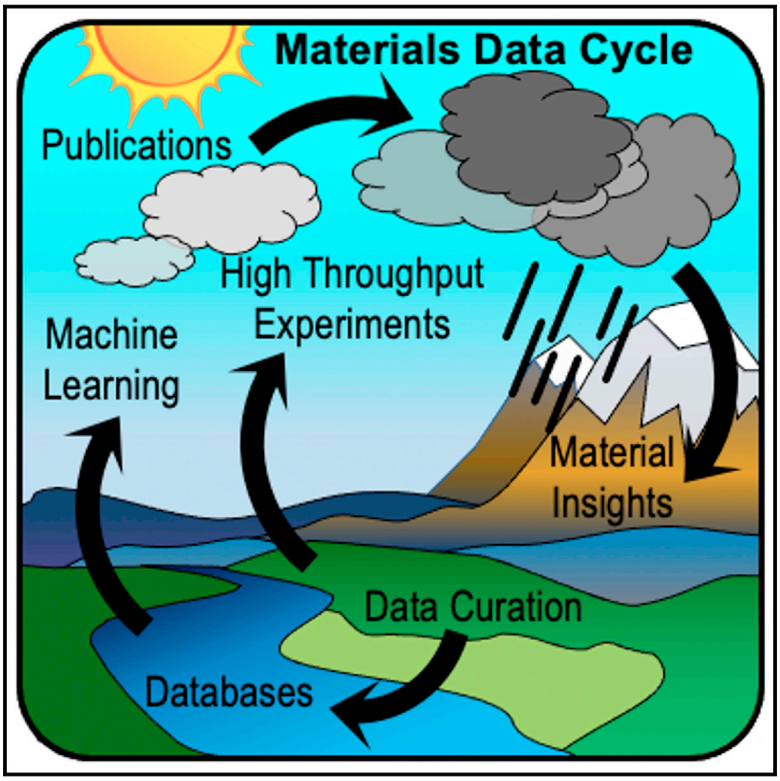
Descriptor

Patterns

Research data infrastructure for high-throughput experimental materials science

Graphical Abstract



Highlights

- Automated curation of experimental materials data
- Integration of data tools into the experimental laboratory
- Simple, effective, and flexible data archival system
- Collection of metadata for enhanced total data value

Authors

Kevin R. Talley, Robert White, Nick Wunder, ..., Kristin Munch, Caleb Phillips, Andriy Zakutayev

Correspondence

andriy.zakutayev@nrel.gov

In brief

This article describes the Research Data Infrastructure (RDI) and its application to create the High-Throughput Experimental Materials Database (HTEM-DB, htem.nrel.gov) at the National Renewable Energy Laboratory (NREL). RDI is a set of custom data tools that collect, process, and store experimental data and metadata, enabling the HTEM-DB repository for inorganic thin-film materials data collected during combinatorial experiments. This coupled experimental and data workflow from the RDI to the HTEM-DB illustrates the best practices currently used for materials data at NREL.

Talley et al., 2021, Patterns 2, 100373
December 10, 2021
https://doi.org/10.1016/j.patter.2021.100373

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

Research data infrastructure for high-throughput experimental materials science

Kevin R. Talley,[1] Robert White,[1] Nick Wunder,[1] Matthew Eash,[1] Marcus Schwarting,[1,2] Dave Evenson,[1] John D. Perkins,[1,3] William Tumas,[1] Kristin Munch,[1] Caleb Phillips,[1] and Andriy Zakutayev[1,4,*]

[1]Materials, Chemical and Computational Science Directorate, National Renewable Energy Laboratory, Golden, CO 80401, USA

[2]Present address: Data Science and Learning Division, Argonne National Laboratory, Lemont, IL 60439, USA

[3]Present address: Applied Chemicals and Materials Division; National Institute of Standards and Technology, Boulder, CO 80305, USA

[4]Lead contact

*Correspondence: andriy.zakutayev@nrel.gov

https://doi.org/10.1016/j.patter.2021.100373

THE BIGGER PICTURE For machine learning to make significant contributions to a scientific domain, algorithms must ingest and learn from high-quality, large-volume datasets. The Research Data Infrastructure (RDI) that feeds the High-Throughput Experimental Materials Database (HTEM-DB, htem.nrel.gov) provides such a dataset from existing experimental data streams at the National Renewable Energy Laboratory (NREL). The described methods for curating experimental data can be applied to other materials research laboratory settings, paving the way for increased application of machine learning to materials science. In turn, the resulting new materials and new knowledge will benefit the society by advancing new technologies in energy, fuels, computing, security, and other important areas.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

The High-Throughput Experimental Materials Database (HTEM-DB, htem.nrel.gov) is a repository of inorganic thin-film materials data collected during combinatorial experiments at the National Renewable Energy Laboratory (NREL). This data asset is enabled by NREL's Research Data Infrastructure (RDI), a set of custom data tools that collect, process, and store experimental data and metadata. Here, we describe the experimental data flow from the RDI to the HTEM-DB to illustrate the strategies and best practices currently used for materials data at NREL. Integration of the data tools with experimental instruments establishes a data communication pipeline between experimental researchers and data scientists. This work motivates the creation of similar workflows at other institutions to aggregate valuable data and increase their usefulness for future machine learning studies. In turn, such data-driven studies can greatly accelerate the pace of discovery and design in the materials science domain.

INTRODUCTION

The High-Throughput Experimental Materials Database (HTEM-DB, htem.nrel.gov)[1] enables the discovery of new materials with useful properties by providing large amounts of high-quality experimental data to the public. The HTEM-DB is one of several commonly used materials databases, with an important distinction that it contains experimental data rather than computational predictions.[2–6] The dataset housed in the HTEM-DB is continuously expanding due to ongoing experiments at the National Renewable Energy Laboratory (NREL). Other related databases contain useful experimental observations[7,8] that are typically focused on a specific collection of results (e.g., crystal structures) from published literature. In HTEM-DB, the complete experimental dataset is made available, including material synthesis conditions, chemical composition, structure, and properties. Similar databases to the HTEM-DB exist for materials science[9] as well as a few other scientific domains,[10] while the advantages and disadvantages of such resources have been discussed in other fields.[11]

HTEM-DB is enabled by the NREL's Research Data Infrastructure (RDI), a modern data management system comparable with a laboratory information management system (LIMS). The RDI is integrated into the laboratory workflow that catalogs

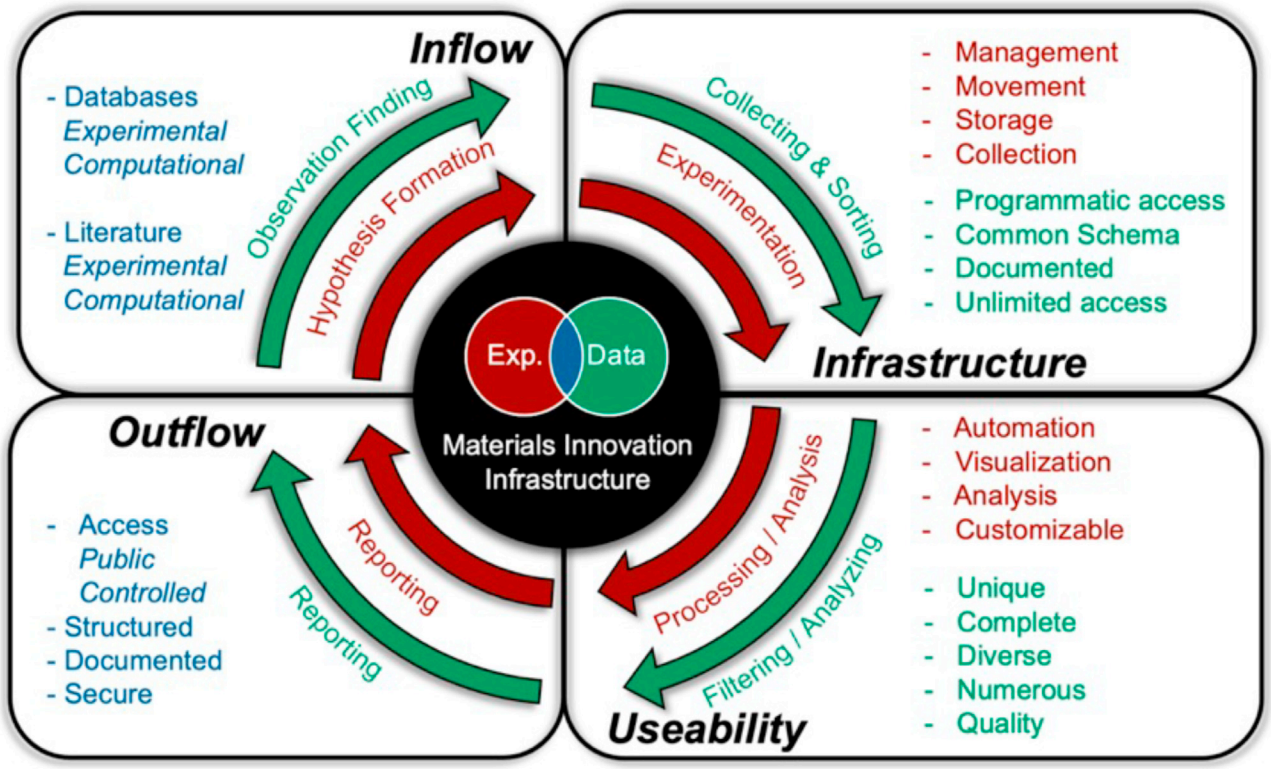


Figure 1. Data management needs for experimental materials research

The experimental (red) and data (green) research workflows have (right) unique infrastructure and usability requirements, but (left) overlapping data inflow and outflow needs. These research workflows are combined within a common RDI and HTEM-DB at NREL, in which the existing experimental data stream is leveraged to develop new materials science insights through machine learning.

experimental data from inorganic thin-film materials experiments at NREL. For the past decade, the RDI has been collecting data from high-throughput experiments (HTEs) across a broad range of thin-film solid-state inorganic materials for various applications, and those data now populate the HTEM-DB. Collecting the results of experimental material synthesis and characterization creates a rich data source for machine learning studies. While the RDI and HTEM-DB workflows discussed here are based in custom data tools, examples of using both custom[12] and off-the-shelf data tools[13] can be found in the material science domain.

Here, we present the RDI that has enabled the HTEM-DB at NREL. This article describes the structural pillars of the RDI, such as raw data collection, metadata collection, data extraction, transformation, loading, and data access, and discusses best practices for future data infrastructure projects of similar scope. After documenting these structural pillars, we discuss the impact of the RDI workflow and the lessons learned during its implementation. While the RDI example described here focuses on high-throughput materials science studies, it is more broadly relevant to any experimental materials science laboratory working to improve their data-related efforts. As such, this article can serve as a blueprint for the future research data infrastructure developments that would increase integration of experimental and data research in the materials science domain.

RESULTS

Overview

To motivate a data infrastructure that collects and augments an experimental data stream for subsequent use by advanced algorithms, first we analyze the needs of experimental and data researchers with respect to their typical workflows, as shown in the example of materials science domain in Figure 1. This analysis identifies both overlapping and unique data infrastructure requirements for high-throughput experimental (HTE) materials researchers and corresponding data researchers. The HTE materials research community begins a study by forming a hypothesis and testing it by experimentation. These data are processed and analyzed, and the results are reported through a peer-reviewed publication on relations between material synthesis, processing, composition, structure, properties, and performance. The materials-data researchers begin a study by identifying a set of relevant data. The dataset is then collected and sorted so that it can be filtered and analyzed for relations between the data, which are the reported results.

Each of these two workflows (experimental and data) has its own requirements, but they can be integrated into a single workflow if the data needs are generalized as inflow, infrastructure, usability, or outflow, as shown in Figure 1. The experimental workflow requires tools for collecting, sorting, and storing newly generated data, whereas the data workflow needs easy access

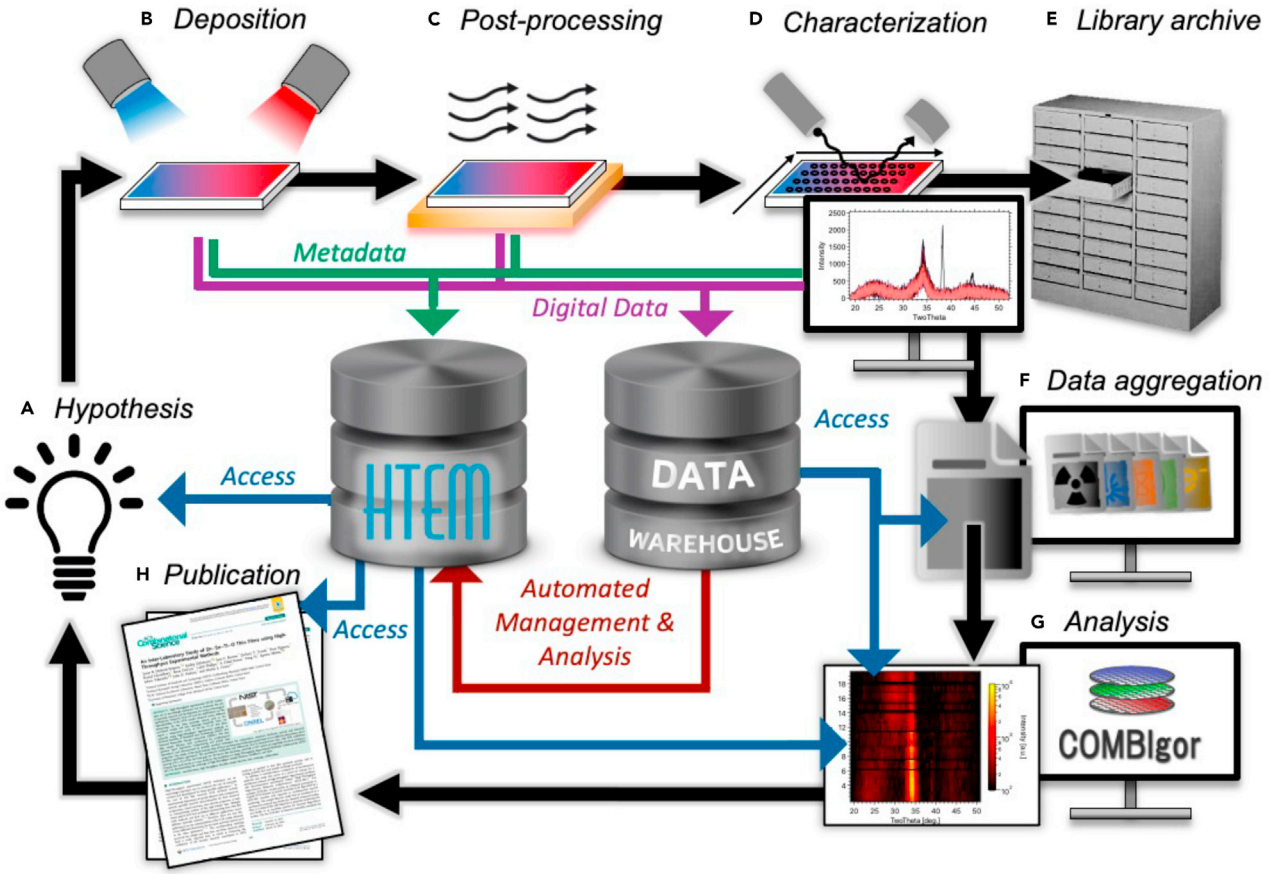


Figure 2. Experimental and data workflows for high-throughput materials research

The workflow starts with (A) experiment design, then material samples are (B) produced, (C) treated, (D) measured, and (E) stored in archives. The measurement data are (F) collected for (G) analysis and presented in (H) a publication, where they inform subsequent experiments. At each step, data tools were developed and implemented for the collection of metadata (green) and measurement (purple) data, automated file management (red), and access (blue).

to the stored data. Furthermore, the experimental researchers need tools to analyze and learn from the data, while the data researchers need large, diverse, high-quality datasets. However, access to previously obtained data and a repository for new data are required for both experimental and data workflows, so there is a strong overlap in their inputs and outputs. These overlapping requirements motivated the creation of the RDI that collects, processes, and stores experimental data and metadata, as well as the HTEM-DB, which provides both a repository for experimental data and a source for the data-driven studies.

The integrated experimental and data workflow that is utilized by numerous researchers in materials discovery area at NREL is illustrated in Figure 2. On the experimental side, NREL possesses a wealth of HTE capabilities and expertise for thin-film materials research (Figures 2A–2E). This experimental research involves depositing and characterizing thin films, often on 50 × 50-mm (2 × 2″) square substrates with a 4 × 11 sample mapping grid, which are common across multiple combinatorial thin-film deposition chambers and spatially resolved characterization instruments at NREL. This experimental workflow at NREL has been benchmarked against other laboratories.[14,15] Other publications demonstrate the range of materials chemistries (e.g., oxides,[16] nitrides,[17] chalcogenides,[18] Li-containing materials,[19] intermetallics)[20] and properties (e.g., optoelectronic,[21] electronic,[22] piezoelectric,[23] photoelectrochemical,[24] thermochemical)[25] to which these HTE methods have been applied.

Each experimental investigation generates large, comprehensive datasets (Figures 2A–2E) that are delivered to the HTEM-DB through the RDI described in this paper (Figures 2F–2H). The RDI was first envisioned almost two decades ago in 2003,[26] then first described in 2014,[27] and briefly summarized in 2018.[1] As a part of the RDI, we also created (2010) and released (2019) COMBIgor (https://www.combigor.com/),[28]an open-source data-analysis package for high-throughput materials-data loading, aggregation, and visualization in combinatorial materials science. Now, after a decade of development, COMBIgor is an integral and useful part of the RDI at NREL. In addition, an early version of COMBIgor has served as a blueprint of parts of the RDI described in this manuscript, such as its extract, transform, and load (ETL) scripts, and the visualization functionality of HTEM-DB.

Another important component of the RDI is the NREL Research Data Network and Data Warehouse (DW) (Figure 2). The DW was first established at NREL in 2010, to manage data

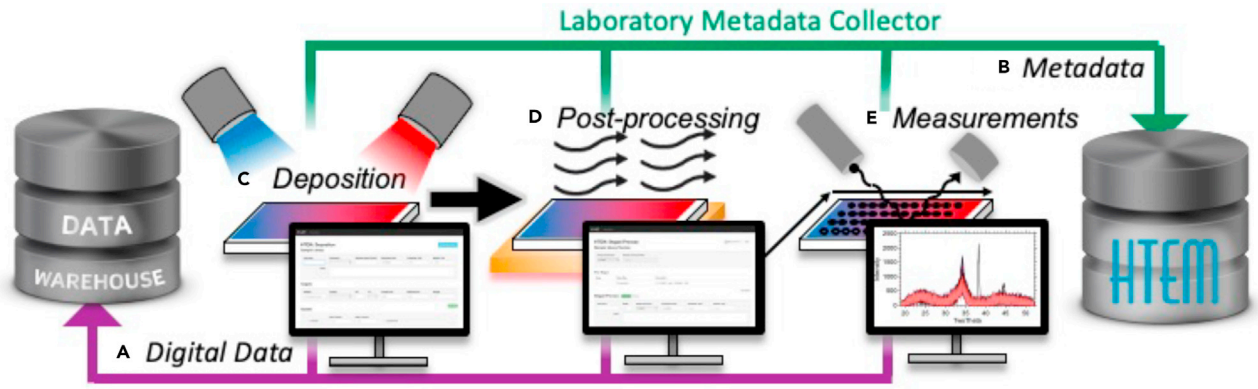


Figure 3. Data collection scheme for high-throughput materials research

(A) Digital data, or raw data files, are collected by the RDI harvesters into the DW, while (B) the metadata are collected by the LMC. Digital data and metadata are collected during three steps of the experimental process: (C) sample library growth, (D) sample post-processing, and (E) materials measurement. These data are combined and entered into the HTEM-DB as a complete sample record. Detailed images of the custom webforms are presented in the supplemental information.

collected from laboratory computers that control experimental instruments. The DW automatically collects data from these tools and makes the files accessible to researchers and to other data tools via the Research Data Network (RDN). For example, the HTEM-DB is populated with measurement data contained in specific high-throughput measurement folders in the DW that are identified by standardized file-naming conventions. Critical metadata from synthesis, processing, and measurement steps are also collected using Laboratory Metadata Collector (LMC) and added to the DW or directly to HTEM-DB, providing experimental context for the measurement results. The data from these files are extracted, transformed, and loaded into the HTEM-DB, which stores processed data for analysis, publication, and data science purposes (Figure 2). The integration of this data workflow was made possible by the RDI that is detailed next.

Components

The individual components of the RDI that facilitate the data workflow presented in Figure 2 form a set of interconnected, custom data tools. This includes tools for data collection (data harvesters, and LMC), data processing (ETL), and storage and access (DW and HTEM-DB). Brief descriptions of each of these data tools are provided below, and additional details can be found in the supplemental information.

Data warehouse

Digital data are the primary source of materials data within this integrated workflow (Figure 2). The software harvests and stores all the digital files that are generated during materials growth and characterization processes (Figures 3 and 4). For this purpose, the harvesting software monitors activity on the instrument computers and automatically identifies target files as they are created or updated. All relevant files on the instrument computers are copied into the data warehouse (DW) archives and processed into the database as necessary.[26] To keep the sensitive research instrumentation segregated from the normal NREL network activity, the computers are connected to the data harvester and archives via a firewall-isolated, specialized subnetwork, called the RDN. The DW currently houses nearly 4 million files harvested from more than 70 instruments across 14 laboratories in four buildings on the NREL campus. This illustrates that the RDI described here is designed for a wide range of experimental material science workflows beyond combinatorial thin-film deposition and spatially resolved characterization.

The DW consists of a back-end relational database (PostgreSQL) and a Qumulo file system housed in the NREL Computational Sciences Data Center. The RDI codebase is built as a series of modules and libraries in C++ and bash scripts but is architected to be modular and allow for easy replacement.[26] The RDI does not require any software to be installed on instrument computers, which is a significant advantage of this architecture. As a result, there are no specific requirements for the individual instrument computers or file types, accommodating a wide range of computer operating systems and ages, typical of an experimental laboratory setting. The DW facilitates easy access to the resulting data files through a custom-built Web application hosted on the internal NREL network, where the aggregated data files are easily and securely accessed by researchers. Thus, the DW serves as both the initial repository and the access gateway to experimental data generated at NREL. More details about the DW are presented in supplemental information (Figure S1).

Laboratory metadata collector

Metadata is one of the most critical types of data in the integrated workflow (Figure 2) because it gives context to the data files collected in the DW. This data stream, although of high importance, is difficult to capture because it requires interaction and input from humans (experimentalists). To simplify metadata collection, the laboratory metadata collector (LMC) was initially prototyped in Python and now has been developed as a Web application (Figures 3 and 4). The LMC includes custom webforms in which users enter and submit sample record information upon completion of an experiment. Each record in the LMC plays a role of a detailed, digital laboratory notebook entry, giving researchers and algorithms access to the experimental variables of interest and making them easy to associate to measurement data, stored alongside metadata in both the DW and HTEM-DB.

The LMC is composed of a front-end, single-page Web application written in JavaScript and a back-end application

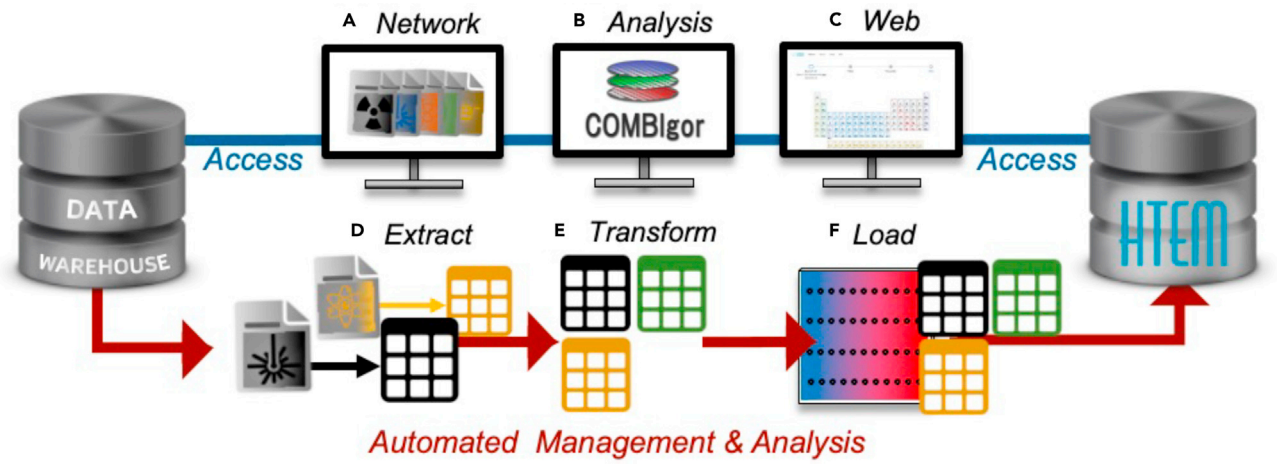


Figure 4. Data management scheme for high-throughput materials research

For experimental researchers, the raw files can be obtained from the DW via (A) network access and can be easily loaded for (B) flexible analysis of the data in COMBIgor. Curated sample records can be searched, filtered, visualized, and downloaded through (C) HTEM-DB Web access. To populate the HTEM-DB with sample records, raw measurement files from the DW are processed through a set of custom ETL scripts. The data are (D) extracted from files in the DW, (E) transformed into useable data, and (F) loaded and post-processed to the corresponding sample library entries. Details and screenshots are presented in the supplemental information.

programming interface (API) written in Node.js. The bulk of the logic is in the front-end Web application, which runs in any modern Web browser. It can be accessed from the NREL network by researchers' laptops and by in-laboratory computers. The application provides model logic for building webforms and conducting form validation. The dynamic view of the webform is adjusted in response to user-entered values and the user's display settings. It presents a preview of the entry upon submission for the user to verify, providing the option for the user to either save the entry or return to editing. The Web front end also provides a searchable interface for finding previous entries. Thus, in the case of a metadata entry error, it can be corrected by opening the erroneous entry, correcting the data, and saving it with the same sample name but a different time stamp. Entry submission and search features are enabled through a back-end API that interacts with the DW. The API simply adds the submitted JSON data to the DW and queries and retrieves past entries.

The LMC webforms collect information from the experimentalist during deposition, post-processing, and measurement of sample libraries. This information is agglomerated in the HTEM-DB for each sample, greatly improving the value of the associated measurement data. Providing additional benefit to the user, COMBIgor plugins port this LMC-generated information directly into the user's Igor Pro experiment via direct JSON file download/import or through the HTEM API (Figures 4A–4C). Easy access to complete experimental records in COMBIgor motivates researchers to participate in metadata reporting through the LMC. As such, the LMC supports the data needs of both the experimental and data science researchers. For the experimentalist, it provides a more efficient, accurate, and accessible record of experimental variables for experimental analysis, compared with typical handwritten laboratory notebooks. For the data science researchers, the LMC offers reliable and complete sample records for individual material samples, which increases the value of any associated data and the overall usability of the entire HTEM-DB in larger data studies. More details about the LMC are presented in supplemental information (Figure S2).

Extract, transform, load

Custom extract, transform, and load (ETL) scripts port raw data from files in the DW into the HTEM-DB (Figures 4D–4F). First, (1) the relevant folders of high-throughput data are copied from the DW and filed into the HTEM-DB repository, based on file name conventions, using a Python script. Information extracted from the raw files, including both open-source (e.g., ASCII, HDF) and reverse-engineered proprietary file formats (e.g., various .raw files), is placed into tables that correspond to the standard 4 × 11 library mapping grid from HTE studies at NREL, or other custom grids (e.g., 176-point grid used by National Institute of Standards and Technology).[14] Next, (2) the data extracted from the files are transformed from original (sometimes proprietary) to final format. The extraction method varies depending on the specific file type and, in some cases, utilizes additional Python libraries or programming languages (e.g., Ruby, C, R). Finally, (3) individual pieces of new data for a given sample are identified and loaded to the database. In this way, these data are correlated to other descriptive metadata pertaining to the same sample. These custom ETL scripts have been designed in a similar way to our open-source data-analysis package COMBIgor,[28] so the interested reader is referred to its public GitHub repository (https://github.com/NREL/CombIgor) for full scripts (e.g., see "Instruments" subfolder) and test datasets (e.g., see "Example Files" subfolder).

Collectively, the custom ETL process produces a continuous flow of new information to the HTEM-DB and supports both experimental and data science researchers. These scripts, written primarily in Python, run on the NREL high-performance computing system, and are automatically executed daily. As part of the ETL workflow, certain data types may be combined

to generate additional, useful data types. For example, sheet resistance and thickness may be combined to calculate resistivity, or optical transmission, reflection, and thickness may be combined to calculate the absorption spectrum and to determine the optical bandgap.[29] Thus, this ETL infrastructure funnels and correlates datasets from raw files in the DW into structured entries in the HTEM-DB that are easily accessed by experimental and data-focused researchers. More details about the custom implementation of the ETL process, including key lines of the script and the screenshot of the code repository, are presented in supplemental information (Figure S3).

HTEM-DB

The HTEM-DB, including its structure, content, and applications, is detailed in a previous publication.[1] Here, we present the HTEM-DB in a level of detail similar to the other workflow components in Figure 4 for completeness. The HTEM is a PostgreSQL database that is a repository for the incoming data from experimental workflows. It is housed in the same NREL data center as the DW, LMC, and ETL scripts. HTEM-DB functions as an access point for both experimental and data science researchers, with two main points of access. Experimentalists typically use the Web interface (https://htem.nrel.gov), which provides the user with quick and effective sample search for retrieval of sample and library information of all data types, which can then be viewed using the built-in data visualization features. Data scientists typically use the API (https://htem-api.nrel.gov/api), which enables machine learning by providing algorithms with programmatic access to the entire HTEM and is also used to load experimental data into COMBIgor. Thus, the HTEM-DB supports both experimental and data researchers by providing tailored data inflow and outflow access and by providing features that enhance the usability of the data that it contains.

Access to the data that flow through RDI provides researchers with an opportunity to interact with them and learn from them. For researchers conducting experimental studies, the RDI provides improved efficiency and increased accuracy of experimental research data handling. Easy access to aggregated data for samples of interest in ongoing research projects is achieved through downloading files directly from the DW, loading library information into COMBIgor directly from the HTEM-DB, or using data-analysis tools built into the HTEM-DB Web interface for finding, filtering, visualizing, and downloading sample records (see "Help" information at the Web page). For the data scientists, the RDI provides a large dataset to investigate by machine learning methods by enabling access to curated, structured, and complete data records for a wide range of materials, which is achieved through the HTEM API or Web page for the publicly available data. This release of the data to the public on HTEM-DB is made under a Creative Commons license (Attribution 4.0 International license) once the manuscript describing the data is published, or a decision is made not to pursue the manuscript publication. Access to some of the HTEM content, including specific chemical elements in a sample library, is restricted in the interest of the stakeholders, including the public or private funding sources that support specific research projects and demand exclusive access to the resulting data.

The HTEM-DB is large and diverse (Figure 5). Due to a rich history of HTE materials studies conducted at NREL, the internal version of the database has been populated with more than 320,000 unique samples from more than 7,300 sample libraries as of September 2020. Of the unique samples within the HTEM-DB, more than half (174,000) have correlated characterization data. These samples cover a wide range of compositions, with more than 33 elements quantified in samples with composition measurements. These characteristics are desirable for materials-data studies and are important factors if such studies are to be realistic and useful. The HTEM-DB will continue to grow due to ongoing experiments and continuing efforts to build additional data collection pathways, providing an ever-increasing potential for useful knowledge extraction. Another interesting future direction is connecting the HTEM DB[1] with some of the computational material databases,[2–6] for co-displaying experimental and theoretical data for related database entries, or for joint experimental/theoretical data analysis. More details about the HTEM-DB are presented in a previous publication.[1]

DISCUSSION

The effort to develop the RDI at NREL has resulted in a more valuable product than initially envisioned. By establishing various RDI tools, integrating them together, and providing them to researchers, a complete data workflow has been implemented that supports the existing experimental research workflow while curating valuable data for future use in machine learning studies. These RDI tools are tailored to the specific experimental setting at NREL (Figure 2) but are broadly designed to meet the needs outlined in Figure 1. Thus, this RDI example should be applicable to other materials science laboratory settings with a somewhat standardized research workflow, like that shown in the example of high-throughput materials research (Figure 3). The data tools that form the RDI (Figure 4)—specifically the harvesters, DW, the ETL process, the LMC, as well as the resulting HTEM-DB (Figure 5)—are all critical to the success of the RDI. The hope is that the designs and features of the data tools described here will serve as examples of best practices for other institutions that require similar RDI for their own data workflows.

Historically, the RDI at NREL has been constructed from the bottom up, with multiple contributions from many people over the time span of more than a decade. For example, the initial version of RDI has been prototyped as a part of a laboratory design and construction project funded by the US Department of Energy (DOE), and more recently supported by NREL internal research data infrastructure funding. The resulting bottom-up RDI products are functional, and can be used to make a blueprint for better RDI planning and construction in the future. In a similar way, bottom-up design and construction of the HTEM-DB has been an indirect outcome of addressing individual data challenges in a collaborative project between materials science and data science. The resulting materials-data relations has shown promise as a valuable contributor to science, and, as such, encourages investments in this area to design and build similar data workflows and databases at other institutions.

Ideally, it is our opinion that these types of RDI systems should be engineered from the top down, leveraging the lessons learned from the prior bottom-up efforts discussed above. Such next-generation RDI built to collect, maintain, and access the data should strike a fine balance between being simple, to encourage contributions from individual researchers, and flexible/scalable,

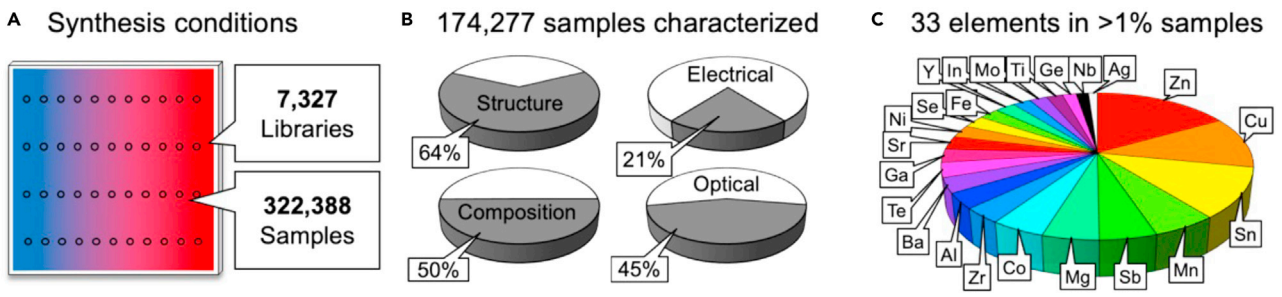


Figure 5. HTEM-DB statistics

As of September 2020, the internal version of the database is (A) large, with over 300,000 unique samples on more than 7,000 libraries; (B) complete, with either structure, composition, and/or property data for more than 170,000 samples; and (C) chemically diverse, shown as percentage of samples containing a given element.

to meet the needs of future research directions. However, building a complete RDI from scratch would require a substantial upfront financial investment in hardware, network installation, and software development, as well as sustained investment in maintenance and improvement, both of which are not always easy to obtain. This is further complicated by the fact that the modern materials science laboratory is a complex and ever-changing terrain where experimental equipment with all types of software, hardware, age, and access is encountered. Thus, a functional RDI for a modern materials science laboratory requires recruiting and retaining skilled and dedicated personnel in material science, data science, and software engineering, which is not always easy, especially in a small laboratory.

An interesting intermediate option between the historical bottom-up and the idealistic top-down approaches discussed above is the one where a general RDI framework is developed from the top down at a large institution with significant resources dedicated to this effort and then customized from the bottom up to be most useful for each individual research laboratory. As a first step toward a more generalizable RDI at NREL, we have implemented and are currently testing an internal user authentication scheme for HTEM-DB that may enable in the future direct data contributions or corrections from *internal* users, rather than just data harvesting from the measurement instruments. However, opening up HTEM-DB for *external* contributions, as well as increasing the open-source fraction of the RDI code beyond COMBIgor, would in turn require further RDI developments, which should be supported by corresponding external funding. A successful prior example of such external data contribution can be found in the MPContribs[30] framework for the Materials Project[3] externally funded by DOE, although with a primary focus on computational rather than experimental data. Other emerging examples are the data hubs of several energy material networks (EMNs) funded by DOE, based on easily deployed cloud technology and open-source software frameworks.[31]

The most successful RDI components to prompt the materials researcher engagement at NREL were built through tight collaborative design efforts between material researchers, data scientists, and software engineers. The RDI constructed by software engineers and presented here is useful to the experimentalist and, as a result, has been widely adopted by the materials researchers at NREL. New materials researchers that join HTE efforts at NREL are encouraged to use the various components of the RDI, including mining of the prior relevant data through HTEM-DB, entering the synthesis conditions using LMC, and remotely collecting the relevant files through the DW for COMBIgor analysis. The motivation for continued used of the RDI is provided by both beneficial functionality of the RDI components that accelerate materials research and by the recent increase in public data requirements of funding agencies and journal publishers.

As more materials researchers use the RDI, the HTEM-DB continues to collect, preserve, and provide materials data. In turn, the large, curated set of samples synthesized, characterized, and captured by the RDI is primed and ready for exploration by machine learning algorithms by the data scientists. To ensure long-term success of this interaction between materials researchers and data scientists, various software components of the RDI have been documented in internal GitHub repositories and inside of the scripts contained therein, and parts of the RDI source code, such as COMBIgor, have been made publicly available[28] for other software engineers to modify and reuse. This RDI and HTEM-DB are just one example of how the individual materials science and data science efforts at NREL provide much greater value when combined to work together. Similar systems brought online at other institutions, producing a broader set of materials data to explore, would further increase the potential of this materials-data relationship to advance science.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Andriy Zakutayev (andriy.zakutayev@nrel.gov).

Materials availability

This study did not generate new material samples.

Data and code availability

The data in HTEM-DB can be accessed at https://htem.nrel.gov/ (using Web interface) and https://htem-api.nrel.gov/ (using API), as described in NREL Data Catalog under DOI: 10.7799/1407128. The COMBIgor code is available on GitHub at https://github.com/NREL/Combigor and under DOI: 10.5281/zenodo.5539029 Additional information about other parts of the RDI described in this publication is available from the lead contact upon request.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100373.

ACKNOWLEDGMENTS

This work was authored at the NREL, operated by Alliance for Sustainable Energy, LLC, for the DOE under contract no. DE-AC36-08GO28308. Financial support for the RDI operation and improvements, including LMC, is provided by NREL indirect funding. Funding HTEM-DB development was provided by NREL's Laboratory Directed Research and Development (LDRD). HTEM data curation efforts were funded by the DOE, Office of Science. DW prototyping was supported by the DOE, Energy Efficiency and Renewable Energy (EERE). A portion of the research was performed using computational resources sponsored by the DOE EERE and located at NREL.

AUTHOR CONTRIBUTIONS

Conceptualization, A.Z. and C.P.; software, R.W., M.S., N.W., C.P., D.E., M.E., and A.Z.; resources, C.P., K.M., J.P., and A.Z.; data curation, R.W., M.S., C.P., and A.Z.; writing, K.R.T., C.P., and A.Z.; visualization, K.R.T.; supervision, K.M., C.P., and A.Z.; project administration, K.M, C.P., J.P., W.T., and A.Z.; funding acquisition, J.P., K.M., C.P., W.T., and A.Z.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 23, 2021
Revised: August 13, 2021
Accepted: September 30, 2021
Published: December 10, 2021

REFERENCES

1. Zakutayev, A., Wunder, N., Schwarting, M., Perkins, J.D., White, R., Munch, K., Tumas, W., and Phillips, C. (2018). An open experimental database for exploring inorganic materials. Sci. Data *5*, 180053. https://doi.org/10.1038/sdata.2018.53.
2. Stevanovic, V., Lany, S., Zhang, X., and Zunger, A. (2012). Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. Phys. Rev. B *85*, 115104. https://doi.org/10.1103/PhysRevB.85.115104.
3. 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). The Materials Project: a materials genome approach to accelerating materials innovation. APL Mater. *1*, 011002. https://doi.org/10.1063/1.4812323.
4. Curtarolo, S., Setyawan, W., Hart, G.L.W., Jahnatek, M., Chepulskii, R.V., Taylor, R.H., Wang, S., Xue, J., Yang, K., Levy, O., et al. (2012). AFLOW: an automatic framework for high-throughput materials discovery. Comput. Mater. Sci. *58*, 218–226. https://doi.org/10.1016/j.commatsci.2012.02.005.
5. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). JOM *65*, 1501–1509. http://doi:10.1007/s11837-013-0755-4.
6. Haastrup, S., Strange, M., Pandey, M., Deilmann, T., Schmidt, P.S., Hinsche, N.F., Gjerding, M.N., Torelli, D., Larsen, P.M., Riis-Jensen, A.C., et al. (2008). The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals. 2D Mater. *5*, 042002. https://doi.org/10.1088/2053-1583/aacfc1.
7. Mariette, H. (2004). The inorganic crystal structure database (ICSD) - present and future. Crystallogr. Rev. *10*, 17–22. https://doi.org/10.1080/08893110410001664882.
8. Xu, Y., Yamazaki, M., and Villars, P. (2011). Inorganic materials database for exploring the nature of material. Jpn. J. Appl. Phys. *50*, 11. 11RH02. https://doi.org/10.1143/jjap.50.11rh02.
9. Stein, H.S., Soedarmadji, E., Newhouse, P.F., Guevarra, D., and Gregoire, J.M. (2019). Synthesis, optical imaging, and absorption spectroscopy data for 179072 metal oxides. Sci. Data *6*, 1–5. https://doi.org/10.1038/s41597-019-0019-4.
10. Morrell, W.C., Birkel, G.W., Forrer, M., Lopez, T., Backman, T.W.H., Dussault, M., Petzold, C.J., Baidoo, E.E.K., Costello, Z., Ando, D., et al. (2017). The experiment data depot: a web-based software tool for biological experimental data storage, sharing, and visualization. ACS Synth. Biol. *6*, 2248–2259. https://doi.org/10.1021/acssynbio.7b00204.
11. Williams, A.J., Ekins, S., and Tkachenko, V. (2012). Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov. Today *17*, 685–701. https://doi.org/10.1016/j.drudis.2012.02.013.
12. Statt, M., Rohr, B.A., Brown, K.S., Guevarra, D., Hummelshøj, J.S., Hung, L., Gregoire, J., and Suram, S. (2021). ESAMP: event-sourced architecture for materials provenance management and application to accelerated materials discovery. https://doi.org/10.26434/chemrxiv.14583258.v1.
13. Banko, L., and Ludwig, A. (2020). Fast-track to research data management in experimental material science–setting the ground for research group level materials digitalization. ACS Comb. Sci. *22*, 401–409. https://doi.org/10.1021/acscombsci.0c00057.
14. Hattrick-Simpers, J.R., Zakutayev, A., Barron, S.C., Trautt, Z.T., Nguyen, N., Choudhary, K., DeCost, B., Phillips, C., Kusne, A.G., Yi, F., et al. (2019). An Inter-Laboratory Study of Zn–Sn–Ti–O thin films using high-throughput experimental methods. ACS Comb. Sci. *21*, 350–361. https://doi.org/10.1021/acscombsci.8b00158.
15. Hattrick-Simpers, J.R., DeCost, B., Kusne, A.G., Joress, H., Wong-Ng, W., Kaiser, D.L., Zakutayev, A., Phillips, C., Sun, S., Thapa, J., et al. (2021). An open combinatorial diffraction dataset including consensus human and machine learning labels with quantified uncertainty for training new machine learning models. Integrat. Mater. Manufact. Innovation *10*, 311. https://doi.org/10.1007/s40192-021-00213-8.
16. Bikowski, A., Holder, A., Peng, H., Siol, S., Norman, A., Lany, S., and Zakutayev, A. (2016). Synthesis and characterization of (Sn, Zn) O alloys. Chem. Mater. *28*, 7765–7772. https://doi.org/10.1021/acs.chemmater.6b02968.
17. Bauers, S.R., Mangum, J., Harvey, S.P., Perkins, J.D., Gorman, B., and Zakutayev, A. (2020). Epitaxial growth of rock salt $MgZrN_2$ semiconductors on MgO and GaN. Appl. Phys. Lett. *116*, 102102. https://doi.org/10.1063/1.5140469.
18. Siol, S., Holder, A., Steffes, J., Schelhas, L.T., Stone, K.H., Garten, L., Perkins, J.D., Parilla, P.A., Toney, M.F., Huey, B.D., et al. (2018). Negative-pressure polymorphs made by heterostructural alloying. Sci. Adv. *4*, EAAQ1442. https://doi.org/10.1126/sciadv.aaq1442.
19. Xu, Y., Wood, K., Coyle, J., Engtrakul, C., Teeter, G., Stoldt, C., Burrell, A., and Zakutayev, A. (2019). Chemistry of electrolyte reduction on lithium silicide. J. Phys. Chem. C *123*, 13219–13224. https://doi.org/10.1021/acs.jpcc.9b02611.
20. Zakutayev, A., Zhang, X., Nagaraja, A., Yu, L., Lany, S., Mason, T.O., Ginley, D.S., and Zunger, A. (2013). Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach. J. Am. Chem. Soc. *135*, 10048–10054. https://doi.org/10.1021/ja311599g.
21. Welch, A.W., Baranowski, L.L., Peng, H., Hempel, H., Eichberger, R., Unold, T., Lany, S., Wolden, C., and Zakutayev, A. (2017). Trade-offs in thin film solar cells with layered chalcostibite photovoltaic absorbers. Adv. Energy Mater. *7*, 1601935. https://doi.org/10.1002/aenm.20160193.
22. Roberts, D.M., Bardgett, D., Gorman, B.P., Perkins, J.D., Zakutayev, A., and Bauers, S.R. (2020). Synthesis of tunable SnS-TaS_2 nanoscale superlattices. Nano Lett. *20*, 7059–7067. https://doi.org/10.1021/acs.nanolett.0c02115.
23. Talley, K.R., Millican, S.L., Mangum, J., Siol, S., Musgrave, C.B., Gorman, B., Holder, A.M., Zakutayev, A., and Brennecka, G.L. (2018). Implications of heterostructural alloying for enhanced piezoelectric performance of (Al, Sc) N. Phys. Rev. Mater. *2*, 063802. https://doi.org/10.1103/PhysRevMaterials.2.063802.

24. Peng, H., Ndione, P.F., Ginley, D.S., Zakutayev, A., and Lany, S. (2015). Design of semiconducting tetrahedral $Mn_{1-x}Zn_xO$ alloys and their application to solar water splitting. Phys. Rev. X *5*, 021016. https://doi.org/10.1103/PhysRevX.5.021016.

25. Heo, S.J., Sanders, M., O'Hayre, R.P., and Zakutayev, A. (2021). Double-site substitution of Ce into (Ba, Sr)MnO_3 perovskites for solar thermochemical hydrogen production. ACS Energy Lett. *2021*, 3037–3043. https://doi.org/10.1021/acsenergylett.1c01214.

26. Nelson, B., Friedman, D., Geisz, J., Albin, D., Benner, J., and Wang, Q. (2003). To data management and beyond… for photovoltaic applications. MRS Online Proc. Libr. *804*, 54–59. https://doi.org/10.1557/PROC-804-JJ11.3.

27. White, R.R., and Munch, K. (2014). Handling large and complex data in a photovoltaic research institution using a custom laboratory information management system. MRS Online Proc. Libr. *1104*, 1–12. https://doi.org/10.1557/opl.2014.31.

28. Talley, K.R., Bauers, S.R., Melamed, C.L., Papac, M.C., Heinselman, K., Khan, I., Roberts, D.M., Jacobson, V., Mis, A., Brennecka, G.L., et al. (2019). COMBIgor: data analysis package for combinatorial materials science. ACS Comb. Sci. *21*, 537–547. https://doi.org/10.1021/acscombsci.9b00077.

29. Schwarting, M., Siol, S., Talley, K., Zakutayev, A., and Phillips, C. (2017). Automated algorithms for band gap analysis from optical absorption spectra. Mater. Discov. *10*, 43–52. https://doi.org/10.1016/j.md.2018.04.003.

30. Huck, P., Gunter, D., Cholia, S., Winston, D., N'Diaye, A.T., and Persson, K. (2016). User applications driven by the community contribution framework MPContribs in the materials project. Concurrency Comput. Pract. Exp. *28*, 1982–1993. https://doi.org/10.1002/cpe.3698.

31. White, R.R., Munch, K., Wunder, N., Guba, N., Sivaraman, C., Van Allsburg, K.M., Dinh, H., and Pailing, C. (2021). Energy material network data hubs: software platforms for advancing collaborative energy materials research. Int. J. Adv. Comput. Sci. Appl. *12*. https://doi.org/10.14569/IJACSA.2021.0120677.