Harnessing X-ray Absorption Spectroscopy Data through Multimodal Mining of Battery Literature

Tanjin He[1*], Aikaterini Vriza[2], Logan Ward[1,3], Xu Huang[4],
Yiming Chen[2], Anubhav Jain[5], Gerbrand Ceder[4],
Rajeev S. Assary[6], Ian T. Foster[1,7*], Maria K. Y. Chan[2*]

[1] Data Science and Learning Division, Argonne National Laboratory, Lemont, IL 60439, USA

[2] Center for Nanoscale Materials, Argonne National Laboratory, Lemont, IL 60439, USA

[3] NVIDIA, Santa Clara, CA 95051, USA

[4] Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA

[5] Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[6] Materials Science Division, Argonne National Laboratory, Lemont, IL, 60439, USA

[7] Department of Computer Science, The University of Chicago, Chicago, IL, 60637, USA

* Corresponding authors: Tanjin He (the@anl.gov), Ian Foster (foster@anl.gov), Maria K. Y. Chan (mchan@anl.gov)

Abstract

X-ray absorption spectroscopy (XAS) is central to understanding the local electronic and atomic structure of materials, yet most published spectra remain inaccessible to data-driven analysis because they are embedded in figures and described through fragmented textual context in the literature. Here, we use multimodal (image and text) literature mining to transform this dispersed knowledge into an AI-ready experimental data resource. We developed a scalable spectroscopy data digitization pipeline that identifies XAS figures in full-text articles, digitizes spectral curves, and links each spectrum to accompanying metadata on the measured edge and material. Applying this pipeline to the battery literature produced an open dataset of 13,740 XAS spectra, spanning 66 absorbing elements and diverse battery chemistries, with expert validation confirming accurate extraction of spectral and metadata information. By converting literature-embedded spectra into structured numerical data, this dataset provides a foundation for large-scale XAS analysis, cross-laboratory comparison, high-throughput characterization, and autonomous discovery of advanced materials.

Background & Summary

The explosive growth of AI/ML in materials science necessitates high-quality, large-scale, labeled, experimental and computational datasets for training. While there are ample computational data, experimental data availability is far more limited. The vast amount of data embedded in the scientific literature represents an important, largely untapped source for enriching experimental datasets.

Large-scale spectroscopy datasets can power materials characterization, autonomous discovery, and cross-laboratory standardization. First, they provide references for identifying structural and compositional information in complex materials, extending existing databases [1, 2, 3, 4, 5] toward broader coverage of materials and experimental conditions. Second, they supply the training data needed to develop the high-throughput characterization algorithms that enable self-driving laboratories [6, 7, 8, 9, 10, 11, 12]. Finally, by spanning many research groups, they can help establish standardized experimental procedures and reporting ontologies for cross-laboratory alignment.

Scientific publications have accumulated a large amount of spectroscopy data that offer a natural complementary source for growing curated reference databases. However, these data are typically presented as scientific figures with textual descriptions scattered throughout the paper, and this unstructured, multimodal format hinders their use in data-driven research. Wang et al. [13] compiled a valuable dataset of 1,652 X-ray absorption spectroscopy (XAS) spectra of iron-containing proteins from the literature, but the manual curation involved is labor-intensive and limits scalability. Automated literature mining is becoming more tractable with advances in natural language processing (NLP) and computer vision (CV), and especially with the growing capabilities of large language models (LLMs) and vision language models (VLMs). For example, He et al. [14] and Huo et al. [15] extracted materials synthesis recipes from the literature for predictive synthesis. In a related vein, Leong et al. [16, 17] extracted chemical reaction information for organic synthesis and photocatalysis. Beyond reaction data, Zheng et al. [18] mined the reticular chemistry literature. Dong et al. [19] extracted semiconductor band-gap data from text. Liu et al. [20] extracted causal mechanisms for materials design and performance optimization. Siegel et al. [21] trained a

neural network to detect bounding boxes of scientific figures. Park et al. [22] finetuned an embedding model to find explanatory sentences most similar to XAS figure captions. More recently, Zhang et al. [23] targeted multimodal information on hydrogen storage materials.

The most challenging part of this multimodal data extraction task is extracting data from figures, where most characterization and performance data reside [24]. Current VLMs can reliably interpret object-level information in figures, such as categorizing an image [18], recognizing text within a figure [16, 17], or generating a qualitative description of the image [20], but they still struggle to capture pixel-level information such as precise curve coordinates, which matter most because they represent the actual spectroscopy data. For example, MatGD [25] reported only 66% accuracy for data-line separation, with particular difficulty when data lines share similar colors or overlap significantly. Similarly, DIVE [23] found that reading key points from data curves with a VLM (Gemini-2.5-Flash + Deepseek-R1) required up to a 50% relative-error tolerance because of visual reading noise. Circi et al. [26] likewise reported that the precision of o1 in extracting tabular data from figures ranged from only 23% to 45%. Typical errors included hallucinated points at convenient x-values, trend-like but imprecise coordinates, and missed tightly clustered or overlapping points. Our own tests directly applying VLMs (GPT-5.2 and Opus 4.6) to extract data from XAS figures show similar issues (Supplementary Figs. S1–S4). Together, these studies highlight substantial room for improvement in multimodal data extraction. These limitations indicate that dedicated models are needed to complement current VLMs and to form a pipeline that balances precise and flexible extraction. Leong et al. [16], for instance, demonstrated the advantages of combining a chemical-structure-recognition model, RxnScribe [27], with a VLM to extract multimodal reaction information. Spectroscopy data extraction, however, remains underexplored, and it is the focus of this work.

In this work, we provide a fully auto-generated, open-source dataset of 13,740 XAS spectra retrieved from 3,510 battery-related papers. We target XAS because it is a powerful element-specific probe of oxidation state and local atomic structure, yet costly to measure. Building on our previous work, including EXSCLAIM [28], Plot2Spectra [29], and figure extraction for optical emissivity [30], we developed a scalable spectroscopy data digitization pipeline (Fig. 1) that searches for and downloads full-text papers, classifies images by

spectroscopy relevance, as well as orchestrates multimodal agents for curve separation, axis and legend recognition, and contextual information extraction. Starting from 485,628 battery-related papers and 4,112,327 figures, the pipeline filtered the collection down to XAS-relevant figures and successfully digitized 13,740 XAS curves into machine-readable numerical spectral data, each accompanied by metadata on the measured edge and material. The dataset is publicly available in JSONL format, joining the community effort to grow curated databases and make more XAS data findable and reusable. By making these data readily accessible, the digitized dataset lowers the barrier to data-driven analysis of XAS measurements and facilitates the characterization of materials and electrochemical systems for autonomous discovery.

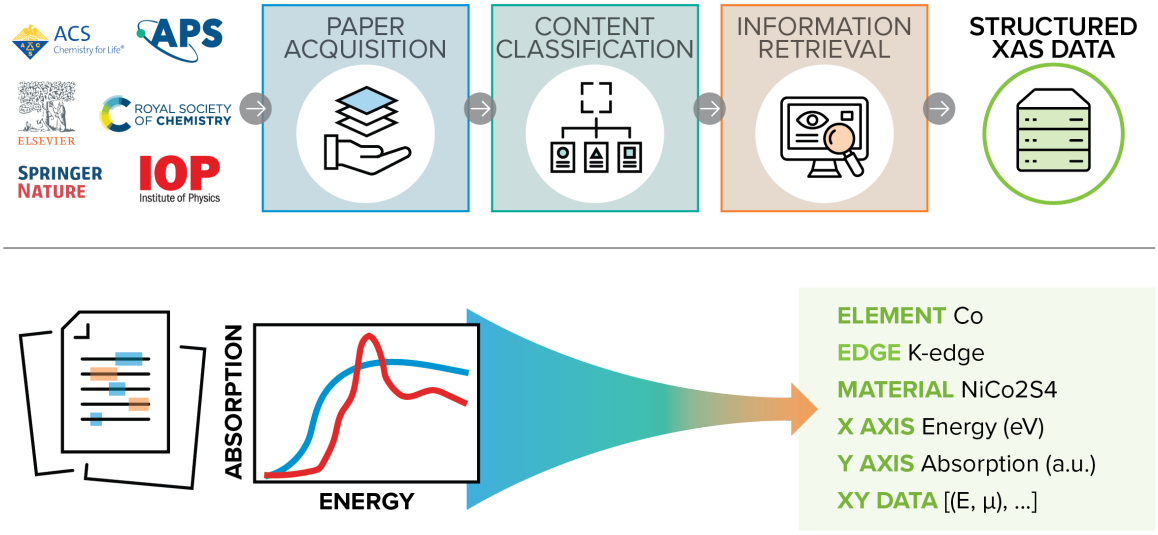

Figure 1. Schematic of the spectroscopy data digitization pipeline proposed in this work. Top panel: the pipeline converts publisher articles into a structured XAS dataset via paper acquisition, content classification, and information retrieval. Bottom panel: example of processing an XAS figure and its contextual text into a machine-readable record containing the absorbing element, absorption edge, material, axis labels, and energy-dependent absorption coordinates.

Methods

Paper acquisition

To keep the scope manageable, we focused on battery-related papers in this work rather than the entire body of scientific publications, as batteries constitute a field in which advanced characterization techniques in general, and XAS in particular, are widely applied to diverse materials. The publications used in this work are journal articles from Springer Nature, Elsevier, the Royal Society of Chemistry (RSC), IOP Publishing (IOP), and the American Physical Society (APS), for which we obtained text and data mining permissions, together with open-access papers from the American Chemical Society (ACS). Using EXSCLAIM to leverage each publisher's search engine, we identified papers containing the keywords "battery" or "batteries" and retrieved their Digital Object Identifiers (DOIs). We then used Playwright [31] and the publishers' APIs to download the full-text articles in HTML/XML format. High-resolution images were also downloaded as needed for downstream processing. Because full-text articles published before 1995 are mostly in PDF format, which complicates data processing, we restricted our search to papers published from 1995 onward. All data are stored and managed in an internal document-oriented database implemented in MongoDB. Finally, the HTML/XML papers were parsed into natural-language paragraphs and figure captions using LimeSoup [32], which removes irrelevant markup while preserving the paper structure and section headings.

Content classification

Figures in scientific papers are highly flexible and vary widely in content. In this work, we focus on XAS figures for data extraction because XAS measurements are costly, typically requiring competitively allocated synchrotron beamtime. Its element specificity and sensitivity to local electronic and atomic structure make it especially well suited for tracking oxidation-state and coordination changes in battery research. Recovering XAS data from the published literature is therefore particularly valuable. To identify XAS figures of interest, we use a multi-stage filtering process in which inexpensive filters are applied first, followed by more sophisticated filters once the dataset has been reduced to a more manageable size.

The filtering proceeds in three stages. We first classify figures based on their captions using a self-hosted open-source LLM, Mixtral-8x22B-Instruct-v0.1, to determine whether each figure contains spectroscopy results and, specifically, whether it corresponds to XAS. We then download only the XAS figures to conserve download bandwidth. Because the downloaded figures may be either single-panel figures or multi-panel composites of several subfigures, we use FigureSeparator [33] in EXSCLAIM to separate multi-panel figures into individual subfigures. Finally, we apply a VLM, GPT-5.2 (gpt-5.2-2025-12-11), to classify each subfigure according to whether it shows typical X-ray absorption near edge structure (XANES) spectra, presented as line plots of absorption intensity as a function of photon energy. Subfigures showing non-XAS data, Fourier-transformed (FT) XAS data (FT magnitude versus radial distance), extended X-ray absorption fine structure (EXAFS) oscillations (signal versus wavevector in k-space), or in-situ XAS visualizations (waterfall or colormap plots) are excluded.

Information retrieval

For each XAS figure, our information-retrieval workflow consists of multiple agents that analyze the figure, caption, and paper text to extract the XAS results and the necessary metadata for identifying the XAS measurement, including the axis information, precise curve coordinates, curve legend, absorbing element, absorption edge, measured material and constituent elements. More details are provided in the following subsections. GPT-5.2 is used as the VLM model in the information-retrieval workflow.

Axis understanding

Axis labels and tick information are extracted using a combination of optical character recognition (OCR) and a VLM. Because the VLM is adept at capturing text and understanding its role in a figure, we first use it to determine which text corresponds to the axis labels, i.e., the text identifying the physical quantity and unit of each axis. The VLM does not, however, provide accurate text positions; we therefore employ an OCR model, PP-OCRv5 [34], to capture the values and positions of the axis tick labels, i.e., the numerical values marked along each axis. We retain only those tick labels whose values are cross-checked by the VLM.

The resulting tick label positions and values are used to construct a linear transformation that converts the pixel positions of curve data points into energy-dependent absorption-intensity coordinates.

Curve segmentation

To capture the precise curve coordinates of XAS results, we developed a two-step curve segmentation algorithm based on connectivity and color decomposition. Because a typical XAS figure contains multiple curves that overlap, intersect, and span the full width of the plot, we use color signals to identify the color-consistent portions of a target curve and connectivity information to recover the remaining parts, which are connected to these portions but may be occluded by other curves. While this algorithm mitigates confusion from elements disconnected from the curves, such as annotations and legends, it cannot recover dashed or dotted lines, which we leave for future work.

We first convert the plot into a line-segment connectivity graph, in which each segment is a connected, non-background pixel region between two adjacent intersection points of the curves. The curve segmentation task then reduces to finding the combination of line segments that best reconstructs a single complete curve trace corresponding to one XAS measurement. To identify this combination, we apply the BIRCH algorithm [35] to decompose the plot into clusters of similarly colored pixels, and we assign higher reward to combinations with higher color consistency, i.e., those whose pixels predominantly belong to a single cluster. We further introduce a roughness penalty that suppresses the sharp transitions arising when segments from different curves are joined, thereby favoring combinations that trace a smooth path. The reward function is defined in Eq. 1:

$$\text{reward}(\ell) = \frac{|\ell \cap C_{\text{color}}|}{W\,\hat{w}} - \frac{\lambda}{N}\sum_{i=1}^{N} \left|f''(x_i)\right|, \tag{1}$$

where ℓ denotes a candidate line (combination of segments), C_{color} is the set of pixels in a color-decomposition cluster, W is the image width, $\hat{w}$ is the estimated line width of ℓ, f is the centerline trajectory of the finite-width plotted curve ℓ, f'' is its second derivative, $\{x_i\}_{i=1}^{N}$ are the N points along f, and λ controls the roughness penalty. The value of λ

was chosen empirically (as 0.01) so that the roughness penalty takes effect only when the color-consistency term alone cannot distinguish between candidate lines. With the curves separated, we convert each centerline trajectory from pixel positions to energy-dependent absorption-intensity coordinates using the linear transformation constructed from the axis tick label positions and values in Section Axis understanding.

Legend recognition

Based on the curves separated in Section Curve segmentation, we attribute the legends in the figure to their corresponding curves. For each separated curve, we synthesize a side-by-side highlighting visualization that places the original figure on the left and an inpainted version on the right, in which the target curve is emphasized while the other visual components are faded. We then ask GPT-5.2 to interpret this highlighting visualization and recognize the legend corresponding to the emphasized curve. In practice, an ensemble approach that queries the VLM multiple times improves the stability and accuracy of inference. We therefore repeat the procedure twice using two slightly different inpainting strategies: in one, the highlighted curve is rendered with the average color of its pixels, while in the other it retains its original pixel-level color variation. We accept a legend assignment only when the two inferences agree, thereby trading recall for precision.

Contextual metadata extraction

To contextualize each XAS measurement, we extract the necessary metadata from the figure caption and paper text, ensuring that the resulting data record is self-contained and interpretable without reference to the original paper.

Figure legends often contain only the minimal information needed to distinguish curves within a single figure, such as sample identifiers or the values of key experimental variables; additional context is therefore required to link each legend entry to the corresponding explanations in the caption and the detailed descriptions in the main text. Because the full paper text is typically lengthy and is reused for extracting different types of metadata, we first use GPT-5.2 to condense it by extracting verbatim descriptions relevant to the figure, including discussions of the figure itself; explanations of abbreviations, sample identifiers, and

terminology; and descriptions of the experimental conditions associated with the measurement. We then provide GPT-5.2 with the highlighting visualization of the target curve, its associated legend, the figure caption, and the extracted figure-related descriptions, and ask it to identify the absorbing element, the absorption edge, and the material being measured together with its constituent elements.

Dataset generation

Using the spectroscopy data digitization pipeline, we searched and found 485,628 battery-related papers and downloaded full text for 460,440 of them. The downloaded papers contain 4,112,327 figure captions. After content classification, 6,055 figures were identified as figures of interest that contain line plots of XAS spectra and were downloaded for downstream processing. Through information retrieval, 4,312 of the 6,055 figures yielded at least one complete curve-level record, resulting in 13,740 digitized XAS curves with machine-readable numerical spectral data and accompanying metadata on the measured edge and material.

Data Records

We provide the complete dataset of 13,740 XAS spectra as a single JSONL file, which will be available at Figshare upon publication. Each record corresponds to a single XAS curve, together with metadata describing its source and measurement details. The detailed data schema is shown in Table 1, and a complete example record is provided in Supplementary Fig. S5. The visual components extracted from each figure are stored in the `xy_data`, `x_axis_label`, `y_axis_label`, and `legend` fields. The spectral data points are stored in `xy_data` as a list of [x, y] pairs, where x is the photon energy and y is the absorption intensity, as indicated by the x- and y-axis labels. The `x_axis_label` and `y_axis_label` fields preserve the original axis text from the figure. The former contains the energy unit, typically eV and occasionally keV. The latter either labels the absorption intensity explicitly as "a.u." or leaves it unspecified, as the intensity is given in arbitrary units and only its relative magnitude is meaningful. The `legend` field stores the original text of the figure legend corresponding to the spectrum, serving as an anchor that links the spectrum to the figure caption and the

paper text for further details.

In addition to the visual components, metadata describing each spectrum are stored in the `measured_element`, `measured_edge`, `material`, and `elements` fields. The `measured_element` field contains the chemical symbol of the absorbing element, and `measured_edge` specifies the absorption edge of the spectrum, typically denoted K, L_2, L_3, etc. The `material` field reports the name or chemical formula of the measured material, drawn from the figure legend and supplemented with contextual information from the paper. For example, if the legend gives only a sample code, this field provides a concise chemical definition of the code together with the relevant experimental conditions. The `elements` field lists the chemical symbols of the constituent elements of the material, enabling element-based queries across the dataset.

To support source traceability and access to additional details when needed, the dataset provides the `doi` and `figure_label` fields. The DOI identifies the original paper, and the figure label locates the corresponding figure within that paper. The original figures and paper text are not included in the dataset, in accordance with text and data mining agreements.

Data description	**Data Key Label**	**Data Type**
DOI of the source paper	`doi`	*string*
Figure label within the source paper	`figure_label`	*string*
Legend of spectrum in the source figure	`legend`	*string*
x-axis label of the spectrum	`x_axis_label`	*string*
y-axis label of the spectrum	`y_axis_label`	*string*
Spectrum data points, shape $(N, 2)$	`xy_data`	*list* of [*float*, *float*]
Absorbing element	`measured_element`	*string*
Absorption edge (e.g. K, L_2)	`measured_edge`	*string*
Material name or chemical formula	`material`	*string*
Constituent elements of the material	`elements`	*list* of *strings*

Table 1. Schema of each dataset record, with fields for source traceability (`doi`, `figure_label`), figure-derived spectral content (`legend`, `x_axis_label`, `y_axis_label`, `xy_data`), and contextual measurement metadata (`measured_element`, `measured_edge`, `material`, `elements`).

Technical Validation

We validate the dataset through manual annotation to quantify extraction accuracy and through statistical analyses to characterize its diversity and coverage.

Extraction accuracy

To evaluate the accuracy of the extracted data, we randomly sampled 100 records from the dataset for verification by materials science experts (annotation interface in Supplementary Fig. S6). For the figure components, the experts visually inspected each figure to confirm that the extracted axis labels, tick labels, and legend entries matched the corresponding text in the original figure. Tick-label placement was verified by overlaying the extracted tick values on the original figure at their extracted positions and checking their alignment with the corresponding tick labels. The tick labels of an axis are counted as correct if every extracted label has the correct value and position and at least two are extracted, the minimum needed for the linear transformation to data coordinates (Section Axis understanding). Curve separation was assessed by overlaying the extracted curve on the original figure and verifying that it followed the original trace without visible discrepancy. To calibrate this visual criterion, we manually annotated 10 curves point by point; curves judged to be in visual agreement had deviations averaging $<1\%$ of the intensity range and falling below 3% for 99% of points.

For the metadata, the experts checked the measured element, edge, material, and constituent elements against the legend and paper text to confirm consistency with the original figure and accompanying text. Each record was expected to specify one elemental absorption edge, such as Fe K or Ni L, for the extracted curve. The `material` field was expected to resolve the legend into a material name or chemical formula, especially when the legend mentioned only an experimental variable and omitted the chemical identity, which then had to be supplemented from the paper text. The `elements` field was evaluated against the material composition, with missing or extra elements counted as errors.

The accuracy results are summarized in Table 2. Axis information is extracted with essentially perfect accuracy, reaching 100% for both labels and ticks. Legend recognition

reaches 97% accuracy. This task is more challenging than simply reading text from a figure, because the VLM must distinguish individual curves and establish a one-to-one correspondence between data curves and legend entries. Legend errors occur primarily when neighboring curves are visually similar or overlap heavily, leading the model to assign an incorrect legend entry to a curve. Curve separation achieves 91% accuracy and is the most challenging task. Because our manual evaluation requires each extraction to capture the entire curve, even a single local error is counted as an invalid extraction. In the failure cases, the discrepancy is typically localized to a small region where multiple curves intersect or overlap extensively, while the majority of the curve remains well captured (examples in Supplementary Fig. S7). Metadata extraction is nearly perfect, reflecting the proficiency of GPT-5.2 on textual tasks. Accuracy is 100% for both the measured element and the edge, 97% for the material, and 98% for its constituent elements. Errors in the material field are correlated with legend recognition, since an incorrect legend can lead to an incorrect material name or a misspecification of the experimental conditions. Constituent-element errors are similarly linked to material-field errors, occurring when the chemical portion of the extracted material is incorrect. Taken together, 89% of the sampled records have all figure components and metadata extracted correctly.

(a) Figure components		**(b) Metadata**	
Attribute	**Accuracy** (%)	**Attribute**	**Accuracy** (%)
Axis labels	100	Measured element	100
Axis ticks	100	Measured edge	100
Legend	97	Material	97
Curve separation	91	Elements	98

Table 2. Accuracy of the spectroscopy data digitization pipeline, evaluated on 100 records randomly sampled from the dataset and verified by materials science experts.

Dataset mining

To assess the diversity and coverage of the dataset, we conducted statistical analyses on its temporal, application, compositional, and inter-laboratory aspects.

We first examined the distribution of the source papers by publication year and application background (Fig. 2). After the digitization pipeline, the 13,740 spectra in the final dataset originate from 3,510 papers. The distribution by publication year is shown in Fig. 2(a). The number of source papers grows approximately exponentially over time, consistent with the general trend of scientific publishing. Because our collection ends in mid-2025, the bar for 2025 is truncated. The continued growth of relevant papers offers an opportunity to build a larger dataset by reapplying the same pipeline over a longer collection period.

The battery applications mentioned in the introduction sections of the source papers are summarized in Fig. 2(b). Because a given paper may address more than one application, papers may be counted under multiple battery types. The source papers span a broad range of battery chemistries—including Li-, Na-, K-, Zn-, and Mg-based systems—along with more general mentions that do not specify a particular chemistry, such as "solid-state battery" or "metal-air battery." Because the dataset is drawn from the recent characterization literature, it is weighted toward chemistries under active investigation; mature commercial systems such as lead-acid and nickel–metal hydride (NiMH) are sparsely represented. This broad coverage demonstrates the effectiveness of the literature-mining approach in capturing diverse domain-specific information, illustrated here by battery applications. Li-based batteries account for the largest share, consistent with their commercial success and sustained prominence in battery research. Zn-based batteries also represent a notable portion of this XAS-focused dataset, largely owing to studies of Zn–air batteries, a widely investigated metal–air system in which XAS is routinely used to probe transition-metal species involved in the oxygen reduction and oxygen evolution reactions (ORR/OER) [36, 37].

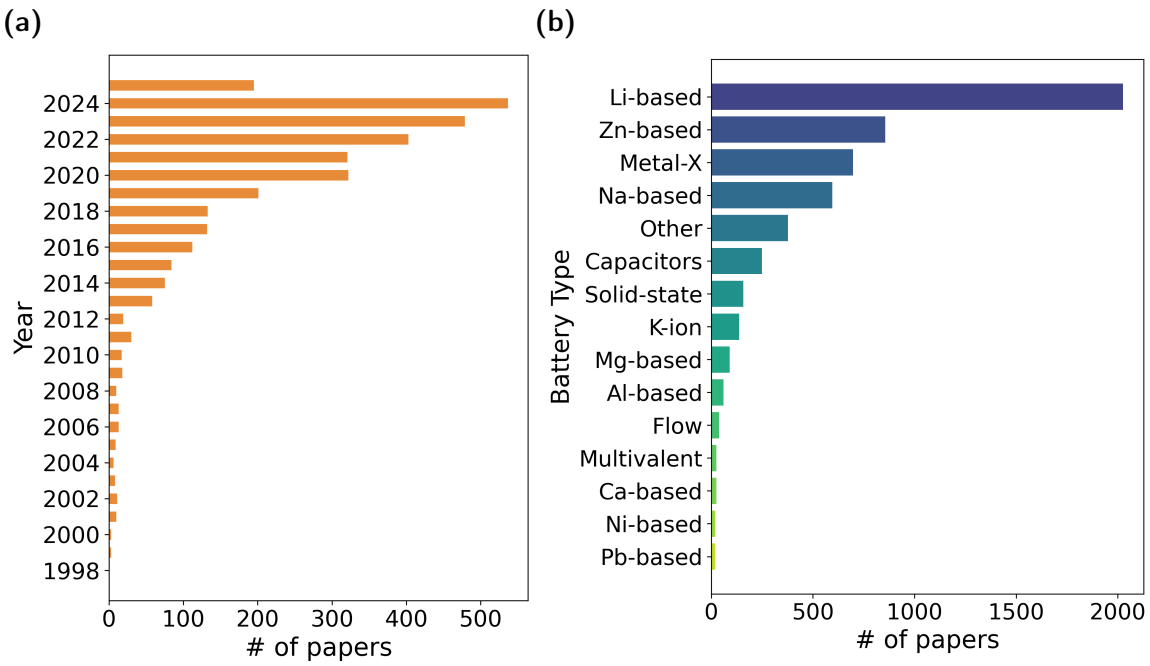

Figure 2. Distribution of source papers from which XAS spectra were extracted. **(a)** Paper counts by publication year. **(b)** Paper counts by battery application. General mentions without a specific metal are grouped as Metal-X (X = air, sulfur, CO_2, etc.), Solid-state, and Multivalent.

Next, we evaluate the chemical space covered by the dataset. The number of XAS spectra measured for each element is mapped onto the periodic table in Fig. 3 using a white-to-blue color gradient, with the count shown in each element box. The dataset spans a broad chemical space of 66 elements. The most frequently measured elements are Fe, Co, Ni, and Mn, reflecting the predominance of cathode materials based on $LiFePO_4$ and lithium nickel cobalt manganese oxide (NCM). Many moderately measured elements are associated with substitution or doping strategies within these frameworks, such as Al substitution and Zr, W, or B doping for enhanced structural stability and cycling performance. Beyond these conventional frameworks, alternative cathode chemistries such as disordered rock-salt (DRX) materials are represented in the dataset. DRX materials combine redox-active metals (e.g., Mn, V, Cr) with high-valent d^0 stabilizers (e.g., Ti^{4+}, Nb^{5+}, Mo^{6+}), reducing dependence on Co and Ni while achieving high capacity. The inclusion of these elements makes the dataset a useful reference for resolving XAS spectral diversity arising from multiple oxidation states

and varied local environments. The dataset also spans a wide energy range, from hard X-rays for heavy metals, such as Pb or Bi, to soft X-rays for nonmetal elements, such as C or O.

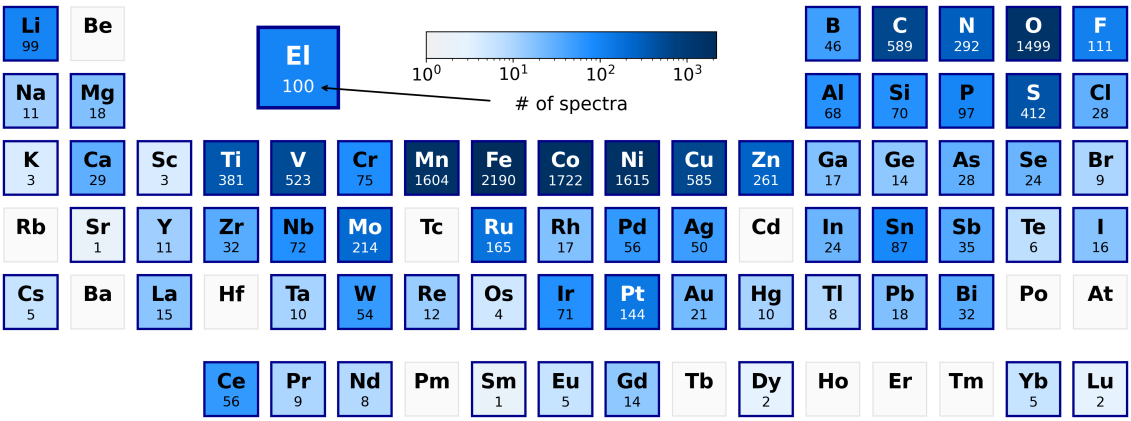

Figure 3. Elemental coverage of the dataset, presented as a periodic-table heatmap of the number of XAS spectra measured for each absorbing element. Counts are reported within each element box and encoded using a logarithmic white-to-blue color scale.

We also examine the chemical complexity of the materials in the dataset. As shown in Fig. 4(a), the dataset spans a broad range of material classes, including oxides and hydroxides, composites and mixtures, metals and alloys, non-oxide anion compounds, and framework or molecular materials. Oxides are the most frequently measured class, comprising both binary and multi-cation oxides, reflecting their widespread use as cathodes, solid electrolytes, and certain anodes in batteries. Composites, which are engineered architectures of two or more materials, also constitute a notable portion of the dataset, consistent with their use as electrocatalysts (e.g., Fe–N–C in Zn–air batteries), heterostructured electrodes (e.g., NCM@rGO), and coated materials (e.g., $LiFePO_4$/C). Beyond simple materials such as elemental metals and binary oxides, the dataset includes multi-component compounds, such as ternary, quaternary, quinary, and high-entropy oxides (Fig. 4(b)). The inclusion of these complex materials provides reference spectra for resolving the local atomic environments of advanced material systems.

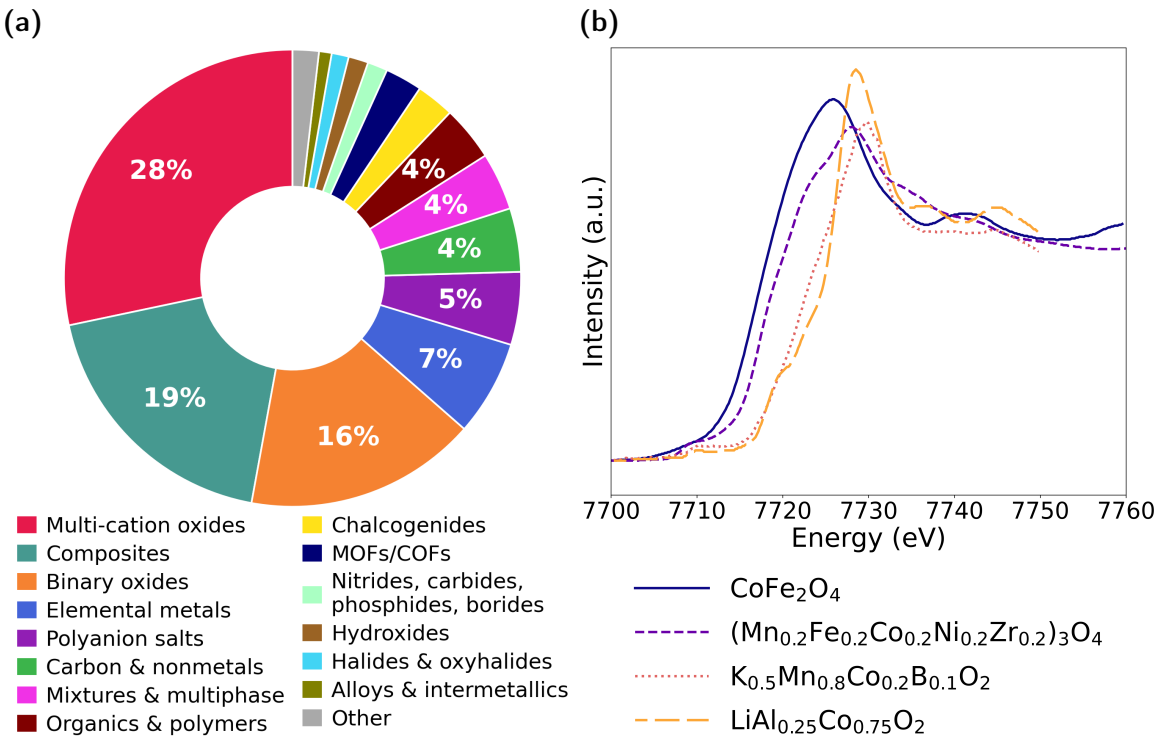

Figure 4. Chemical diversity and complexity of the materials in the dataset. **(a)** Distribution of material classes by composition. **(b)** Representative Co K-edge XAS spectra of multi-component oxide materials beyond binary oxides, including ternary, quaternary, quinary, and high-entropy oxides.

Finally, the literature-mined dataset allows us to evaluate how measurements of the same material vary across laboratories, providing a valuable basis for standardization efforts. As an example, we overlay 146 Co K-edge XAS spectra of CoO in Fig. 5(a). Because XAS intensities are commonly reported in arbitrary units, each spectrum is vertically rescaled to set its highest peak to a common reference intensity, defined as the median peak intensity of comparable, consistently normalized literature spectra. Inspection of the curves that deviate markedly from the typical pattern reveals a few anomalous spectra. For instance, one reports the Co(II) K-edge within 7100–7180 eV, likely a plotting error [38], while another corresponds to CoO after electrochemical cycling and exhibits a post-edge structure more consistent with mixed Co^{2+}/Co^{3+} states [39]. Spectra with clearly identifiable errors are excluded from Fig. 5, whereas the remaining outliers are retained. This example illustrates that aggregating

spectra from many sources facilitates the identification of anomalies and outliers.

For each spectrum, the edge energy is defined as the energy at which the first derivative of the spectrum with respect to energy reaches its maximum within the edge region. The distribution of edge energies (Fig. 5(b)) has a median of 7720.8 eV and an interquartile range (IQR; the spread between the 25th and 75th percentiles) [40] of 1.8 eV. The median and IQR are reported here because they are more robust to outliers than the mean (7720.5 eV) and standard deviation (2.4 eV). In addition, we compute the energy spacing between two adjacent post-edge peaks (Fig. 5(c)). This internal energy difference is less sensitive to absolute energy-calibration shifts while retaining structural relevance, as post-edge features primarily reflect multiple-scattering effects and the medium-range order around the absorber atom [41]. The spacing has a median of 32.7 eV and an IQR of only 0.5 eV. The smaller spread in this internal spacing suggests that the apparent variation in edge energy could be reduced through improved cross-laboratory alignment of the reference energy and corresponding shifts in the edge positions. Because individual measurements are subject to fluctuations, statistics aggregated over many results provide a more reliable basis for identifying peaks and trends. The dataset thus offers a quantitative reference for expected discrepancies across laboratories and informs the development of standards for cross-lab alignment and improved consistency.

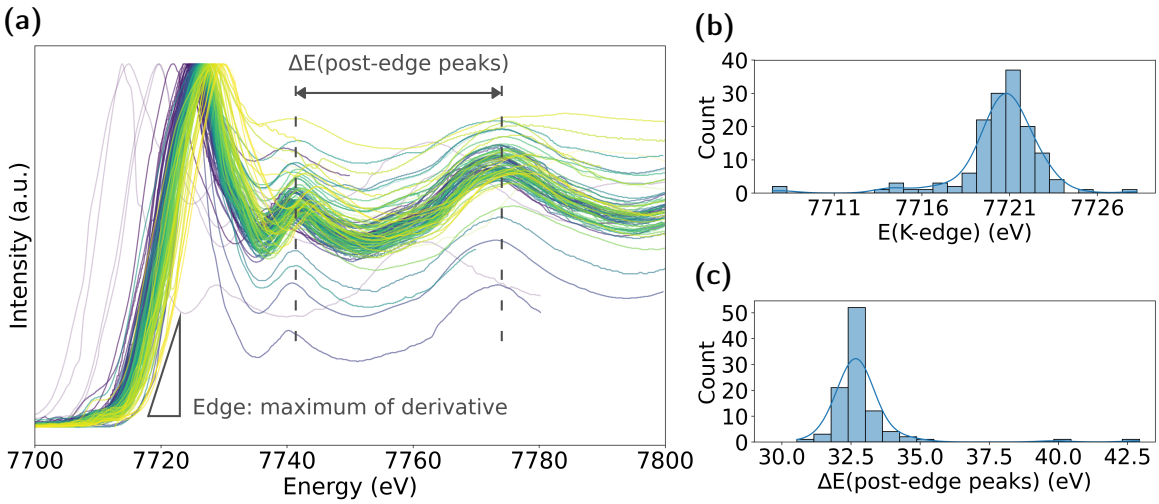

Figure 5. Cross-laboratory comparison of 146 Co K-edge XAS spectra of CoO. **(a)** Overlay of the 146 spectra, each vertically rescaled to set its highest peak to the same reference intensity. **(b)** Histogram of edge energies, defined by the maximum first-derivative position in the edge region. **(c)** Histogram of the energy spacing between the two adjacent post-edge peaks for spectra in which both peaks are resolved.

Overall, we have constructed a large-scale, multimodal XAS dataset from the battery literature, pairing digitized spectra with essential measurement metadata as an AI-ready resource for foundation-model-scale analysis. As a reference library, the dataset complements existing curated databases [1, 2, 3, 4, 5] with broader materials coverage for identifying trends in oxidation state, local coordination, and composition. As training and evaluation data, the spectra and associated metadata can support machine-learning models for automated XAS interpretation and retrieval systems that connect new spectra to prior measurements. Spanning thousands of source-traceable studies, the dataset also enables assessing experimental variability, harmonizing reporting practices, and linking spectroscopy to materials and their broader experimental context in multimodal knowledge bases.

Usage Notes

The dataset is provided as a single JSONL file, readable with the standard libraries of major programming languages, including Python, MATLAB, R, and Wolfram Mathematica, with

no additional dependencies. For the spectral XY data, the values and units are preserved as reported in the original figure. The energy unit of each record is given in its `x_axis_label` field; x values in keV can be converted to eV by multiplying by 1,000. As absorption intensities are in arbitrary units, some figures omit numerical y-axis ticks; for these records, the y values are given in image coordinates, increasing from top to bottom of the image. When plotting such spectra, the y-axis can be inverted such that intensity increases upward, as conventionally displayed. Each record can be traced back to its source paper and figure through the `doi` and `figure_label` fields, if details beyond the dataset are needed.

Data Availability

The dataset described in this Data Descriptor will be available at Figshare upon publication.

Code Availability

The code used for XAS data digitization will be available on GitHub upon publication.

References

[1] International X-ray Absorption Society. XASLIB: X-ray absorption data library. https://xaslib.xrayabsorption.org (2026). Accessed: 2026-06-28.

[2] Ishii, M. *et al.* Global cross-database search system for X-ray absorption spectra **32**, 661–668. URL https://journals.iucr.org/s/issues/2025/03/00/ing5006/.

[3] Spasyuk, D. XASDB – Design and Implementation of an Open-Access Spectral Database. URL http://arxiv.org/abs/2509.13566. 2509.13566.

[4] Paripsa, S. *et al.* RefXAS: An open access database of X-ray absorption spectra **31**, 1105–1117. URL https://journals.iucr.org/s/issues/2024/05/00/up5002/.

[5] Kieffer, I. & Testemale, D. SSHADE/FAME: "French absorption spectroscopy beamline in material and environmental science" database service. SSHADE (OSUG Data Center). Service/Database. https://doi.org/10.26302/SSHADE/FAME (2016).

[6] Chen, Y. *et al.* Robust machine learning inference from X-ray absorption near edge spectra through featurization. *Chemistry of Materials* **36**, 2304–2313 (2024). URL https://doi.org/10.1021/acs.chemmater.3c02584.

[7] Jia, H. *et al.* Revealing local structures through machine-learning-fused multimodal spectroscopy. *ACS Nano* **20**, 4228–4240 (2026). URL https://doi.org/10.1021/acsnano.5c16942.

[8] Liu, S. & Cole, J. M. Automated determination of the molecular substructure from nuclear magnetic resonance spectra using neural networks. *Journal of Chemical Information and Modeling* **65**, 8435–8447 (2025). URL https://doi.org/10.1021/acs.jcim.5c00499.

[9] Fei, Y. *et al.* Agentic llm reasoning in a self-driving laboratory for air-sensitive lithium halide spinel conductors. *arXiv preprint arXiv:2604.11957* (2026).

[10] Huang, X. *et al.* Cascade: Cumulative agentic skill creation through autonomous development and evolution. *arXiv preprint arXiv:2512.23880* (2025).

[11] Huang, X., Chen, J., Schwaller, P. & Ceder, G. Skillpuzzler: A self-evolving agentic framework for materials and chemistry research with minimal reliance on predefined tools. In *NeurIPS 2025 AI for Science Workshop*.

[12] Penfold, T. *et al.* Machine-learning strategies for the accurate and efficient analysis of x-ray spectroscopy. *Machine Learning: Science and Technology* **5**, 021001 (2024).

[13] Wang, Y. *et al.* An x-ray absorption spectrum database for iron-containing proteins. *arXiv:2504.18554v1* (2025).

[14] He, T. *et al.* Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science Advances* **9**, eadg8180 (2023). URL https://www.science.org/doi/10.1126/sciadv.adg8180.

[15] Huo, H. *et al.* Machine-learning rationalization and prediction of solid-state synthesis conditions. *Chemistry of Materials* **34**, 7323–7336 (2022). URL https://doi.org/10.1021/acs.chemmater.2c01293.

[16] Leong, S. X., Pablo-García, S., Wong, B. & Aspuru-Guzik, A. MERMaid: Universal multimodal mining of chemical reactions from PDFs using vision-language models. *Matter* **8** (2025). URL https://www.cell.com/matter/abstract/S2590-2385(25)00374-1.

[17] Leong, S. X., Pablo-García, S., Zhang, Z. & Aspuru-Guzik, A. Automated electrosynthesis reaction mining with multimodal large language models (MLLMs). *Chemical Science* **15**, 17881–17891 (2024). URL https://pubs.rsc.org/en/content/articlelanding/2024/sc/d4sc04630g.

[18] Zheng, Z. *et al.* Image and data mining in reticular chemistry powered by GPT-4V. *Digital Discovery* **3**, 491–501 (2024). URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d3dd00239j.

[19] Dong, Q. & Cole, J. M. Auto-generated database of semiconductor band gaps using ChemDataExtractor. *Scientific Data* **9**, 193 (2022).

[20] Liu, Y. *et al.* A multimodal dataset of causal mechanisms in materials science literature. *Scientific Data* **13**, 269 (2026). URL https://www.nature.com/articles/s41597-026-06598-5.

[21] Siegel, N., Lourie, N., Power, R. & Ammar, W. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries*, 223–232 (2018).

[22] Park, G. & Pouchard, L. Advances in scientific literature mining for interpreting materials characterization. *Machine Learning: Science and Technology* **2**, 045007 (2021).

[23] Zhang, D. *et al.* "DIVE" into hydrogen storage materials discovery with AI agents. *Chemical Science* **17**, 3031–3042 (2026). URL https://pubs.rsc.org/en/content/articlelanding/2026/sc/d5sc09921h.

[24] Sayeed, H. M., Smallwood, W., Baird, S. G. & Sparks, T. D. NLP meets materials science: Quantifying the presentation of materials data in literature. *Matter* **7**, 723–727 (2024). URL https://www.cell.com/matter/abstract/S2590-2385(23)00645-8.

[25] Lee, J., Lee, W. & Kim, J. MatGD: Materials Graph Digitizer. *ACS Applied Materials & Interfaces* **16**, 723–730 (2024). URL https://doi.org/10.1021/acsami.3c14781.

[26] Circi, D. *et al.* Information extraction from diverse charts in materials science. In *LLM for Scientific Discovery: Reasoning, Assistance, and Collaboration* (2025).

[27] Qian, Y., Guo, J., Tu, Z., Coley, C. W. & Barzilay, R. RxnScribe: A sequence generation model for reaction diagram parsing. *Journal of Chemical Information and Modeling* **63**, 4030–4041 (2023). URL https://doi.org/10.1021/acs.jcim.3c00439.

[28] Schwenker, E. *et al.* EXSCLAIM!: Harnessing materials science literature for self-labeled microscopy datasets. *Patterns* **4** (2023). URL https://www.cell.com/patterns/abstract/S2666-3899(23)00222-2.

[29] Jiang, W. *et al.* Plot2Spectra: An automatic spectra extraction tool. *Digital Discovery* **1**, 719–731 (2022). URL https://pubs.rsc.org/en/content/articlelanding/2022/dd/d1dd00036e.

[30] Baibakova, V., Elzouka, M., Lubner, S., Prasher, R. & Jain, A. Optical emissivity dataset of multi-material heterogeneous designs generated with automated figure extraction. *Scientific Data* **9**, 589 (2022). URL https://www.nature.com/articles/s41597-022-01699-3.

[31] Microsoft. Playwright: Fast and reliable end-to-end testing for modern web apps. https://playwright.dev (2026). Accessed: 2026-06-28.

[32] Kononova, O. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data* **6**, 203 (2019). URL https://www.nature.com/articles/s41597-019-0224-1.

[33] Jiang, W. *et al.* A two-stage framework for compound figure separation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1204–1208 (IEEE, 2021). URL https://ieeexplore.ieee.org/document/9506171.

[34] Cui, C. *et al.* PaddleOCR 3.0 Technical Report. *ArXiV 2507.05595* (2025). URL https://arxiv.org/abs/2507.05595.

[35] Zhang, T., Ramakrishnan, R. & Livny, M. Birch: an efficient data clustering method for very large databases. *ACM SIGMOD Record* **25**, 103–114 (1996).

[36] Chen, J., Huang, X., Hua, C., He, Y. & Schwaller, P. A multi-modal transformer for predicting global minimum adsorption energy. *Nature Communications* **16**, 3232 (2025).

[37] Chen, J., Huang, X., Hua, C., He, Y. & Schwaller, P. Adsgt: Graph transformer for predicting global minimum adsorption energy. In *NeurIPS 2023 AI for Science Workshop* (2023).

[38] Chen, Y. *et al.* Precise solid-phase synthesis of CoFe@FeOx nanoparticles for efficient polysulfide regulation in lithium/sodium-sulfur batteries **14**, 7487. URL https://www.nature.com/articles/s41467-023-42941-9.

[39] Wu, M. *et al.* Cobalt (II) oxide nanosheets with rich oxygen vacancies as highly efficient bifunctional catalysts for ultra-stable rechargeable Zn-air flow battery **79**, 105409. URL https://www.sciencedirect.com/science/article/pii/S2211285520309861.

[40] Tukey, J. W. *Exploratory Data Analysis* (Addison-Wesley, Reading, MA, 1977).

[41] Fdez-Gubieda, M. L., García-Prieto, A., Alonso, J. & Meneghini, C. X-ray absorption fine structure spectroscopy in Fe oxides and oxyhydroxides. In Faivre, D. (ed.) *Iron Oxides: From Nature to Applications*, 397–422 (Wiley-VCH Verlag GmbH & Co. KGaA, 2016). URL https://doi.org/10.1002/9783527691395.ch17.

Author Contributions

Conceptualization: T.H., L.W., I.T.F., and M.K.Y.C. Methodology: T.H. and A.V. Investigation: T.H., A.V., and X.H. Visualization: T.H. and Y.C. Supervision: L.W., A.J., G.C., R.S.A., I.T.F., and M.K.Y.C. Manuscript writing: T.H., I.T.F., and M.K.Y.C. drafted the manuscript; A.V., L.W., and A.J. provided critical revisions; all authors reviewed and contributed to the manuscript.

Competing Interests

The authors declare no competing interests.

Acknowledgements

We thank the Argonne Research Library for assistance in obtaining Text and Data Mining agreements with the specified publishers. We thank W. Jiang, K. Chard, R. Underwood, C. Siebenschuh, J. Huang, and H. Zheng for valuable discussions, and G. Burns III, T. Cohron, and A. Avarca for technical support with data management.

Funding

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Basic Energy Sciences, Scientific Discovery through Advanced Computing (SciDAC) program under the FORUM-AI project. Work performed at the Center for Nanoscale Materials, a U.S. Department of Energy Office of Science User Facility, was supported by the U.S. DOE, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. Y.C. and M.K.Y.C acknowledge support by the Energy Storage Research Alliance "ESRA" (DE-AC02-06CH11357), an Energy Innovation Hub funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences. T.H. acknowledges the support from Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. This research used resources of the Argonne Leadership Computing Facility and the National Energy Research Scientific Computing Center, which are U.S. Department of Energy Office of Science User Facilities.