Materials Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mater. Adv.*, 2025, 6, 2543

Advanced scientific information mining using LLM-driven approaches in layered cathode materials for sodium-ion batteries†

Youwan Na,‡ Jeffrey J. Kim,‡ Chanhyoung Park, Jaewon Hwang, Changgi Kim, Hokyung Lee and Jehoon Lee *

Received 3rd January 2025,
Accepted 27th February 2025

DOI: 10.1039/d5ma00004a

rsc.li/materials-advances

Materials informatics (MI) has emerged as a powerful paradigm for accelerating materials discovery and development through data-driven approaches. The scarcity of structured materials data, however, remains a critical bottleneck in minimizing the error between experimental and predicted values. Here, we present an advanced large language model (LLM) framework for building a comprehensive materials database of layered metal oxide (LMO) cathode materials in sodium-ion batteries (SIBs). By implementing optimized advanced retrieval-augmented generation techniques, including the tree of clarity (ToC) methodology, our system achieved an accuracy of 0.8861 and an *F*1-score of 0.9371 in extracting structured materials data from open-source publications. The framework successfully processed 312 publications, rapidly extracting 945 data points related to material composition, crystallinity, operating voltage, and electrode composition at approximately 20 seconds per paper. This automated approach to materials data acquisition demonstrated here is expected to significantly accelerate the development of comprehensive materials databases and enable rapid materials discovery through MI.

Introduction

Energy storage systems are an essential part of modern society. Among various types, lithium-ion batteries (LIBs) are widely used due to their intrinsically high energy density and long lifespan.[1–3] However, challenges still lie in the cost and sustainability of LIBs because of the scarcity of lithium and unequal distributions of global lithium reserves.[4–6] In the light of such concerns, sodium-ion batteries (SIBs) have attracted attention as an alternative to LIBs, given the abundant resource, low cost, and eco-friendly nature of sodium.[6–8] Research on SIBs should be accelerated in order to advance the sustainable and price-competitive energy storage system.

However even to this day, the majority of materials research follows traditional methodologies that rely on the knowledge and experimental experience of individual researchers, placing a significant limitation on research performance.[9,10] To advance the research progress at an unprecedented rate, it is important to rapidly collect and integrate internal data from individual researchers and external data published by others. Therefore, an innovative methodology which efficiently collects, integrates, and applies the latest research data is necessary.

In recent years, there has been burgeoning interest in searching for methodologies that can accelerate materials research by implementing programming and artificial intelligence (AI) technologies.[11–13] One of the main areas of focus is the extraction of specific material information from published papers and its conversion into structured database (DB).[14,15] However, these innovative methods still face challenges. For instance, rule-based methods are largely effective but encounter difficulties as they require technical knowledge in fields such as chemical engineering, resulting in complications when attempting to apply and expand into other domains.[16] On the other hand, natural language processing (NLP) technologies have made tremendous advancements, but each specific research domain requires its own datasets and corresponding trained algorithms, presenting obstacles to their universal and widespread application.[17,18]

In the midst of these challenges, the emergence of high-performance large language models (LLMs) such as GPT-3 brought a revolutionary advancement.[19] LLMs have immense potential in the field of materials science to innovate methods by which scientific papers are collected, analyzed, and utilized.[20,21] This technology can extract and convert insightful data into the DB from the massive amount of published data without the need for complex code development or model trainings,

Technology Planning Department, LG Chem Ltd., 30 Magokjungang 10-ro, Gangseo-gu, Seoul, 07796, Republic of Korea. E-mail: jehoonlee@lgchem.com

† Electronic supplementary information (ESI) available: Constructed database. See DOI: https://doi.org/10.1039/d5ma00004a

‡ Youwan Na and Jeffrey J. Kim contributed equally to this work.

making it incomparably more effective than the traditional methods mentioned earlier. The utilization of LLMs will not only address the limitations of traditional methodologies but also lead to innovative advancement in materials science by accelerating research and enabling researchers to make more systematic decisions.[22,23]

In this study, we propose an innovative framework of utilizing LLMs to build a materials DB, specifically focusing on layered metal oxide cathode materials (LMOs), which are some of the key cathode active materials for SIBs. In particular, we address classification and data extraction issues by applying various retrieval augmented generation (RAG) techniques.[24] The main objective of this LLM approach is to overcome the methodological challenges of the traditional rule-based and NLP techniques and to provide widely scalable and efficient solution that could accelerate research. Through the implementation of LLMs, we anticipate a future where materials research becomes not only faster but also more accessible and impactful.

Experimental

Scientific research in batteries and electrochemistry demands systematic analysis due to the complex interplay of multiple parameters that determine the optimal performance. These studies require comprehensive examination across various aspects including electrochemical analysis, materials characterization, and system parameters. Electrochemical analysis encompasses characterization through various techniques and performance evaluation, while materials characterization involves synthesis optimization, structural analysis, and surface chemistry studies. System parameters focus on electrolyte composition, electrode formulation, and testing conditions, all of which significantly influence the overall performance.

The corpus of literature selected in this study encompasses the field of LMOs for SIBs. Journal metadata in excel format and full documentation of each paper were downloaded manually from Scopus, one of the widely recognized academic search engines.[25] The search term "Layered metal oxide cathode material for sodium ion battery" was used to search papers within "Article title, Abstract, Keywords" option and was limited to "all open access" publications, resulting in a total of 312 open access papers with diverse experimental results. All of the papers found were downloaded and utilized as the fundamental data for conducting analysis and evaluation throughout this study. Furthermore, to obtain objectivity and credibility of this investigation, 33 papers were randomly selected among 312 papers and stored as a set of test data for the evaluation of LLM performance in various tasks.

All documents underwent a conversion process from PDF to text format with GROBID (GeneRation Of BIbliographic Data), which is the library specifically designed to convert unstructured scholarly PDF documents into well-structured XML/TEI formats, systematically identifying and extracting key document components.[26] The text pre-processing stage involved excluding information such as copyright notices and page information from the PDF documents where target data were not present. The detailed description of GROBID is presented in S1 (ESI†).

To extract key information from the pre-processed text, we utilized an LLM-powered method called RAG (retrieval augmented generation), as shown in Fig. 1. RAG consists of three main stages for the data extraction process. In the first stage, prompt engineering involves designing systematic queries for information extraction and establishing a framework for data structuring. The detailed prompts utilized in this process are described in S2 (ESI†). In the second stage, information retrieval implements automated methods to extract and integrate information from multiple sources. In the final stage, answer generation systematically organizes the extracted data, analyzing composition–performance relationships, and derives insights for optimization. The key features to be extracted are as follows: (1) purpose, (2) strategy, (3) composition, (4) crystal, (5) coating, (6) morphology, (7) upper voltage, (8) lower voltage, (9) active material content, (10) conductive additive, (11) conductive additive content, (12) binder and (13) binder content.

The performance and reliability of the proposed LLM framework were evaluated using a comprehensive methodology consisting of four key dimensions: confusion matrix for quantitative measurement, economic efficiency analysis for resource utilization assessment, reliability and consistency evaluation for stability testing, and hallucination detection through the RAGAS framework.[27] The detailed evaluation methods and results for each dimension are provided in S3 (ESI†).

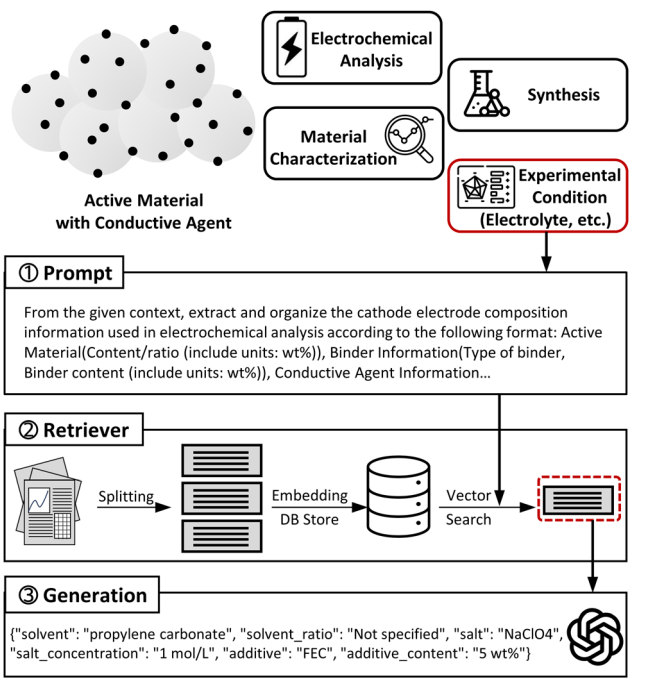

Fig. 1 Schematic illustration of scientific information extraction using the RAG method.

Results and discussion

The optimization of chunk size, which is a token length used to divide large document into segments of equal length, in the implementation of RAG systems using LLMs serves as a critical parameter determining system performance and reliability. Particularly in materials engineering, where accurate extraction of material properties, experimental conditions, physical data, and maintenance of contextual continuity significantly impacts the system performance, this can be achieved through optimizing the token length.

Inappropriate token lengths can lead to two major issues. First, an excessively short token length may fragment single experimental conditions or property data into multiple chunks, compromising information continuity. Second, overly long contexts may reduce model accuracy in identifying and extracting key information. This optimization becomes particularly crucial in the materials engineering domain, where the interconnectivity between material structure–properties and processing–performance relationships is highly intricate. Against this background, this study quantitatively evaluated the model performance by extracting material properties and experimental conditions from specialized literature in materials and electrochemistry fields under various chunk size conditions.

The analysis of performance evaluation using the GPT-4o[28] model showed a general trend of improved performance as the chunk size increased. All of the evaluations were based on extracting 13 features without separate analysis for each feature. As shown in Fig. 2a, in the confusion matrix, optimal performance was achieved at chunk size 2000, where the GPT-4o model recorded a Precision of 0.9335 and a Recall of 0.9111. Notably, it achieved an *F*1-score of 0.9221 and an Accuracy of 0.8555, demonstrating the excellent performance of this large language model and detailed evaluation results are provided in S4 (ESI†).

In terms of economic efficiency, the GPT-4o model showed a notable increase in cost and processing time as the chunk size increased. While input token usage increased dramatically from 590 to 7805, output token usage showed a relatively modest increase from 81 to 145. The processing time remained under 1 second for most intervals but reached its longest duration of 1.1230 seconds at 5000. In terms of cost, there was a consistent increase with chunk size, rising to $2.0958 per 100 questions at full context.

The reliability metrics of the GPT-4 model consistently demonstrated stable performance. Consistency remained at very high levels between 0.97 and 0.99 across all chunk sizes, while self-confidence improved with increasing chunk size, reaching 0.9241 at 2000. Semantic similarity showed particularly notable performance, maintaining above 0.9 across all intervals and reaching 0.9529 at 2000.

When analyzing these results comprehensively, the GPT-4o model showed optimal balance points across most performance metrics at chunk size of 2000. Faithfulness was recorded as 0.8717 in this interval, representing a highly stable performance level for an LLM. However, this excellent performance comes with the trade-off of increased costs and processing time, indicating that improving the overall cost efficiency remains an important challenge for the GPT-4o model.

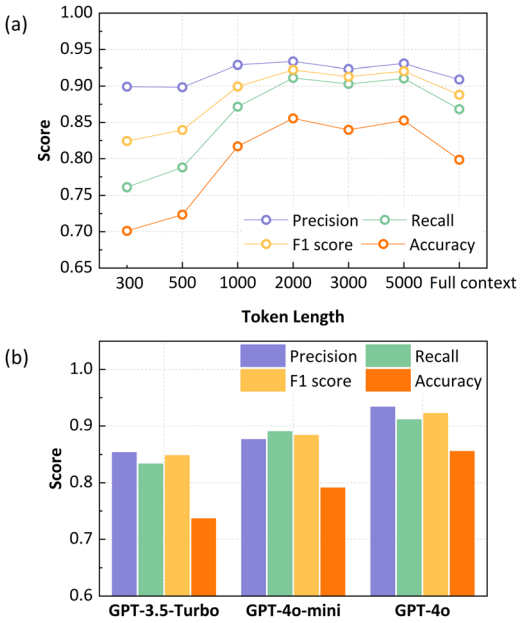

Fig. 2 Performance analysis of GPT models using confusion matrices. (a) Confusion matrix showing classification performance across different token lengths for the GPT-4 model. (b) Comparative confusion matrix analysis of various GPT models at a fixed chunk size of 2000 tokens.

As depicted in Fig. 2b, in comparison, both GPT-4o-mini and GPT-3.5-turbo models also showed an optimal performance at a chunk size of 2000. In the confusion matrix, GPT-4o-mini recorded a precision of 0.8766, a recall of 0.8900, an *F*1-score of 0.8832, and an accuracy of 0.7909 (S5, ESI†), while GPT-3.5-turbo achieved a precision of 0.8635, a recall of 0.8333, an *F*1-score of 0.8482, and an accuracy of 0.7363 (S6, ESI†). Both models demonstrated excellent performance in reliability metrics, with semantic similarity showing very similar levels: 0.9496 for GPT-4o-mini and 0.9476 for GPT-3.5-turbo.

The three models showed interesting differences in economic efficiency. While GPT-4o-mini took 0.8524 seconds of processing time and cost $0.0384 per 100 questions, GPT-3.5-turbo achieved faster processing at 0.5154 seconds with a similar cost of $0.0376 per 100 questions. Notably, in the consistency metric, GPT-4o-mini demonstrated superior stability with 0.9500 compared to GPT-3.5-turbo's 0.8427.

Considering the unique requirements of materials engineering, accuracy serves as the critical performance indicator over cost efficiency. Therefore, despite the trade-offs in processing time and cost, we chose to adopt the GPT-4o model with chunk size 2000 as the optimal parameter setting, given its superior performance across various metrics. However, recognizing that the current performance levels do not fully meet the stringent accuracy requirements of materials engineering, we applied a set of various techniques known as advanced RAG to enhance the model's performance.

Advanced RAG[29] is a sophisticated information retrieval and generation system designed to overcome the limitations of

naive RAG systems. While conventional RAG simply retrieves relevant documents and generates responses based on them, advanced RAG integrates various advanced techniques to produce more accurate and reliable outputs. We adopted several innovative approaches that demonstrated effectiveness in previous studies. The HyDE approach,[30] which generates hypothetical documents using LLMs before performing the actual search, showed significant performance improvements in zero-shot environments. The fusion RAG,[31] which expands queries into multiple sub-queries and performs re-ranking based on reciprocal rank fusion, substantially enhanced retrieval accuracy.

Furthermore, we implemented tree of clarification (ToC)[32] and Self-RAG,[33] which are respectively a recursive clarification approach through hierarchical sub-question generation and a framework for improving retrieval efficiency through dynamic search execution. ToC automatically prunes unnecessary clarification paths, while Self-RAG utilizes reflection tokens to perform retrieval only when necessary and evaluates the factuality and relevance of generated outputs.

Experimental results revealed that most advanced RAG techniques failed to achieve significant performance improvements compared to the baseline Naive RAG approach, as demonstrated in Fig. 3a. Quantitative evaluation using a confusion matrix showed that while HyDE achieved a Precision of 0.9409, its *F*1-score of 0.9223 was comparable to that of Naive RAG (0.9221). Similarly, RAG-Fusion and Self-RAG recorded *F*1-scores of 0.9213 and 0.9281, respectively, failing to demonstrate expected performance enhancements. Detailed evaluation metrics for each advanced RAG technique are provided in S7 (ESI†).

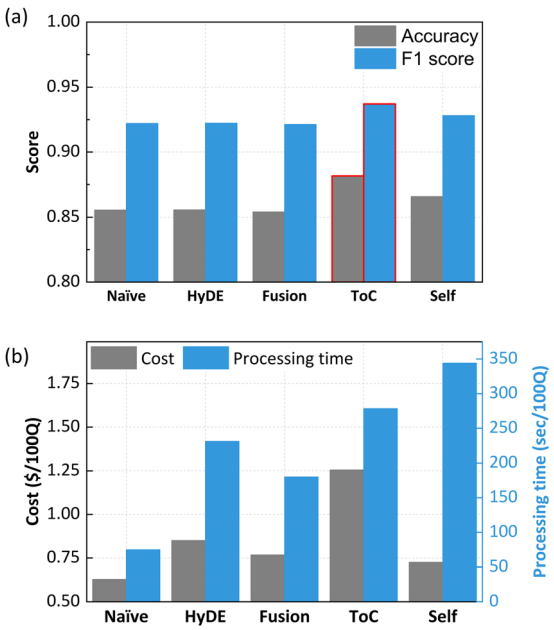

Fig. 3 Comparative analysis of advanced RAG techniques. (a) Performance evaluation showing accuracy and *F*1. (b) Efficiency metrics displaying computational cost and processing time requirements.

However, the tree of clarity (ToC) technique uniquely demonstrated notable performance improvements. ToC achieved an *F*1-score of 0.9371, marking a significant 1.5% improvement over Naive RAG, with balanced enhancements in both Precision (0.9481) and Recall (0.9264). This performance superiority can be attributed to ToC's structural characteristics aligning well with the specificity of chemical materials literature. ToC's recursive clarification mechanism effectively decomposed complex chemical compound queries, while its BFS-based exploration enabled comprehensive mapping of various chemical properties and relationships. Additionally, its automatic pruning functionality for efficient information filtering and long-form response generation capability contributed to accurately capturing complex chemical material characteristics.

Conversely, other advanced RAG techniques struggled to effectively handle the domain specificity of chemical materials literature. HyDE's virtual document generation approach faced challenges in ensuring accuracy for specialized information such as chemical structures and properties, while RAG-Fusion's query expansion strategy showed limitations in maintaining specialized context in the chemical materials field. Self-RAG's selective search mechanism also risked omitting critical information in this specialized domain.

Although these advanced RAG techniques showed varying levels of effectiveness, computational efficiency analysis revealed significant overhead across all approaches. As illustrated in Fig. 3b, processing times increased substantially compared to that of Naive RAG (0.7473s); HyDE required 2.3102s, RAG-Fusion required 1.8s, required ToC 2.7831s, and Self-RAG required 3.44s. Cost analysis also showed higher requirements compared to Naive RAG ($0.6287 per 100queries), with ToC particularly demanding the highest cost ($1.2556 per 100queries) and maximum token usage (averaging 3831 tokens).

These analytical findings suggest the necessity for careful consideration when implementing advanced RAG methodologies in specialized domains such as chemical materials literature. Particularly, methodologies excluding ToC demonstrated minimal performance improvements despite increased computational costs and slightly diminished self-confidence, indicating that Naive RAG optimization might represent a more effective approach for practical system implementation. However, in our chemical material information extraction task, where accurate information extraction took precedence over computational cost and processing time considerations, we proceeded with the final implementation using the ToC-based RAG system, which demonstrated superior performance. This decision was predicated on the paramount importance of accurate capture of chemical material complex characteristics and relationships and reliable information provision, despite elevated computational costs.

As shown in Fig. 4, we demonstrated the practical application and effectiveness of the ToC RAG system. In a large-scale information extraction task targeting 312 open-source publications, we extracted a total of 945 data points related to material composition, crystallinity, operating voltage, and electrode composition. The extraction process was conducted systematically, first

Index	Meta data				Specific Strategy		Material & Performance Data						Electrode composition				
	Title	Year	Journal	DOI	Purpose	Strategy	Composition	Crystal	Coating	Morphology	Upper voltage	Lower voltage	Active material content	conductive additive	Conductive additive content	Binder	Binder content
1	Impact of Pe	2024	Advanced F	10.1002/adfi	Cyclability	Doping	K0.7Fe0.5Mn0.5O2				4	1.5	80	uper P carbo	10	PTFE	10
2	Impact of Pe	2024	Advanced F	10.1002/adfi	Cyclability	Doping	Na0.7Fe0.5Mn0.5O2				4.2	1.6	80	uper P carbo	10	PTFE	10
3	Impact of Pe	2024	Advanced F	10.1002/adfi	Cyclability	Doping	K0.5Co0.33Mn0.67O2	layered			4.2	1.6	80	uper P carbo	10	PTFE	10
4	Impact of Pe	2024	Advanced F	10.1002/adfi	Cyclability	Doping	K0.5Co0.33Mn0.67O2*0.5H2O	layered			4.2	1.6	80	uper P carbo	10	PTFE	10
5	Impact of Pe	2024	Advanced F	10.1002/adfi	Cyclability	Doping	K0.1Na0.7Co0.8Ti0.2O2	P63/mmc			4.2	1.6	80	uper P carbo	10	PTFE	10
6	High-Entropy	2024	Nano-Micro	10.1007/s408	power	Composition	Na0.95Li0.06Ni0.25Cu0.05Fe0.15Mn0.49O2	O3		spherical	4.2	2					
7	High-Entropy	2024	Nano-Micro	10.1007/s408	power	Composition	Na0.67Mn0.5Fe0.5O2				4.2	2					
8	High-Entropy	2024	Nano-Micro	10.1007/s408	power	Composition	Na0.67Ni0.33Mn0.67O2				4.2	2					
9	A novel hier	2023	RSC Advan	10.1039/d2ra	Cyclability	Morphology	Na0.7MnO2.05	hexagonal		book-like	4.5	1.5	75	Ketjen black	15	vinylidene flu	10
10	A novel hier	2023	RSC Advan	10.1039/d2ra	Cyclability	Morphology	Na0.7MnO2.05@NaPO3	P2	NaPO3	book-like	4.5	1.5	75	Ketjen black	15	vinylidene flu	10

Fig. 4 A partial view of the database constructed from LMO cathode materials literature, extracted from scientific papers using LLMs. Complete database is available in the ESI.†

extracting material compositions, then generating customized questions to extract the operating voltage and crystallinity of each material, followed by electrode composition information. All of the extracted data and detailed processing workflows are provided in the ESI.† The total processing time was less than 2 hours, averaging 20 seconds per paper. This represents a significant improvement in efficiency compared to the traditional manual extraction method, which takes approximately 20 minutes per paper.

Using the extracted data, we constructed a database that serves as a crucial asset for material composition design and development. For example, it can be utilized to train machine learning models for proposing new LMO compositions with specific electrochemical properties. Additionally, through structure–property relationship analysis, in-depth analysis of the correlation between the crystal type and electrochemical performance of LMO materials becomes possible, providing important insights into structural design to enhance properties such as high energy density, long lifespan, and fast charging capabilities.

Conclusions

In this study, LLMs were employed to classify papers, extract key parameters and the main approach of the studies, and successfully build a DB. This methodology is expected to significantly accelerate the research development of SIBs. Moreover, it is not limited to SIBs but has the potential to be expanded to studies on different materials development.

However, this extraction methodology, based solely on text, still has limitations in fully capturing all the key information because visual elements such as figures and charts,[34] which often contain important data, were excluded. Therefore, it is essential to process visual elements in order to improve the accuracy of this data extraction methodology. To address this existing issue, many researchers are continuously working to improve the performance of vision models. It is anticipated that these advancements will ultimately lead to the ability to extract information from a paper in any format.

Furthermore, the use of external LLMs significantly limited the utilization of the internal data due to security concerns in industry.[35] This may be a major bottleneck in research dealing with internal information of an organization. However, it is anticipated that these concerns will be somewhat resolved with the emergence of open-source LLMs such as Meta's Llama.[36] Such open-source models can be built and fine-tuned within each organization, making it possible to utilize internal data while maintaining security.[37] This is expected to be a critical part of the roadmap towards reducing data security risk and enabling broader applications

Author contributions

Youwan Na: investigation, formal analysis, visualization, conceptualization, and writing – original draft. Jeffrey J. Kim: investigation, data curation, validation, formal analysis, and writing – original draft. Chanhyoung Park: visualization and validation, Jaewon Hwang: data curation and investigation, Changgi Kim: validation and investigation. Hokyung Lee: supervision and project administration, Jehoon Lee: conceptualization, supervision, project administration, and writing – original draft. All authors have given approval to the manuscript.

Data availability

The data used for this study is available in the ESI.†

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

This research was supported by LG Chem. Ltd.

References

1 B. Diouf and R. Pode, *Renewable Energy*, 2015, **76**, 375–380.
2 M. Li, J. Lu, Z. Chen and K. Amine, *Adv. Mater.*, 2018, **30**, 1800561.
3 C. P. Grey and D. S. Hall, *Nat. Commun.*, 2020, **11**, 1–4.
4 T. Prior, P. A. Wäger, A. Stamp, R. Widmer and D. Giurco, *Sci. Total Environ*, 2013, **461**, 785–791.
5 B. J. Skinner, *Studies in Environmental Science*, Elsevier, Amsterdam, 1979, vol. 3, pp. 559–575.
6 J. Speirs, M. Contestabile, Y. Houari and R. Gross, *Renewable Sustainable Energy Rev.*, 2014, **35**, 183–193.
7 C. Vaalma, D. Buchholz, M. Weil and S. Passerini, *Nat. Rev. Mater.*, 2018, **3**, 1–11.

8 J. Y. Hwang, S. T. Myung and Y. K. Sun, *Chem. Soc. Rev.*, 2017, **46**, 3529–3614.
9 M. Miremadi, C. Musso and J. Oxgaard, *McKinsey Chem.*, 2013, **2**, 3–12.
10 L. Himanen, A. Geurts, A. S. Foster and P. Rinke, *Adv. Sci.*, 2019, **6**, 1900808.
11 B. L. DeCost, J. R. Hattrick-Simpers, Z. Trautt, A. G. Kusne, E. Campo and M. L. Green, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 033001.
12 J. Li, K. Lim, H. Yang, Z. Ren, S. Raghavan, P. Y. Chen, T. Buonassisi and X. Wang, *Matter*, 2020, **3**, 393–432.
13 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, *Digital Discovery*, 2023, **2**, 1233–1250.
14 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *arXiv*, 2024, preprint, arXiv:2407.16867, DOI: **10.48550/arXiv.2407.16867**.
15 P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang and R. Ramprasad, *npj Comput. Mater.*, 2023, **9**, 52.
16 B. Waltl, G. Bonczek and F. Matthes, Rule-based information extraction: Advantages, limitations, and perspectives, Technical University of Munich, 2018, Wa18b.
17 Y. Gou, Y. Zhang, J. Zhu and Y. Shu, *Sci. Data*, 2024, **11**, 372.
18 M. Munjal, T. Prein, V. Venugopal, K. J. Huang and E. Olivetti, AI for Accelerated Materials Design-NeurIPS 2023 Workshop, 2023.
19 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, *Adv. Neural Inf. Process Syst.*, 2020, **33**, 1877–1901.
20 M. P. Polak and D. Morgan, *Nat. Commun.*, 2024, **15**, 1569.
21 T. Gupta, M. Zaki, N. M. A. Krishnan and Mausam, *npj Comput. Mater.*, 2022, **8**, 102.
22 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *arXiv*, 2024, preprint, arXiv:2407.16867, DOI: **10.48550/arXiv.2407.16867**.
23 E. Eigner and T. Händler, *arXiv*, 2024, preprint, arXiv:2402.17385, DOI: **10.48550/arXiv.2402.17385**.
24 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel and D. Kiela, *Adv. Neural Inf. Process Syst.*, 2020, **33**, 9459–9474.
25 Scopus, **https://www.scopus.com/**, 2024, Accessed 27 Dec 2024.
26 P. Lopez, *Research and Advanced Technology for Digital Libraries*, Springer, Berlin, Heidelberg, 2009, 5714.
27 S. Es, J. James, L. Espinosa-Anke and S. Schockaert, *arXiv*, 2023, preprint, arXiv:2309.15217, DOI: **10.48550/arXiv.2309.15217**.
28 OpenAI, GPT-4o, **https://openai.com/index/hello-gpt-4o/**, 2024.
29 W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T. Chua and Q. Li, Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6491–6501.
30 L. Gao, X. Ma, J. Lin and J. Callan, *arXiv*, 2022, preprint, arXiv:2212.10496, DOI: **10.48550/arXiv.2212.10496**.
31 Z. Rackauckas, *arXiv*, 2024, preprint, arXiv:2402.03367, DOI: **10.48550/arXiv.2402.03367**.
32 G. Kim, S. Kim, B. Jeon, J. Park and J. Kang, *arXiv*, 2023, preprint, arXiv:2310.14696, DOI: **10.48550/arXiv.2310.14696**.
33 A. Asai, Z. Wu, Y. Wang, A. Sil and H. Hajishirzi, *arXiv*, 2023, preprint, arXiv:2310.11511, DOI: **10.48550/arXiv.2310.11511**.
34 Y. Hu, Z. Zhang and L. Zhao, *arXiv*, 2023, preprint, arXiv:2310.04944, DOI: **10.48550/arXiv.2310.04944**.
35 Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun and Y. Zhang, *High-Confid Comput.*, 2024, **4**, 100211.
36 H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave and G. Lample, *arXiv*, 2023, preprint, arXiv:2302.13971, DOI: **10.48550/arXiv.2302.13971**.
37 K. V. H. Warrier and Y. Gupta, *arXiv*, 2024, preprint, arXiv:2404.10779, DOI: **10.48550/arXiv.2404.10779**.