh Phuong[k], Naoto Saito[a], Yoshihiro Sakamoto[c], Atsumi Tanaka[a], Dewi Yana[a], Kaoru Kimura[g], Koji Tsuda[a,c,l] and Masahiko Demura[a]

[a]Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [b]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan; [c]RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan; [d]Sakura Internet Research Center, Sakura Internet Inc., Osaka, Japan; [e]Institute for Integrated Radiation and Nuclear Science, Kyoto University, Osaka, Japan; [f]Research Center for Structural Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [g]Department of Advanced Data Science, The Institute of Statistical Mathematics (ISM), Research Organization of Information and Systems, Tokyo, Japan; [h]Department of Physical and Environmental Sciences, University of Toronto Scarborough, Toronto, Canada; [i]Zhongshan Institute of Advanced Cryogenic Technology, Zhongshan, China; [j]Guangdong Green Peak Energy Technology Corporation Limited, Zhongshan, China; [k]Research Center for Magnetic and Spintronic Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [l]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

ABSTRACT

We have developed the Starrydata2 web system, an open, web-based database for collecting and organizing experimental material property data from the literature. It assists users worldwide in extracting and sharing curve data from plot images in published papers, along with relevant sample information such as chemical compositions and fabrication methods. Starrydata2 streamlines the manual data collection process through partial automation. Currently, Starrydata encompasses over 194,000 curves extracted from more than 82,000 physical samples, as reported in over 13,000 publications on functional inorganic materials, including thermoelectric and magnetic materials. All data in Starrydata are openly accessible to the public for both commercial and non-commercial purposes. In this paper, we introduce the web interface, data curation workflow, data structure, and system architecture of Starrydata2. We then described in detail the datasets currently included in Starrydata2 and discuss their use cases. We also present the methods for applying the collected dataset, including a unique large-scale data representation method called 'all-data plots', which provides a comprehensive overview of the entire dataset. Finally, we report on how the collected datasets are being utilized in data-driven materials research through machine learning, modelling and simulation.



IMPACT STATEMENT

Starrydata2 web system is the first open platform dedicated to digitizing and sharing experimental property curves from published materials science literature, enabling preservation of curve shapes for over 194,000 datasets.

ARTICLE HISTORY

Received 14 February 2025
Revised 24 April 2025
Accepted 12 May 2025

KEYWORDS

Materials informatics; database; data mining; plot digitization; open data; literature data; inorganic materials; functional materials; thermoelectric materials; magnetic materials

CONTACT Yukari Katsura KATSURA.Yukari@nims.go.jp Materials Modelling Group, Data-driven Materials Research Field, Center for Basic Research on Materials, National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

1. Introduction

Materials informatics, recognized as the fourth paradigm of materials science [1,2], relies heavily on the availability of comprehensive and accessible datasets [3,4]. Data- driven approaches have made significant progress in both computational predictions and experimental research, yet the popular datasets in material properties are either theoretical [5–11] or obtained from high-throughput experiments [12]. Projects such as the Materials Project [5] and other computational platforms [6,7] have exemplified the power of open data in materials science through their combination of comprehensive coverage, clear licensing, and convenient data access. However, the vast amount of experimental data accumulated in scientific literature, which contains records of the most successful materials developed to date, remains a largely untapped resource.

Incorporating these proven high-performance materials into databases would be invaluable for accelerating future materials development, yet systematic efforts to collect and organize this knowledge face significant challenges.

The CRC Handbook of Chemistry and Physics [13] and the numerous volumes of Landolt-Börnstein [14] (now distributed in Springer Materials [15]) represent the classical approach of curated experimental datasets in chemistry and physics, initially distributed as books before transitioning to digital formats. This transition to digital platforms has led to the emergence of comprehensive proprietary databases, such as the Pauling File [16,17] and the related databases [18–20], which compile experimental physical properties from selected literature, assigned to ideal inorganic crystal structures. The success of computational studies in materials informatics has demonstrated how open data access and convenient bulk download capabilities can accelerate research progress, suggesting similar advances could be achieved with experimental data if made more freely accessible.

The digital transformation has enabled new approaches to experimental data collection and curation. While tools for automatic data extraction from literature, such as ChemDataExtractor [21,22], have made significant advances in processing scientific texts at scale, manual curation remains crucial for ensuring data quality and completeness. The Perovskite Database Project [23], focused specifically on halide perovskite solar cell devices, exemplifies this through their systematic manual extraction of over 42,400 device records from approximately 7,400 experimental papers, representing an intensive effort of 5,000–10,000 person-hours. Similarly, NanoMine [24–26] has established a structured framework for nanocomposite materials data submission through standardized templates. The Materials Project's MPContribs [5,27] provides a platform for tabular experimental data submission, accumulating diverse datasets through researcher contributions worldwide. While these examples represent some of the notable efforts in the field, numerous other valuable initiatives and databases exist across various subfields of materials science.

The field of thermoelectric materials has emerged as a frontier for materials informatics approaches, likely due to its inherent complexity in materials development. In theoretical studies, the TE Design Lab [28], launched in 2015, provided an open GUI (Graphical User Interface) platform for exploring electronic structure parameters such as band gaps, effective masses, and band degeneracies for over 2,000 crystal structures, creating a foundation for researchers to efficiently explore new thermoelectric materials. Ricci et al. further expanded this computational approach in 2018 by publishing a dataset covering 48,000 compounds [29]. These theoretical efforts have been complemented by various experimental data collection initiatives. A pioneering effort was made by Gaultois et al. who published the UCSB thermoelectric database in 2013, providing the first open platform to visually explore thermoelectric properties with approximately 1,000 data points covering four temperature points per sample for over 300 samples from more than 100 papers [30]. Following this path, Na et al. have published ESTM dataset covering thermoelectric properties of 880 published samples [31]. Zang et al. have reported GPTArticleExtractor [32] to apply large language models to extract numeric values from the texts in the published papers, to create databases of 7,123 thermoelectric compounds [33] and 26,706 magnetic materials [34].

Lee et al. have created a data sharing platform named TeXplorer.org to gather over 1,000 experimental data of thermoelectric property curves from the laboratory [35].

While many of these databases and data collection projects primarily gather discrete numerical values from text and tables or manually digitize selected points from plots, some platforms can store curve-based data (collections of x-y coordinate pairs that form experimental curves). Nanomine [24–26], TeXplorer [35] along with data repositories and platforms like Figshare [36], allow users to submit curve-based datasets. However, there have been few initiatives specifically focused on systematically extracting and preserving complete curve data from published plots as their primary objective.

Starrydata is our ongoing project to collect materials data from scientific literature. Following the proof-of-concept demonstration of the original version [37], we developed Starrydata2, a web-based system that enables materials scientists to collaborate globally in collecting materials data from literature. First

introduced in our research paper [38], this system streamlines the extraction of digital data from graphical plots representing materials properties and helps users organize the extracted data by indexing it with details from the original publications. To prevent redundant data extraction efforts, Starrydata2 web system shares the data, along with sample descriptions extracted by users, with all participants. Its data structure is flexible, using a JSON (JavaScript Object Notation)-like document database that allows researchers to customize the initial data structure for sample descriptions, including material classifications, fabrication methods, and experimental techniques.

The name 'Starrydata' was inspired by the night sky, where both bright and dim stars together reveal the true structure of the Milky Way – similarly, our database aims to collect not only well-known, high-quality data but also lesser-known results, as this comprehensive view reveals the true landscape of materials science.

In the following sections, we detail the design and implementation of Starrydata2 web system, the collaborative data collection process, and the integration of advanced tools to aid in data extraction. We also discuss the benefits of open data sharing, the sustainability of our approach, and the potential impact on the field of materials informatics.

2. Data collection process in Starrydata

2.1. Overview of data collection process

Figure 1 illustrates the overview of the paper data collection process in the Starrydata2 web system. This figure represents the typical flow from project planning to data collection and sharing.

The data collection process begins with project design, where users in Starrydata can freely determine the optimal data collection rules for their research. For large-scale data collection projects, it is crucial to decide on key items with the starting members.

First, they select common template plots frequently used in the target field as the target plots. Next, they identify papers likely to include these target plots, referred to as target papers. Finally, they define sample descriptors, which specify how sample information is presented. Setting minimal data collection rules helps in constructing efficient and large-scale datasets.

Following the project design, users select figures from papers. They list potential target papers using a paper search system and send the DOI list to the Starrydata2 web system. The CrossRef [39] API (Application Programming Interface) automatically retrieves bibliographic information, generating a paper list. Users access the full-text PDF of the papers and take screenshots of the target plots when found.

For initial data extraction, users load the plot images into a browser-based digitizer that processes images locally without transmitting them to any external servers. The system performs axis calibration by having users specify reference points on the x- axis and y-axis, and input the physical quantities, units, and scales for each axis. Data points are detected either automatically or manually, and only the extracted numerical values, after conversion to the SI unit system, are transmitted to and saved in the Starrydata database.

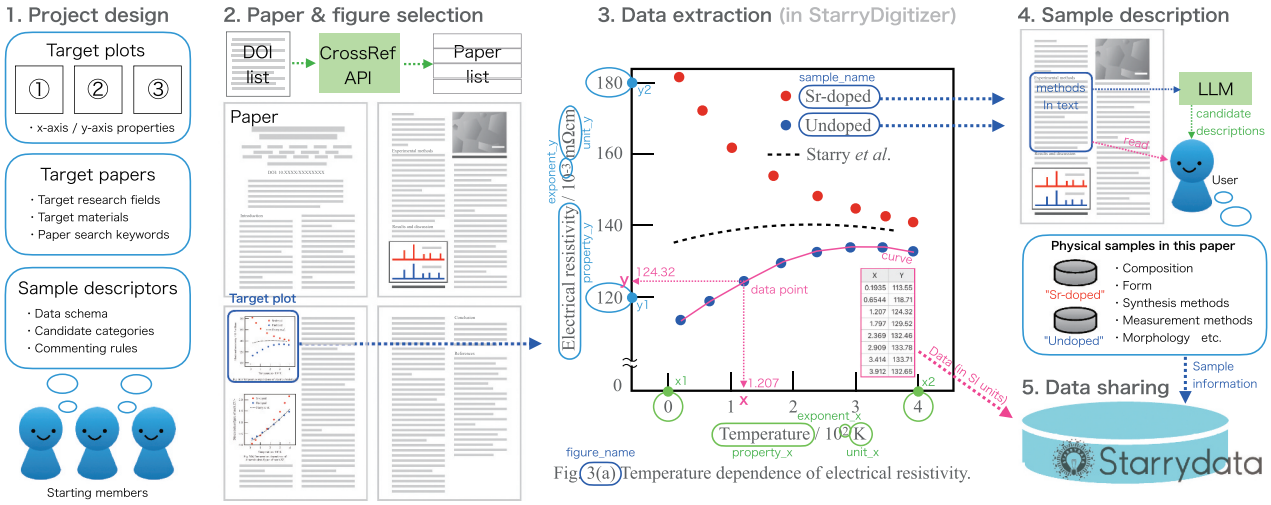


Figure 1. Overview of the paper data collection process in the Starrydata2 web system, illustrating the workflow from project planning through data extraction to data sharing. The process consists of five main stages: (1) project design, where users establish data collection rules and identify target plots, papers, and sample descriptors; (2) figure selection, involving paper search and plot image capture; (3) data extraction, where plot digitization and axis calibration convert visual data to numerical values; (4) sample information entry, which includes detailed sample documentation and entity linking; and (5) data sharing, enabling open access to collected datasets for materials informatics research.

The final stage of collection involves sample information entry. Since sample information is often abbreviated in the plot, users need to read and extract detailed information from the paper's text. Multiple plots in materials science papers often involve the same sample, so these plots are linked to a single sample entity. To assist in this process, the system can suggest sample information entries by analyzing user- provided text from the paper using a LLM (Large Language Model) via external API services.

2.2. Data extraction procedure

This section explains the detailed steps of data extraction and entry using Figures 2 and 3. Figure 2 provides the workflow flowchart for data input, while Figure 3 illustrates the relationships between the various GUIs in Starrydata2, highlighting the Paper GUI as the central hub and showing connections to different pages for detailed data entry.

Starrydata2 employs two types of digitizers to facilitate the extraction of data points from plot images: WebPlotDigitizer [40] and the custom-developed StarryDigitizer. While WebPlotDigitizer offers a broader range of features, StarryDigitizer [41] includes specific functionalities tailored for our data collection needs, such as simplified axis calibration and spline interpolation for accurate curve tracing. Our evaluation using a scatter plot with ~ 300 known data points showed that WebPlotDigitizer's semi-automatic extraction achieved 0.30% precision relative to graph width. The curator optimized parameters

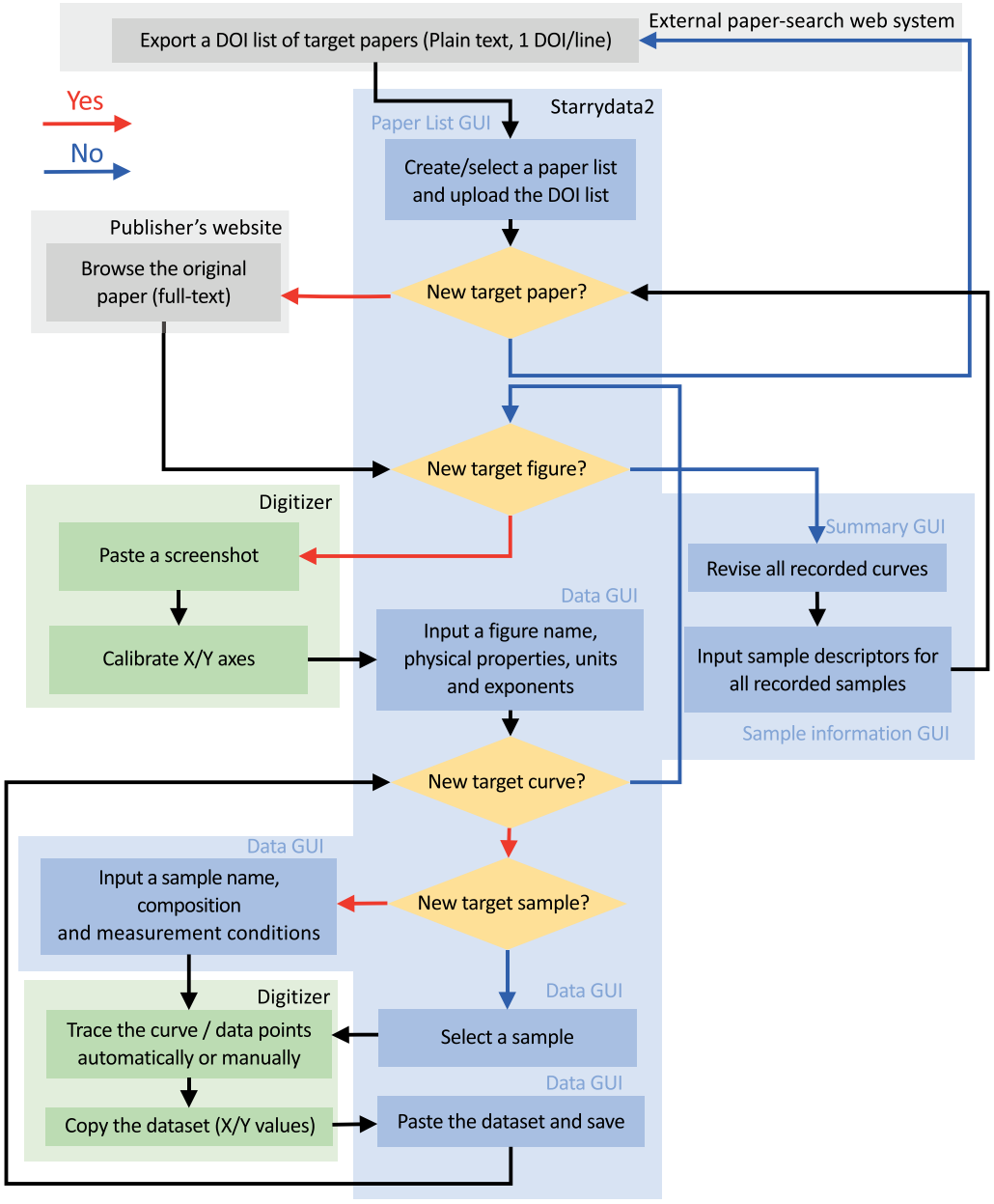


Figure 2. Workflow flowchart for data input. This flowchart illustrates the steps involved in capturing screenshots of target plots, performing axis calibration, entering data points, and saving the data in the starrydata database. Tasks performed in external systems are highlighted in gray, tasks within the Starrydata2 web system GUI in blue, and tasks within the digitizer in green.

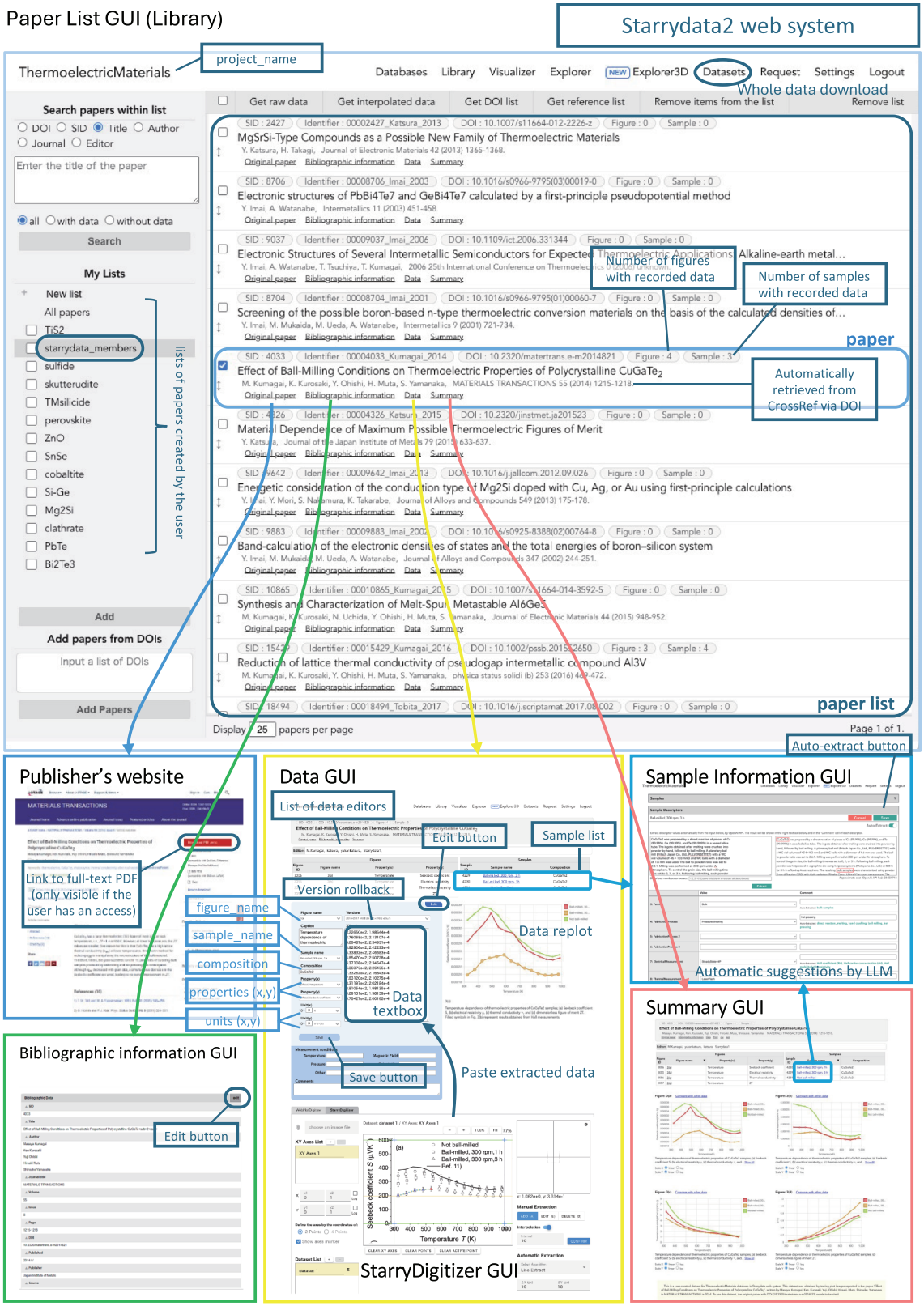


Figure 3. Schematic representation of the GUI interconnections in Starrydata2, with the paper GUI as the central hub. The diagram shows navigation paths between different interfaces and pages for comprehensive data entry, including sample information, bibliographic data, and related functionalities.

(foreground color, color recognition distance, averaging window size) while excluding legends, with minimal errors from overlapping points. Four curators performing manual extraction by simple clicking achieved similar 0.30% precision. Manual adjustment after clicking improved precision to 0.10%, though points near $x = 0$ or $y = 0$ in linear plots occasionally showed errors exceeding 10% of original values.

The data input workflow, as illustrated in Figure 2, involves capturing screenshots of target plots from

papers' PDFs, loading them into the digitizer, and performing axis calibration. Users specify reference points on both axes and input the physical quantities, units, and scales to ensure accurate data representation. The extracted data points are then saved in the Starrydata database after conversion to the SI unit system.

The Paper GUI serves as the central hub for the data collection process, as shown in Figure 3. From this interface, users can access various linked pages to enter detailed sample information, make bibliographic corrections, and input other relevant data. The figure visually represents the flow of navigation, indicating which components to click to move between different pages and functions within the system.

The final stage involves detailed sample information entry, where data points are linked to their corresponding samples and figures. Since sample information is often abbreviated in the plots, users must read the paper's text to gather complete information. Multiple plots involving the same sample are linked as the same entity in the database.

Recent developments include the use of large language models (LLM) to assist in this task, improving the accuracy and efficiency of data entry.

Through this structured process, the collected data becomes available as open data to other users, preventing redundant data collection and promoting various materials informatics research. While dedicated curators currently handle most data collection, voluntary contributions from users are essential for expanding the dataset. Data saved by external users are immediately incorporated into our open dataset without mandatory prior review. We do not require our curators to verify all saved data because comprehensive checking would significantly slow down the data collection process.

While we trust users to add and modify data honestly, we have implemented a robust version control system that allows us to track and revert any inappropriate changes made by users with malicious intent. For quality assurance, we have established additional review processes specifically targeting outlying data. Our approach represents a carefully considered balance between data collection speed and quality.

While our system is primarily designed for collecting multiple data points from plot images, it can also accommodate single data points from tables or text. For table data, users may enter the table name in the 'Figure name' field. Single data points can also be registered as custom sample descriptors. This flexibility allows researchers to incorporate various types of published information from scientific papers into our database.

We designed our data collection workflow to allow the uses of various semi- automatic tools while maintaining the requirement for data curators to verify outputs before data storage. We have implemented a textbox in the sample descriptor page where users can paste relevant paragraphs describing experimental methods to get suggestions from an LLM. Currently, users can add their personal OpenAI API key to enable suggestions for descriptor entries using the latest model (without fine-tuning). However, given the rapid evolution of LLM technologies, we avoid including specific details, as such information would likely become outdated in the short term.

Furthermore, we must consider that some major publishers including ACS [42], Elsevier [43], and Wiley [44] currently restrict AI use on their provided content. While anticipating changes to these restrictions in future, we currently maintain our approach where researchers and curators read papers for their own projects and record extracted data in our web system, which is then shared with other users as open data.

3. System architecture and data management

3.1. System architecture

This section provides an overview of the Starrydata2 web system architecture. Figure 4 illustrates the comprehensive structure, including the various GUIs and databases within Starrydata2 (shown in blue), external web systems (shown in gray), and human involvement such as data curators (shown in green).

The left half of Figure 4 outlines the steps from data extraction to storage in the database, represented by numbers 1–19. The data in Starrydata2 is stored in a document- oriented database, MongoDB [45,46]. Tables such as Paper, Figure, Sample, Sample Information, and Data are linked via IDs, functioning similarly to a relational database. Flexible data structures defined by users, such as Sample Information and Measurement Conditions, are stored in JSON format to maintain flexibility. Copyrighted content like full-text papers and plot images are never sent to the system. The database stores bibliographic information, image names, numerical data extracted by users, summarized sample information, selected figure descriptions, and other metadata that objectively describe the research content.

The right half of Figure 4 illustrates the data viewing and output workflow. Data curators can review multiple datasets through the Summary, Data, and Visualizer GUIs, with interactive visualizations powered by Chart.js [47], Bokeh [48], and Beautiful Soup [49]. The menu bar provides access to our external data explorers developed using Streamlit [50] and Plotly [51].

The backend system was built with Django [52] and its extensions (Django- nonrel [53], Django MongoDB

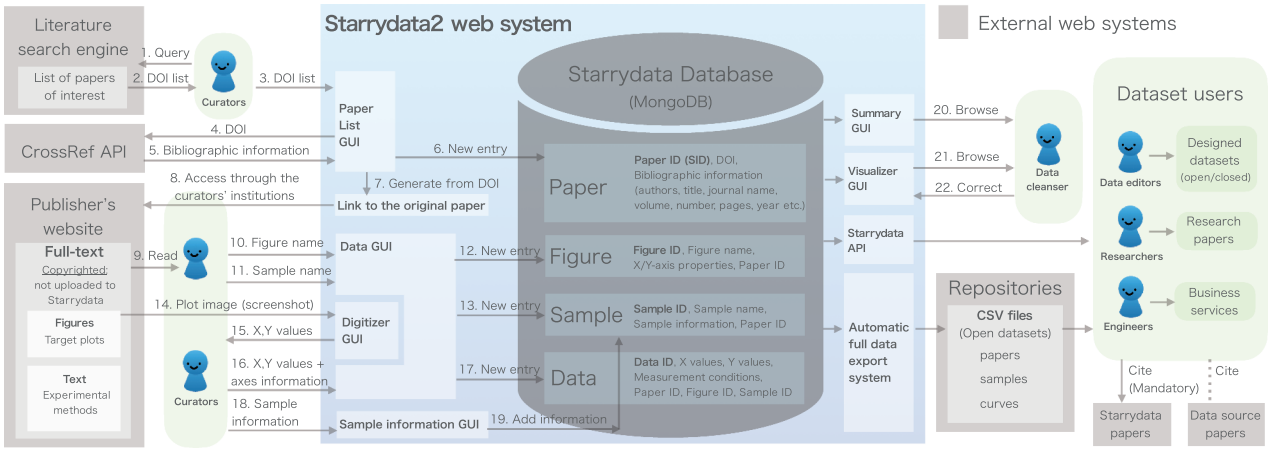


Figure 4. System architecture of Starrydata2, showing the complete workflow from data input to output visualization. The diagram illustrates three main components: (1) the internal Starrydata2 system components including GUIs and MongoDB databases (blue), (2) external web systems (grey), and (3) human interactions (green). The left side depicts the data collection and storage process, while the right side shows the data visualization and output workflow through the summary and visualizer GUIs. The system employs a document-oriented database structure with ID-linked collections and JSON-formatted flexible data schemas, storing bibliographic information, extracted numerical data, summarized metadata including sample information and selected figure descriptions, while ensuring copyright compliance by excluding copyrighted content like full-text papers and plot images.

Engine [54], and Djangotoolbox [55]), while the frontend utilized Vue.js [56] and Vue Select [57]. Server-side data processing and validation rely on Python libraries including NumPy [58,59], pandas [60], and pymatgen [61,62].

Administrators can perform direct data modifications through a PyMongo-based command-line interface [63]. System configuration and libraries may be updated in future refactoring.

3.2. Data management

Starrydata2 provides multiple methods to access its data: a REST (Representational State Transfer) API for flexible queries, bulk downloads for comprehensive datasets, and a Python package for easy integration into data analysis workflows.

The REST API enables users to retrieve data by Paper ID (SID: Starrydata ID), Figure ID, or Sample ID in JSON format. For example, to search for samples containing specific elements like Mg and Si, users can make a request to: https://www.starrydata2.org/api/sample?atom=Mg,Si.

This returns a list of Sample IDs. Detailed information for each sample can then be retrieved using these IDs: https://www.starrydata2.org/api/sample/2212.

The same query pattern applies to papers and figures by replacing 'sample' with 'paper' or 'figure' in the URI.

For materials informatics research requiring large-scale data analysis, Starrydata provides bulk download options accessible through the 'Datasets' link in the menu bar. This leads to a GitHub repository page containing links to daily-updated datasets on Google Drive and monthly archived versions on Figshare37, facilitating efficient access to comprehensive data snapshots.

To further simplify data access, Starrydata offers a Python package that handles the data loading process. After installation via pip, users can load the data into pandas dataframes with just a few lines of code:

```
import starrydata as sd
import pandas as pd
sd_dataset = sd.load_dataset(date='20250201')
df_curves = pd.read_csv(sd_dataset.curves_csv)
df_samples = pd.read_csv(sd_dataset.samples_csv)
df_papers = pd.read_json(sd_dataset.papers_json)
```

4. Collected data and examples

4.1. Overview of data collection projects

Table 1 provides a summary of the open data collection projects actively operating within Starrydata. These projects, referred to as 'Databases' in the Starrydata2 GUI, are established in collaboration with the Starrydata operating team. For each project, the table shows the numbers of papers, figures, samples and curves collected so far, along with the top six plot types presented as combinations of physical quantities for x and y axes.

The General DB is a generic project that provides access to all collected data. Other projects, such as the ThermoelectricMaterials and MagneticMaterials projects, use customized sample information templates with specific descriptors (such as Form and Synthesis method) that can be populated through dropdown

Table 1. Summary of data records in major active collection projects in Starrydata as of January 15, 2025. For each project, the table shows the number of curves and physical quantity combinations for the top six plot types.

	Number of records			Top 6 plot types		Number of figures		Number of curves	
Project name	Record type	Total	Unique	*x*-axis	*y*-axis	Total	Unique	Total	Unique
GeneralDB	Papers	13,210	53	Temperature	Seebeck coefficient	8,845	52	33,941	201
	Figures	54,642	88	Temperature	Thermal conductivity	6,152	2	24,340	6
	Samples	82,006	170	Temperature	*ZT*	4,671	1	19,590	5
	Curves	194,053	277	Temperature	Electrical resistivity	5,395	4	18,880	5
				Temperature	Electrical conductivity	4,330	2	17,098	5
				Temperature	Power factor	3,831	2	16,277	4
Thermoelectric	Papers	9,004	8,032	Temperature	Seebeck coefficient	8,821	8,157	33,845	30,968
Materials	Figures	37,970	34,049	Temperature	Thermal conductivity	5,566	5,015	22,289	19,936
	Samples	52,326	46,749	Temperature	*ZT*	4,662	4,219	19,564	17,458
	Curves	147,993	132,314	Temperature	Electrical resistivity	4,926	3,883	17,338	15,330
				Temperature	Electrical conductivity	4,188	4,235	16,661	15,044
				Temperature	Power factor	3,828	3,533	16,254	14,908
Magnetic	Papers	1,770	1,761	Magnetic field strength (H)	magnetization_per_weight	1,828	1,818	6,042	6,019
Materials	Figures	3,770	3,748	Magnetic field	magnetization_per_weight	567	564	3,626	3,616
	Samples	13,011	12,965	Magnetic field strength (H)	Magnetization	452	448	1,269	1,259
	Curves	13,293	13,233	Magnetic field strength (H)	magnetization_per_volume	409	409	1,040	1,040
				Temperature	magnetization_per_weight	215	215	629	629
				Magnetic field	magnetization_per_volume	100	100	268	268
Battery	Papers	1,418	1,346	Discharge capacity	Voltage	2,336	2,234	5,721	5,480
Materials	Figures	10,697	10,350	Charge capacity	Voltage	2,298	2,196	5,653	5,411
	Samples	10,796	10,436	Cycle number	Discharge capacity	2,401	2,326	5,118	4,957
	Curves	25,108	24,308	C rate	Discharge capacity	1,049	992	2,402	2,269
				Cycle number	Voltage	820	820	1,849	1,849
				Time	Voltage	805	805	1,832	1,832
Condensed	Papers	794	342	Temperature	Electrical resistivity	810	329	2,553	1,147
Matter	Figures	1,856	776	Temperature	Seebeck coefficient	168	47	677	257
	Samples	4,168	2,261	Temperature	Thermal conductivity	104	48	409	192
	Curves	6,572	2,840	Temperature	Power factor	83	69	364	185
				Temperature	*ZT*	75	34	326	133
				Temperature	Electrical conductivity	110	21	312	111
HighThermal	Papers	349	238	Temperature	Thermal conductivity	471	328	1,663	1,165
Conductivity	Figures	702	357	Temperature	Seebeck coefficient	59	15	250	31
Materials	Samples	1,722	1,170	Temperature	*ZT*	40	3	173	15
	Curves	2,586	1,231	Temperature	Electrical conductivity	25	8	112	10
				Temperature	Electrical resistivity	29	1	110	5
				Temperature	Power factor	26	1	104	4
LowThermal	Papers	268	134	Temperature	Thermal conductivity	319	169	1,225	618
Conductivity	Figures	822	205	Temperature	Seebeck coefficient	107	6	473	17
Materials	Samples	1,601	647	Temperature	*ZT*	101	5	464	15
	Curves	3,451	705	Temperature	Power factor	67	2	321	10
				Temperature	Electrical conductivity	67	2	303	10
				Temperature	Lattice thermal conductivity	54	7	241	9
Hypermaterial	Papers	239	195	Temperature	Electrical resistivity	176	138	488	387
	Figures	682	482	Temperature	Electrical conductivity	88	70	298	238
	Samples	1,223	951	Temperature	Seebeck coefficient	68	59	253	180
	Curves	2,046	1,398	Temperature	Thermal conductivity	87	44	240	124
				Temperature	Magnetic susceptibility	59	22	180	90
				Temperature	*ZT*	20	16	79	44

The GeneralDB project encompasses all data across projects. Since a single record may be associated with multiple projects, the number of unique records (those belonging exclusively to a single project) is also indicated.

categories and detailed comments. Each paper in Starrydata is linked to one or more projects, which determines its visibility within those projects. When users open a specific project and select 'All Papers' or perform a search within that project, they see only the papers linked to that project. This project-based filtering helps researchers focus on relevant papers within their field while maintaining access to all associated experimental data, including samples, figures, and digitized curves from each visible paper.

New projects are created by Starrydata's administrative engineers when research projects focusing on specific functional materials begin, whether funded by public grants or industry collaborations. While most projects are openly accessible, Starrydata also supports closed data collection on separate servers for certain collaborative projects, particularly with industry partners, with the understanding that the data will eventually be made public. Data curators and users can add papers to existing projects through the 'Add papers' function by inputting DOI lists. When adding papers by DOI, if a paper already exists in the database, it automatically becomes linked to the new project, while 'Search papers' function is used only for browsing existing content.

While papers are generally linked to relevant projects through their DOIs (with rare exceptions due to DOI duplications in CrossRef), it is common for users to add papers to projects regardless of their thematic relevance, and this practice is not restricted by the system administrators.

Each project's status is displayed, including the number of all registered papers (including the papers waiting for data collection), papers with collected data, figures, and samples, as well as the number of paper lists created by users within the project. Projects with at least one paper list created by the user are marked as active, and the user's frequently used projects are shown at the top.

4.2. Case study: thermoelectric materials project

The ThermoelectricMaterials Project aims to collect and organize comprehensive experimental data on thermoelectric materials. This project is one of the largest within Starrydata, focusing on thermoelectric properties such as Seebeck coefficient, electrical conductivity (resistivity), thermal conductivity, power factor, and the dimensionless figure of merit (ZT).

The initial set of target papers was obtained by searching Scopus using the keyword 'thermoelectric properties', specifying the research field as 'Materials Science'. We retrieved the search results as CSV file, including bibliographic information, to obtain the DOI of candidate papers. Subsequently, data curators have attempted various search keywords to obtain efficient lists of papers, to find additional papers containing the target plots.

Figure 5 provides an example of a specific paper on thermoelectric materials and the data collected from it in the ThermoelectricMaterials Project. The left side of Figure 5 shows the original PDF of the paper, which is not stored in Starrydata. Instead, the bibliographic information is stored in the Paper table, including details such as authors and titles, mostly obtained directly from CrossRef.

The data collected from each paper is organized into several interconnected tables in the Starrydata database. The properties table contains standardized information about the physical quantities (parameters and properties) used throughout Starrydata.

The figure table contains information about each figure extracted from the papers, including a unique

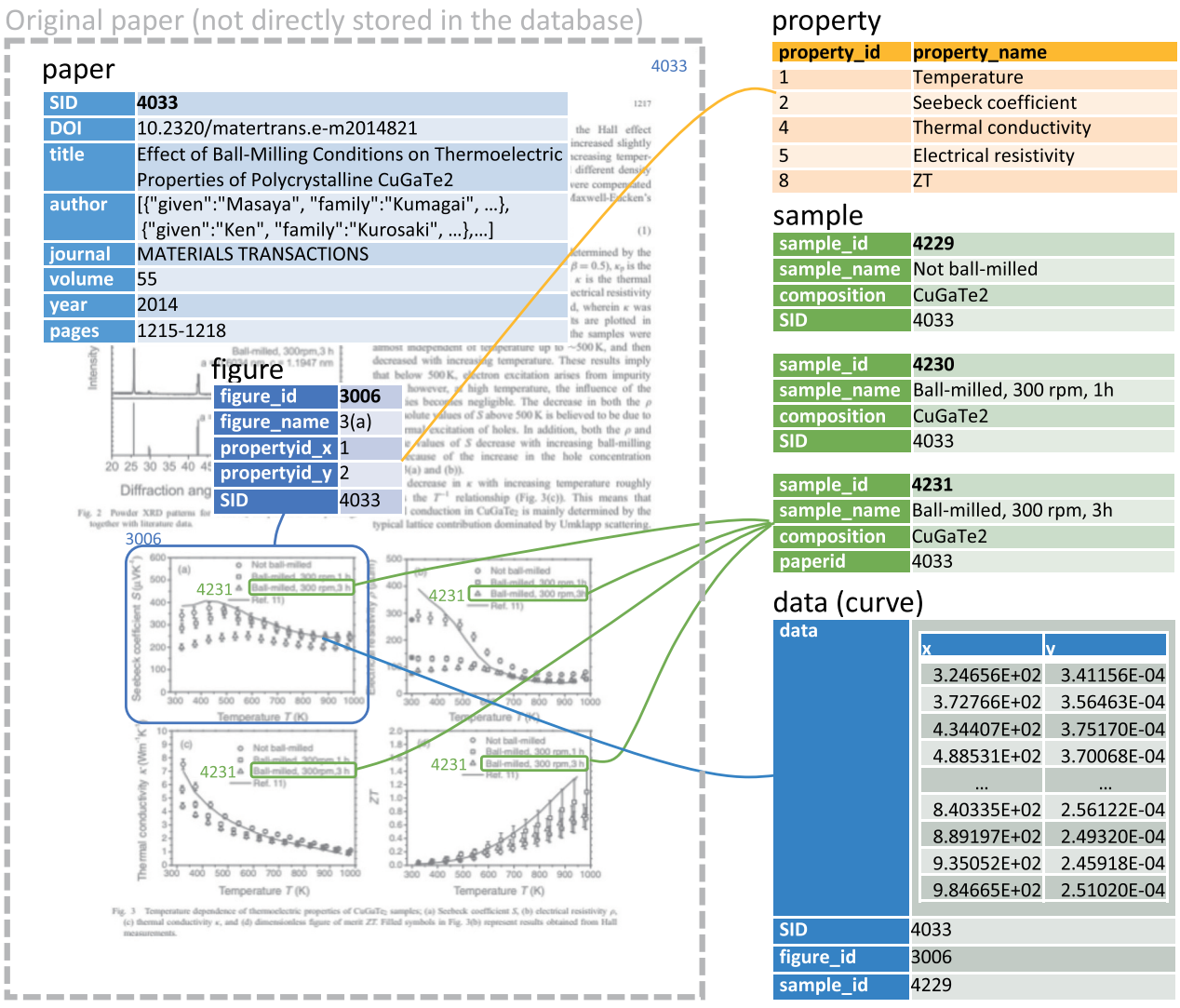


Figure 5. Example of a specific paper on thermoelectric materials [64] and the data collected from it in the ThermoelectricMaterials project. The figure shows the connections between different tables in the database and the process of data extraction and entry. Reproduced with permission from thermoelectrics society of japan.

figure ID, figure name (based on the figure number in the original paper), and the physical quantities represented on the x and y axes. The sample table contains detailed information about the samples described in the papers, including sample ID, sample name (based on the nomenclature used in the paper), and the comprehended molar composition. The data table contains the 2-dimensional numerical data points extracted from the curves in the figures. Each curve is linked to a specific figure ID and sample ID. The data is stored in SI units, without prefixes, as defined in the dimensions in the properties table.

5. Application of the dataset

5.1. Data preprocessing

The Starrydata2 web system collects diverse experimental data, and we suggest several preprocessing steps to help users effectively utilize this dataset. While these steps are flexible and can be adapted to specific research needs, they represent our recommendations based on experience with the data.

First, we recommend filtering the data by project names and axis quantities to focus on curves relevant to your analysis. Next, carefully examine potential outliers. Our experience shows that the most critical outliers often arise from mistakes in units or exponents. While we continuously review and revise the registered data, we occasionally find cases where calculation results were mistakenly registered as experimental data, or where unit and notation errors in original papers passed through peer review processes unnoticed.

Chemical compositions in Starrydata are provided as molar ratio strings, which our data curators have converted from various original formats including mass ratios, volume ratios, and general formulas where possible. Users can parse these strings into structured data using libraries like pymatgen [62] for systematic analysis and classification of material systems.

When comparing datasets with different x-value sampling points (whether from different publications or different measurements within the same paper), we recommend users carry out appropriate interpolation as part of their analysis workflow. Depending on the data characteristics, users might select polynomial, spline, or other specialized interpolation equations that capture the behavior of physical properties. This interpolation transforms array-type curve datasets into grid-structured tabular datasets, making them significantly more amenable to data analysis, visualization, and machine learning applications.

Finally, users can enrich their analysis by combining information from our supplementary files. The starrydata_samples.csv contains categorical and free-text metadata about samples, while starrydata_papers.csv provides bibliographic context. While some fields may be empty due to our optional input policy, this additional information can provide valuable context for research. The latest dataset is available for download from the Datasets link on the Starrydata web system.

These preprocessing steps are intended to help transform the experimental data into a more analyzable format, supporting users in their research objectives.

5.2. Visualization example: all-data plots

Figure 6 showcases an unprecedented visualization approach, overlaying thousands of measurement curves standardized from diverse original units and scales. Each material system forms characteristic bundle-like distributions, revealing the inherent property variations in materials science.

The color-coding in these plots reflects material classifications based on compositional thresholds. For thermoelectric materials, we grouped samples using atomic ratios, allowing for site substitution within element families (e.g. S+Se+Te > 0.3 for chalcogenides) and accounting for dopants by setting thresholds below unity.

Similarly, magnetic materials were classified using criteria such as Fe > 0.6 for Fe-rich compounds and B > 0.02 for borides. This classification approach effectively captured material families while accommodating chemical variations and doping.

Figures 6(a–e) present thermoelectric transport properties color-coded by anion types and major material systems. The electrical conductivity plots (a) show that tellurides and antimonides tend to exhibit higher conductivity, while oxides show lower values, with sulfides/selenides falling in between. The Seebeck coefficient plot (b) separates p-type (positive values) and n-type (negative values) materials, revealing distinct material tendencies: while Bi_2Te_3 and PbTe systems show both *p*- and n-type behavior, tellurides and antimonides are predominantly p-type, possibly due to their defect formation tendencies. Notably, p-type Bi_2Te_3 systems show exceptionally large Seebeck coefficients near room temperature, which appears to be the key factor in their superior performance rather than their moderate electrical and thermal conductivity values. Sulfides/selenides and silicides tend toward n-type behavior. Among oxides, Co- O systems form a consistent cluster in the p-type region, while n-type oxides show greater diversity. The thermal conductivity (c) is generally lower in tellurides and antimonides.

These characteristics combine to influence the power factor (d) and figure of merit ZT (e). Bi_2Te_3-based materials demonstrate superior performance near room temperature, with measurements typically stopping around 500 K to avoid thermal decomposition. In higher temperature ranges, compounds meeting our Tellurides (Te

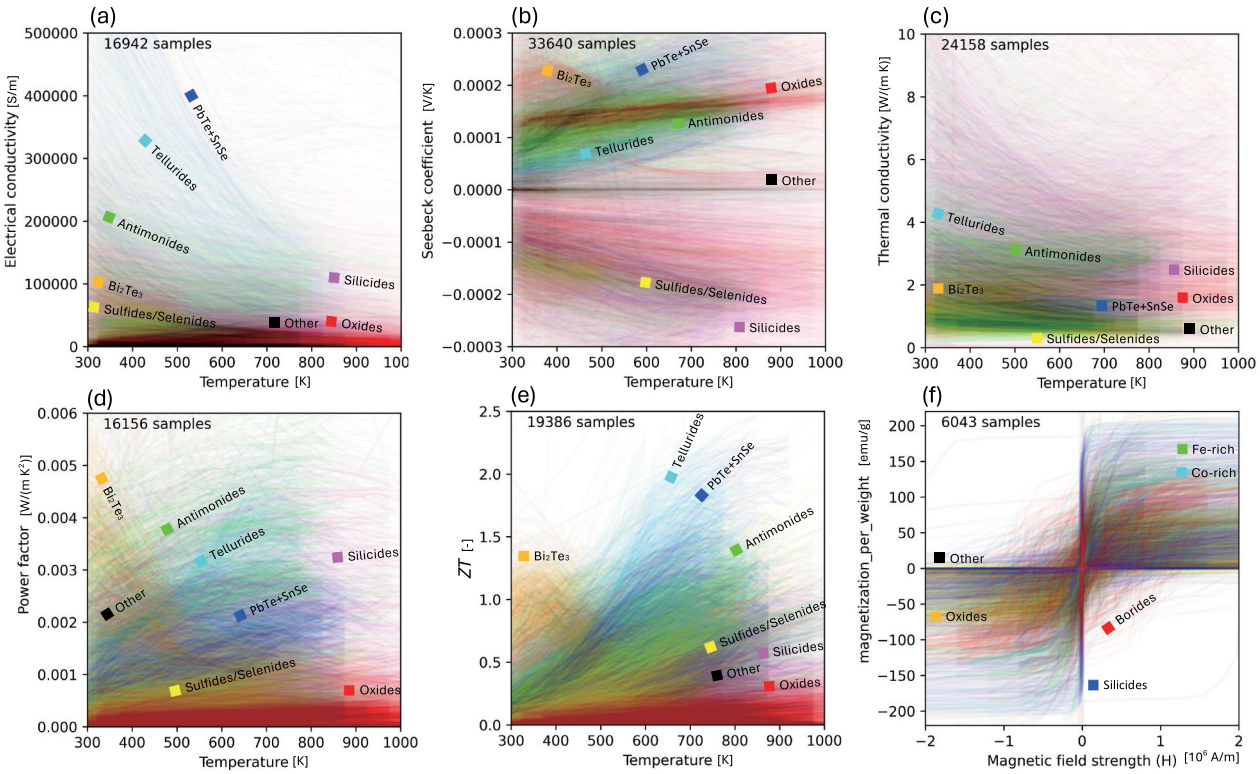


Figure 6. 'All-data plots' showing large-scale visualization of transport and magnetic properties through standardized plotting of experimental data. (a-e) temperature dependence of thermoelectric properties: (a) electrical conductivity, (b) Seebeck coefficient, (c) thermal conductivity, (d) power factor, and (e) dimensionless figure of merit *ZT*. Materials are classified based on atomic ratios: antimonides (Sb > 0.3, lime), Silicides (Si+Ge+Sn > 0.3, magenta), chalcogenides (S+Se+Te > 0.3, cyan), Bi_2Te_3 (Bi+Sb > 0.35 and S +Se+Te > 0.55, orange), PbTe (Pb+Sn > 0.45 and S+Se+Te > 0.45, blue), Sulfides/Selenides (S+Se > 0.3, yellow), and oxides (O > 0.3, red). (f) Magnetization hysteresis curves for magnetic materials, classified as: oxides (O > 0.2, yellow), non-oxides (O < 0.01, blue), Fe-rich (Fe > 0.6, lime), Co-rich (Co > 0.6, cyan), borides (B > 0.02, red), and nitrides (*N* > 0.02, magenta).

> 0.3) and PbTe classification criteria (Pb+Sn > 0.45 and S+Se+Te > 0.45) exhibited the highest performance, though it should be noted that this PbTe group includes both PbTe-based materials and structurally distinct SnSe-based systems [65,66].

The magnetization hysteresis curves (f) reveal that oxide-based materials, particularly ferrites and spinels, dominate the dataset despite their moderate saturation magnetization. Silicide-based materials like Mn-Si show soft magnetic properties with narrow coercivity (distance between x-intercepts) and high magnetization. While fewer in number, boron-containing magnets (Sm-Co-B and Nd-Fe-B systems) registered as magnetization per weight demonstrate notably high coercivity.

These distributions, visualized here at an unprecedented scale, represent the true nature of materials properties – not as single definitive values, but as ranges that emerge from the complex interplay of synthesis conditions, measurement methods, and intrinsic material characteristics.

5.3. Machine learning

Machine learning applications using our dataset have been pursued in various ways to predict candidate materials and improve methodologies [67–78]. Numerous studies have utilized these datasets for supervised learning tasks, such as predicting thermoelectric properties by using the chemical composition of samples as features and thermoelectric characteristics as target variables. By inputting a list of candidate compositions into a trained model, it is possible to select compositions that can achieve high performance, although the accuracy may vary depending on the distribution of the training data.

These selections mimic the intuitive judgment of veteran researchers, allowing novices in the field to gain an understanding of which compositions appear promising.

As illustrated in the previously mentioned all-data plots, the reproducibility of thermoelectric properties is low, and the relationship between composition and properties is not one-to-one. Therefore, attempts to synthesize compositions predicted by machine learning models to evaluate prediction accuracy or to improve prediction accuracy by including them in the training data may not be considered very meaningful. In such cases, the main contribution lies in providing a catalog for experimental materials scientists to select candidate materials of interest. However, discussions and the development of evaluation methods regarding the accuracy of these machine learning

models are progressing, and if superior prediction methods are developed in the future, it may be possible to overcome reproducibility issues.

The combination of our dataset with other informative approaches is expected to lead to more precise predictions of new materials. For instance, Ren et al. used our dataset on Zintl phases in conjunction with DFT-based calculations, resulting in the experimental discovery of new low-thermal-conductivity Zintl phases [77].

5.4. Modeling and simulations

Starrydata's extensive dataset has enabled innovative applications in modelling and simulations. For instance, Snyder et al. developed model equations to infer electronic structure parameters, such as weighted mobility [79] and effective mass [80], using readily measurable properties like the Seebeck coefficient and electrical resistivity.

They leveraged Starrydata's comprehensive database to validate these theoretical models against experimental results. In another significant application, Ryu et al. developed a precise integral-form model for calculating thermoelectric module efficiency and systematically evaluated maximum achievable efficiencies using various combinations of p-type and n-type materials from the Starrydata database [81]. These examples demonstrate how an open data platform like Starrydata empowers researchers to explore materials science from novel perspectives, enabling innovative theoretical approaches and accelerating materials development through data-driven insights.

5.5. Development of new tools and datasets

Starrydata2's web system serves as a platform that enables researchers to collect, analyze, and redistribute literature data in new ways. We encourage the research community to create and share specialized datasets using our open data. In our research group, Fujita et al. demonstrated this approach by creating HYPOD-X [82], an open dataset focused on quasicrystal-related materials. This specialized dataset has facilitated the analysis of unconventional electrical resistivity patterns and the identification of candidate materials for novel thermal diode applications [83].

To demonstrate approaches for data visualization, we have developed several complementary tools that build upon our open datasets. The Starrydata Sample Explorer, which currently focuses on thermoelectric materials, provides an interactive interface for two-dimensional visualization, sample searching, clustering analysis, and scatter plot generation based on a specific snapshot of the thermoelectric materials database. Starrydata Explorer 3D serves as a curated gallery of pre-generated three- dimensional scatter plots and composition distribution maps for thermoelectric materials, using a different subset of the database to offer unique perspectives on material properties. For comprehensive data exploration, the Visualizer GUI generates interactive plots automatically for most possible combinations of x and y variables from the daily-updated database, providing users with a versatile and dynamic way to explore relationships between various material properties. All these visualization tools are readily accessible through Starrydata2's menu bar, and we are actively expanding their functionalities to provide users with increasingly powerful analytical capabilities.

While detailed descriptions of these visualization tools are beyond the scope of this paper, their development exemplifies how Starrydata2's open architecture can catalyze the creation of specialized analytical tools. We believe that the continued community-driven development of both curated datasets and analytical tools will significantly accelerate materials research by providing researchers with increasingly sophisticated ways to explore, analyze, and utilize materials data.

6. Conclusion

Starrydata2 has established a robust framework for collecting, managing, and sharing experimental materials data extracted from plot images in published literature. Rather than relying on fully automated extraction, our web-based system facilitates comprehensive interaction between data curators and source papers, which has proven effective in streamlining data handling processes, expanding dataset coverage, and ensuring high data quality. While collaborative projects with industry partners and academic institutions have been the primary drivers of database growth thus far, Starrydata2 is designed to embrace contributions from individual users who can add data from previously unrecorded papers, as well as organized data collection initiatives. We anticipate that this community-driven approach, combined with our existing partnerships and ongoing development of data collection tools, will significantly expand the breadth and depth of our dataset coverage. The open datasets provided through Starrydata2 have been successfully utilized in various materials informatics projects, enabling researchers to develop new concepts, gain comprehensive field overviews, and implement machine learning models for predicting experimental properties. This demonstrates Starrydata2's significant contribution to advancing data-driven materials research.

Acknowledgements

We express our gratitude to Masayuki Fujimoto, Sakiko Gunji, Yoji Imai, Takushi Kodani, Shunji Kohri, Hideyasu Ouchi, and Kazuki Tobita for their contributions to data curation. We also thank the anonymous crowd workers and the voluntary users of Starrydata for sharing their extracted data. We deeply thank all the collaborators from academia and industries for financial supporting our staffs including the data curators.We are grateful to all those who provided valuable advice and suggestions throughout this work, and to those who offered opportunities for disseminating our research.

Author contributions

Y. Katsura designed the concept of the Starrydata project, managed the team and analysed the dataset. She wrote the main manuscript under the guidance of M. Demura. Starrydata2 and the related web systems were developed by M. Kumagai, T. Mato, and Y. Takada. Data collection and supportive analyses were performed by Y. Ando, E. Fujita, F. Hosono, E. Koyama, F. Mudasar, T. N. T. Phuong, N. Saito, Y. Sakamoto, A. Tanaka, and D. Yana. Project management by Katsura was supported by A. Tanaka, K. Kimura, K. Tsuda, and M. Demura.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by 'Materials Research by Information Integration' Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub from the Japan Science and Technology Agency (JST), JST-CREST (Grant JPMJCR19J1), KAKENHI (Grants 16K14379, 19K04999, 19H05818, and 19H05820), Watanabe Memorial Foundation, Research Association of Automobile Internal Combustion Engines, the Kazuchika Okura Memorial Foundation, and our industry partners.

Data availability statement

The experimental data obtained from the publications can be freely downloaded from the link in the menu bar of our Starrydata2 web system at http://www.starrydata2.org.

References

[1] Hey T. In: Tansley S, and Tolle K, editors. The fourth paradigm: data-intensive Scientific discovery. 1st ed. Redmond (WA): Microsoft Research; 2009.
[2] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater [Internet]. 2016 [cited 2024 Jun 6];4(5). doi: 10.1063/1.4946894
[3] Ottomano F, De Felice G, Gusev V, et al. Not as simple as we thought: a rigorous examination of data aggregation in materials informatics. ChemRxiv [Internet]. 2023 [cited 2023 Nov 8]. doi: 10.26434/chemrxiv-2023-r9n12
[4] Xu P, Ji X, Li M, et al. Small data machine learning in materials science. Npj Comput Mater. 2023;9(1):1–15. doi: 10.1038/s41524-023-01000-z
[5] Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater [Internet]. 2013 [cited 2024 Jun 6];1(1). doi: 10.1063/1.4812323
[6] Curtarolo S, Setyawan W, Hart GLW, et al. AFLOW: an automatic framework for high-throughput materials discovery. Comput Mater Sci. 2012;58:218–226. doi: 10.1016/j.commatsci.2012.02.005
[7] Saal JE, Kirklin S, Aykol M, et al. Materials design and discovery with high- throughput density functional theory: the open quantum materials database (OQMD). JOM. 2013;65(11):1501–1509. doi: 10.1007/s11837-013-0755-4
[8] Pizzi G, Cepellotti A, Sabatini R, et al. AiiDA: automated interactive infrastructure and database for computational science. Comput Mater Sci. 2016;111:218–230. doi: 10.1016/j.commatsci.2015.09.013
[9] Draxl C, Scheffler M. The NOMAD laboratory: from data sharing to artificial intelligence. J Phys Mater. 2019;2(3):036001. doi: 10.1088/2515-7639/ab13bb
[10] Stevanović V, Lany S, Zhang X, et al. Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. Phys Rev B Condens Matter. 2012;85(11):115104. doi: 10.1103/PhysRevB.85.115104
[11] Choudhary K, Garrity KF, Reid ACE, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. Npj Comput Mater. 2020;6(1):1–13. doi: 10.1038/s41524-020-00440-1
[12] Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials. Sci Data. 2018;5(1):180053. doi: 10.1038/sdata.2018.53
[13] Weast RC. CRC handbook of chemistry and physics. 1973.
[14] Landolt-Bornstein. Numerical data and functional relationships in science and technology. Berlin (Germany): Springer-Verlag; 1961.
[15] SpringerMaterials – properties of materials [Internet]. [cited 2024 Jun 19]. Available from: https://materials.springer.com/
[16] Villars P, Berndt M, Brandenburg K, et al. The pauling file. Mater Sci For. 2004;443–444:357–360. doi: 10.4028/www.scientific.net/MSF.443-444.357
[17] Villars P, Phases Data System M, Cenzual K, et al. PAULING FILE - towards a holistic view. Chem Met Alloy. 2018;11(3/4):43–76.
[18] Xu Y, Yamazaki M, Villars P. Inorganic materials database for exploring the nature of material. Jpn J Appl Phys. 2011;50(11S):11RH02. doi: 10.1143/JJAP.50.11RH02
[19] AtomWork-Adv [Internet]. [cited 2024 Jun 19]. https://atomwork- adv.nims.go.jp/
[20] Materials Platform for Data Science [Internet].MPDS.IO. [cited 2024 Jun 19]. Available from: https://mpds.io/tutorial/
[21] Swain MC, Jm C. ChemDataExtractor: a toolkit for automated extraction of chemical information from

the scientific literature. J Chem Inf Model. 2016;56 (10):1894–1904. doi: 10.1021/acs.jcim.6b00207
[22] Mavračić J, Court CJ, Isazawa T, et al. ChemDataExtractor 2.0: autopopulated ontologies for materials science. J Chem Inf Model. 2021;61 (9):4280–4289. doi: 10.1021/acs.jcim.1c00446
[23] Jacobsson TJ, Hultqvist A, García-Fernández A, et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nat Energy. 2021;7(1):107–115. doi: 10.1038/s41560-021-00941-3
[24] Zhao H, Li X, Zhang Y, et al. Perspective: NanoMine: a material genome approach for polymer nanocomposites analysis and design. APL Mater. 2016;4 (5):053204. doi: 10.1063/1.4943679
[25] Zhao H, Wang Y, Lin A, et al. NanoMine schema: an extensible data representation for polymer nanocomposites. APL Mater. 2018;6(11):111108. doi: 10.1063/1.5046839
[26] Brinson LC, Deagen M, Chen W, et al. Polymer nanocomposite data: curation, frameworks, access, and potential for discovery and design. ACS Macro Lett. 2020;9(8):1086–1094. doi: 10.1021/acsmacrolett.0c00264
[27] Huck P, Gunter D, Cholia S, et al. User applications driven by the community contribution framework MPContribs in the materials project: MPCONTRIBS- DRIVEN USER APPLICATIONS. Concurr Comput. 2016;28(7):1982–1993. doi: 10.1002/cpe.3698
[28] Gorai P, Gao D, Ortiz B, et al. TE design lab: a virtual laboratory for thermoelectric material design. Comput Mater Sci. 2016;112:368–376. doi: 10.1016/j.commatsci.2015.11.006
[29] Ricci F, Chen W, Aydemir U, et al. An ab initio electronic transport database for inorganic materials [Internet]. cited 2024 Jun 19]. Available from: http://datadryad.org/stash/dataset/doi%253A10.5061%252Fdryad.gn001
[30] Gaultois MW, Sparks TD, Borg CKH, et al. Data-driven review of thermoelectric materials: performance and resource considerations. Chem Mater. 2013;25(15):2911–2920. doi: 10.1021/cm400893e
[31] Na GS, Chang H. A public database of thermoelectric materials and system- identified material representation for data-driven discovery. Npj Comput Mater. 2022;8(1):1–11. doi: 10.1038/s41524-022-00897-2
[32] Zhang Y, Itani S, Khanal K, et al. GPTArticleExtractor: an automated workflow for magnetic material database construction. J Magn Magn Mater. 2024;597(172001):172001. doi: 10.1016/j.jmmm.2024.172001
[33] Itani S, Zhang Y, Zang J. Large language model-driven database for thermoelectric materials [Internet]. arXiv [cond-mat.mtrl-sci]. 2024 [cited 2025 Jan 7]. Available from: http://arxiv.org/abs/2501.00564
[34] Itani S, Zhang Y, Zang J. Northeast materials database (NEMAD): enabling discovery of high transition temperature magnetic compounds [Internet]. arXiv [cond-mat.mtrl-sci]. 2024 [cited 2025 Jan 23]. Available from: http://arxiv.org/abs/2409.15675
[35] Lee YL, Lee H, Jang S, et al. Texplorer. org: thermoelectric material properties data platform for experimental and first-principles calculation results. APL Mater [Internet]. 2023;11(4). doi: 10.1063/5.0137642
[36] Leach-Murray S. Figshare—get credit for your research: Figshare.com. Tech Serv Q. 2016;33 (1):98–99. doi: 10.1080/07317131.2015.1093855
[37] Katsura Y, Kumagai M, Gunji S, et al. Development of "starry data" web system for data curation of published experimental thermoelectric properties. J Jpn Soc Powder Powder Metall. 2017;64(8):467–470. doi: 10.2497/jjspm.64.467
[38] Katsura Y, Kumagai M, Kodani T, et al. Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. Sci Technol Adv Mater. 2019;20(1):511–520. doi: 10.1080/14686996.2019.1603885
[39] Hendricks G, Tkaczyk D, Lin J, et al. Crossref: the sustainable source of community-owned scholarly metadata. Quant Sci Stud. 2020;1(1):414–427. doi: 10.1162/qss_a_00022
[40] Marin F, Rohatgi A, Charlot S. WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: application to ultraviolet spectropolarimetry Available from: http://arxiv.org/abs/1708.02025
[41] Mato T, Takada Y. StarryDigitizer [Internet]. 2024. Available from: https://digitizer.starrydata.org/.
[42] American Chemical Society. Terms of Use [Internet]. [cited 2025 Apr 15]. Available from: https://www.acs.org/terms.html
[43] Elsevier. Terms and conditions [Internet]. www.elsevier.com. [cited 2025 Apr 15]. https://www.elsevier.com/legal/elsevier-website-terms-and- conditions
[44] Wiley Online Library. Terms of Use [Internet]. [cited 2025 Apr 15]. Available from: https://onlinelibrary.wiley.com/terms-and-conditions
[45] Banker K, Garrett D, Bakkum P, et al. MongoDB in action: covers MongoDB version 3.0. London, (UK): Simon and Schuster; 2016.
[46] Merriman D, Horowitz E, Ryan K. MongoDB [Internet]. 2016. Available from: https://www.mongodb.com/
[47] Downie N. Chart.js [Internet]. 2024. Available from: https://www.chartjs.org/
[48] Bokeh Development Team. Bokeh: Python library for interactive visualization [Internet]. 2022. Available from: https://bokeh.org/
[49] Richardson L. Beautiful soup [Internet]. 2023. Available from: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[50] Streamlit • A faster way to build and share data apps [Internet]. [cited 2025 Apr 15]. Available from: https://streamlit.io/
[51] Plotly Technologies Inc. Collaborative data science. Montréal, QC: Plotly Technologies Inc.; 2015.
[52] Django Software Foundation. Django [Internet]. 2014. Available from: https://www.djangoproject.com/
[53] Django non-rel developers. Django non-rel [Internet]. 2013. Available from: https://github.com/django-nonrel/django
[54] Haag J. Django MongoDB Engine [Internet]. 2015. Available from: https://github.com/django-nonrel/mongodb-engine
[55] Ruszczewski W. Djangotoolbox [Internet]. 2015. Available from: https://github.com/django-nonrel/djangotoolbox
[56] You E, Vue J [Internet]. 2023. Available from: https://vuejs.org/.

[57] Vue Select SJ [Internet]. 2019. Available from: https://vue-select.org/
[58] NumPy Developers NumPy [Internet]. 2017. Available from: https://numpy.org/
[59] Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585 (7825):357–362. doi: 10.1038/s41586-020-2649-2
[60] The pandas development team. pandas [Internet]. 2017. Available from: https://pandas.pydata.org/
[61] Sp O. pymatgen [Internet]. 2023. Available from: https://pymatgen.org/
[62] Ong SP, Richards WD, Jain A, et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci. 2013;68:314–319. doi: 10.1016/j.commatsci.2012.10.028
[63] Dirolf M, Pymongo [Internet]. 2016. Available from: https://pymongo.readthedocs.io/
[64] Kumagai M, Kurosaki K, Ohishi Y, et al. Effect of Ball-Milling Conditions on Thermoelectric Properties of Polycrystalline $CuGaTe_2$. Mater Trans. 2014;55 (8):1215–1218. doi: 10.2320/matertrans.E-M2014821
[65] Zhao L-D, Lo S-H, Zhang Y, et al. Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals. Nature. 2014;508(7496):373–377. doi: 10.1038/nature13184
[66] Zhou C, Lee YK, Yu Y, et al. Polycrystalline SnSe with a thermoelectric figure of merit greater than the single crystal. Nat Mater. 2021;20(10):1378–1384. doi: 10.1038/s41563-021-01064-6
[67] Shimizu N, Kaneko H. Direct inverse analysis based on Gaussian mixture regression for multiple objective variables in material design. Mater Des. 2020;196:109168. doi: 10.1016/j.matdes.2020.109168
[68] Kaneko H. Lifting the limitations of Gaussian mixture regression through coupling with principal component analysis and deep autoencoding. Chemom Intellig Lab Syst. 2021;218:104437. doi: 10.1016/j.chemolab.2021.104437
[69] Yoshihama H, Kaneko H. Design of thermoelectric materials with high electrical conductivity, high Seebeck coefficient, and low thermal conductivity. Anal Sci Adv. 2021;2(5–6):289–294. doi: 10.1002/ansa.202000114
[70] Kumagai M, Ando Y, Tanaka A, et al. Effects of data bias on machine-learning– based material discovery using experimental property data. Sci Technol of Adv Mater: Methods. 2022;2(1):302–309. doi: 10.1080/27660400.2022.2109447
[71] Borg CKH, Muckley ES, Nyby C, et al. Quantifying the performance of machine learning models in materials discovery. Digit Discov. 2023;2(2):327–338. doi: 10.1039/D2DD00113F
[72] Kaneko H. Interpretation of machine learning models for data sets with many features using feature importance. ACS Omega. 2023;8(25):23218–23225. doi: 10.1021/acsomega.3c03722
[73] Jia X, Aziz A, Hashimoto Y, et al. Dealing with the big data challenges in AI for thermoelectric materials. Sci China Mater. 2024;67(4):1173–1182.
[74] Sun Y, Kumagai M, Jin M, et al. A multiclass classification model for predicting the thermal conductivity of uranium compounds. J Nucl Sci Technol [Internet]. 2024;61(6):778–788. doi: 10.1080/00223131.2023.2269974
[75] Kaneko H. Evaluation and optimization methods for applicability domain methods and their hyperparameters, considering the prediction performance of machine learning models. ACS Omega. 2024;9 (10):11453–11458. doi: 10.1021/acsomega.3c08036
[76] Kaneko H. Clustering method for the construction of machine learning model with high predictive ability. Chemometr Intell Lab Syst. 2024;246 (105084):105084. doi: 10.1016/j.chemolab.2024.105084
[77] Ren Q, Chen D, Rao L, et al. Machine-learning-assisted discovery of 212-Zintl- phase compounds with ultra-low lattice thermal conductivity. J Mater Chem A. 2024;12(2):1157–1165. doi: 10.1039/D3TA05690B
[78] Sun Y, Miyawaki Y, Kumagai M, et al. Thermophysical characterization of UFe3B2 and USiNi: An experimental study. J Nucl Mater. 2024;595(155048):155048. doi: 10.1016/j.jnucmat.2024.155048
[79] Snyder GJ, Snyder AH, Wood M, et al. Weighted Mobility. Adv Mater. 2020;32(25):e2001537. doi: 10.1002/adma.202001537
[80] Snyder GJ, Pereyra A, Gurunathan R. Effective mass from Seebeck coefficient. Adv Funct Mater. 2022;32 (20):2112772. doi: 10.1002/adfm.202112772
[81] Ryu B, Chung J, Kumagai M, et al. Best thermoelectric efficiency of ever-explored materials. iScience. 2023;26(4):106494. doi: 10.1016/j.isci.2023.106494
[82] Fujita E, Liu C, Ishikawa A, et al. Comprehensive experimental datasets of quasicrystals and their approximants. Sci Data. 2024;11(1):1–9. Article no. 1211. doi: 10.1038/s41597-024-04043-z
[83] Kurono T, Zhang J, Kamimura Y, et al. Large-scale database analysis of anomalous thermal conductivity of quasicrystals and its application to thermal diodes. Sci Technol of Adv Mater [Internet]. 2025;5(1). Article no. 2444866. doi: 10.1080/27660400.2024.2444866