Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 328

Harnessing GPT-3.5 for text parsing in solid-state synthesis – case study of ternary chalcogenides

Maung Thway, †[a] Andre K. Y. Low, †[ab] Samyak Khetan,[c] Haiwen Dai,[a] Jose Recatala-Gomez,[a] Andy Paul Chen[a] and Kedar Hippalgaonkar *[ab]

Optimally doped single-phase compounds are necessary to advance state-of-the-art thermoelectric devices which convert heat into electricity and *vice versa*, requiring solid-state synthesis of bulk materials. For data-driven approaches to learn these recipes, it requires careful data curation from large bodies of text which may not be available for some materials, as well as a refined language processing algorithm which presents a high barrier of entry. We propose applying Large Language Models (LLMs) to parse solid-state synthesis recipes, encapsulating all essential synthesis information intuitively in terms of primary and secondary heating peaks. Using a domain-expert curated dataset for a specific material (Gold Standard), we engineered a prompt set for GPT-3.5 to replicate the same dataset (Silver Standard), doing so successfully with 73% overall accuracy. We then proceed to extract and infer synthesis conditions for other ternary chalcogenides with the same prompt set. From a database of 168 research papers, we successfully parsed 61 papers which we then used to develop a classifier to predict phase purity. Our methodology demonstrates the generalizability of Large Language Models (LLMs) for text parsing, specifically for materials with sparse literature and unbalanced reporting (since usually only positive results are shown). Our work provides a roadmap for future endeavors seeking to amalgamate LLMs with materials science research, heralding a potentially transformative paradigm in the synthesis and characterization of novel materials.

Received 8th October 2023
Accepted 21st December 2023

DOI: 10.1039/d3dd00202k

rsc.li/digitaldiscovery

Introduction

Solid-state synthesis is a pivotal method towards the discovery of inorganic materials for thermoelectric applications. The formation of high-quality crystalline materials depends heavily on the recipe, which in the most general terms comprises heating, cooling, and densification steps. Structural or chemical deviations from the intended structure and composition could happen under the synthesis conditions, which leads to deviations (some favorable, others not) in the materials' electronic and thermal transport properties.[1] Therefore, predicting recipes that produce phase-pure materials and engineering dopability is a key and unsolved challenge.

Much work has been performed in applying thermodynamic calculations and reaction networks[2] to rationally determine appropriate heating curves (time–temperature profiles) in solid state synthesis extracted from literature for a variety of chemical systems,[3–8] requiring a combination of domain expertise and trial-and-error which can be inaccessible when dealing with new materials.[9] While materials informatics and data-driven materials research through machine learning and AI has recently emerged as a new paradigm for materials research; in thermoelectric materials, the goal of materials-by-design would be augmented with a systematic development of synthesis recipes.[10] Complementarily, molecular retrosynthesis planning and reaction pathway prediction has had some success and therefore provides[11] a different perspective to solve this challenge.

However, data-driven approaches require (1) tedious manual extraction and (2) cleaning from the corpus of research publications, which is inefficient. To this endeavor, Natural Language Processing (NLP) algorithms stand out as efficient means of automating this process. Work by Ceder *et al.* demonstrated such a use-case in solid-state synthesis,[12] built upon techniques such as name entity recognition[13] and process classifications.[14] Yet, because NLPs are domain-specific, and demand not only extensive domain knowledge but also data curation for appropriate parsing,[15] deployment of NLPs to one's own research field is challenging. Therefore, there is an emerging need to develop a recipe extraction process with low transferable cost.

Language Models (LLMs) have recently emerged as an alternative tool to extract knowledge from scientific literature,

[a]*School of Materials Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore. E-mail: kedar@ntu.edu.sg*

[b]*Institute of Materials Research and Engineering (IMRE), Agency for Science, Technology and Research (A*STAR), Singapore 138634, Singapore*

[c]*Department of Metallurgical Engineering and Materials Science, Indian Institute of Technology Bombay, Maharashtra 400076, India*

† Equal contribution.

enabling contextualizing and summarizing of information efficiently and robustly. This has been demonstrated across different materials science fields; chemistry,[12,16–21] polymers,[22] general materials,[23–27] optical materials,[28] crystal structures,[29] and even other fields such as medicine.[30–32] We then ask the question – can LLMs be used to parse information specific to fields with sparse literature and strong reporting bias, to extract synthesis recipes, but also produce a machine learning readable dataset?

Such an approach bypasses the need for specialized NLP tools, offering a streamlined method for text parsing that is more accessible to the scientific community. To further illustrate the applicability of GPT parsing, we focus on ternary chalcogenide-based materials because they are the state-of-the-art thermoelectric materials at intermediate temperatures,[33] where the availability of synthesis literature is relatively smaller in size compared to the examples cited previously, meaning that the ability to tune the LLM is also limited. We consider a similar prompt engineering strategy reported by Zheng *et al.*[34] to refine this workflow, which we describe further below.

Relying on human domain expertise, we first craft a "Gold Standard" dataset based on a publication set of 21 papers for CuInTe/Se, a well-studied mid-temperature (400–600 K) range thermoelectric material. The Gold Standard was then used to optimize a GPT-3.5 prompt set to automatically generate a second dataset, which we call the "Silver Standard". The workflow was then used to automatically extract synthesis recipes in ABX_2 and Tl-based chalcogenide systems[35] directly from over 100 PDF documents. An illustration of this workflow is shown in Fig. 1.

In our results, we show that our method was able to capture relevant information effectively and infer details that were not provided with a reasonable measure of success, demonstrating potential for efficient data mining that translates to time savings. Additionally, our work showcases the generalizability of applying GPT-3.5 for parsing solid-state synthesis recipes in thermoelectrics.

Methods

We first undertook a comprehensive manual download of all papers published between 2000 and 2023 that discussed solid-state synthesis recipes of $CuInTe_2$, excluding methods such as solution synthesis and the Bridgman method, which is typically used for single crystal growth. This forms the Gold Standard for prompt refinement. We provide this dataset in our GitHub repository.

Using domain expertise, we determined that the following key aspects were most crucial to attaining pure compounds: primary heating, secondary heating, annealing, and densification.[36] In papers reported, an intermediate reaction between some reactants was purposely introduced to control the reaction kinetics, such as preparing binary precursors for ternary synthesis. Primary heating is therefore defined to capture such delicate heating information which is the first temperature held for an extended period, especially as we expect that thermodynamic driving forces are likely to be caused *via* pairwise reactions anyway.[37] Secondary heating refers to the final temperature where reactants tend to melt and diffuse unless annealing is explicitly mentioned.

To facilitate the extraction of these details, we designed specific prompts that would output the captured information in a tabular (comma separated value, .csv) format. While other studies often employ JSON[12] for data extraction due to its capability to handle nested lists and multi-dimensional data, we opted for a tabular approach for ease of interpretation and prompting. This is especially pertinent for small, focused datasets such as ours, since we expect that additional

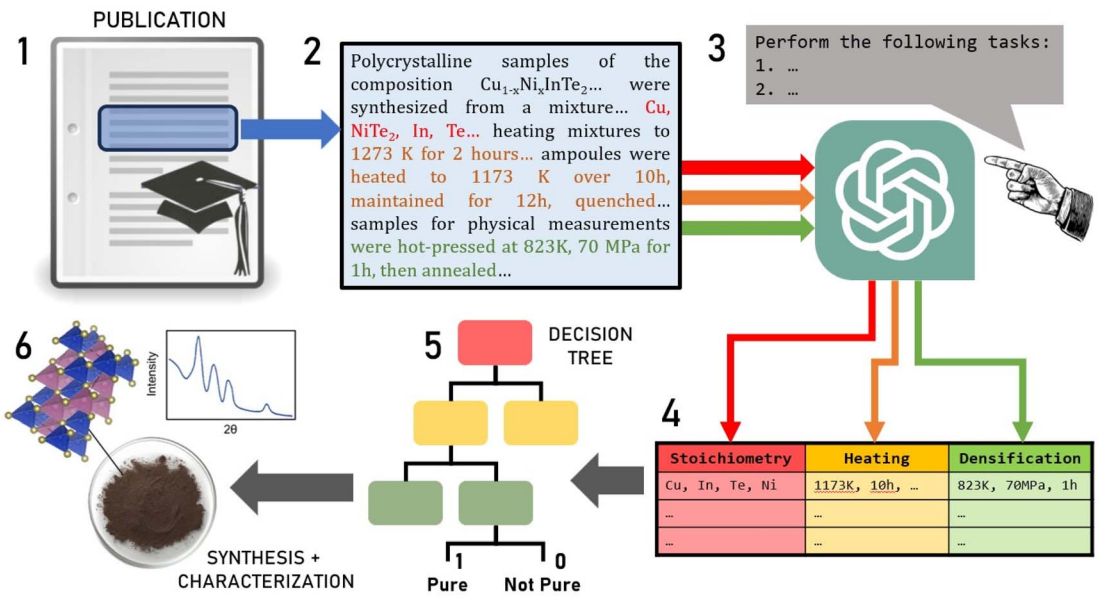

Fig. 1 An illustration of the automated text extraction and parsing procedure using GPT-3.5. The domain-specific information is broken down into specific sub-tasks for GPT-3.5 to extract and collate into a table, whereby further analysis and machine learning can be performed.

complexity does not yield better results. We noted for the paper referenced which used JSON, it was challenging to quickly make sense of the dataset for visualization and analysis.

Following this, we refined a set of prompts for GPT-3.5 to extract the same information, taking note to logically infer information when not provided, giving examples from the Gold Standard. The prompt set was optimized iteratively based on the following principles:

(1) All questions put together in a single prompt, without any standard formatting and based on human intuition.

(2) All questions put together in a single prompt, with standard formatting.

(3) Questions broken up into a sequence of prompts, without standard formatting.

(4) Questions broken up into a sequence of prompts, with standard formatting.

We noticed that the LLM has a hard time trying to reason/extract information from a paragraph that require human intuition. The answer gets more consistent when we provide appropriate examples in the prompt. However, when we extract too much information in one go, we found that sometimes the LLM misses certain information and other times it 'misbehaves' with unexpected behaviour, without adhering to the formatting instructions in our prompt set. Overall, we found that sequentially extracting information one by one with standard answers gives the best results. The iterative process is reported in prompt_engineering_progress.ipynb in the GitHub repository.

It was observed that the use of simple questions and a restriction to no more than two questions per prompt contributed to improved accuracy in information extraction. The initial question is aimed at the identification of synthesis information within the paper. If such information is absent in the paper, it is skipped, and the next paper is then processed. Once the synthesis paragraph is detected by the program, the subsequent question is employed to extract details regarding the base compound and dopant. Following that, the next question pertains to the temperature profile mentioned within the synthesis paragraph. However, from time to time, we observed that the output doesn't follow the format that is given in the prompt, especially for prompts which require multiple outputs in the same response. Therefore, it was necessary to include formatting checks in the sequence.

Presented below is a series of eight questions within six prompts:

(1) Does it include description of synthesis information?

(2) Does the experiment result in pure phase formation of crystal?

(3) What is the base compound used in the experiment? Exclude dopant and do not include "x" when you mention the base compound.

(4) What is the dopant used in the experiment to dope the base compound? Generally, it is written before "x". The dopant is not included in the base compound. Write chemical symbol (*e.g.* C for carbon).

(5) What is the temperature profile of the experiment? Answer in a tabular format.

(6) Is the given data following this format? If not re-format.

(7) Choose one of the cooling types whether it is left in the room, in water or immersed in something cold, or left in the furnace: "room" or "quenching" or "furnace".

(8) Choose what is the densification technique used to densify the powder: "hot press" or "sintering" or "NA".

Our engineered prompts aim to provide a cost-efficient method for parsing solid state synthesis recipes, which we did so with the total budget of all prompt refining experiments and actual text parsing using GPT-3.5 being within 50 SGD (~36 USD). Therefore, the effective cost per PDF is around 0.29 SGD (~0.20 USD) per PDF. While GPT-4 allows for higher accuracy in certain scenarios and enables more functionalities, we focused on GPT-3.5 instead as: (1) GPT-3.5 is accessible compared to GPT-4 which requires a subscription (2) a lower API cost compared to GPT-4 and (3) the parsing accuracy of 3.5 and 4 from text were found to be similar for such literature. A preliminary comparison between both models is reported in the Github repository as conversation histories.

The re-generated set of the same dataset of papers is hereby named the Silver Standard. This prompt set is paired with the PyPDF library to then convert PDFs to machine readable form, where we broke down each PDF into text chunks to fit into the token limit. For splitting the text string into chunks, we use PyPDFLoader.load_and_split function. The function in turns uses RecursiveCharacterTextSplitter, which has a default maximum chunk size of 4000 with an overlap of 200 between each chunk.

Additionally, we consider a secondary and larger dataset of solid-state synthesis, extended to ABX_2 and Tl-based chalcogenide systems. Similarly, we performed comprehensive manual download of English-based research papers published between 2000 and 2023 that discussed solid-state synthesis recipes of ABX_2 compounds including AgInTe/Se_2, CuGaTe/Se_2, AgInTe/Se_2, $TlSbTe_2$, $TlGdTe_2$, $TlBiTe_2$, and $KGdTe_2$, excluding methods such as solution-based synthesis (too many precursors and generally speaking, lower phase purity) or the Bridgman method, which is for single crystal growth. The same set of 21 CuInTe/Se synthesis papers were used constructing the Gold Standard, and subsequently parsing the Silver Standard. Additionally, a total of 168 papers from other ternary chalcogenide compounds were compiled, but only 61 were successfully parsed by GPT-3.5; the rest failed the first prompt (did not contain synthesis information, or PyPDF failed to format it). Table 1 below provides a list of the datasets applied in this work.

We propose evaluating the accuracy of text parsing by reporting the fraction of correct labels, and the overall error (inaccurate parsing) rate. In total, there are four possible situations:

Results and discussion

We first consider the comparison between Gold and Silver Standard, which are based on the same set of CuInTe/Se papers. The GPT-based Silver Standard achieves a 73% overall accuracy as shown in Fig. 2. In general, the highest accuracies were seen for all heating temperatures and time, base compound, and densification techniques, which are among the most important

Table 1 Names and description of each dataset

Name	Description
Gold Standard	Manually extracted dataset of 21 CuInTe/Se papers from human expertise
Silver Standard	GPT prompted dataset from the same 21 CuInTe/Se papers, with prompts optimized by comparing to Gold Standard
Expanded Chemical Space (ExChSp)	GPT prompted dataset of other ternary chalcogenide compounds using same prompt set as Silver Standard, with 61 successful papers out of 168 total parsed

information towards high purity products. We observe that the errors in base compound and dopant are often due to cases where papers discuss multiple types of compounds, or when the reactants reported are based on ternary compounds rather than base elements, which leads to confusion in parsing by GPT-3.5.

When applied on the expanded chemical space (ExChSp), >60% accuracy was achieved to correctly parse the dopants from complex chemical formulae. Our developed approach demonstrates that even without complicated NLP tuning, information embedded in chemical formulae can successfully be extracted *via* optimized GPT-based prompting.

Additionally, being able to extract sequential heating stages is important for further material engineering, such as crystallinity or crystal structure, as it corresponds to the time–temperature profile. In the Gold Standard where information was manually parsed, we inferred the ramping rate and cooling type based on the technique used, and phase purity *via* the diffraction graph, which is obviously not contained directly in the text. Most notably, details on secondary melt, cooling type and dopant are often not explicitly reported in the text but could be easily inferred by a human expert. Consequently, these categories reported significantly poorer accuracies.

To conduct a thorough analysis of the extracted data, we next employ a simple machine learning model, a decision tree classifier, to identify how temperature profiles may impact the production of the pure-phase compound. This model is trained on both the Gold Standard and ExChSp datasets, and the results are reported in Fig. 3a and b respectively. Referring to the 4 possible conditions listed in Table 2, we manually compute the accuracy metric for each entry (every detail for each paper).

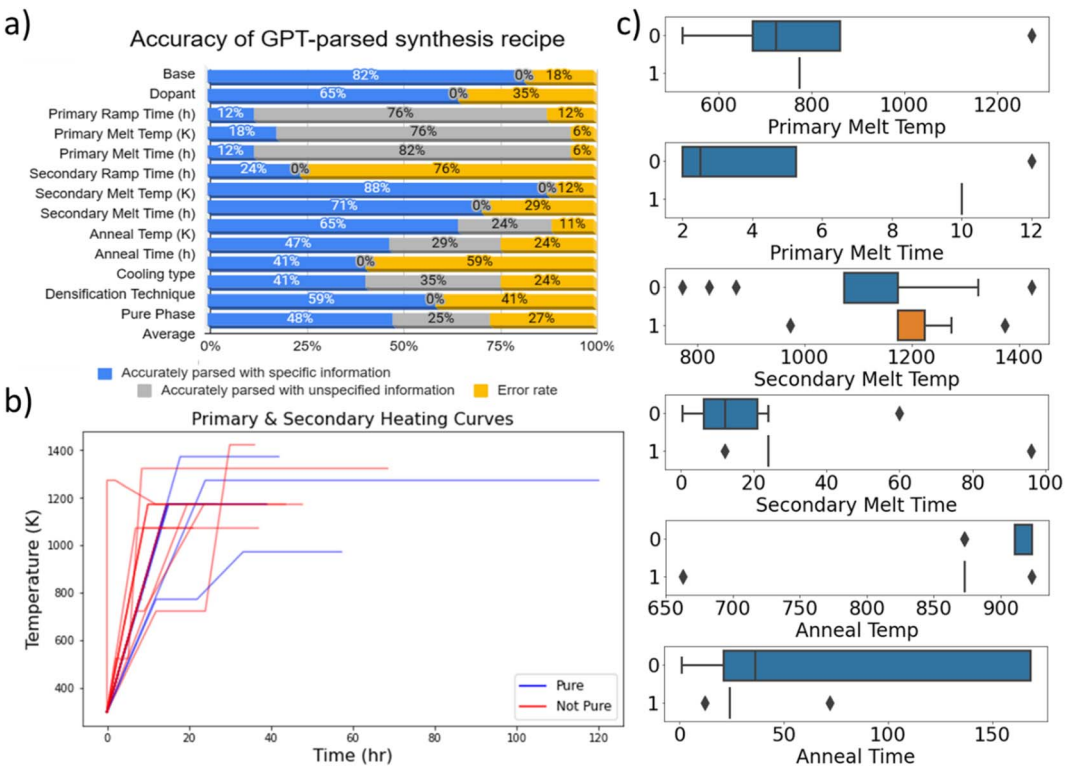

Fig. 2 Details on the Gold and Silver Standard. (a) Accuracy of GPT-3.5 extracted Silver Standard comparing against manually obtained Gold Standard, considering accuracy of both specified (dark blue) and unspecified details (light blue), as well as overall percentage of wrong details (orange). (b) Heating curves reported for the Gold Standard dataset. (c) Box charts for heating information in Gold Standard with respect to phase purity (1 refers to pure, 0 to not).

Even though trained with a lightweight dataset and model, we are able to achieve a 83% training and 61.5% of test accuracy for the ExChSp dataset to predict phase purity. According to the feature importance analysis in Fig. 3b, secondary temperature is the most important factor followed by the annealing and primary heating stage. Moreover, in the expanded dataset ExChSp, we observe that the varied formatting of scientific papers makes it challenging to extract useable data consistently, as only about 30% of the examined papers yielded relevant information. As a potential solution, in future work, one could shift the focus to creating a literature search tool that efficiently summarizes synthesis procedures for target compounds, a more achievable goal that still holds value.

It is clear that there is a strong reporting bias in the Gold Standard where most papers report a similar temperature profile as they follow an already established synthesis recipe, which leads to lack of distribution in other factors asides from secondary ramping time. This leads to overestimated importance of the secondary ramping time, which we understand to have limited impact on phase purity according to our domain expertise. Therefore, Fig. 3c and d reports our findings on applying leave-one-out (LOO) validation instead. In doing so, the model was able to capture better variation in the features. The error lines in Fig. 3(c) and (d) are generated by running the LOO strategy 200 times. It can be observed that the ramp times, which has the most variation, dominates the other features in this analysis.

As an alternative means of analysis, an XGBoost classifier was implemented to derive SHAP values of the features. Further analysis on both datasets yielded information that can infer

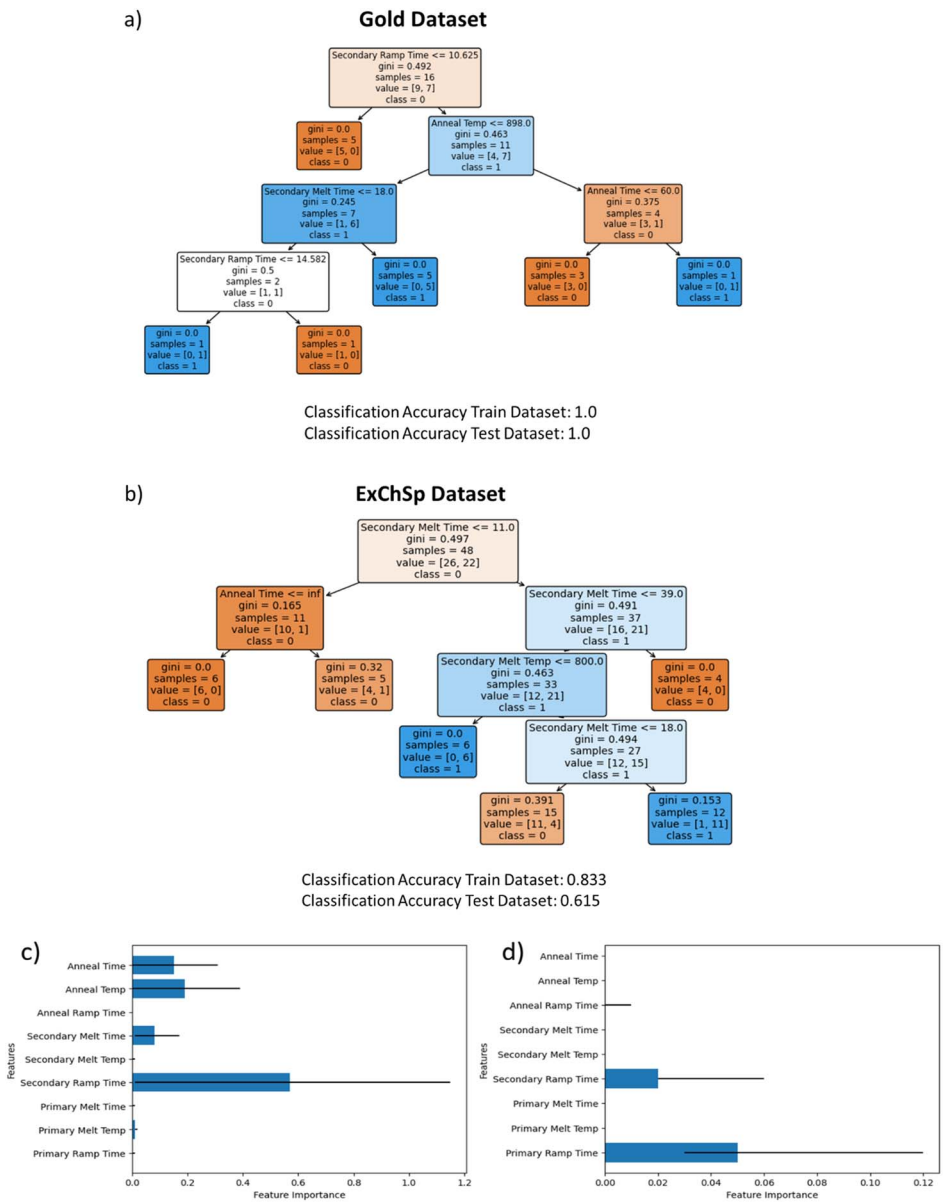

Fig. 3 Decision tree classifier results. The decision tree architecture and accuracy reported for (a) Gold Standard dataset (b) ExChSp dataset. Feature importance of both decision trees for leave-one-out strategy are reported in (c) and (d) respectively.

Table 2 Accuracy metric for specified and unspecified information

Original text	GPT-3.5 parsing	Accurate (1) or not (0)
Information specified	Same as specified	1
Information specified	Different from specified	0
Information unspecified	NA	1 (tagged as unspecified in Fig. 2a)
Information unspecified	Specified otherwise, not NA	0

how heating curves contribute to phase purity. Based on SHAP analysis reported in Fig. 4(a and b), a general trend of feature importance could be found: secondary heating > annealing > primary heating. This can be explained simply by how we defined these heating stages.

For primary heating, although a kinetic intermediate step was purposely included, here the reaction typically follows thermodynamic pathways especially as primary heating occurs at high temperatures where driving forces are strong.[38] The significance of secondary heating stage is closely related with the dominance of melting temperature over thermodynamic driving forces influencing diffusion and crystallization.[39]

Annealing stage, typically being included in the synthesis steps allows for solid state recrystallization and growth and has been known to have less influence on phase purity unless impurities or structure inhomogeneity remains after crystallization during secondary heating. Paradoxically, *via* closer observation of the SHAP values from secondary and annealing temperatures in Fig. 4b, higher temperature is seen to inversely contribute to phase impurity. Such a phenomenon could be related with limited accuracy (61.5%) in phase purity information extraction, which could invert the axis of the SHAP plot. As a caveat, the SHAP analysis is also model-dependent, and could change if the final optimized machine learning model is improved. Finally, there could be several 'unknown unknowns' which are not directly extracted from the synthesis text such as oxidation, sample preparation and handling, *etc.*

Further discussion

Our GPT-based framework's implications reach beyond just solid-state synthesis recipes or thermoelectric materials. It showcases the adaptability of LLMs to handle niche domains with limited literature and not requiring highly tuned models with extensive data curation. Traditional NLP models are often closely linked and tuned based on their training data, risking a drop in performance when tasked with new domains. In contrast, we were able to successfully perform text extraction with very minimal initial training as shown on the results for Gold and Silver Standard. The feature importance reported in Fig. 3 and 4 suggest that secondary temperature is the most crucial step for solid-state synthesis, which would help scientists in developing temperature profiles.

Leveraging upon the ChExSp dataset, we tested the possibility to interpolate and extrapolate synthesis conditions for $AgInTe_2$ (part of the dataset), as well as extrapolate for $AgSbTe_2$, both of which are chemically similar to the material studied in the Gold Standard, *i.e.*, $CuInTe_2$ provided as a contextual prompt. We anticipate that GPT-3.5 has no knowledge of their synthesis conditions, as they would only be found in subscription-based scientific journals, and not an open-source dataset.

According to Table 3, the synthesis temperature and time for $AgInTe_2$ are reasonable, with a ~1200 K melting stage and a ~700 K annealing temperature, no primary melting. One interesting fact is that the interpolation is able to suggest quenching as cooling state which is related to phase precipitation tendency during the synthesis.[40] For extrapolating to predict a synthesis recipe for $AgSbTe_2$, the GPT-3.5 model was not able to yield a proper synthesis recipe, by merely guessing that the recipe for $AgSbTe_2$ is similar to that of $AgInTe_2$ or $AgInSe_2$. This result is only because the ChExSp dataset was provided to the GPT-3.5 API as an input – else, GPT-3.5 responds with response with three sequential melt times going from low to high temperatures, which we know is inaccurate based on domain expertise. The suggested sequentially increasing temperature stages are different from domain expert recipes where a secondary high-temperature melting stage happens before a mid-temperature annealing stage to allow for melt crystallization followed by phase homogeneity.

We posit that GPT-3.5 which is trained on a corpus of mainly non-scientific text, contains incomplete information

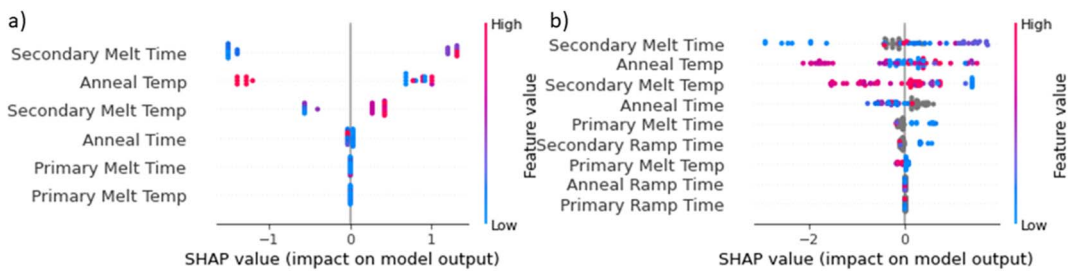

Fig. 4 SHAP analysis on (a) CuInTe/Se gold dataset, and (b) secondary ExChSp dataset of chalcogenides.

Table 3 Prompting GPT-3.5 to predict a synthesis recipe for phase purity with the ChExSp dataset as an input

	$AgInTe_2$ (interpolation)	$AgSbTe_2$ (extrapolation)
Primary melting temp	298 K	Same as $AgInTe_2$ considering their chemical similarity
Secondary melt temp	1273 K	
Secondary melt time	24 hours	
Anneal temp	773 K	
Anneal time	72 hours	
Cooling type	Quenching	
Densification technique	Hot press	

on solid-state synthesis recipes. It is possible that the Gold Standard used to prompt GPT-3.5 is contradictory to its knowledge, which we suggest is likely since it is paired with the fact that we implemented the responses with a temperature of zero (*i.e.*, no creativity, since we deemed this as a non-creative writing task).

We also expect that our workflow can be improved, possibly by developing a custom LLM model with greater capabilities and without budget constraints, such as building it on AWS EC2 to allow for unlimited API calls with predictable monthly fees. In practice, we find that the time and effort invested into refining the prompts for GPT-3.5 as well as the time taken to generate the ExChSp dataset is longer than what would have taken for one or two domain experts to do so manually – this is probably due to the small dataset (61 entries) relative to other areas of materials science, for this domain. The automated approach might work better if an order (or more) of research papers is available for a user's specific domain. This paves the way for more advanced scientific information extraction depending on the project's objectives.

One improvement from recent developments includes the ability to develop multi-modal models to extract data from plots and figures. For construction of the Silver Standard dataset using GPT-3.5 to extract information instead, heuristic estimations based on text-available information is more challenging since GPT has not been trained as a chemist. Analyzing diffraction information automatically requires further expertise in automated plot digitization[41] and refinement,[42] which is outside the scope of this work. Following the development of GPT-4 and further, one could anticipate its development to deal with more complicated data including images, illustration, plots, and learning domain expertise in the near future. We expect that this will only augment our general approach to provide higher quality data to materials synthesis experimentalists.

We also acknowledge the paucity of literature in this specific field. Even though we searched and extracted knowledge from 3 decades of literature (with a total of 162 research papers, although only 61 were successfully extracted), the dataset is severely biased towards positive results. Hence, we would emphasize upon the community that there is a pressing need for balanced datasets where negative experimental results are also reported. We hope that combining our framework with a domain-specific Gold Standard is the first step towards a transferable approach, applicable across different realms of materials science, that enables users to conduct text mining and corpus curation without developing specific NLP algorithms.

Apart from the need of high-quality and balanced datasets, the outlook of this work includes further refinement in prompt engineering and chain of thought inference to better tune responses for a given model. Further on, fine-tuning base models or even training new models on a sufficiently large corpus of scientific text is a more ambitious task. Finally, we also propose sequencing another LLM to cross check on the data extraction work for better reliability of results, and to help compute the accuracy metric that we use above.

Conclusion

In conclusion, our introduction of a GPT-based framework bridged the gap between literature and the lack of domain synthesis database. Parsing 173 research articles and processing 61 out of those to create an Extended Chemical Space as a thermoelectric synthesis database from existing scientific literature signifies a shift towards more accessible, adaptable, and scalable data extraction tools. The framework is able to achieve an average of 73% extraction accuracy, enabling accelerated statistical visualization from text-based literature. We further demonstrate the possibility of feeding automated extracted data as feed for machine learning, *via* decision tree and XGBoost, which were able to learn and extrapolate the contributing factors towards phase purity, a common target for solid state chemists. As the landscape of research becomes increasingly vast and fragmented, tools like these will be instrumental in ensuring that the collective knowledge of the scientific community remains accessible and actionable.

Data availability

Details and generated data of our implementation can be found in our group repository: **https://github.com/Kedar-Materials-by-Design-Lab/Harnessing-GPT-3.5-for-Text-Parsing-in-Solid-State-Synthesis-case-study-of-ternary-chalchogenides**.

Author contributions

Conceived the research: KH. Developed the GPT3.5 prompts: MT, SK, AL, HD, JRG. Extracted data: MT, SK, AL, HD, JRG, APC. Wrote the manuscript: MT, AL, DH, KH with input from all co-authors.

Conflicts of interest

KH owns equity in a startup focused on applying Machine Learning for Materials.

Acknowledgements

The authors acknowledge funding from AME Programmatic Funds by the Agency for Science, Technology and Research under Grant (No. A1898b0043). KH also acknowledges funding from the NRF Fellowship (NRF-NRFF13-2021-0011). The manuscript writing was prepared with minimal assistance (to generate a 'first draft' with improved grammar and synthesizing some information for better readability) from ChatGPT.

References

1 H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain and G. Ceder, Machine-learning rationalization and prediction of solid-state synthesis conditions, *Chem. Mater.*, 2022, **34**, 7323–7336.

2 M. J. McDermott, S. S. Dwaraknath and K. A. Persson, A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis, *Nat. Commun.*, 2021, **12**, 3097.

3 A. Miura, C. J. Bartel, Y. Goto, Y. Mizuguchi, C. Moriyoshi, Y. Kuroiwa, Y. Wang, T. Yaguchi, M. Shirai and M. Nagao, Observing and Modeling the Sequential Pairwise Reactions that Drive Solid-State Ceramic Synthesis, *Adv. Mater.*, 2021, **33**, 2100312.

4 M. Bianchini, J. Wang, R. J. Clément, B. Ouyang, P. Xiao, D. Kitchaev, T. Shi, Y. Zhang, Y. Wang and H. Kim, The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides, *Nat. Mater.*, 2020, **19**, 1088–1095.

5 A. Miura, H. Ito, C. J. Bartel, W. Sun, N. C. Rosero-Navarro, K. Tadanaga, H. Nakata, K. Maeda and G. Ceder, Selective metathesis synthesis of MgCr 2 S 4 by control of thermodynamic driving forces, *Mater. Horiz.*, 2020, **7**, 1310–1316.

6 P. K. Todd, M. J. McDermott, C. L. Rom, A. A. Corrao, J. J. Denney, S. S. Dwaraknath, P. G. Khalifah, K. A. Persson and J. R. Neilson, Selectivity in yttrium manganese oxide synthesis via local chemical potentials in hyperdimensional phase space, *J. Am. Chem. Soc.*, 2021, **143**, 15185–15194.

7 A. Wustrow, G. Huang, M. J. McDermott, D. O'Nolan, C.-H. Liu, G. T. Tran, B. C. McBride, S. S. Dwaraknath, K. W. Chapman and S. J. L. Billinge, Lowering ternary oxide synthesis temperatures by solid-State cometathesis reactions, *Chem. Mater.*, 2021, **33**, 3692–3701.

8 H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain and G. Ceder, Machine-learning rationalization and prediction of solid-state synthesis conditions, *Chem. Mater.*, 2022, **34**, 7323–7336.

9 C. N. R. Rao, H. S. S. R. Matte, R. Voggu and A. Govindaraj, Recent progress in the synthesis of inorganic nanoparticles, *Dalton Trans.*, 2012, **41**, 5089–5120.

10 K. Hippalgaonkar, Q. Li, X. Wang, J. W. Fisher III, J. Kirkpatrick and T. Buonassisi, Knowledge-integrated machine learning for materials: lessons from gameplaying and robotics, *Nat. Rev. Mater.*, 2023, **8**, 241–260.

11 Y. Shen, J. E. Borowski, M. A. Hardy, R. Sarpong, A. G. Doyle and T. Cernak, Automation and computer-assisted planning for chemical synthesis, *Nat. Rev. Methods Primers*, 2021, **1**, 23.

12 O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan and G. Ceder, Text-mined dataset of inorganic materials synthesis recipes, *Sci. Data*, 2019, **6**, 203.

13 T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari and G. Ceder, Similarity of precursors in solid-state synthesis as text-mined from scientific literature, *Chem. Mater.*, 2020, **32**, 7861–7873.

14 H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan and G. Ceder, Semi-supervised machine-learning classification of materials synthesis procedures, *npj Comput. Mater.*, 2019, **5**, 62.

15 O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti and G. Ceder, Opportunities and challenges of text mining in materials research, *iScience*, 2021, **24**, 102155.

16 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary and D. Circi, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, *Digital Discovery*, 2023, **2**, 1233–1250.

17 A. M. Bran, S. Cox, A. D. White and P. Schwaller, ChemCrow: Augmenting large-language models with chemistry tools, *arXiv*, 2023, preprint, arXiv:2304.05376, DOI: **10.48550/arXiv.2304.05376**.

18 G. M. Hocky and A. D. White, Natural language processing models that automate programming will transform chemistry research and teaching, *Digital Discovery*, 2022, **1**, 79–83.

19 A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner and H. J. Kulik, MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks, *Sci. Data*, 2022, **9**, 74.

20 A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson and A. Jain, Structured information extraction from complex scientific text with fine-tuned large language models, *arXiv*, 2022, preprint, arXiv:2212.05238, DOI: **10.48550/arXiv.2212.05238**.

21 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062, DOI: **10.1021/jacs.3c05819**.

22 C. Xu, Y. Wang and A. Barati Farimani, TransPolymer: a Transformer-based language model for polymer property predictions, *npj Comput. Mater.*, 2023, **9**, 64.

23 M. Yoshitake, F. Sato, H. Kawano and H. Teraoka, Materialbert for natural language processing of materials science texts, *Sci. Technol. Adv. Mater.: Methods*, 2022, **2**, 372–380.

24 T. Gupta, M. Zaki, N. M. A. Krishnan and Mausam, MatSciBERT: A materials domain language model for text mining and information extraction, *npj Comput. Mater.*, 2022, **8**, 102.

25 Z. Hong, A. Ajith, J. Pauloski, E. Duede, K. Chard and I. Foster, The Diminishing Returns of Masked Language Models to Science, in *Findings of the Association for Computational Linguistics: ACL 2023*, ed. A. Rogers, J. Boyd-Graber and N. Okazaki, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1270–1283.

26 M. P. Polak and D. Morgan, Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering-Example of ChatGPT, *arXiv*, 2023, preprint, arXiv:2303.05352, DOI: **10.48550/arXiv.2303.0535**.

27 I. Beltagy, K. Lo and A. Cohan, SciBERT: A Pretrained Language Model for Scientific Text, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng and X. Wan, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620, DOI: **10.18653/v1/D19-1371**.

28 J. Zhao, S. Huang and J. M. Cole, OpticalBERT and OpticalTable-SQA: Text-and Table-Based Language Models for the Optical-Materials Domain, *J. Chem. Inf. Model.*, 2023, **63**, 1961–1981.

29 L. M. Antunes, K. T. Butler and R. Grau-Crespo, Crystal structure generation with autoregressive large language modeling, *arXiv*, 2023, preprint, arXiv:2307.04340, DOI: **10.48550/arXiv.2307.04340**.

30 I. S. Fins, H. Davies, S. Farrell, J. R. Torres, G. Pinchbeck, A. D. Radford and P.-J. Noble, Evaluating ChatGPT text-mining of clinical records for obesity monitoring, *Vet. Rec.*, 2024, DOI: **10.1002/vetr.3669**.

31 Q. Chen, H. Sun, H. Liu, Y. Jiang, T. Ran, X. Jin, X. Xiao, Z. Lin, Z. Niu and H. Chen, A comprehensive benchmark study on biomedical text generation and mining with ChatGPT, *bioRxiv*, 2023, preprint, DOI: **10.1101/2023.04.19.537463**.

32 R. Nadkarni, D. Wadden, I. Beltagy, N. A. Smith, H. Hajishirzi and T. Hope, Scientific Language Models for Biomedical Knowledge Base Completion: An Empirical Study, in *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual*, ed. D. Chen, J. Berant, A. McCallum and S. Singh, 2021, DOI: **10.24432/C5QC7V**.

33 S. Roychowdhury, T. Ghosh, R. Arora, M. Samanta, L. Xie, N. K. Singh, A. Soni, J. He, U. V Waghmare and K. Biswas, Enhanced atomic ordering leads to high thermoelectric performance in AgSbTe2, *Science*, 2021, **371**, 722–727.

34 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.

35 J. P. Heremans, V. Jovovic, E. S. Toberer, A. Saramat, K. Kurosaki, A. Charoenphakdee, S. Yamanaka and G. J. Snyder, Enhancement of thermoelectric efficiency in PbTe by distortion of the electronic density of states, *Science*, 2008, **321**, 554–557.

36 S. Y. Tee, D. Ponsford, C. L. Lay, X. Wang, X. Wang, D. C. J. Neo, T. Wu, W. Thitsartarn, J. C. C. Yeo and G. Guan, Thermoelectric Silver-Based Chalcogenides, *Adv. Sci.*, 2022, **9**, 2204624.

37 A. Miura, C. J. Bartel, Y. Goto, Y. Mizuguchi, C. Moriyoshi, Y. Kuroiwa, Y. Wang, T. Yaguchi, M. Shirai and M. Nagao, Observing and Modeling the Sequential Pairwise Reactions that Drive Solid-State Ceramic Synthesis, *Adv. Mater.*, 2021, **33**, 2100312.

38 P. K. Todd and J. R. Neilson, Selective formation of yttrium manganese oxides through kinetically competent assisted metathesis reactions, *J. Am. Chem. Soc.*, 2019, **141**, 1191–1195.

39 P. Shewmon, *Diffusion in solids*, Springer, 2016.

40 V. Meschke, L. C. Gomes, J. M. Adamczyk, K. M. Ciesielski, C. M. Crawford, H. Vinton, E. Ertekin and E. S. Toberer, Designing for dopability in semiconducting AgInTe 2, *J. Mater. Chem. C*, 2023, **11**, 3832–3840.

41 F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian and G. Romano, Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks, *npj Comput. Mater.*, 2019, **5**, 60.

42 P. Baptista de Castro, K. Terashima, M. G. Esparza Echevarria, H. Takeya and Y. Takano, XERUS: An Open-Source Tool for Quick XRD Phase Identification and Refinement Automation, *Adv. Theory Simul.*, 2022, **5**, 2100588.