Machine Learning Inorganic Solid-state Synthesis from Materials Science Literature

By

Tanjin He

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Materials Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gerbrand Ceder, Chair
Professor Sandrine Dudoit
Professor Kristin Persson

Spring 2023

Machine Learning Inorganic Solid-state Synthesis from Materials Science Literature

Tanjin He

Abstract

Machine Learning Inorganic Solid-state Synthesis from Materials Science Literature

by

Tanjin He

Doctor of Philosophy in Engineering – Materials Science and Engineering

University of California, Berkeley

Professor Gerbrand Ceder, Chair

Solid-state synthesis prediction is a key accelerator for the rapid design of advanced inorganic materials. However, determining synthesis variables such as the choice of precursor materials is challenging for inorganic materials because the sequence of reactions during heating is not well understood. To achieve predictive synthesis for the desired material, one potential approach is to learn synthesis design patterns from a large volume of experimental synthesis procedures. Nevertheless, a comprehensive, large-scale database of structured synthesis procedures for inorganic materials does not exist. Provided the ability of converting unstructured text to structured information, the decades of solid-state chemistry literature constitutes a treasure trove of synthesis data. Therefore, this study aims at: (1) developing natural language processing (NLP) algorithms to text mine a large-scale inorganic synthesis dataset from materials science literature, and (2) developing machine learning algorithms for precursor selection in solid-state synthesis based on the text-mined dataset.

Although many general-purpose NLP methods exist, text mining for inorganic synthesis requires dedicated development of models for information retrieval (Chapter 2). During the development of a text-mining pipeline, one major problem is the difficulty of identifying which materials from a synthesis paragraph are precursors or are target materials. In this study, we developed a two-step Chemical Named Entity Recognition (CNER) model to identify precursors and targets, based on information from the context around material entities. By integrating our information retrieval model for precursors and targets, and also the ones for other synthesis variables, we established a fully automated text-mining pipeline that extracts the structured data of synthesis procedures from the literature. Starting from 4,973,165 materials science papers, we applied our text-mining pipeline and successfully extracted 33,343 solid-state synthesis procedures. The quality of the text-mined synthesis dataset is validated by the high accuracy of 93% at the chemistry level, where each extracted reaction has the target and precursor materials consistent with the original literature report.

This dataset for inorganic solid-state synthesis is currently the largest of its kind and paves the way toward the development of data-driven approaches for rational synthesis design.

Using the extracted data, we conducted a meta-analysis to study the similarities and differences between precursors in the context of solid-state synthesis (Chapter 3). To quantify precursor similarity, we built a substitution model to calculate the viability of substituting one precursor with another while retaining the target. From a hierarchical clustering of the precursors, we demonstrate that the "chemical similarity" of precursors can be extracted from text data, without the need to include any explicit domain knowledge. Quantifying the similarity of precursors offers a reference for suggesting candidate reactants when researchers alter existing recipes by replacing precursors. The capability of creating alternative recipes constitutes an important step toward developing a predictive synthesis model.

While the selection of alternative precursors is enabled by the similarity of precursors, it is limited to existing materials. To learn which precursors to recommend for the synthesis of a novel target material, we further developed a representation learning model to evaluate the similarity of targets (Chapter 4). The data-driven approach learns "chemical similarity" of target materials and refers the synthesis of a new target to precedent synthesis procedures of similar target materials, mimicking human synthesis design. When proposing five precursor sets for each of 2,654 unseen test target materials, our recommendation strategy achieves a success rate of at least 82%. Our approach captures decades of heuristic synthesis data in a mathematical form, making it accessible for use in recommendation engines and autonomous laboratories.

Overall, this study contributes a valuable large-scale synthesis dataset and interpretable precursor selection algorithms to the materials science community, representing a step forward in the prediction of solid-state synthesis.

Contents

List of Figures

List of Tables

Acknowledgments

Finally, I completed my dissertation, albeit in the very last month before my graduation. I am not sure when the experience of scrambling to meet deadlines became a familiar aspect of my daily Ph.D. journey. Although the journey was not smooth, I was fortunate enough to thrive. As someone with a mechanical engineering background at the beginning of my Ph.D. journey, I am now comfortable stating my proficiency in both materials science and computer science. This is largely because of the exceptional support I have received from a multitude of benevolent individuals. Although it is difficult to list everyone, I would like to express my gratitude to every individual who has assisted me over the years.

I thank my advisor, Prof. Gerbrand Ceder, for his invaluable support and guidance. I vividly recall the words from his offer email, stating that "we can do great things together." Indeed, he taught me to think big and instructed me to do big things. I also remember an inspirational moment in one of my presentations when he mentioned the word "similarity". That conversation led to the formulation of the core idea in this dissertation. Undoubtedly, he is a role model from whom I am constantly learning.

I thank my qualifying exam and thesis committee members, specifically Prof. Gerbrand Ceder, Prof. Kristin Persson, Prof. Sandrine Dudoit, Prof. Daryl Chrzan, and Dr. Anubhav Jain, for all their valuable feedback and assistance throughout my Ph.D. program.

I also thank our collaborators at MIT, particularly Prof. Elsa Olivetti, Edward Kim, and Zachary Jensen, for their engaging discussion and inspirations on this project.

I was also fortunate to have many good mentors, colleagues, and friends. I am grateful to Ziqin Rong, Wenhao Sun, and Bin Ouyang. They helped me to get started in different areas of my research. I express my gratitude to Haoyan Huo, Olga Kononova, Chris Bartel, Vahe Tshioyan, Amalie Trewartha, Tiago Botari, Zheren Wang, and Kevin Cruse. They are the gold teammates making our text mining for synthesis project a success. I would also like to thank my nearest neighbors, Peichen Zhong, Bowen Deng, Zhuohan Li, Xinye Zhao, Shashwat Anand, Yunyeong Choi, and Fengyu Xie, for the many thought-provoking discussions. I also thank the A-lab team, Yan Zeng, Nathan Szymanski, Bernardus Rendy, Yuxing Fei, and David Milsted, for the wonderful collaboration. In addition, I want to thank the whole Ceder group family for all the precious memories.

Lastly, I would like to extend my gratitude to my parents and girlfriend for their unwavering support and encouragement throughout this endeavor.

Chapter 1

Challenges and opportunities for materials synthesis and text mining

1.1 Predictive solid-state synthesis

Materials discovery has become significantly facilitated and accelerated by high-throughput *ab initio* computations. Thanks to the gigantic leaps in computational power and focused big-data-driven efforts such as the Materials Genome Initiative [1, 2], the ability to rapidly design interesting novel compounds has displaced the materials innovation bottleneck to the development of synthesis routes for the desired material [3]. In other words, we are "discovering" exciting new materials, but we are unable to produce them. Understanding how to synthesize the desired compounds is a grand challenge in the development of novel materials.

Solid-state synthesis is the prevailing approach for making inorganic materials [4]. In a typical solid-state synthesis experiment for a target material, precursor materials are mixed to obtain a homogeneous mixture. The mixture is then heated at a high temperature for a specified period of time. The complexity of synthesis mainly originates from the interactions of many design variables, including the diversity of precursor candidates for each element in the target material (oxides, hydroxides, carbonates, etc.), the experimental conditions (temperature, atmosphere, etc.), and the chronological organization of operations (mixing, firing, reducing, etc.). Properly selecting the combination of experimental variables is crucial and demanding for successful synthesis [5–7].

Because of the lack of a general theory for how phases evolve during heating, solid-state synthesis design is mostly driven by heuristics and basic chemical insights. Unlike the success of retrosynthesis [8] and automated design for organic materials based on the conservation and transformation of functional groups [9–11], the mechanisms underlying inorganic solid-

state synthesis are not well understood [10, 12–14]. Researchers are trying to tackle this challenge from different perspectives, including *in situ* experiments [6, 7, 15], thermodynamic analysis [16–21], and machine learning-guided synthesis parameter search [22–24].

To achieve predictive synthesis for the desired material, one potential approach is to learn synthesis design patterns from a large volume of experimental synthesis procedures. Successful machine-learning methods to retrosynthesis in organic chemistry [11, 25–27] have been enabled by organic chemistry reaction databases, such as Reaxys [28], which include over 12 million single-step reactions. However, two questions arise when applying machine learning to inorganic solid-state synthesis. The first question is how to obtain a large volume of inorganic synthesis data. Prior to our study [29, 30], a comprehensive, large-scale database of structured synthesis procedures for inorganic materials did not exist. The second question is how to devise an interpretable learning strategy that rationalizes inorganic synthesis design. Inorganic synthesis is mostly a one or few-step process, while retrosynthesis for organic materials usually consists of a sequence of steps, where each step corresponds to an independent chemical reaction. Benefiting from existing knowledge in organic retrosynthesis, designing an algorithm that decomposes the entire organic synthesis into multiple chemical steps for the prediction of retrosynthesis is not only natural but also effective [11]. In contrast, the understanding of inorganic synthesis is relatively poor, and it remains unclear which algorithm is most effective for predicting inorganic synthesis.

In this work, we aim to address these two questions by employing interdisciplinary methods that combine materials science, natural language processing (NLP), and machine learning. First, we develop NLP algorithms to extract a large-scale inorganic synthesis dataset from materials science literature. Although there is no structured synthesis dataset available, decades of synthesis data are locked up in written natural language in papers published by materials researchers. Recent progress in NLP [31–35] has made it possible to retrieve structured data from unstructured textual descriptions of synthesis procedures. Based on our text-mined dataset, we leverage the concept of similarity to mimic how experimental researchers select precursor materials during synthesis design. We create machine learning models to evaluate the similarity of precursor and target materials. Given the similarity of materials, new synthesis reactions can be suggested by adapting precedent synthesis procedures.

1.2 Progress in NLP

The development of NLP dates back to the 1940s, but the most significant progress has occurred in the past two decades (Figure 1.1 [31]). The progress of NLP can be roughly divided into two main parts.

The first part is the breakthroughs of fundamental architectures. In 2001, Bengio et al. [36]

proposed the concept of a feed forward neural network to learn the joint probability function of sequences of words. In 2008, Collobert et al. [37] proposed the application of multi-task learning in the field of NLP. Nowadays, the concept of multi-task learning is widely used in large language models (LLMs) [38] . In the early 2010s, more advanced neural networks were applied to NLP problems. At first, the attention was on convolutional neural network (CNN) [39] because of its success in the field of computer vision. Subsequently, researchers adopted the recurrent neural network (RNN) to capture the context of a word with respect to surrounding words in the sentence. Long Short-Term Memory (LSTM) [40] neural network and gated recurrent unit (GRU) [41] are the most successful variants of RNN because they are good at learning long-term historical information. Another important breakthrough was the neural networks using attention mechanisms such as Transformer [42]. Attention mechanisms are influential in NLP as they facilitate the modeling of distant dependencies similar to LSTM while enabling better training parallelization. As a result, they can scale up to larger datasets more effectively than LSTM. The Transformer is the most widely used architecture in state-of-the-art LLMs.

The other part is the breakthroughs of word embeddings. Mikolov et al. [43, 44] proposed a word embedding process where the dense vector representation of text was addressed. Their Word2Vec model automatically learns a vector embedding by predicting the context for each word. The success of Word2Vec opened up the possibility of automated feature engineering with the use of neural networks. Pre-trained language models, such as BERT [38] and GPT [45], are able to produce contextual word embeddings that move beyond global word representations like Word2Vec and achieve ground-breaking performance on a wide range of NLP tasks [46]. The technique for training NLP models has evolved from training separate models to pre-training an LLM and fine-tuning it for specific tasks. Pre-trained LLMs have also opened up a new paradigm known as "prompt-based learning" [47], enabling model prediction to be guided by different prompts and achieving greater flexibility.

Thanks to these breakthroughs, state-of-the-art NLP models, such as ChatGPT [48], have demonstrated superiority over humans in many general-purpose tasks [49]. This study leverages advanced NLP methods to extract inorganic synthesis data from materials science literature.

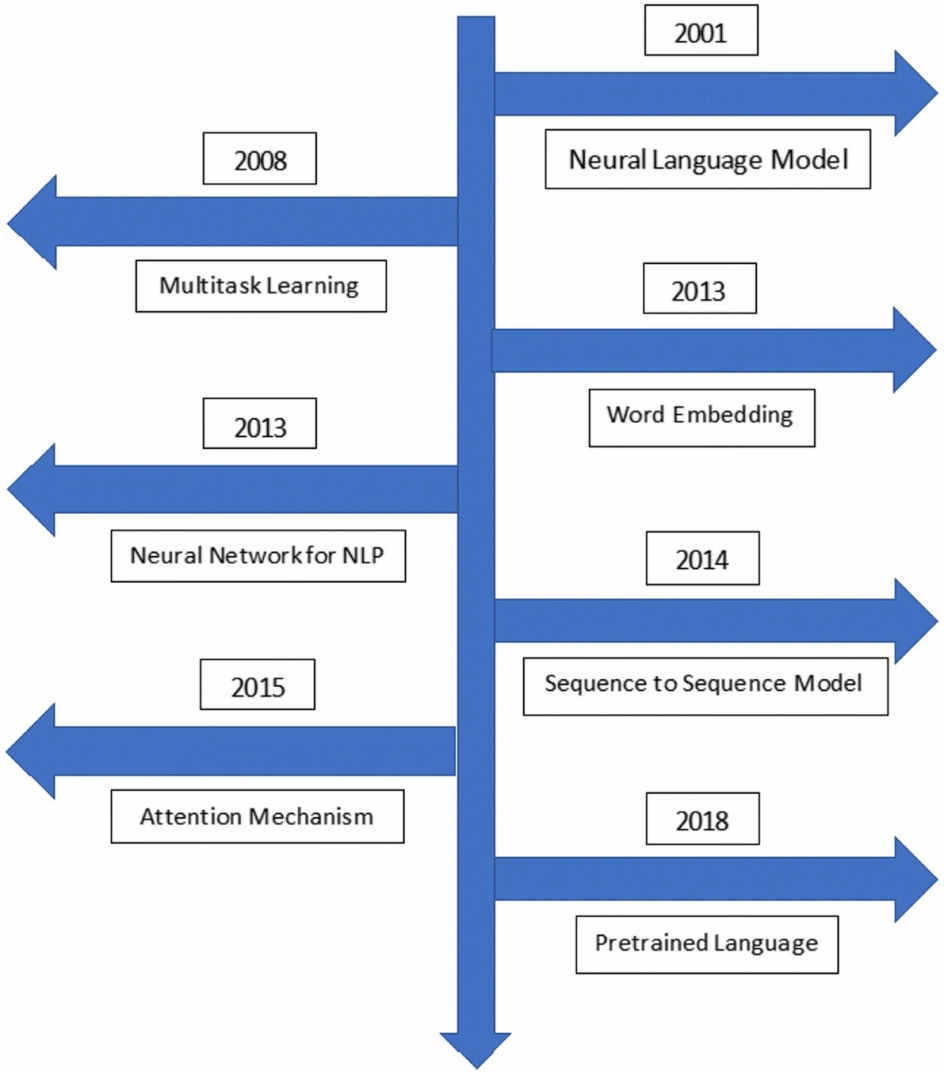

Figure 1.1: Important NLP breakthroughs in the past two decades [31].

1.3 Text mining for materials science

The significant progress in NLP provides new opportunities for data-centric materials science research, such as the Materials Genome Initiative [1]. Our analysis of the papers indexed in the Web of Science repository shows that since the beginning of the 2000s, the number of publications in different fields of materials science has increased exponentially (Figure 1.2). The increasing number of published papers is making it difficult for individuals to access the knowledge contained within them. Text mining will become an indispensable method for curating reported but uncollected materials science data.

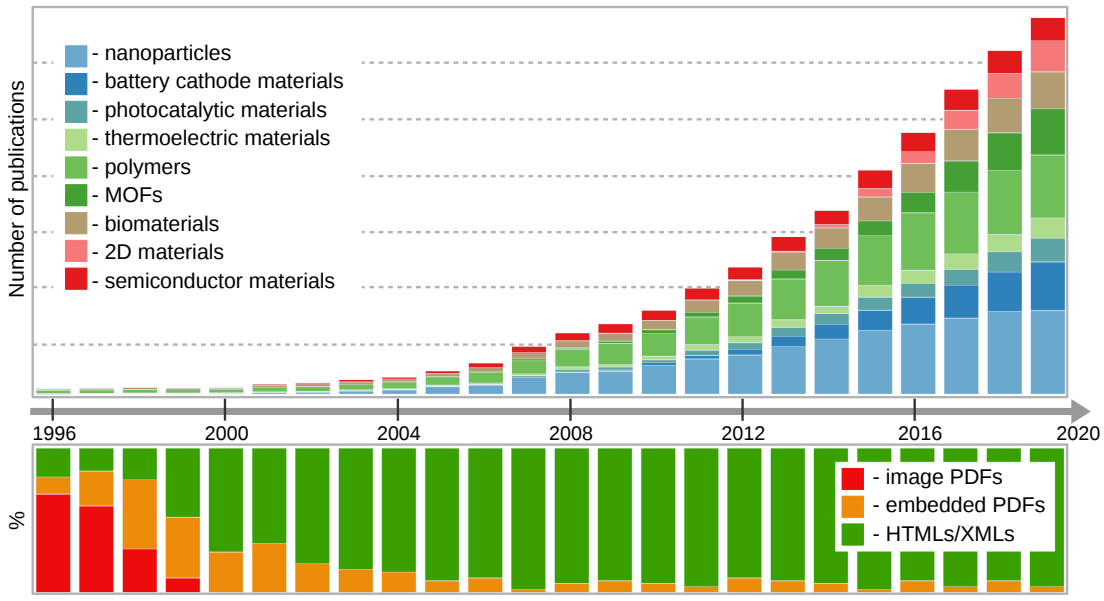

Figure 1.2: Publication trend from 1996 to 2020 in materials science. *Top panel:* The growing number of annual publications in various areas of materials science. Data was collected through manual queries in the Web of Science publication database. The analysis includes only research articles, communications, letters, and conference proceedings. The publication count is in the order of 10^3. *Bottom panel:* A relative comparison of the proportion of scientific papers accessible online as image PDF or embedded PDF versus those in HTML/XML format. The grey arrow indicates time intervals for both the top and bottom panels.

Previous research has demonstrated the effectiveness of text mining for obtaining materials insights. Young et al. [50] developed a semi-automated text-mining pipeline to extract and analyze the growth conditions for four different oxide materials synthesized with a pulsed laser deposition technique. They were able to obtain the range of growth temperatures and pressures, and predict the relative values of critical temperatures by applying a decision tree classifier. Court et al.[51] used the records of Curie and Néel temperatures text-mined from the scientific literature [52] to reconstruct the phase diagrams of magnetic and superconducting materials. Kim et al. [23] explored the parameters of hydrothermal and calcination reactions for metal oxides by analyzing the data extracted from 22,065 scientific publications. A decision tree model applied to predict synthesis routes for titania nanotubes identified the concentration of NaOH and synthesis temperature as the most important factors that lead to nanotube formation. Jensen et al. [53] used a similar approach to predict the density of germanium-containing zeolite frameworks and to uncover their synthesis parameters. Tshitoyan et al. [54] applied the Word2Vec model [43] to 3 million abstracts to learn the word embeddings for materials and application areas. Interestingly, their model was able to not

only learn some aspects of the chemistry underlying the relations between materials but also to draw a similarity between materials for different applications. In particular, it was demonstrated that such similarity could be used to predict novel thermoelectric materials.

These successful examples across diverse fields of materials science confirm the broad applicability of text mining. They also highlight an important aspect of scientific text mining: its capability to uncover latent knowledge about a subject by comprehending a large amount of unstructured data – a task that is not possible for a human. Similarly, text mining holds great potential for contributing novel data and insights to the field of solid-state synthesis science.

1.4 Structure of this thesis

This work aims at two objectives: (1) developing NLP algorithms and an automated text-mining pipeline to extract large volumes of structured inorganic synthesis data from material science literature, and (2) developing algorithms for precursor selection in solid-state synthesis based on the text-mined dataset.

In Chapter 2, we first introduce the problems to solve for text mining of inorganic synthesis procedures. Then, we present our algorithm for the identification of precursor and target materials in the text description of experimental details and the full text-mining pipeline for the extraction of structured synthesis data from materials science papers. We also evaluate the quality of our text-mined dataset and demonstrate its possible usage with a series of meta-analyses.

In Chapter 3, we first present the variety of the extracted precursors. Then, we touch on the problem of precursor selection by exploring how frequently researchers substitute one precursor with another while retaining the target. Next, We formalize the similarity of precursors based on the substitution of precursors and synthesis temperatures. At last, we validate whether this text-mined similarity of precursors is a reasonable metric for the creation of alternative recipes.

In Chapter 4, we first discuss the challenges of precursor selection for novel materials and the limitation of the approach in Chapter 3. Then, we propose a precursor recommendation strategy based on the similarity of target materials. At last, we evaluate the performance of our precursor recommendation pipeline through comparison with baseline models and case studies.

In Chapter 5, we summarize this work and outline future directions.

Chapter 2

Text mining for inorganic solid-state synthesis

Data-driven approaches could offer a promising solution to address the challenge of inorganic solid-state synthesis. However, the availability of a comprehensive database that includes experimental procedures for a wide range of materials is crucial. In this chapter[1], we present a text-mining pipeline to extract synthesis procedures from millions of materials science papers. The large-scale dataset of "codified recipes" could open a new avenue for understanding and predicting inorganic solid-state synthesis.

We first point out the promise and also potential problems associated with text mining for materials synthesis. Next, we provide an in-depth explanation of identifying precursor and target materials in the text description of experimental details, as selecting the proper precursors for various targets is a core problem in materials synthesis. By compiling the algorithms together, including the one for extraction of precursors and targets and the ones for extraction of other synthesis variables, we established a fully automated text-mining pipeline that extracts the structured data of synthesis procedures from the literature. Based on the text-mined synthesis set, we conducted a series of meta-analyses to exemplify how to use this dataset to acquire knowledge on synthesis. Digitizing and systematizing the large corpus of existing solid-state chemistry literature paves the way toward the development of data-driven approaches for understanding inorganic materials synthesis [29, 30, 35].

[1] This Chapter incorporates sections from three previously published papers with permission from the authors: (1) Tanjin He, Wenhao Sun, Haoyan Huo, Olga Kononova, Ziqin Rong, Vahe Tshitoyan, Tiago Botari, and Gerbrand Ceder. "Similarity of precursors in solid-state synthesis as text-mined from scientific literature." *Chemistry of Materials* 32, no. 18 (2020): 7861-7873 [29]; (2) Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A. Olivetti, and Gerbrand Ceder. "Opportunities and challenges of text mining in materials research." *Iscience* 24, no. 3 (2021): 102155 [35]; and (3) Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. "Text-mined dataset of inorganic materials synthesis recipes." *Scientific data* 6, no. 1 (2019): 203 [30].

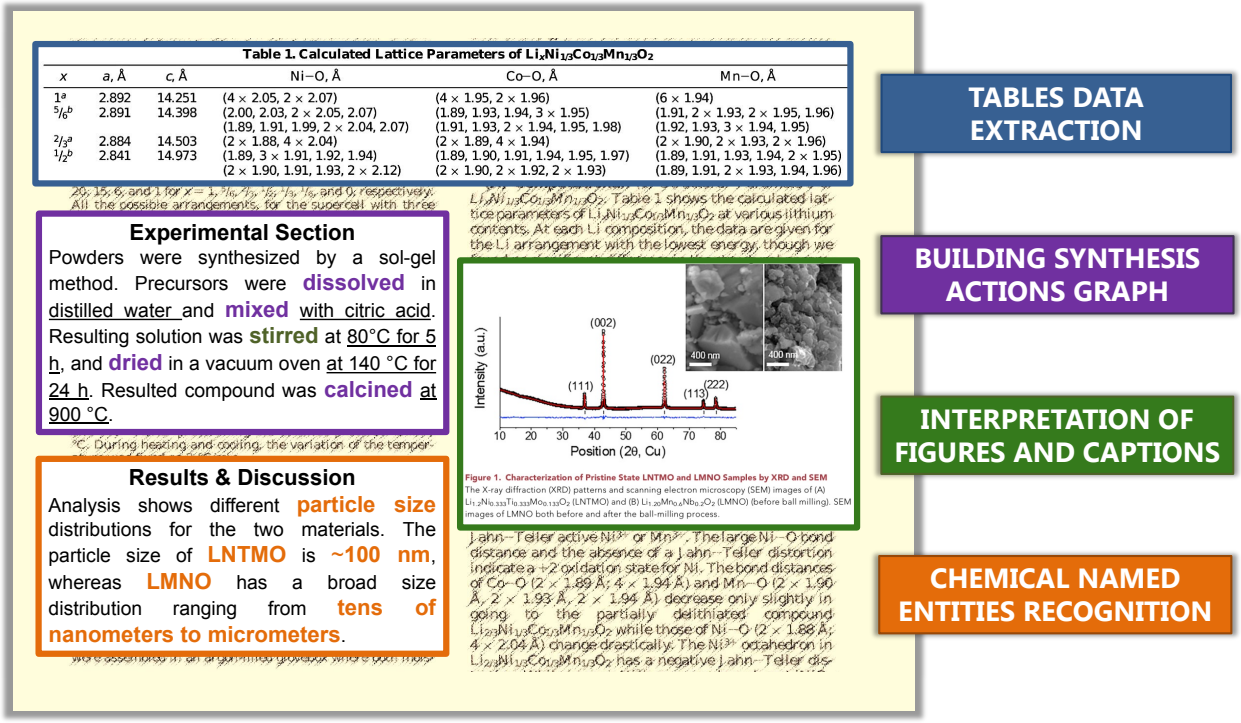

Figure 2.1: Schematic representation of various information types that can be extracted from a typical materials science paper.

2.1 Problems in text mining for materials synthesis

A scientific paper usually contains a wealth of information, including text, table, and figures, as shown in Figure 2.1. Various types of data can be extracted from materials science papers, including synthesis experimental procedures, crystal structure parameters, diffraction patterns, morphology information, materials properties, and so forth. The diversity of available data creates a substantial opportunity for data-centric research in materials science. This study aims to extract data related to synthesis, which is commonly found as natural language in the experimental section of materials science papers.

In order to take advantage of decades of valuable synthesis knowledge buried in the literature, the text data must be converted from an unstructured to a structured form. To achieve this goal, a range of NLP challenges must be tackled, such as acquiring the text corpus, transforming it into raw text, text segmentation, text representation modeling, text classification, and information retrieval. Although many general-purpose NLP approaches exist for these problems, the application to the chemical or materials science domain requires adaptation of both methods and models, as well as the development of an adequate training set that

complies with the goal of the specific project.

Given the crucial nature of precursors and targets in synthesis data, we intend to expand on a discussion about material entity recognition (MER), i.e., identifying precursor and target materials in the text description of experimental details. The extraction of precursors and targets from written text is difficult because of the complexities of natural language for inorganic synthesis.

First, a material entity can be written in various complicated forms; they can be represented as chemical formulas such as "Al2O3" and "AxB1-xC2-δ", chemical terms such as hafnium oxide, acronyms such as "PZT" for "Pb(Zr0. 5Ti0. 5)O3", and even more complicated notations for composites and doped materials such as "Si3N4-30wt%ZrB2" and "Zn3Ga2Ge2-xSixO10: 2. 5mol%Cr3+". Translating this knowledge into explicit rules for Chemical Named Entity Recognition (CNER) [55] is difficult.

Second, material entities can play different roles in synthesis experiments such as targets, reagents, reaction media, and so forth. While this can usually be recognized easily by researchers based on their domain-specific knowledge and grammar comprehension, such an implicit assignment of meaning is much harder in computational algorithms. One naïve approach could be to use multiple rules to distinguish between targets and precursors. For example, assign a simple material (e.g., "TiO2)" as a precursor and a complex material (e.g., "Pb(Zr0.5Ti0.5)O3") as a target, because researchers usually use simple materials to synthesize a complex one. However, there are many cases that do not follow this rule: the same material zirconia can be a precursor for a Zr-based complex oxide, an auxiliary component as a grinding media, or even a target in the synthesis of stabilized or doped zirconia [56]. In order to correctly identify the role of a material, one needs to read the context of the sentence or entire paragraph, in addition to finding the material expressions. Hardcoding all possible rules would require an enormous amount of human effort.

In the following sections, we first tackle the problem of MER and further present the full text-mining pipeline, as well as the synthesis dataset extracted using these algorithms.

2.2 Extraction of precursor and target materials

In this section, our focus is specifically to identify precursor and target materials in inorganic solid-state synthesis text for further studying the relations between various precursors and correlating them with targets. We describe the Synthesis Materials Recognizer (SMR) model for this MER task. By comparing with a baseline model, we explain how the SMR model works and its advantages and limitations.

Algorithm Design and Execution

To identify and extract precursor and target materials from a synthesis paragraph, a two-step model (SMR) based on bi-directional long-short term memory (Bi-LSTM) neural network [40] was implemented. The SMR model recognizes context clues provided by the words around the precursors/targets in the sentence. The identification of material entities in the text and their subsequent classification as targets, precursors, or something else were performed in two steps, as shown in Figure 2.2: first we identified all material entities present in a sentence; next we replaced each material with a keyword "<MAT>" and classified it as a "Target", "Precursor", or just "Material". Each step was executed by a different Bi-LSTM neural network with a conditional random field (CRF) [57] layer on top of it (Bi-LSTM-CRF) [40, 58].

For the first step, each input word was represented as the combination of a word-level embedding and a character-level embedding via an embedding layer. The word-level embeddings, which are vectors of real numbers representing the words, were trained using the Word2Vec approach [43, 44] with ∼33,000 paragraphs on solid-state synthesis to capture the semantic and syntactic similarity of words in synthesis text. In this embedding layer, the characters of each word were converted into an embedding vector using another Bi-LSTM to learn the character-level features such as the prefix and suffix information. The character embedding was concatenated with the pre-trained word embedding and input into a Bi-LSTM to capture the left and right context at every word. Finally, the output from Bi-LSTM was combined with a CRF model, which characterized the transition probability from one tag to another to produce the final prediction.

For the second step, a Bi-LSTM with a similar structure to that in the first step was used but the inputs were different. All the materials in the input sentences were replaced with the word "<MAT>" so that the role of a material in synthesis is predicted mainly based on the surrounding context. We found this to be more effective than directly using the specific materials words as input to the Bi-LSTM, because such a direct model tries to store the mapping information from each different material to the particular role this material is mostly used for, which brings in bias for frequently appearing materials. For example, as "zirconia" often describes the balls used in ball milling, it is difficult for the neural network to deviate from this assignment and treat "zirconia" as a target or precursor. Since all the chemical information about the material is lost by inputting "<MAT>" instead of the materials words, we also included two additional features in the word representation, that is, the number of metal/metalloid elements and a flag indicating whether the material contained C, H, and O elements only. These additional features assist in the differentiation of precursors and targets, as they tend to have different numbers of metal/metalloid elements and are generally not organic compounds in inorganic synthesis. The composition information was obtained by parsing the raw text of the material entities by regular expression comparison [30].

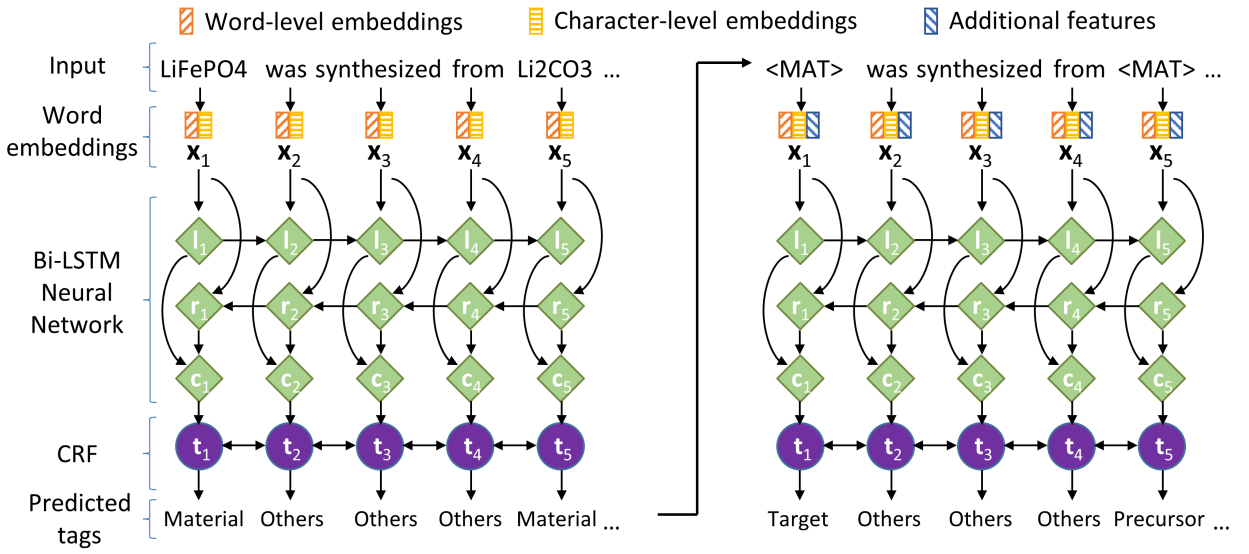

Figure 2.2: Main architecture of the SMR model. $\boldsymbol{x}_i$ is the embedding used as the input for the Bi-LSTM-CRF neural network. $\boldsymbol{l}_i$ represents the i^{th} token and its left context. $\boldsymbol{r}_i$ represents the i^{th} token and its right context. $\boldsymbol{c}_i$ is the combination of $\boldsymbol{l}_i$ and $\boldsymbol{r}_i$. $\boldsymbol{t}_i$ represents the score for different tags.

Bi-LSTM is able to infer the role of materials from context because Bi-LSTM specifies a variable called cell state to store the information about the words around the material. Figure 2.3 shows a typical example of the trained Bi-LSTM cell state continuously changing depending on token context in the example sentence [59] when feeding the tokens into Bi-LSTM one by one. In this study, 100 neurons (cells) were used to represent the context information; Figure 2.3 displays one of the cell states relevant to the context of precursors. To obtain the cell states for the next token, both the next token and the current cell states are input to the network. Hence, after seeing the sequence of tokens "was prepared from" in the example sentence, the network predicts from the context that the tokens following this phrase most likely refer to a precursor(s). Likewise, the network predicts that the tokens following "at 700 ° C for" most likely are not precursors.

To train the SMR model, 834 solid-state synthesis paragraphs from 750 papers were tokenized with ChemDataExtractor [60], and each token was manually annotated with tags of "Material", "Target", "Precursor", and "Outside" (not a material entity). In the annotation, a target is defined as a final material obtained through a series of lab operations in the complete synthesis process, and a precursor is defined as a starting reagent involved in the synthesis process through a lab operation and contributing to the target composition. Other materials include media, gas, device materials, and so forth. The annotation dataset contains 8,601 materials, out of which 1,256 are targets and 3,295 are precursors. The anno-

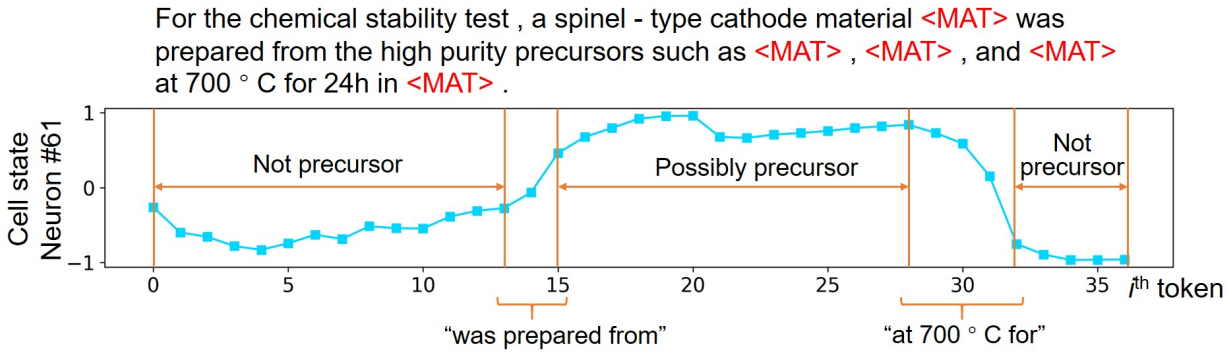

Figure 2.3: Change of one LSTM cell state in different context for precursor classification. The tokens in the example sentence are separated by spaces in the hanging text and represented as the sequence numbers on the x-axis.

tated dataset was randomly split into training/validation/test sets with 500/100/150 papers in each set. Early stopping [61] was used to minimize overfitting by stopping the iterative training when the best performance was achieved on the validation set. To reduce the variance resulting from the limited size of the training set, the six models trained in a six-fold cross-validation process were combined together to make the final decision by voting in the classification. The entire training and test process was repeated 10 times, and the average result of the test sets is reported.

Accuracy and working examples

We first aim to demonstrate that the recognition of context clues is necessary for the MER task by comparing the SMR model with a baseline model based on naïve rules. To build this baseline model, we used ChemDataExtractor [60] to identify and extract materials from the text. Then, inspired by a scientific perspective that researchers usually use simple materials to synthesize a complex one, the precursors and targets were assigned based on the number of elements: materials with only one metal/metalloid element were assumed to be precursors, and materials with at least two metal/metalloid elements were assumed to be targets. This baseline is a least-effort model but provides a quantitative reference for understanding the importance of capturing context information.

In Table 2.1, we compare the performance of the SMR model and the baseline model using F_1 scores, which provides a measure of the accuracy of a binary classification test based on the harmonic mean of the precision and recall. The F_1 scores on the extraction of all materials, precursors, and targets using the SMR model are 95.0%, 90.0%, and 84.5%, respectively. Out of all the extracted entities, 88.9% of precursors and 85.9% of targets in

Table 2.1: Precision, recall, and F_1 scores for the baseline and SMR models to extract materials, precursors, and targets. The type "Materials" include precursors, targets, and all other materials.

Model	Type	Precision (%)	Recall (%)	F_1 score (%)
Baseline	Materials	78.3	68.3	73.0
	Precursors	60.9	82.2	70.0
	Targets	48.5	33.0	32.1
SMR	Materials	94.6	95.3	95.0
	Precursors	88.9	91.2	90.0
	Targets	85.9	83.4	84.5

the test set are correctly identified. These correct cases account for 91.2% and 83.4% of all the precursors and targets which should be extracted, respectively. The possibility of errors increases when multiple precursors and targets are present in the same sentence. Out of all the sentences containing precursors/targets, the rate to successfully retrieve all the precursors and targets in each sentence is 73.4%. Some representative successful examples from the SMR model, such as the recognition of the targets "LiBaBO3:Sm3+" and "(0.725-x)BiFeO3-xBi(Ni0.5Mn0.5)O3-0.275BaTiO3 + 1 mol% MnO2", are shown in Table 2.2.

We interpret the results as follows. In the baseline model, only the information from the material entity itself is used, resulting in low F_1 scores for the extraction of precursors and targets (70.0% and 32.1%, respectively). In contrast, the SMR model achieves better F_1 scores because Bi-LSTM is able to infer the role of materials from the context. For example, as discussed previously, the Bi-LSTM infers from the tokens "was prepared from" to mean that the following tokens probably refer to a precursor(s). Likewise, the network predicts that the tokens following "at 700 ° C for" most likely are not precursors. For a precursor with more than one metal/metalloid element, the baseline model fails to recognize it regardless of the context, while the SMR model can still identify the precursor nature of this material.

To highlight the unique difficulty in the problem of named entity recognition (NER) for inorganic material synthesis, we also compared our SMR model with other publicly available toolkits intended for general NER tasks, such as NLTK [62] and SpaCy [63], and the ones for CNER tasks, such as OSCAR4 [64], tmChem [65], ChemSpot [66] and ChemDataExtractor [60] (Table 2.3). Here we only compared the performance for the identification of material entities because these toolkits are not able to distinguish the different roles of materials, such as precursors and targets. The general-purpose NER toolkits, NLTK and Spacy, only classify text tokens into common categories such as locations, persons, and organizations.

They are expected not to be able to extract any material entities as the second example in Table 2.3. In the first and third examples, NLTK and Spacy even make misleading predictions, such as classifying "CH3COO" as an organization and "Bi2Cu1-xNixO4" as a person. The CNER toolkits, OSCAR4, tmChem, ChemDataExtractor, and ChemSpot perform better than NLTK and Spacy because they were designed to identify chemical terms. Typically, these CNER toolkits work well on simple material entities such as "water" and "Ni(NO3)2·6H2O". Nonetheless, they exhibit an inability to recognize complex material entities such as "lithium, cobalt, and manganese nitrates". Our model outperforms the aforementioned toolkits because of the training set specifically curated for inorganic synthesis. Therefore, the adaption of off-the-shelf NLP methods and models is an essential step toward text mining in materials research.

However, some situations remain difficult for the SMR model:

(1) Some material entities tokenized into multiple tokens are not completely extracted. For example, the incomplete pieces "(Ba1-x(K" and "Na)x/2Lax/2)(Mg1/3Nb2/3)O3" are extracted instead of "(Ba1-x(K or Na)x/2Lax/2)(Mg1/3Nb2/3)O3", as listed in Table 2.2. The identification of these materials is difficult because of the syntactic variability and ambiguity of multiword expressions (MWEs) [67], which might be improved by incorporating recent progress on MWE identification such as the language-independent architecture proposed by Taslimipoor et al. [68]. The number of training sentences containing MWE materials might remain an issue considering the relatively large dataset [69] used by Taslimipoor et al. [68].

(2) Some sentences are ambiguous to the SMR model because of the limitations of the training set. For example, the model correctly classifies "Y2O3" as a precursor in "Y2O3 as a precursor was added" and "Y2O3" as neither target nor precursor in "Y2O3 as a grinding media was added". However, in the sentence "Y2O3 as a donor impurity was added", the model does not understand "donor impurity" and only assigns "Y2O3" as an ordinary material rather than a precursor. This situation might be improved by including more contextual information in the input, such as the sentence embeddings [70] of previous and next sentences, and contextualized word embeddings trained on a much larger corpus (e.g. BERT [38] and SciBERT [71]). Future possible directions for research include training these embedding models on papers specifically on materials synthesis, although the training process may require a significant manual time investment and considerable computational resources.

(3) Misclassification can occur when the sentence is written with a complicated structure. For example, the target "Ba0.5Sr0.5CoxFe1-xO3-δ" is misclassified as a precursor when the order of precursors and targets is reversed or closely mixed in the sentence and the materials around this word are all precursors, as shown in Table 2.2. These sentences with a complicated structure must often be treated on a case-by-case basis, and it is

difficult for an NLP model to pick up general rules to correct these errors. A potential solution is to conduct selective sampling to annotate sentences with complex syntax more efficiently, where only the ones that a pre-trained classifier is less confident with will be sampled for annotation [72]. Our current model lays a foundation for selective sampling.

Table 2.2: Representative successful and failed examples from the SMR model in this study.

Example Sentences	Expected	Error in Extraction
Successful		
The LiBaBO3:Sm3+ samples were prepared by solid-state reaction. [73]	Target: LiBaBO3:Sm3+	N/A
Ceramic samples of (0.725-x)BiFeO3-xBi(Ni0.5Mn0.5)O3-0.275BaTiO3 + 1 mol% MnO2 (x = 0-0.08) (BFO-BT-BNM-x) were prepared by the conventional solid-state route using high-purity metal oxides and carbonates as starting materials: Bi2O3 (99 %), Fe2O3 (99 %), BaCO3 (99 %), TiO2 (98 %), NiO (99 %), MnO2 (99.99 %). [74]	Targets: (0.725-x)BiFeO3-xBi(Ni0.5Mn0.5)O3-0.275BaTiO3 + 1 mol% MnO2, BFO-BT-BNM-x Precursors: Bi2O3, Fe2O3, BaCO3, TiO2, NiO, MnO2	N/A
Y2O3 as a precursor was added.	Precursor: Y2O3	N/A
Y2O3 as a grinding media was added.	Material: Y2O3	N/A
Failed		
(Ba1-x(K or Na)x/2Lax/2)(Mg1/3Nb2/3)O3 with $0 \leq x \leq 1$ were synthesized by a conventional solid-state reaction method. [75]	Target: (Ba1-x(K or Na)x/2Lax/2)(Mg1/3Nb2/3) O3	"(Ba1-x(K" and "Na)x/2Lax/2) (Mg1/3Nb2/3)O3" extracted
Y2O3 as a donor impurity was added. [76]	Precursor: Y2O3	Y2O3 extracted as an ordinary material
Required amounts of BaCO3, SrCO3, CoCO3·0.5H2O and Fe2O3 powders for Ba0.5Sr0.5CoxFe1-xO3-δ, Pr6O11, BaCO3, and CoCO3·0.5H2O powders for PrBaCo2O5+δ were mixed and ball-milled for 24h. [77]	Targets: Ba0.5Sr0.5CoxFe1-xO3-δ, PrBaCo2O5+δ Precursors: BaCO3, SrCO3, CoCO3·0.5H2O, Fe2O3, Pr6O11, BaCO3, CoCO3·0.5H2O	Ba0.5Sr0.5CoxFe1-xO3-δ extracted as a precursor

Benefits of the Two-step Model

In the SMR model, the identification of materials entities in the text and their subsequent classification as targets, precursors, or something else were performed in two steps: first we identified all materials entities present in a sentence; next we replaced each material with a keyword "<MAT>" and classified it as a "Target", "Precursor", or just "Material". It is also possible to classify "Target" and "Precursor" materials from initial plain tokens without first finding materials entities and then replacing materials with the keyword "<MAT>" (i.e., make only one-step of the model). However, we found that the two-step model provides improved generality.

The two-step model reduces bias from frequently appearing materials. For example, "zirconia" frequently appears in synthesis as a grinding media, leading the one-step model to classify it as a non-target material based on a word mapping. In the two-step model, this problem is avoided because the term "zirconia" is replaced by the keyword "<MAT>", thus only the context of the token is used for the classification of targets/precursors, rather than the material name. To quantify this finding, we selected 50 "unusual" sentences in which the role of "zirconia" was identified and was classified as a target using the two-step model. Manual inspection indicated that 30 of the classifications were correct and 20 were wrong due to the ambiguity in the sentences (Table 2.4). In this particular situation, both the one-step and two-step models were easily confused, and the accuracy of the one-step model was even slightly higher. However, the one-step model behaved inconsistently. It classified correctly only 17 of the zirconia targets, while in 13 cases the targets were missed because of the tendency to regard "zirconia" as a non-target material. When the word "zirconia" was replaced with "LiMn2O4", the one-step model classified all 50 cases as targets. This experiment indicates that the one-step model leads to inconsistent results and unstable accuracy for different materials in the same context, which can introduce a systematic bias between frequently and infrequently appearing materials.

Since the extraction of materials and classification of targets/precursors are implemented separately, the two-step model can, in principle, be applied to research fields outside of solid-state materials synthesis. For example, consider the sentence "1,3,7-trimethylxanthine was synthesized from uracil" [78]. Here, the target "1,3,7-trimethylxanthine" is an organic material very different from the expressions of inorganic materials. As a result, both the one-step model and the two-step model trained with solid-state synthesis data mistook "1,3,7-trimethylxanthine" for a general English word rather than a material. To understand that "1,3,7-trimethylxanthine" is a target, the one-step model needs to be retrained by adding similar materials to the training data. However, when the two-step model fails in identification of some materials, the second step is still useful because it reserves the capability to classify a material as a precursor or target from context. The first step can be fixed by combining with some popular chemical name databases such as PubChem [79] because it is easy to find "1,3,7-trimethylxanthine" as a material in such databases. After replacing

Table 2.3: Examples of the chemical named entities extracted by the general-purpose NER tools NLTK [62] and SpaCy [63], and the tools trained on chemical corpus OSCAR4 [64], tmChem [65], ChemSpot [66], ChemDataExtractor [60], SMR (this work). For the general-purpose tools, the assigned labels are given in parentheses. For the chemical NERs, only entities labeled as chemical compounds are shown.

An aqueous solution was prepared by dissolving lithium, cobalt, and manganese nitrates in de-ionized water	
NLTK	–
SpaCy	'manganese' (*nationalities or religious or political groups*)
OSCAR4	'aqueous', 'lithium', 'cobalt', 'manganese', 'nitrates', 'water'
tmChem	'lithium', 'cobalt', 'manganese nitrates'
ChemDataExtractor	'lithium', 'cobalt', 'manganese nitrates'
ChemSpot	'lithium', 'cobalt', 'manganese nitrates', 'water'
SMR	'lithium, cobalt, and manganese nitrates', 'water'
A series of Ce3+-Eu2+ co-doped Ca2Si5N8 phosphors were successfully synthesized	
NLTK	–
SpaCy	–
OSCAR4	'Ce3+', 'Eu2+', 'Ca2Si5N8'
tmChem	'Ce3+-Eu2+', 'Ca2Si5N8'
ChemDataExtractor	'Ce3+-Eu2+', 'Ca2Si5N8'
ChemSpot	'Ce3+-Eu2', 'co', 'Ca2Si5N8'
SMR	'Ce3+-Eu2+ co-doped Ca2Si5N8'
High-purity Bi(NO3)3·5H2O, Ni(NO3)2·6H2O and Cu(CH3COO)2·H2O were used as starting materials for Bi2Cu1-xNixO4 powders	
NLTK	'NO3', 'NO3', 'CH3COO' (*organizations*); 'Ni', 'Cu' (*countries, cities, states*)
SpaCy	'Bi2Cu1-xNixO4' (*person*)
OSCAR4	'Bi(NO3)3·5H2O', 'Ni(NO3)2·6H2O', 'Cu(CH3COO)2·H2O'
tmChem	'Bi(NO3)3·5H2O', 'Ni(NO3)2·6H2O', 'Cu(CH3COO)2·H2O', 'Bi2Cu1-xNixO4'
ChemDataExtractor	'Bi(NO3)3·5H2O', 'Ni(NO3)2·6H2O', 'Cu(CH3COO)2·H2O', 'Bi2Cu1-xNixO4'
ChemSpot	'Bi(NO3)3·5H2O', 'Ni(NO3)2·6H2O', 'Cu(CH3COO)2·H2O', 'Bi2Cu1-xNixO4'
SMR	'Bi(NO3)3·5H2O', 'Ni(NO3)2·6H2O', 'Cu(CH3COO)2·H2O', 'Bi2Cu1-xNixO4'

Table 2.4: Analysis of 50 sentences containing the term "zirconia" classified as targets by the two-step model. One-step: Bi-LSTM-CRF was used to classify materials/precursors/targets from plain tokens without replacing materials with the keyword "<MAT>". Two-step: first materials were identified and replaced with keyword "<MAT>", and then these materials were classified as precursors/targets. Notation "T" and "F" represents correct and incorrect classification, respectively.

Role of zirconia	Manual inspection	Two-step (input "<MAT>")	One-step (input "zirconia")	One-step & replace "zirconia" with "LiMn2O4"
Target	30	50 (T 30, F 20)	21 (T 17, F 4)	50 (T 30, F 20)
Non-target	20	0	29 (T 16, F 13)	0

"1,3,7-trimethylxanthine" with the keyword "<MAT>", the two-step model would correctly classify "1,3,7-trimethylxanthine" as a target, because the classifier in the second step is trained to recognize "<MAT>" in different context.

Comparison with BERT

The recently released large language models such as BERT [38] and GPT [45] are attracting extensive attention because of their universality and excellent predictive power. Although trained on a large corpus of over 3 billion words, BERT can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. We fine-tuned a BERT model with the annotated dataset in this paper and the hyperparameter searching recommended by Devlin et al. [38]. The precision, recall, and F_1 scores from the fine-tuned BERT model are similar to our SMR model (Table 2.5). However, the fine-tuned BERT model is 5 times slower than the SMR model in the prediction because BERT employs a huge neural network structure. Nevertheless, BERT is still highly potential because more work inherited from BERT [71, 80] keeps coming out since 2019. Especially, we are interested in retraining and distilling a BERT model with the synthesis papers in our database rather than only fine-tuning, though it requires significant time and computational resources. The utilization of BERT model is one of our future directions.

Table 2.5: Precision, recall, F_1 scores, and time for our SMR model and the fine-tuned BERT model on the same annotation dataset as the manuscript. Timing is conducted on the same computer with an I7-7700K CPU and a GTX 1080 Ti GPU.

Model	Type	Precision (%)	Recall (%)	F_1 score (%)	Time of prediction for 500 paragraphs (s)
SMR	Materials	94.6	95.3	95.0	105.3
	Precursors	88.9	91.2	90.0	
	Targets	85.9	83.4	84.5	
Fine-tuned BERT	Materials	91.9	94.9	93.4	551.7
	Precursors	89.3	91.9	90.5	
	Targets	82.6	86.5	84.4	

2.3 Full text-mining pipeline

To extract "codified recipes" for inorganic synthesis from scientific literature, a range of NLP problems need to be addressed besides MER. Here, we define a recipe to be any structured information about a target material, including the precursors, operations, conditions, and other experimental details. Our full text-mining pipeline (Figure 2.4) breaks down into the following steps: (i) acquisition of documents and conversion from markup languages into plain text; (ii) text pre-processing, i.e. segmentation into sentences and tokens, text normalization and morphological parsing; (iii) text representation modeling and paragraph classification; (iv) information retrieval for various synthesis variables. The resulting collection of synthesis recipes provides a source of data for further scientific analysis and machine learning.

Document acquisition

A total of 4,973,165 papers were scraped from main publishers including Elsevier, Wiley, Springer, the American Chemical Society, the Electrochemical Society, the Royal Society of Chemistry, the American Institute of Physics, and the American Physical Society. For all these publishers, we have received permission to download large amounts of web content in an appropriate manner. We conducted a preliminary screening by manually identifying all journals related to materials science that each publisher offers for download. A web-scraping engine, Borges [81], was built using the *scrapy* [82] toolkit. Since the full-text articles published before the 1990s are mostly in PDF format, which complicates their parsing, we chose to process only papers in HTML/XML format published after the 1990s. The downloaded content includes the text of the article as well as its metadata such as journal name, article

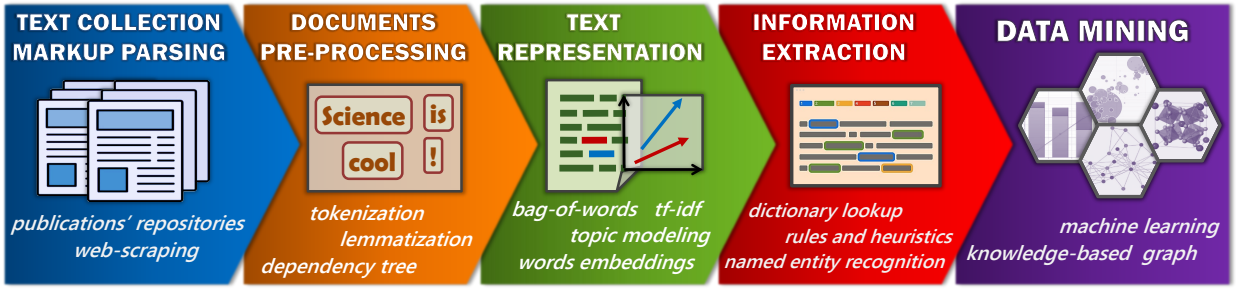

Figure 2.4: Schematic representation of the text mining pipeline for information extraction from the scientific publications.

Table 2.6: Number of journals and papers for each publisher in the database of downloaded articles.

Publisher	Number of journals	Number of papers
Elsevier	2,738	3,435,507
Wiley	1,398	298,088
Royal Society of Chemistry	57	281,168
Nature Publishing Group	26	247,806
American Institute of Physics	9	211,462
Springer	1479	193,839
American Chemical Society	23	140,722
American Physical Society	5	131,676
Electrochemical Society	9	32,897

title, article abstract, authors, and so forth. All data was stored in a document-oriented database implemented using a MongoDB [83] database instance. Because downloaded articles contain irrelevant markups, we developed a customized library, LimeSoup [84], for parsing article markup strings into text paragraphs while keeping the structure of paper and section headings. The number of papers from each publisher is shown in Table 2.6. With the highest number of journals related to materials science, Elsevier was the primary source for ∼70% of the papers in our database of downloaded articles. However, publishers with fewer journals, such as the Royal Society of Chemistry, also contributed a considerable amount of papers. It is important to collect papers from each of these major publishers.

Paragraph segmentation and sentence tokenization

After markup parsing, a downloaded article is converted from its original HTML/XML file into many paragraphs of plain text that can be easily read by humans. However, the computer does not understand the internal structure (e.g. words, sentences) because text is a monotonic sequence of bits representing each character to the computer. Paragraph segmentation and sentence tokenization are required to identify the boundaries of sentences and tokens in a paragraph.

In general, identifying the beginning and end of a sentence segment requires recognition of specific symbolic markers, such as period (“.”), question mark (“?”), and exclamation mark (“!”). The challenge for scientific text lies in the combination of these markers with other meaningful notations. Frequently used expressions, such as “Fig. X”, “et al.,” and periods in chemical formulas often lead to over-segmentation of a paragraph. On the other hand, citation numbers placed at the end of a sentence promote the merging of two sentences. There is no universally accepted solution to this issue, and it typically involves implementing a set of rules tailored to address specific cases [65].

Sentence tokenization, or the process of splitting a sentence into its logical constituents, is vital for further information extraction, as errors generated at this stage tend to propagate throughout the pipeline (Figure 2.4) and negatively impact the accuracy of final results. Extensive research has been conducted on tokenization for general-purpose text, leading to the development of various advanced methods and techniques [85]. However, accurate tokenization in the fields of chemistry and materials science necessitates significant workarounds and revisions to standard approaches. Table 2.7 showcases typical examples of sentence tokenization executed by general-purpose tokenizers, such as NLTK [62] and SpaCy [63]. Similar to sentence segmentation, the primary source of errors stems from the arbitrary usage of punctuation symbols within chemical formulas and other domain-specific terms. The chemical NLP toolkits, such as OSCAR4 [64], ChemicalTagger [86], and ChemDataExtractor [60], address this issue by introducing their own rule- and dictionary-based approaches to solve the over-tokenization problem. In our study, we used ChemDataExtractor for both paragraph segmentation and sentence tokenization.

Classification of synthesis paragraphs

Out of the nearly five million papers possibly related to materials science, only a small portion is relevant to inorganic synthesis. Before using expensive models for information retrieval, we need to first find paragraphs on inorganic synthesis. In this study, we used a two-step paragraph classification approach [87] which consists of an unsupervised algorithm to cluster common keywords in experimental paragraphs into “topics” and generate a probabilistic topic assignment for each paragraph, followed by a random forest (RF) classifier trained on annotated paragraphs. The outcome of the RF is a classification of the synthesis methodology

Table 2.7: Examples of sentence tokenization using various tokenizers. NLTK [62] and SpaCy [63] serve as general-purpose tokenizing toolkits, while ChemDataExtractor [60], OSCAR4 [64], ChemicalTagger [86] are specifically designed for scientific corpora. Tokens are delineated by "|" symbol.

Reagents (NH4)2HPO4 and Sm2O3 were mixed	
NLTK	Reagents | **(** | **NH4** | **)** | **2HPO4** | and | Sm2O3 | were | mixed
SpaCy	Reagents | **(** | **NH4)2HPO4** | and | Sm2O3 | were | mixed
OSCAR4	Reagents | **(NH4)2HPO4** | and | Sm2O3 | were | mixed
ChemicalTagger	Reagents | **(NH4)2HPO4** | and | Sm2O3 | were | mixed
ChemDataExtractor	Reagents | **(NH4)2HPO4** | and | Sm2O3 | were | mixed
We made Eu2+-doped Ba3Ce(PO4)3 at 1200 °C for 2 h	
NLTK	We | made | **Eu2+-doped** | **Ba3Ce** | **(** | **PO4** | **)** | **3** | at | 1200 | **°C** | for | 2 | h
SpaCy	We | made | **Eu2** | **+** | **-doped** | **Ba3Ce(PO4)3** | at | 1200 | **°** | **C** | for | 2 | h
OSCAR4	We | made | **Eu2+** | **-** | **doped** | **Ba3Ce(PO4)3**| at | 1200 | **°C** | for | 2 | h
ChemicalTagger	We | made | **Eu2+-doped** | **Ba3Ce(PO4)3** | at | 1200 | **°C** | for | 2 | h
ChemDataExtractor	We | made | **Eu2+** | **-** | **doped** | **Ba3Ce(PO4)3** | at | 1200 | **°** | **C** | for | 2 | h
Lead-free a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3 ceramics were investigated	
NLTK	**Lead-free** | **a** | **(** | **Bi0.5Na0.5** | **)** | **TiO3-bBaTiO3-c** | **(** | **Bi0.5K0.5** | **)** | **TiO3** | ceramics | was | investigated
SpaCy	**Lead** | **-** | **free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated
OSCAR4	**Lead** | **-** | **free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated
ChemicalTagger	**Lead-free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated
ChemDataExtractor	**Lead-free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated

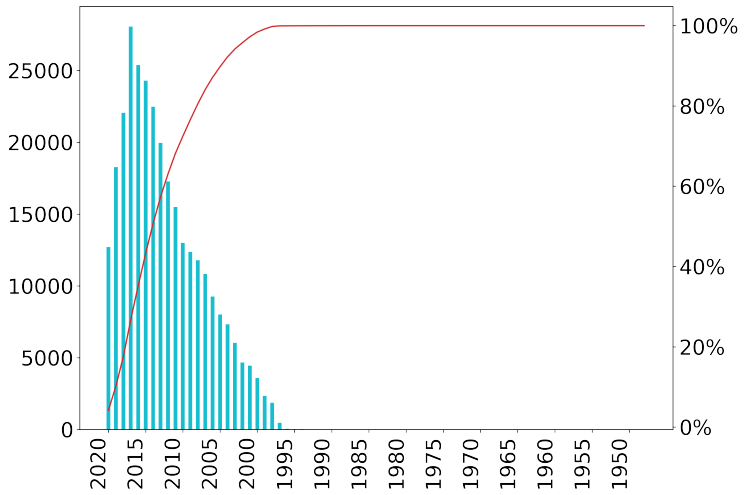

Figure 2.5: Number of papers containing at least one paragraph on inorganic synthesis. The fewer papers from 2018-2020 is mainly because of the termination of subscription with Elsevier by UC [88]. The fewer papers in and before the 1990s is mainly because the HTML/XML format is not available for those papers.

in a paragraph as either solid-state synthesis, sol-gel synthesis, hydrothermal synthesis, co-precipitation synthesis, or "none of the above". As a result, 364,076 paragraphs in the experimental sections, corresponding to 302,000 papers, were found to describe inorganic synthesis, with 89,197 of them corresponding to solid-state synthesis. The number of papers containing at least one paragraph on inorganic synthesis is rapidly increasing with the year as displayed in Figure 2.5. With such a large amount of papers published every year, it is becoming impossible for an individual to track the progress in materials synthesis. An automated data extraction pipeline, such as this study, provides a feasible solution to this challenge.

Retrieval of synthesis information

A typical synthesis procedure in solid-state chemistry literature contains information regarding precursor and target materials, synthesis operations, and associated conditions. Collectively, these elements constitute a materials synthesis "recipe" and can be extracted from a synthesis paragraph. As detailed in Section 2.2, we used the SMR model, a two-step model based on Bi-LSTM neural network, to identify precursor and target materials. Each material entity was processed utilizing a material parser [89], transforming the string representation of the material into a chemical formula, and subsequently splitting it into constituent elements and stoichiometries. Balanced reactions were derived from parsed precursors and target materials by solving a system of linear equations. The variables of these linear equations

represent molar quantities of materials involved in a reaction, with each equation asserting the conservation of a specific chemical element within the reaction. Similarly, we used another Bi-LSTM model [90] to identify the type of synthesis operations, such as `MIXING`, `HEATING`, `SHAPING`, and `COOLING`. For every operation of type `HEATING`, we extracted the values or range of values for time, temperature, and atmosphere corresponding to the operation, if they are mentioned in the same sentence. We applied a regular expression approach to find the values of temperature and time, and a keyword search to find atmosphere. For every operation of type `MIXING`, we extracted corresponding mixing media and type of mixing device, if they are mentioned in the same sentence. We used the list of materials labeled by the SMR model, as well as keyword matching, to find potential devices or media substances. Finally, the precursor and target materials, operations, and conditions were compiled together with the balanced reaction to generate the synthesis dataset.

2.4 Dataset of structured synthesis recipes

Starting from 4,973,165 materials science papers, we applied our text-mining pipeline (Section 2.3) and successfully extracted 33,343 inorganic solid-state synthesis recipes. To our best knowledge, this is the largest dataset publicly available for inorganic synthesis.

Each recipe record corresponds to an individual chemical reaction derived from a paragraph describing inorganic material synthesis, and is represented as a key-value pairs structure within a top-level list. If a paragraph reports the synthesis of multiple materials or a material with variable substituted elements, the corresponding reactions are divided into separate data records. In addition to a balanced chemical equation, the metadata for each reaction includes: the DOI of the paper from which the reaction is extracted, a snippet (the 50 first and 50 last characters to facilitate lookup) of the corresponding synthesis paragraph, chemical information about target and precursor materials involved in the reaction, operations and conditions for heating and mixing steps to synthesize the target material. The data format specifics are provided in Table 2.8.

The chemical equation for the reaction is stored as a string and a list of pairs, including chemical substance (`material`) and stoichiometric coefficient (`amount`). Reactants and products are specified in `left_side` and `right_side`, respectively. In cases where the original paper presents the target compound with variable substituted elements, the specific element used in the particular reaction is provided in `element_substitution`.

The metadata for target and precursor materials utilized in constructing and balancing the chemical equation is represented by a data structure exhibiting the following properties:

- `material_string`: string of the material as given in the original paragraph before being parsed into its chemical composition.

- `material_formula`: the chemical formula corresponding to the material (either provided originally or derived empirically by the parser).
- `composition`: the chemical composition of the material derived from its formula. Aside from single compound materials, a considerable portion of the materials (predominantly target materials) are found to be composites, mixtures, solid solutions or alloys, presented as a sequence of ratio-compound pairs. Consequently, a chemical composition entity is denoted by a list of dictionaries, in which each item corresponds to a compound identified within the material's formula. The ratio of each compound within the material is specified as `amount`, while its chemical composition (i.e. element and fraction) is provided in `elements`. In instances where a material is one compound, the list contains a single item with `amount`=1.0. For hydrate materials, the water is incorporated into the `composition` list, with the `amount` equal to the number of water molecules (if specified).
- `additives`: a list of additive elements, such as those employed for doping, stabilization, and substitution, as resolved from the material string.
- `elements_vars`: lists all variable elements and their associated values identified within the materials.
- `amounts_vars`: lists all variable element ratios and their associated values identified within the material formula. The values for each variable are provided as a structure with `values` listing specific variable's values, and `max_value`/`min_value` values if a range is presented in the paragraph.
- `oxygen_deficiency`: a yes/no attribute indicating whether the material was synthesized with unspecified oxygen stoichiometry.
- `mp_id`: the ID of the lowest-energy polymorph entry in the Materials Project database [2], if specified.

To streamline querying of the dataset, the `targets_string` field contains all target material formulas obtained by substituting `amounts_vars` in the `material_formula`.

The sequence of synthesis steps for the reaction, if detailed in the paragraph, is listed as a data structure with the following fields: the original text token (`token`), its type (`type`) in terms of operation, and conditions employed at this stage (`conditions`). If the synthesis step is classified as `HEATING`, the temperature, time and atmosphere conditions are included in the `conditions` attribute. Temperature and time are specified as `values` if discrete values are available, or `max_value`/`min_value` if a range is provided. If the synthesis step is of the `MIXING` type, the mixing device and mixing media are included in the `conditions` attribute.

Table 2.8: Format of each data record: description, key label, and data type. [a]{`amount`: *float*, `material`: *string*}. [b]{`formula`: *string*, `elements`: {`element`: amount of element}, `amount`: *string*}. [c]{`max_value`: *float*, `min_value`: *float*, `values`: *list* of *floats*}. [d]{`max_value`: *float*, `min_value`: *float*, `values`: *list* of *floats*, `units`: *string*}.

Data description	**Data Key Label**	**Data Type**
DOI of original paper	`doi`	*string*
Snippet of raw text	`paragraph_string`	*string*
Chemical equation	`reaction`	Object (*dict*): - `element_substitution`: *dict* - `left_side`: *list* of Objects[1] - `right_side`: *list* of Objects[1]
Chemical equation in string format	`reaction_string`	*string*
Target material data	`target`	Object (*dict*): - `material_string`: *string*, - `material_formula`: *string*, - `composition`: *list* of Objects[2] - `additives`: *list* of *strings* - `elements_vars`: {`var`: *list* of *strings*} - `amounts_vars`: {`var`: *list* of Objects[3] } - `oxygen_deficiency`: *boolean* - `mp_id`: *string*
List of target formulas obtained after variables substitution	`targets_string`	*list* of *strings*
Precursor materials data	`precursors`	*list* of Objects (See `target`)
Sequence of synthesis steps and corresponding conditions	`operations`	*list* of Objects (*dict*): - `token`: *string*, - `type`: *string* - `conditions`: Object - - `heating_temperature`: *list* of Objects[4] - - `heating_time`: *list* of Objects[4] - - `heating_atmosphere`: *list* of *strings* - - `mixing_device`: *list* of *strings* - - `mixing_media`: *list* of *strings*

Table 2.9: Performance of data extraction for dataset entries.

Data attribute	**Precision**	**Recall**	**F_1 score**
Materials			
- targets	0.97	/	/
- precursors	0.99	0.99	0.99
Operations	0.86	0.95	0.90
Heating conditions			
- temperature	0.85	0.87	0.86
- time	0.90	0.88	0.89
- atmosphere	0.89	0.86	0.87
Mixing conditions			
- mixing media	0.62	0.66	0.64
- mixing device	0.82	0.55	0.66
Balanced reactions	0.95	/	/

To assess the quality of the extracted dataset, we randomly selected 100 data entries and manually verified each extracted field against the original paragraph. The calculated precision, recall, and F_1 score for every attribute of the data entry are provided in Table 2.9. Overall, we achieved high accuracy in extracting targets (precision 97%), precursors (F_1 score 99%), operations (F_1 score 90%), and balanced reactions (precision 95 %). The high accuracy for identifying targets and precursors is attributed to the additional constraints imposed by constructing balanced chemical equations, which helps to reduce potential errors caused by composition parsing. The lower accuracy of heating conditions (F_1 score $< 90\%$) primarily results from cases where the operations extraction algorithm misses the heating step. The retrieval of mixing conditions exhibits relatively poor accuracy with an F_1 score of 65%, largely due to misidentification by MER of the device material or media substance used for mixing and the fact that those conditions are frequently not mentioned in the same sentence as the mixing procedure.

This analysis leads us to conclude that at the chemistry level (correct precursors, targets, reactions), the accuracy of the dataset is 93%. When including all operations and their conditions, the accuracy of correctly extracting and assigning all recipe components (chemistry, operations and attributes of the operations) is 51%, which is low due to poor performance in extracting mixing attributes. For many solid-state recipes, the specifics of mixing precursors are of lesser importance, rendering this extraction failure less critical. When considering only the correctness of the recipe without conditions for heating and mixing (i.e. chemistry, operations, and reactions), the accuracy increases to 64%.

To give a sense of the typical materials being extracted, we list the ten most frequent targets (Table 2.10), precursors (Table 2.11) and reactions (Table 2.12) in the dataset. The target compounds effectively capture the types of materials most frequently investigated via solid-state synthesis over the past two decades. These are lithium-ion battery cathode materials such as $LiFePO_4$, $LiMn_2O_4$, and $LiNi_{0.5}Mn_{1.5}O_4$, in addition to perovskites employed in multiferroics, LEDs, and CMOS applications, such as $BaTiO_3$, $BiFeO_3$, and $SrTiO_3$. The ranking of the most frequent targets is not exactly the same as that of the reactions because the synthesis complexity differs from one target to another. For example, various recipes have been attempted to lower the synthesis cost of the cathode material $LiFePO_4$ because of its high commercial value. $LiFePO_4$ can be synthesized from different Li sources, such as Li_2CO_3 and LiOH, and different Fe sources, such as Fe_2O_3 or FeC_2O_4. While $LiFePO_4$ is the most frequent target in the dataset, the frequency of each specific reaction leading to $LiFePO_4$ is smaller than that of the top ten most frequent reactions. In contrast, the perovskite $BaTiO_3$ is synthesized from $BaCO_3$ and TiO_2 most of the time, "$BaCO_3$ + TiO_2 = $BaTiO_3$ + CO_2" is also the most frequent reaction in the dataset.

Table 2.10: Ten most common targets present in the dataset.

Rank	**Target**	**Rank**	**Target**
1	$LiFePO_4$	6	$SrTiO_3$
2	$LiMn_2O_4$	7	$Li_4Ti_5O_{12}$
3	$BaTiO_3$	8	$Y_3Al_5O_{12}$
4	$BiFeO_3$	9	$CaTiO_3$
5	$CaCu_3Ti_4O_{12}$	10	$LiNi_{0.5}Mn_{1.5}O_4$

Table 2.11: Ten most common precursors present in the dataset.

Rank	**Precursor**	**Rank**	**Precursor**
1	TiO_2	6	Bi_2O_3
2	$SrCO_3$	7	Fe_2O_3
3	$BaCO_3$	8	Nb_2O_5
4	La_2O_3	9	Li_2CO_3
5	$CaCO_3$	10	Na_2CO_3

Table 2.12: Ten most common reactions present in the dataset.

Rank	Reaction
1	$BaCO_3 + TiO_2 = BaTiO_3 + CO_2$
2	$3CuO + 4TiO_2 + CaCO_3 = CaCu_3Ti_4O_{12} + CO_2$
3	$0.5Bi_2O_3 + 0.5Fe_2O_3 = BiFeO_3$
4	$SrCO_3 + TiO_2 = SrTiO_3 + CO_2$
5	$2Li_2CO_3 + 5TiO_2 = Li_4Ti_5O_{12} + 2CO_2$
6	$TiO_2 + CaCO_3 = CaTiO_3 + CO_2$
7	$Nb_2O_5 + ZnO = ZnNb_2O_6$
8	$6Fe_2O_3 + BaCO_3 = BaFe_{12}O_{19} + CO_2$
9	$Li_2CO_3 + TiO_2 = Li_2TiO_3 + CO_2$
10	$0.5Li_2CO_3 + 0.333Co_3O_4 + 0.083O_2 = LiCoO_2 + 0.5CO_2$

2.5 Exploratory data analysis

The large-scale dataset presents significant opportunities for improving our understanding of solid-state synthesis. In this section, we exemplify the typical ways to use this text-mined dataset through exploratory data analysis in different directions, including coverage of chemical space, synthesis temperature, synthesis routes, synthesis time, reaction energy, and the application to a specific chemical system. Subsequent studies may wish to expand on these analyses to explore synthesis design more comprehensively.

Coverage of chemical space

Although over 100 elements exist in the universe, they are not used evenly due to disparities in availability and variations in chemical properties. In this study, we assessed the chemical space covered by the text-mined dataset. For each chemical element, we calculated the number of reactions involving the given element in the target materials. The results are illustrated in Figure 2.6 using a yellow-to-green gradient frame at the top of each element box. The database features a predominance of target materials containing Ti, Sr, Ba, La, and Fe, with over 3,000 reactions involving these elements. This observation is also corroborated by the ten most frequent target materials listed in Table 2.10. Subsequently, the next-most common targets are materials containing Li, Ca, Nb, Mn, and Bi, with these elements involved in 2,000 $\sim$ 3,000 reactions. Conversely, Au, Pt, Os, and Be are the least common elements, appearing in fewer than 13 reactions within the dataset. Notably, rare

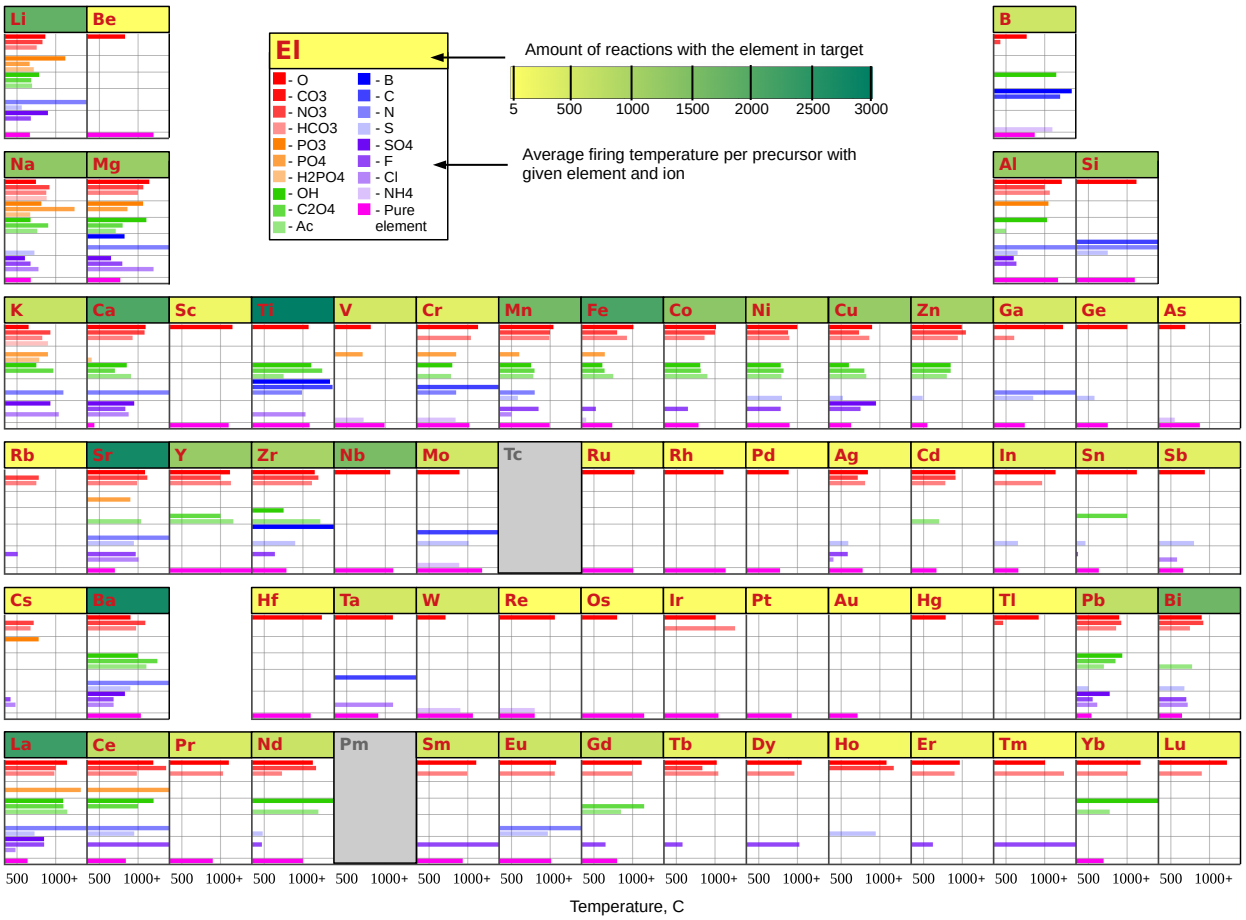

Figure 2.6: Map of chemical space covered by the dataset and average firing temperature for different precursors. For each element, a yellow-to-green gradient frame signifies the cumulative number of reactions yielding target materials containing the element. The bar graph beneath each element illustrates the list of ions paired with the element in precursor compounds, with the length of the bar corresponding to the firing temperature averaged across all reactions employing the given precursor (i.e. element+counter-ion). Elements present in five or fewer targets are displayed in grey. The notation "Ac" represents the acetate anion CH_3COO^- within the compound formula.

and radioactive elements such as francium, radium, technetium, and promethium are absent from the target materials in the dataset.

Synthesis temperature

In addition to exploring the coverage of chemical space, we investigated the co-occurrence of chemical elements and the most prevalent counter-ions in precursor materials, and calculated the average firing temperature associated with each precursor. The firing temperature is operationally defined as the temperature used during the final heating stage in the synthesis process sequence. Figure 2.6 presents the findings in the form of bar graphs for each element, where the bar color represents a particular counter-ion and the pure element as a precursor is depicted in magenta. The bar length signifies the average firing temperature.

In solid-state synthesis, the counter-ion influences the melting or decomposition temperature of the precursor and may determine when the precursor becomes active during synthesis. The distribution of firing temperatures in Figure 2.6 aligns closely with this statement and demonstrates how different precursors are employed in various temperature regimes during solid-state synthesis. For instance, the blue bars generally exhibit greater length (higher average temperature) compared to the red ones, because transition metal borides, carbides and nitrides often require higher reaction temperatures than their corresponding oxides, owing to the refractory nature of their precursors. Conversely, the green bars are comparatively shorter (lower average firing temperature) than the red ones, because, compared to oxides and complex oxide anions (carbonates, phosphates, etc.), synthesis with hydroxides, oxalates, and acetates facilitates lower temperature reactions because of their reduced melting points. This data-driven temperature analysis is based on the precursor, and we acknowledge that reaction temperatures also rely on the thermal stability and reactivity of the target materials. Nevertheless, Figure 2.6 offers a semi-quantitative starting point for researchers: if a target material decomposes at a relatively low temperature, it may be advantageous to select a precursor that tends to become active at lower temperatures.

Synthesis routes

To illustrate the variety of synthesis routes present in the dataset, we organized the sequence of synthesis steps based on the following pre-defined patterns (refer to the table in Figure 2.7):

- *One-step synthesis* involves solely solid mixing/grinding operations and a maximum of one heating step (final firing) without regrinding.
- *Synthesis with grinding in a liquid medium* homogenizes (without dissolution) the starting materials in any liquid medium.
- *Solution-based synthesis* includes any form of dissolution of starting materials in a solvent.

- *Synthesis with intermediate heat* incorporates one or more heating steps (excluding drying after mixing with the liquid part) prior to the final firing of the materials.

First, we found that various synthesis types are almost evenly represented in the database (refer to the top pie-chart in Figure 2.7): 26% of materials are synthesized in one step, 25% of the syntheses routes involve intermediate heating step(s) before the final firing, 21% of the syntheses incorporate grinding (homogenizing) in liquid, and 14% require dissolution of precursors in a solvent. The remainder of the recipes (14%) either lack a detailed synthesis procedure (6%), or the pathway is more intricate (8%).

As the selection of the counter-ion in a precursor frequently relies on the synthesis method, we investigated the prevalent synthesis type for a particular ion in a precursor. We examined a subset of reactions containing the given counter-ion in a precursor compound and determined the proportion of each synthesis type within this subset. The resulting pie-charts are shown in Figure 2.7. The observed pattern aligns with established aspects of solid-state synthesis. For instance, the solution-based synthesis (orange fraction) often employs soluble precursors with nitrates, acetates, and organic (CH-containing) anion groups. Certain counter-ions are more amenable to one-step synthesis than others. For example, chlorides, sulfides, and hydrides do not require much additional processing. Conversely, relatively stable precursors such as oxides and carbonates undergo various processing methods, frequently requiring intermediate heating and grinding. This is likely because of the common formation of reaction impurities and non-equilibrium intermediates during reaction sequences.

Synthesis time

To investigate the typical synthesis time in reported experiments, we show the histogram of heating time for all reactions in the text-mined dataset in Figure 2.8. We observed that most successful syntheses are conducted within 24 h. It is worth noting that heating time is usually reported as an upper bound rather than the actual reaction duration. Several spikes (red color) at 12, 24, 48, and 72 h are present in Figure 2.8, because time is frequently determined in an arbitrary way. In these cases, the actual reaction duration should be even shorter. Synthesis experiments lasting more than three days are rare. That is mainly because of the practical factor: shorter turnaround times are necessary to accommodate more experimental attempts given the limited access to lab equipment. The limited time for synthesis imposes an additional constraint on the synthesis process. In addition to thermodynamics, practical synthesis requires a sufficient rate of chemical kinetics to form the target material in a matter of days. The availability of data over time can assist in the study of chemical kinetics involved in the solid-state synthesis process.

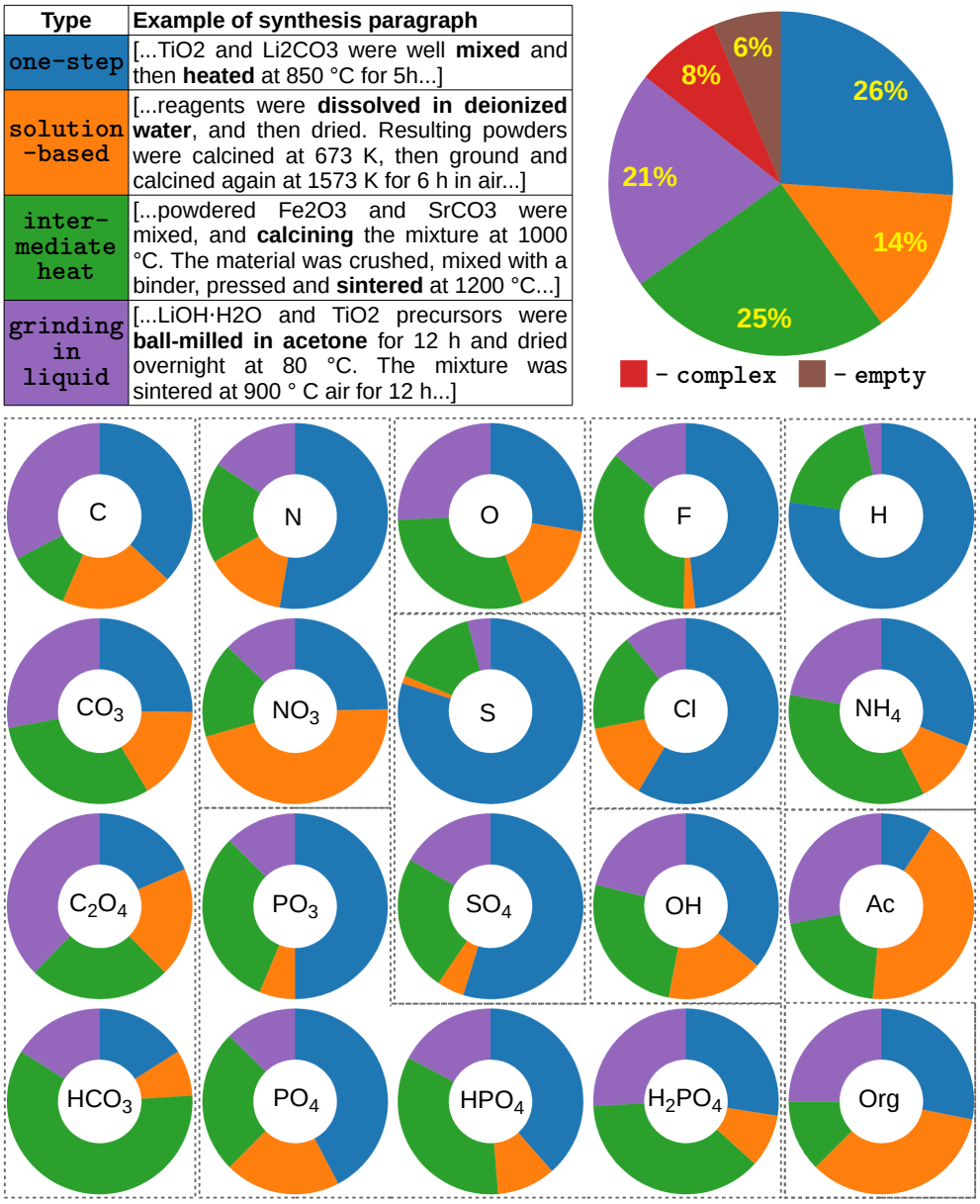

Figure 2.7: Association between the choice of synthesis route and precursors counter-ions. The top table exemplifies the four defined synthesis types: one-step synthesis, solution-based, synthesis with intermediate heating steps, and synthesis involving grinding of precursors in liquid media. The pie charts on the right illustrate the proportion of each synthesis route within the dataset. The donut-like charts depict the fractions of the four synthesis routes (outlined in the table) for each counter-ion used in precursors. "Ac" denotes the acetate anion CH_3COO^- in the compound formula, while "Org" represents the organic anion group (-CH-) in the compound formula.

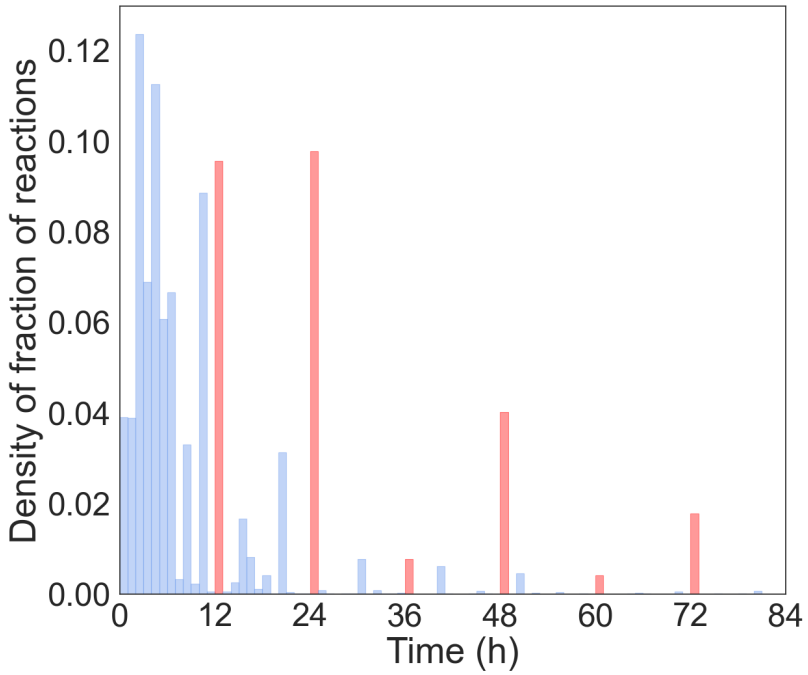

Figure 2.8: Distribution of reported heating time in the dataset. The longest time is adopted when multiple values are reported in the same paper.

Energetic analysis

The energy of a material is one of the most important material properties because it enables further analysis of thermodynamics and kinetics. However, the energy is usually not present in the text description related to synthesis experiments in a research paper. To address this situation, we assume that the composition in the text corresponds to the lowest-energy structure with the identical chemical composition in the Materials Project database [2]. For the materials not found in the Materials Project database, their energies are interpolated using a linear combination of known compounds. Since only energy at 0 K is available from the Materials Project database, the finite-temperature Gibbs free energy is estimated with the method by Bartel et al. [91].

With the capability to estimate the finite-temperature Gibbs free energy of materials in our dataset, we investigated the typical distribution of reaction Gibbs energy in solid-state synthesis (Figure 2.9). Specifically, for each reaction, we calculated the difference of Gibbs energy normalized by the number of atoms per target formula at 1000 °C between products and reactants. For gas phase species such as O_2 and CO_2, the temperature-dependent enthalpy and entropy were from FREED [92] and NIST [93] experimental databases. For materials with CO_3^{2-} anions, an empirical correction of -1.2485 eV/CO_3^{2-} from fitting exper-

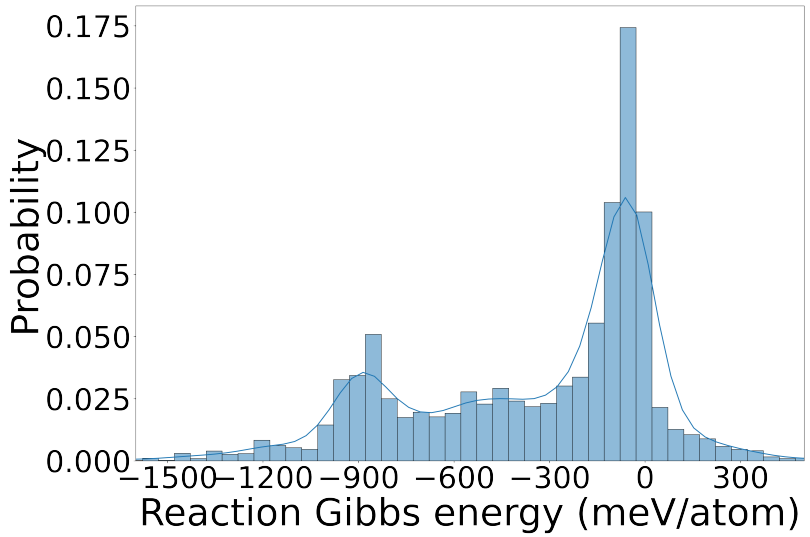

Figure 2.9: Distribution of reaction Gibbs energy in the dataset.

imental enthalpies for common carbonates was applied [94]. The reaction Gibbs energy is negative for most of the reactions, agreeing with the known thermodynamic rule that the driving force of a synthesis reaction comes from the lower energy of the product side compared to the reactant side. It is also notable that a small portion (∼12%) of reactions exhibit a positive Gibbs energy change from the reactants to the products, possibly because of the uncertainties arising from the series of assumptions made during the estimation of material energetics. Despite the uncertainties, estimating material energetics can still be potentially useful in capturing trends when analyzing synthesis thermodynamics and kinetics.

Application to a specific chemical system

The text-mined dataset is particularly useful for a rapid literature review of different synthesis procedures within a single chemical space. Figure 2.10 displays a birds-eye perspective of the various solid-state synthesis routes to target materials in the Li-Mn-O (LMO) chemical system using the dataset. By querying the dataset, a researcher not familiar with the syntheses in the LMO system can quickly acquire the pertinent knowledge of synthesizing LMO materials.

For example, the number of reaction records for each material is shown by the circle size and color in Figure 2.10. The material with the most number of studies in the LMO system is $LiMn_2O_4$, which is related to its potential applicability in batteries. Besides $LiMn_2O_4$ itself, many materials with close stoichiometry in the $Li_{1+x}Mn_{2-x}O_4$ family, such as $Li_4Mn_5O_{12}$ and $Li_5Mn_7O_{16}$, are also synthesizable. The synthesis temperature for Li_2MnO_3

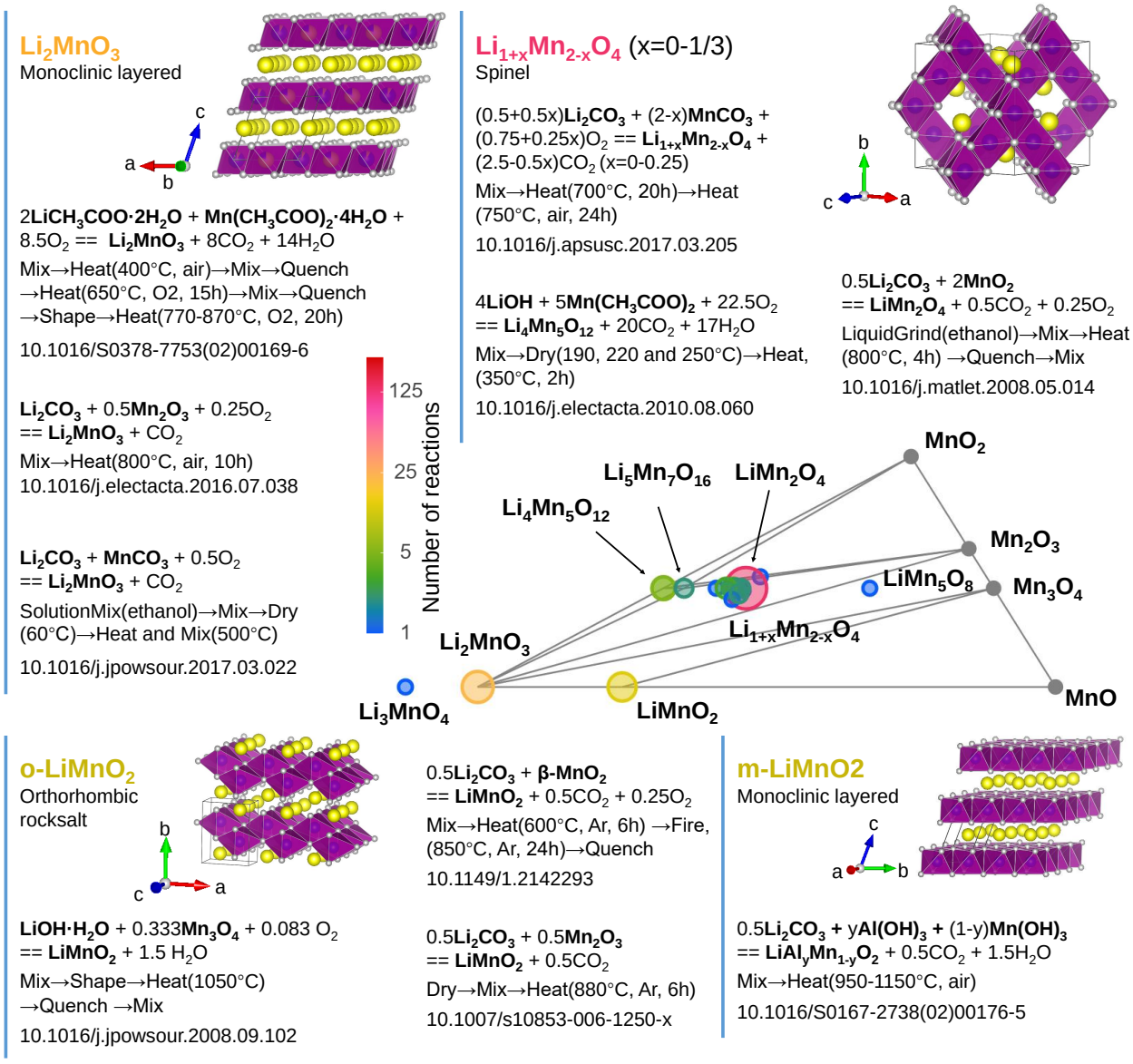

Figure 2.10: Graphical representation of dataset entries queried for the Li-Mn-O system. Examples of the subset entries: target LMO material, synthesis reaction and route. The DOIs are provided for reference. The triangle shows the distribution of the LMO materials on the phase diagram. The circles' size and color are scaled according to the number of reactions in the dataset with the given target material.

and $Li_{1+x}Mn_{2-x}O_4$ is typically lower than 900 °C, while the synthesis of $LiMnO_2$ polymorphs requires either higher heating temperature [95, 96] or reducing atmosphere [97, 98]. While these details are also available by manually reading the literature, the text-mined dataset saves time for a researcher of reading and summarizing hundreds of papers.

2.6 Conclusion

We have developed a fully automated text-mining pipeline for inorganic materials synthesis science. Starting from 4,973,165 materials science papers, we applied our text-mining pipeline and successfully extracted 33,343 solid-state synthesis recipes. Each recipe is the structured data of a synthesis procedure, including the target material, the precursors, the chemical reaction, the experimental operations, and the conditions associated with `HEATING` and `MIXING` operations. Detailed discussions on extracting precursor and target materials demonstrate the challenges and solutions for information retrieval in the field of material science. The usage of the text-mined synthesis dataset is exemplified by exploratory data analysis with various aspects, including coverage of chemical space, synthesis temperature, synthesis routes, synthesis time, reaction energy, and the application to a specific chemical system.

Chapter 3

Similarity of precursor materials for alternative recipes

In order to understand and eventually predict solid-state synthesis recipes, one of the important questions is how to select precursors [5–7]. Knowledge of which precursors to use is often achieved by an individual's experience. In this chapter[1], we present a data-driven approach to assess the similarities and differences between precursors in solid-state synthesis by conducting a meta-analysis with the extracted data. The similarity could help guide the selection of precursors when researchers alter existing recipes by replacing precursors.

We first present the variety of the extracted precursors. Quantitative analysis of this large-scale dataset indicates that the most common precursors for each element are usually the oxides, carbonates, or hydroxides stable at ambient environment. An intriguing question is how frequently researchers substitute one precursor with another while retaining the target, which sheds light on how similarly these precursors behave in a solid-state reaction. We utilized a substitution model based on the work of Hautier et al. [100] and Yang et al. [101] to quantify the probability that two precursors are interchangeable. Combining the substitution probability and the distribution of synthesis temperatures, we define a multi-feature distance metric to characterize the similarity of precursors. A hierarchical clustering of precursors based on this metric demonstrates that the "chemical similarity" can be extracted from text data, without the need to include any explicit domain knowledge. The quantitative similarity metric offers a reference to rank precursor candidates and constitutes an important

[1] This Chapter incorporates sections from a previously published paper and one in press, with permission from the authors: (1) Tanjin He, Wenhao Sun, Haoyan Huo, Olga Kononova, Ziqin Rong, Vahe Tshitoyan, Tiago Botari, and Gerbrand Ceder. "Similarity of precursors in solid-state synthesis as text-mined from scientific literature." *Chemistry of Materials* 32, no. 18 (2020): 7861-7873 [29]; and (2) Tanjin He, Haoyan Huo, Christopher J. Bartel, Zheren Wang, Kevin Cruse, and Gerbrand Ceder. "Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature." *Science Advances* (2023), in press [99].

step toward developing a predictive synthesis model [29].

3.1 Common and uncommon precursors

The solid-state synthesis dataset extracted from the literature (Chapter 2) contains 71 different metal/metalloid elements and 1,619 distinct precursors. Some precursors are rarely used. Restricting the statistics to precursors used at least 30 times, there are 58 metal/metalloid elements and 182 precursors.

To visualize the variety of precursors, the precursors for each metal/metalloid element are categorized by the anion (group) class and counted by the number of corresponding reactions in which they are used. The frequency of each anion class normalized by the total number of reactions for an element is shown in Figure 3.1. One precursor is usually used much more frequently than other precursors for the same element, which we denote as the *common precursor*. Figure 3.1 shows that for alkali and alkaline earth elements, the common precursors are carbonates, except for MgO which is the typical source for Mg. For transition metals and other main group elements, the common precursors are oxides except for $B(OH)_3$ for B. In general, the common precursor tends to be the compound that is stable under ambient conditions, which is beneficial to the purity and accurate weighting in experiments [102]. Our observation on the common precursors suggests that laboratory chemists will prioritize shelf stability of precursors; although we note that more reactive precursors can help to facilitate synthesis reactions.

However, the predominant use of the common precursor does not mean precursor selection is a trivial problem. In our text-mined synthesis dataset, we find that approximately half of the target materials were synthesized using at least one uncommon precursor. Figure 3.2 presents the fraction of targets in the text-mined dataset that can be achieved as one increases the number of available precursors. The precursors on the x-axis are ordered by the relative frequency with which they are used to bring a specific element into a synthesis target. Sometimes, the decision to use an uncommon precursor is motivated by an interesting advantage for a specific nontraditional precursor. For example, in some cases precursors can function as morphology templates; Zhao et al. reported that γ-MnOOH nanorods were used to obtain $LiMn_2O_4$ nanorods, whereas $LiMn_2O_4$ from electrolytic MnO_2 (EMD) only consisted of many irregular and aggregated particles [103]. The use of a lower-melting-point precursor can result in a target with a smaller particle size; Liu et al. adopted $Sr(NO_3)_2$ instead of $SrCO_3$ to synthesize $SrTiO_3$ nanocrystals [104]. An amorphous precursor can facilitate the reaction process and minimize the possibility of forming chemical segregations; Mercury et al. utilized amorphous $Al(OH)_3$ rather than Al_2O_3 in the synthesis of $Ca_3Al_2O_6$ [105]. In these examples, there were strategically designed precursors in order to achieve a particular synthesis result. Collecting these individual-use cases provides interesting insights into synthesis design.

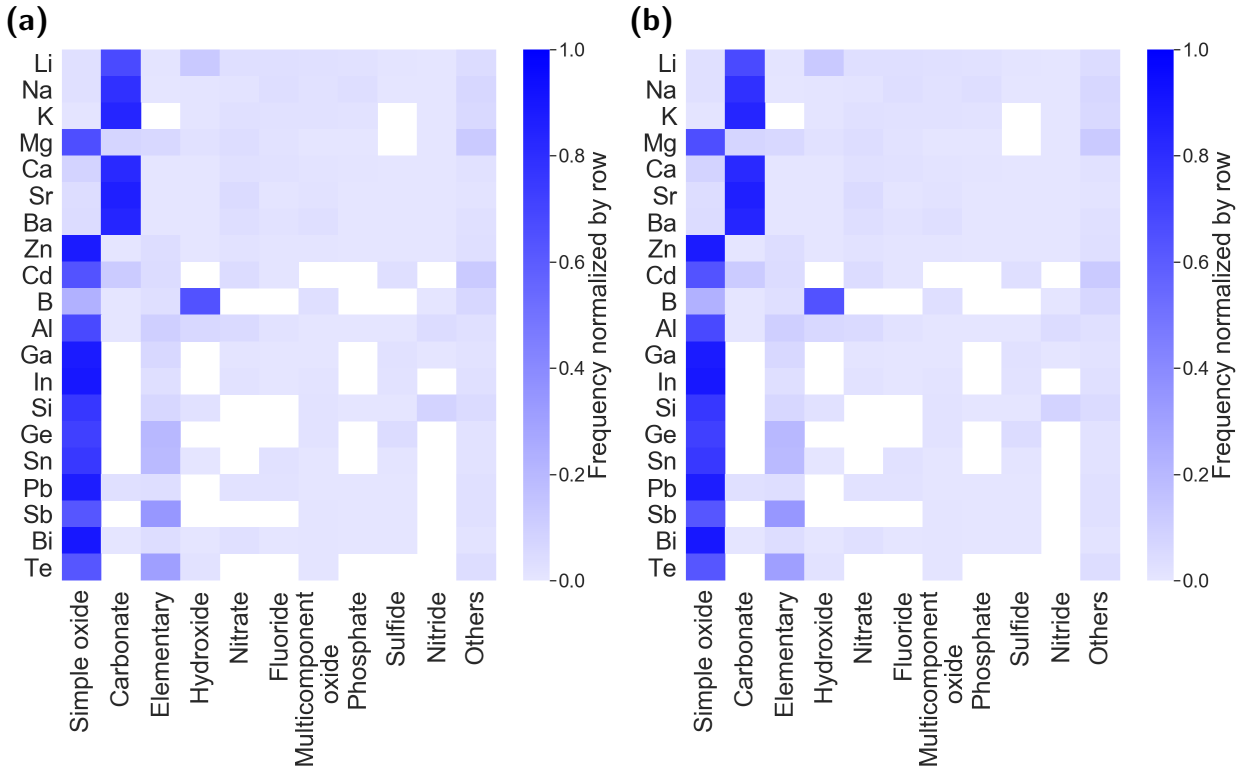

Figure 3.1: Fraction of different classes of precursors corresponding to each element: **(a)** main group elements and **(b)** transition metal elements.

3.2 Substitution model for precursors

The large number of reactions we obtained gives us the opportunity to understand to what extent precursors are interchangeable. To measure the probability that one precursor can be substituted by another while retaining the target, we utilized a substitution model similar to the one developed by Hautier et al. [100] and used by Yang et al. [101] for structure prediction. For each pair of precursors, the model counts the number of occurrences where the same targets can be synthesized from either of the precursors. The more frequently the two precursors are interchanged, the more similar they are.

In the following part, we define the substitution model in a mathematical form, and express the probability of finding a substitutional precursor pair $P_{sub}(p_i^{j,1}, p_i^{j,2})$ as a sigmoid with unknown parameter $\boldsymbol{\lambda}$. Assuming the independence of substitutions, we deconvolute the probability of finding substitution between two lists of precursors $P_{sub}(R_X, R'_X)$ into the product of $P_{sub}(p_i^{j,1}, p_i^{j,2})$. At last, we maximize $P_{sub}(R_X, R'_X)$ over substitution observations to solve $\boldsymbol{\lambda}$ and use it to calculate substitution probability.

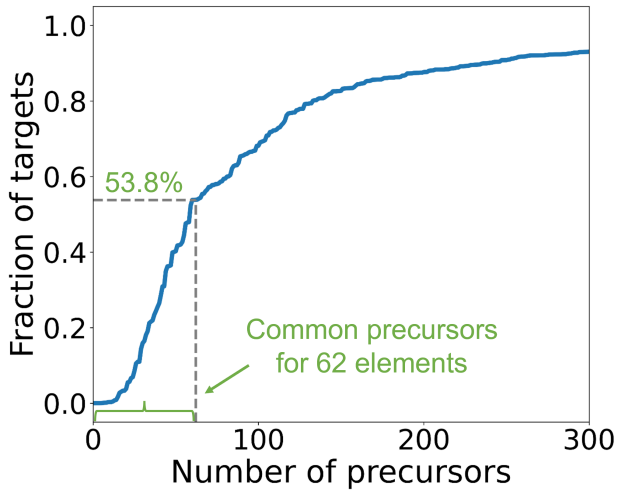

Figure 3.2: Fraction of targets that can be synthesized with limited number of available precursors. The precursors are ordered by relative frequency per metal/metalloid element. Precursors for 62 elements are considered. A new target is included if at least one reported reaction for that target was performed with the available precursors.

First, we define precursor substitution in a mathematical form. Let $E = (e_1, e_2, \cdots, e_n)$ be a pre-defined ordered list of all the metal/metalloid elements given in the periodic table. We assume each precursor contributes one metal/metalloid element to targets. For the target R_{Tar} in a reaction synthesis $R = (R_{Tar}, R_X)$, define the precursor list as $R_X = (p_1, p_2, \cdot, p_n)$, where p_i is the precursor for element e_i present in R_{Tar}; otherwise p_i is null. For a pair of reaction $\{R, R'\}$ if $R_{Tar} = R'_{Tar}$ and $R_X \neq R'_X$, we say precursor substitution occurs. Through iterating over all the possible combinations of any two reactions, we obtain a collection of N reaction pairs where precursor substitution occurs, denoted as the data $D = \{\{R, R'\}^1, \{R, R'\}^2, \cdots, \{R, R'\}^N\}$. Our objective is now to find the values of the pairwise precursor substitutions that maximize the likelihood of D.

Next, we define the potential substitutional precursor pairs. For element e_i, denote the list of candidate precursors as $(p_i^1, p_i^2, \cdots, p_i^{m_i})$, where m_i is the total number of unique precursors. We assume that potentially every precursor $p_i^{\tau_1}$ can be substituted by any other one $p_i^{\tau_2}$, forming a substitutional pair $\{p_i^{\tau_1}, p_i^{\tau_2}\}$ where $1 \leq \tau_1 < \tau_2 \leq m_i$. In total, there can be up to $M_i = \binom{m_i}{2}$ such pairs for element e_i. For simplicity, we assemble all substitutional pairs for all elements into one list and renumber the pairs as $\{p_i^{j,1}, p_i^{j,2}\}$ where $j = 1, \cdots, \sum_{i=1}^{n} |M_i|$. Although the index i is not necessary, we retain it for clarity to distinguish between elements. The probability that the pair $\{p_i^{j,1}, p_i^{j,2}\}$ can be found as a substitution occurs is written as

$$P_{sub}(p_i^{j,1}, p_i^{j,2}) = \text{sigmoid}(\lambda_j), \tag{3.1}$$

where λ_j is a parameter to be optimized. Assuming all substitutional precursor pairs are independent of each other, the probability that the pair of precursor lists $\{R_X, R'_X\}$ can be found as a substitution occurs is

$$P_{sub}(R_X, R'_X) = \frac{e^{\sum_j \lambda_j \mathbf{I}_j(R_X, R'_X)}}{Z}, \tag{3.2}$$

where

$$\mathbf{I}_j(R_X, R'_X) = \begin{cases} 1, & \{R_{X,i}, R'_{X,i}\} = \{p_i^{j,1}, p_i^{j,2}\} \\ 0, & \text{otherwise} \end{cases}, \tag{3.3}$$

and Z is the partition function for normalization, given by

$$Z = \prod_j (1 + e^{\lambda_j}). \tag{3.4}$$

The value of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots)$ is obtained by maximizing the likelihood over the data D:

$$\boldsymbol{\lambda}^* = \text{argmax}_{\boldsymbol{\lambda}} \sum_{t=1}^{N} \log P_{sub}((R_X, R'_X)^t | \boldsymbol{\lambda}) \tag{3.5}$$

For those substitutional pairs not found in D, the value of λ_j will be set to a common low value such that $P_{sub}(p_i^{j,1}, p_i^{j,2})$ in Eq. 3.1 is close to zero.

Finally, we define the substitution probability. Here we discuss one substitutional pair $\{p_i^{j,1}, p_i^{j,2}\}$ and omit the index j for simplicity. For a given reaction using precursor p_i^1, the probability that p_i^1 is substitutable by p_i^2 is

$$P(p_i^2 | p_i^1) = P(p_i^1 \text{ substituted}) \frac{P_{sub}(p_i^1, p_i^2)}{\sum_{k \neq 1} P_{sub}(p_i^1, p_i^k)}, \tag{3.6}$$

where $P(p_i^1$ substituted) is a prior probability of p_i^1 being substitutable and is calculated as the number of reactions with the substituted precursor p_i^1 divided by the total number of all reactions using p_i^1. The fractional part in the right-hand side accounts for the conditional probability that p_i^1 is substitutable by p_i^2 when substitution occurs, which can be calculated with Eq. 3.1. A small fraction of reactions ($\sim 5\%$) which included multiple metal/metalloid elements in the same precursors or used multiple precursors for the same element were not considered in this model.

3.3 Cross-validation of substitution model

We evaluated the predictive power of the substitution model by performing a cross-validation test on the generation of alternative precursor lists. Cross-validation consists in training the model on part of the available data (the training set) and predicting back the remaining data

(the validation set). Given a target R_{Tar} and an existing precursor list R_X in the training set, we can propose an alternative precursor list R'_X to synthesis the same target by replacing the precursors in R_X with different ones. With the substitution probability defined in Eq. 3.6, the conditional probability of R_X being substitutable by R'_X is given by

$$P(R'_X|R_X) = \prod_{p_i^1 \in R_X, p_i^2 \in R'_X, p_i^1 \neq p_i^2} P(p_i^2|p_i^1). \tag{3.7}$$

If $P(R'_X|R_X)$ is higher than a given threshold, the proposed R'_X will be accepted as a positive prediction of an alternative precursor list. Otherwise, R'_X will be rejected as a negative prediction. Applying this procedure on all possible R'_X, we obtain all the positive and negative predictions and compare with the validation set for evaluation. Two-thirds of the reactions were used as the training set and the remaining one-third of the data were used as the validation set. For example, $La_{0.7}Ca_{0.3}MnO_3$ is synthesized from La_2O_3, $CaCO_3$, and MnO_2 [106] in the training set. As a true positive prediction, the substituted precursor list La_2O_3, CaO, and $Mn(Ac)_2$ (Ac stands for acetate anion CH_3COO^-) [107] was also found in the validation set. The true positive rate (TPR) and false positive rate (FPR) were used as metrics to evaluate the performance. The TPR and FPR of the prediction vary with the probability threshold, as shown in Figure 3.3. Overall, the TPR is higher than the FPR, indicating that the substitution model has a predictive power in the selection of alternative precursors and can effectively distinguish between the substitutions leading to existing precursor lists and those leading to nonexistent ones. Higher threshold values lead to fewer false alarms but imply fewer true hits. An adequate threshold can be found by selecting the one resulting in relatively higher TPR and lower FPR.

3.4 Substitution probability

The probability $P(B|A)$ that a precursor A is substituted by another precursor B for the same metal/metalloid element (Eq. 3.6) is displayed as a heatmap in Figure 3.4, where the rows are A and the columns are B. The color represents the probability of substitution defined in Eq. 3.6, as shown by the colorbar. For each element, the precursors are ordered by the number of reactions using it from the most to the least, that is, the first precursor is the common precursor for each element. For the sake of simplicity, we merged the precursor in its hydrated form and its anhydrous form, for example $LiOH{\cdot}H_2O$ and LiOH, based on the assumption that water will evaporate early on during the solid-state heating process. The rows for the common precursors usually display relatively high substitution probability, which implies that many uncommon precursors can be replaced with the common precursors. Note that our analysis only indicates that substitution can lead to the same target compound under similar reaction conditions. The choice of different precursors can still be justified as they might infer different properties on the compound. For example, in the battery chemistry,

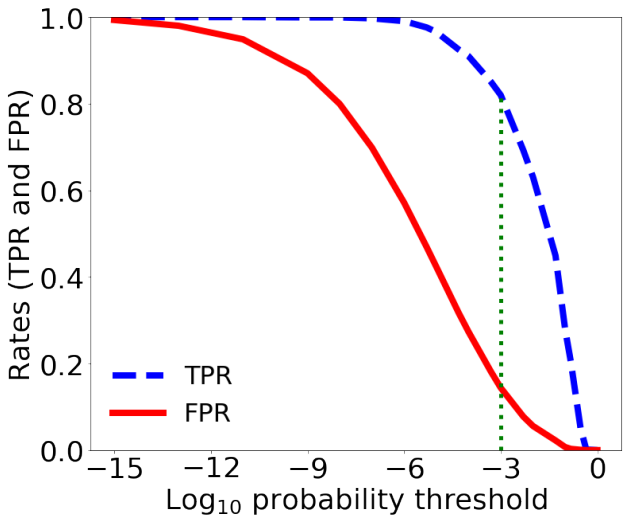

Figure 3.3: TPR and FPR with varying probability threshold in the prediction of alternative precursor list. The green dashed line indicates where the largest difference between the TPR and FPR was observed.

LiOH is sometimes preferred over Li_2CO_3 as it leaves less carbonate residual on the surface of the particles.

Intuitively, hydroxides are similar to oxides; however, Figure 3.4 also captures some differences in this similarity for different elements. For example, the common precursor for Al is the oxide, whereas that for B is the hydroxide. Furthermore, the probability of substitution between $Al(OH)_3$ and Al_2O_3 is considerably higher than between $B(OH)_3$ and B_2O_3. The number of reactions using Al_2O_3, $Al(OH)_3$, $B(OH)_3$, and B_2O_3 are 1,606, 148, 705, and 252, respectively, indicating that this difference is not due to limited data. The reason behind this is possibly correlated with the unique bonding in B_2O_3; B is highly hybridized with O in B_2O_3, much more than Al with O in Al_2O_3. This creates strong units in B_2O_3 held together by relatively weak forces [108] accounting for its low melting point and high glass-forming ability [109]. Although nitrates are often used in solution-based synthesis, the chance to use nitrates in solid-state synthesis is also considerable. Figure 3.4 shows that for elements Ca, Ba, Al, and Fe, nitrates frequently replace the common oxide or carbonate precursors. For example, the probability of substituting Fe_2O_3 with $Fe(NO_3)_3$ is high. The nitrates are used in various ways such as in conventional solid-state synthesis [110], modified solid-state synthesis [111], combustion synthesis [112], and sol-gel synthesis [113]. Although carbonates appear interchangeable with oxides, the metals in them might not occupy the same valence state. The probability of substitution between $MnCO_3$ and MnO_2 is higher than that between $MnCO_3$ and MnO, indicating that MnO_2 is more similar to $MnCO_3$ than MnO.

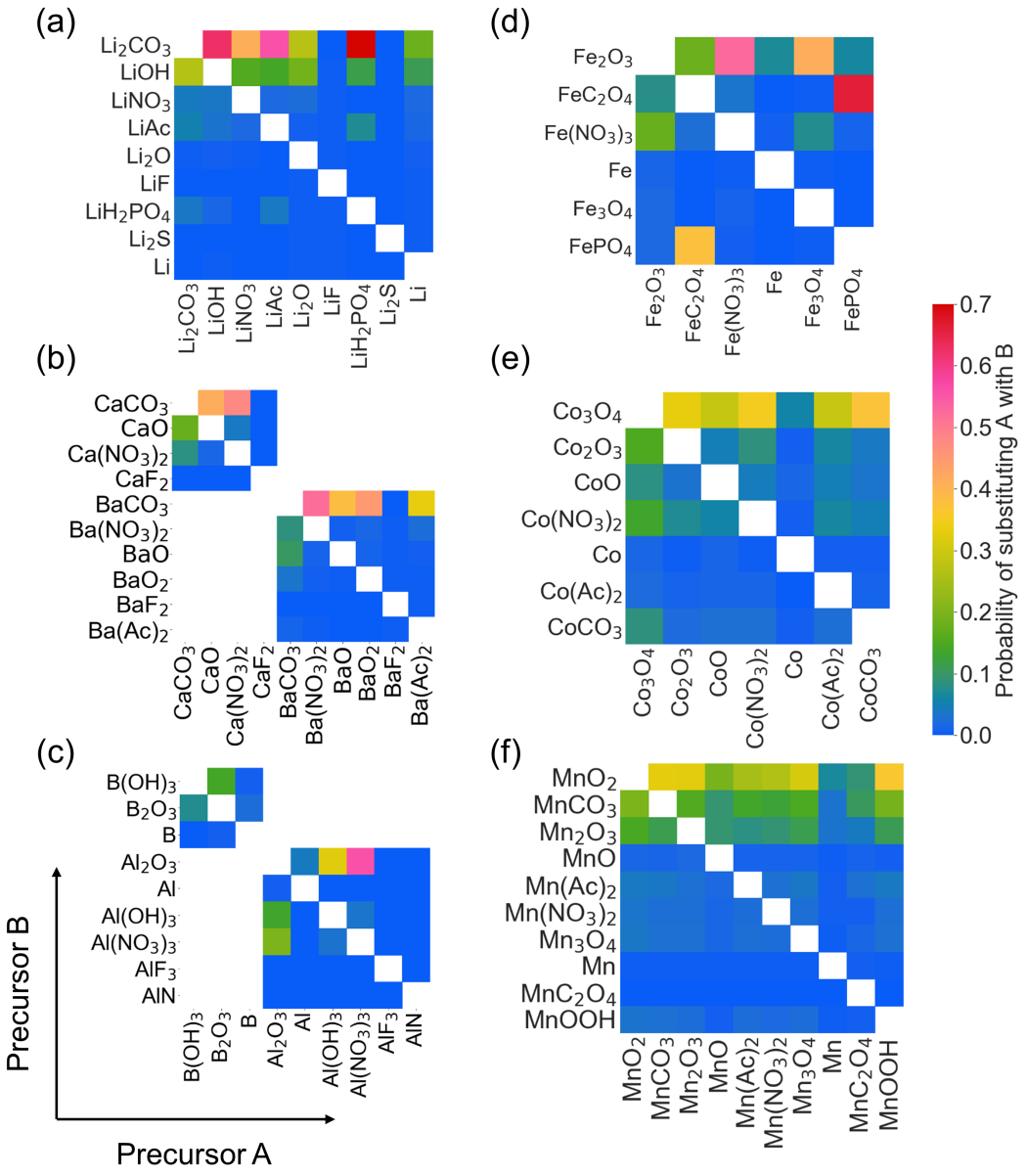

Figure 3.4: Substitution probability $P(B|A)$, which is the probability that the precursor A on the x-axis is substituted with precursor B on the y-axis: (a) Li, (b) Ca and Ba, (c) B and Al, (d) Fe, (e) Co, (f) Mn. For example, we found that in 15% of reactions that use $CaCO_3$, it could also be substituted with another precursor to introduce Ca into the same targets; in 73% of the substitutions, the other precursor is CaO. The joint probability that $CaCO_3$ is substituted and the substitute is CaO is 11%. Because CaF_2 is exclusively used for the synthesis of fluorine-containing compounds, the probability that CaF_2 is substituted to synthesize the same target is zero.

To better understand how precursors are chosen for elements with variable valence, for each Mn precursor with reasonable frequency of use, we plot in Figure 3.5 the distribution of valence states for Mn in the targets synthesized from that precursor. The valence of Mn in the target compound was determined by iterating all possible combinations of valence states and finding the one resulting in the charge neutrality for the compound [114]. The width of each violin plot is proportional to the probability density for different valence states; the total area is proportional to the number of reactions using the corresponding precursor. The adoption of MnO, Mn_2O_3, and MnO_2 is preferred in the literature to synthesize targets with similar valence states, that is, most Mn ions in targets from MnO, Mn_2O_3, and MnO_2 correspond to 2+, 3+, and 3+$\sim$4+, respectively. Different from the oxides, the valence states in targets from $MnCO_3$ and $Mn(Ac)_2$ are more evenly distributed, indicating that the use of $MnCO_3$ and $Mn(Ac)_2$ is less dependent on the valence states in the targets. This appears reasonable given the ease by which $MnCO_3$ and $Mn(Ac)_2$ decompose when heated and Mn^{2+} can be oxidized to whatever is stable in the high-component solid under proper oxygen chemical potential. This observation is consistent with the higher probability of substitution between $MnCO_3$ and MnO_2 as aforementioned. By comparing the number of reactions using different precursors, it should be noted that the most frequently used Mn precursor to synthesize targets with Mn valence states lower than 3+ remains MnO_2, which is the common precursor for Mn, even though MnO_2 is more frequently used to synthesize targets with Mn valence states between 3+ and 4+. One possible reason is that Mn at high temperature can rapidly reduce or oxidize driven by the extent of entropic stabilization of O_2 on the right-hand side of the reaction $MnO_2 + \Delta H \rightleftharpoons MnO_{2-x} + \frac{x}{2}\ O_2$. In other words, the metal valence state in the precursor does not necessarily impose the valence state in the target in solid-state synthesis.

3.5 Similarity of precursors

While *substitutionability*, discussed in the previous section, indicates that a solid-state reaction to the target is possible with the substitutional precursors, it makes no statement as to whether the reaction condition needs to be modified. In the following section we define the *similarity* of precursors based on the *substitutionability* as well as the extent to which the reaction conditions are similar. At this point, we only use temperature to describe the reaction condition considering the amount of effort, but one could extend this concept to capture other synthesis info such as atmosphere, time, number of operations, milling speed, and so forth.

Metric for similarity

Two features, the substitution probability and the distribution of synthesis temperatures of the reactions that use a particular precursor, were utilized to characterize the *similarity* of precursors.

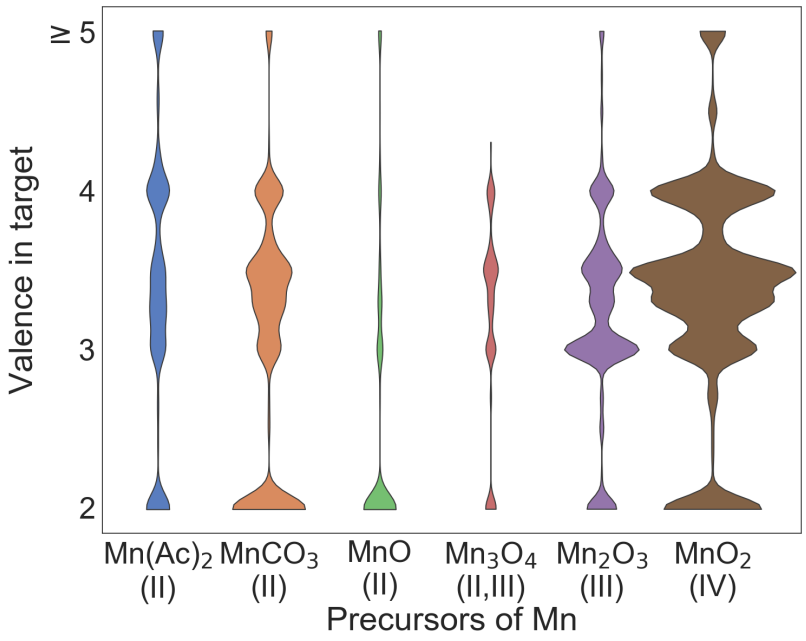

Figure 3.5: Mn valence states in targets from $Mn(Ac)_2$ (manganese acetate), $MnCO_3$, MnO, Mn_3O_4, Mn_2O_3, and MnO_2. The width in each violin plot is proportional to the probability density for valence at different values. The total area of each violin plot is proportional to the number of reactions using the corresponding precursor.

As introduced in Section 3.2, a precursor p_i^1 is substituted by another precursor p_i^2 with the probability $P(p_i^2|p_i^1)$. We use the geometric average of $P(p_i^2|p_i^1)$ and $P(p_i^1|p_i^2)$ to balance the asymmetric situations where p_i^1 or p_i^2 is substituted. The distance accounting for the substitution probability is defined as

$$d_{sub}(p_i^1, p_i^2) = 1 - \sqrt{(P(p_i^1|p_i^2)P(p_i^2|p_i^1))}, \tag{3.8}$$

where p_i^1 and p_i^2 are two precursors for element e_i.

A different precursor can be used with a different synthesis temperature. As an example, the distribution of the highest firing temperature used in synthesis reactions with two different Fe or Ca precursors is presented in Figure 3.6. The temperatures were extracted by regular expression matching in the corresponding synthesis paragraphs. For example, Figure 3.6 shows that the typical firing temperature is much lower when FeC_2O_4 is used as a precursor than when Fe_2O_3 is, whereas the firing temperature for CaO is comparable to that for $CaCO_3$. Utilizing the overlap between the distributions of temperatures for two precursors,

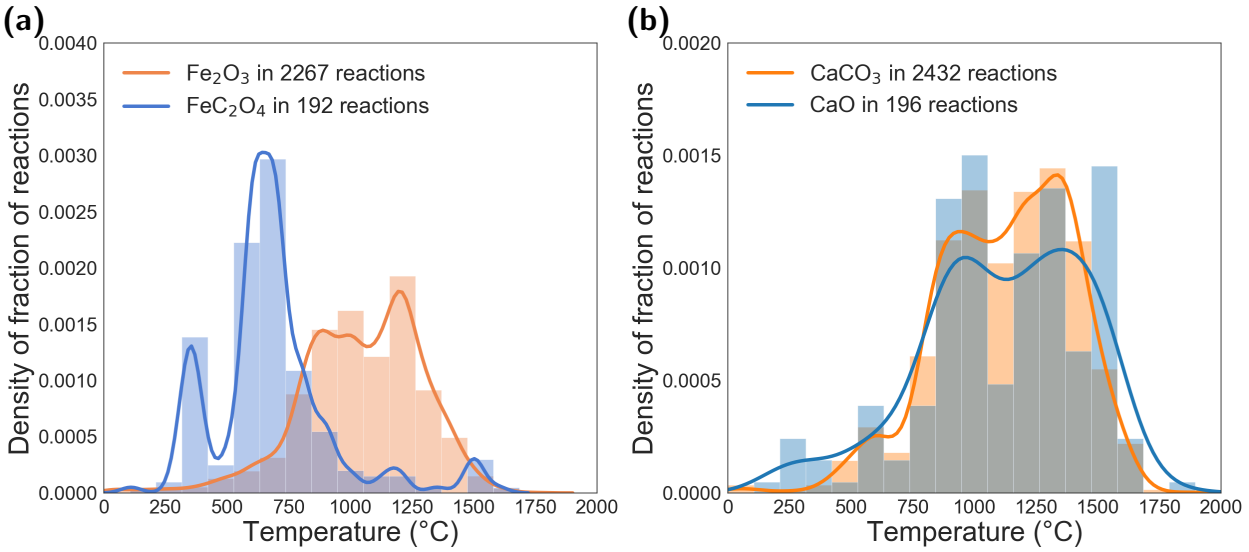

Figure 3.6: Highest firing temperature in the synthesis process for: **(a)** Fe_2O_3 and FeC_2O_4 and **(b)** $CaCO_3$ and CaO.

a distance is defined as follows to describe the similarity between the two precursors.

$$d_{temp}(p_i^1, p_i^2) = 1 - \frac{\text{overlapping area of two temperature distribution}}{\text{total area of two temperature distribution}}. \tag{3.9}$$

Both d_{sub} in Eq. 3.8 and d_{temp} in Eq. 3.9 satisfy the property that $0 \leq d_i \leq 1$. We utilized the Euclidean distance to define a multi-feature distance metric [115, 116] to combine the two features together. The distance between a pair of precursors for the same element is defined as

$$D(p_i^1, p_i^2) = \sqrt{d_{sub}^2(p_i^1, p_i^2) + d_{temp}^2(p_i^1, p_i^2)}. \tag{3.10}$$

The multi-feature aspect of this distance metric is general; it is straightforward to include additional features into this distance metric as new relevant features are considered. The current two representative features are selected because the substitution probability reflects the comparison of overall reactions in synthesis, and temperature is the most important parameter to activate these reactions. Finally, to visualize the similarity of precursors for the same element, we performed hierarchical clustering based on the pairwise distance $D(p_i^1, p_i^2)$ using Ward's minimum variance method [117]. The hierarchical clustering method iteratively identifies two nearest clusters and merges them until only one supercluster is left.

Clustering precursors by similarity

Based on the distance defined in Eq. 3.10, precursors for the same elements were hierarchically clustered, and the similarities between them are displayed as dendrograms in Figure 3.7. The vertical axis represents the distance between two precursors or the distance between two clusters. In general, similar precursors will be drawn closer to each other on the horizontal axis.

Generally, the cluster with the smallest internal distance includes the common precursors, indicated using bold fonts in Figure 3.7. Simple binary fluorides and sulfides are far away from the common precursors and are typically used as source of F and S in target materials so that HF and H_2S can be avoided. Metals are sometimes used as precursors directly; however, they are far away from the common precursors, indicating that metals and metal oxides tend to be used as precursors for different classes of materials. There is a trend that precursors are clustered following the order: oxide, carbonate, nitrate, and acetate, where the adjacent precursors are more similar (e.g., carbonate and oxide, or carbonate and nitrate), and the nonadjacent precursors are less similar (e.g. oxide and acetate) though there are variations to this for some elements. When the common precursor is a carbonate, the order may change to nitrate, carbonate, oxide, and acetate (e.g., Ba), where the carbonate and the nitrate are more similar than the carbonate and the oxide, but the carbonate still sits between the nitrate and the oxide. The similarity between different classes is possibly correlated with the different bonding strength between the cations and anions, which can be indicated by the order of melting points, namely, oxide > carbonate > nitrate/acetate.

However, there are also some observations that are not easy to immediately rationalize. For Li, it is the hydroxide rather than oxide or nitrate closest to the carbonate, whereas for Ca and Ba, the hydroxides are even absent, which means $Ca(OH)_2$ and $Ba(OH)_2$ are rarely used. This difference may originate from the methods used to prepare these precursors being different, resulting in different availabilities. One practical clue is that Li_2O is more expensive than LiOH; Li_2O ($\geq$95% purity) is \$378.00 for 100 g (\$8.10/g of Li), while lithium hydroxide monohydrate ($\geq$95% purity) is \$181.00 for 2 kg (\$0.54/g of Li) from the chemical supplier Strem Chemicals [118]. It is also observed that LiAc and LiH_2PO_4, as well as FeC_2O_4 and $FePO_4$, are clustered together, because they are frequently used to synthesize the extensively studied cathode material $LiFePO_4$, reflecting possible application bias in the data. In addition, oxides are similar to each other for variable valence elements, but the most similar precursor to the common oxide is not necessarily an oxide. For example, the oxides of Mn are clustered together, ranging from MnO_2 to MnO. However, the most similar precursor to MnO_2 is $MnCO_3$, as discussed in Section 3.4. Similarly, the nitrate $Fe(NO_3)_3$ is more similar to Fe_2O_3 than the mixed-valence oxide Fe_3O_4 to Fe_2O_3. There are many factors in the selection of precursors, including both scientific reasons such as bonding, reactivity, and melting point, and anthropogenic reasons [119] such as literature success, convenience, applications, price, and human bias. The data in this work are a reflection of all those

factors; it is not entirely clear how to deconvolute all these issues. An interesting scientific advance would be to identify the precursors that are chemically compelling while avoiding the implicit anthropogenic biases. This work provides a historical statistical analysis to serve as a baseline comparison.

The similarity could help guide the selection of precursors when researchers alter existing recipes by replacing precursors. For a starting experiment, it might be profitable to pick precursors similar to what has been tried before. On the other hand, when the synthesis is not going well, it is best to use a very different precursor in order to diversify the synthesis space. If there are many possible combinations of precursors, the quantitative value of the similarity could also serve as a reference to rank them. Currently, the creation of new recipes is in principle limited to targets already in our dataset. Therefore, it is also important to develop similarity among targets. In that way, it would be possible to predict synthesis recipes for new target materials by evaluating the similarity with targets for which synthesis is known, a process that is very similar to the current literature-based approach for the synthesis of novel materials.

3.6 Conclusion

Using the solid-state synthesis data extracted from materials science literature, we conducted a meta-analysis on the similarities and differences between precursors. The statistics on the frequency to use different classes of precursors shows that each element usually has a common precursor to bring it into a target compound. A substitution model is used to quantify the probability of substituting one precursor with another while the target remains unchanged. By establishing distance metrics from the substitution model and the distribution of synthesis temperature, precursors for the same element were clustered to show the similarities between these precursors. This hierarchical clustering demonstrates that chemical domain knowledge of solid-state synthesis can be captured from text mining and provides a foundation for developing a predictive synthesis model.

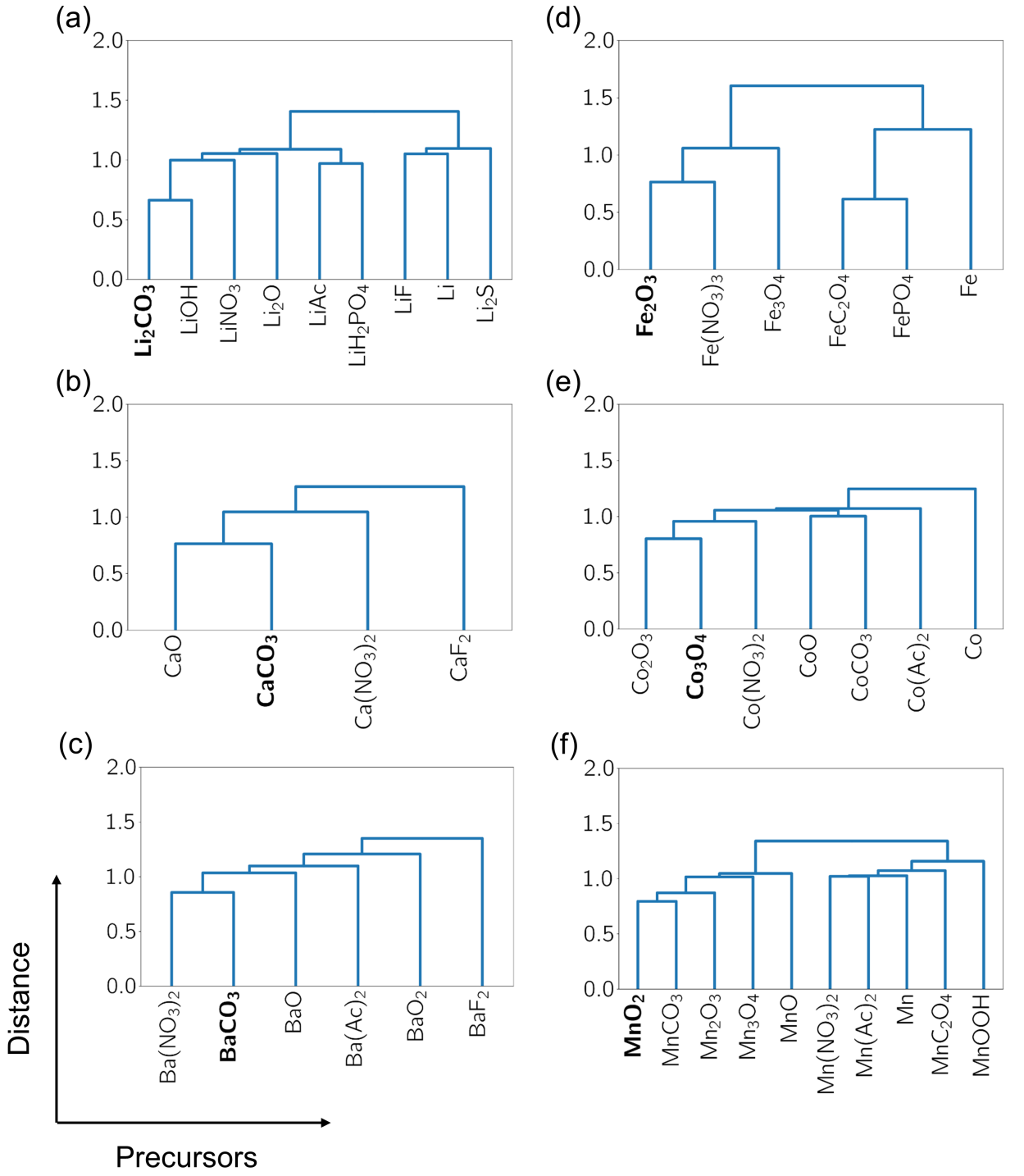

Figure 3.7: Clusters of precursors for (a) Li, (b) Ca, (c) Ba, (d) Fe, (e) Co, and (f) Mn by similarity. The common precursors are indicated using bold fonts.

Chapter 4

Target similarity for predictive synthesis of new materials

The similarity of precursors introduced in Chapter 3 provides a meaningful metric to create alternative recipes for existing targets, but it cannot be used to predict synthesis recipes for new target materials. Experimental researchers usually approach a new inorganic synthesis by manually looking up similar materials in the literature and repurposing precedent recipes for a novel material. However, deciding what materials are similar and thus where to look is often driven by intuition and limited by individuals' personal experience in specific chemical spaces, hindering the ability to rapidly design syntheses for new chemistries. With a large-scale materials synthesis dataset from text-mining efforts, it is possible to statistically learn the similarity of materials and the correlation of their synthesis variables in a more systematic and quantitative fashion, and provide such tools as a guide to scientists when approaching the synthesis of novel compounds [99].

In this chapter[1], we propose a precursor recommendation strategy (Fig. 4.1) based on machine-learned similarity of materials to automate the literature-based approach used by experimental researchers. Inspired by natural language processing (NLP) models [38, 43, 44], we designed an encoding neural network to learn the vectorized representation of a material based on its corresponding precursors for the quantification of materials similarity. Assuming that the target material can be synthesized using an experimental design adapted from a similar material, synthesis variables such as precursors, operations, and conditions can be proposed and ranked by querying the knowledge base of previously synthesized materials. We applied the recommendation strategy to predict precursors for 2,654 test target materials in a historical validation. Learning from a knowledge base of 29,900 synthesis reactions text-

[1] This Chapter incorporates sections from a paper in press with permission from the authors: Tanjin He, Haoyan Huo, Christopher J. Bartel, Zheren Wang, Kevin Cruse, and Gerbrand Ceder. "Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature." *Science Advances* (2023), in press [99].

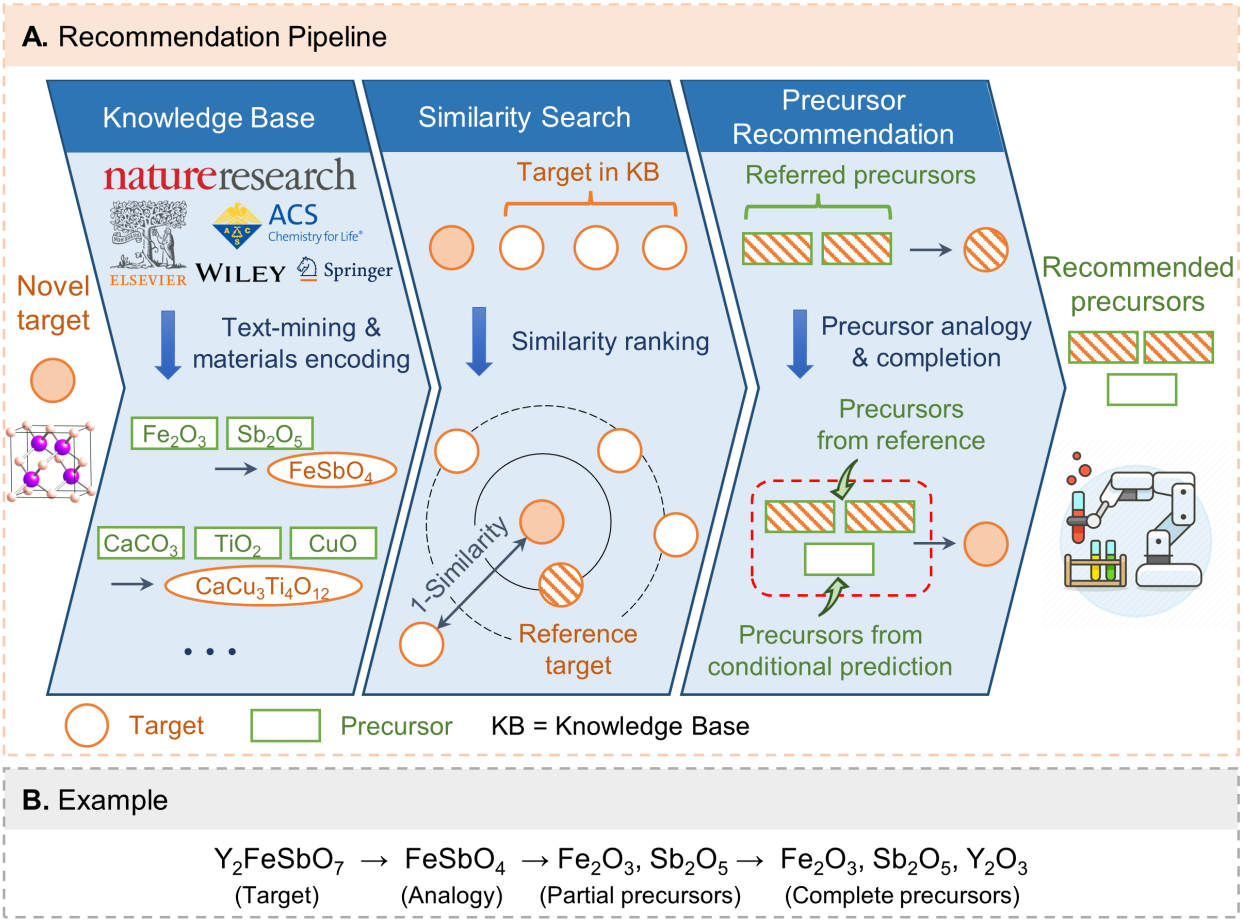

Figure 4.1: Precursor recommendation strategy. **(A)** Pipeline for precursor recommendation consisting of three steps: (1) digitize target materials in the synthesis knowledge base text-mined from scientific literature, (2) rank target materials in the knowledge base according to the similarity to the novel target, and (3) recommend precursors based on analogy to the most similar target. **(B)** An example of precursor recommendation for Y_2FeSbO_7 by referring to the synthesis of $FeSbO_4$.

mined from the scientific literature, we demonstrate that the algorithm can acquire chemical knowledge on materials similarity via self-supervised learning, and make promising decisions on precursor selection.

Here, we begin with statistical insights from our text-mined solid-state synthesis dataset to better understand the problem of precursor selection (Section 4.1). Because a universal model for solid-state synthesis has not yet been established, we use a data-driven method to recommend potential precursor sets for the given target material (Figure 4.1). The rec-

ommendation pipeline consists of three steps: (i) an encoding model to digitize the target material as well as known materials in the knowledge base (Section 4.2), (ii) similarity query based on the materials encoding to identify a reference material that is most similar to the target (Section 4.3), and (iii) recipe completion to (a) compile the precursors referred from the reference material and (b) add any possibly missed precursors if element conservation is not achieved using conditional predictions based on referred precursors (Section 4.4).

4.1 Problem of precursor selection

In the solid-state synthesis of inorganic materials, precursor selection plays a crucial role in governing the synthesis pathway by yielding intermediates that may lead to the desired material or alternative phases [5, 6].

For each metal/metalloid element, one precursor is often used predominantly over all others, which we denote as the common precursor in Chapter 3. However, we also find that approximately half of the target materials were synthesized using at least one uncommon precursor (Figure 3.2). Uncommon precursors may be used for a variety of reasons including synthetic constraints (e.g., temperature and time), purity, morphology, and anthropogenic factors [5, 29, 119].

In addition, a probability analysis of the text-mined dataset indicates that precursors for different chemical elements are not randomly combined. The joint probability to select a specific precursor pair (A_i, B_i) can be compared to the marginal probability to select A_i for element Ele_a and B_i for Ele_b. If the choices of A_i and B_i are independent, the joint probability should equal the product of the marginal probabilities, namely, $P(A_i, B_i) = P(A_i)P(B_i)$. In Chapter 3, we assume the substitutional precursor pairs are independent of each other as a simplification, where the equality between the joint and marginal probabilities holds true. However, inspection of 6,472 pairs of precursors from our text-mined dataset (Figure 4.2) reveals that many show a strong dependency on each other (i.e., $P(A_i, B_i)$ deviating significantly from $P(A_i)P(B_i)$). A well-known example is that nitrates such as $Ba(NO_3)_2$ and $Ce(NO_3)_3$ tend to be used together, likely because of their solubility and applicability for solution processing (e.g. slurry preparation). Unfortunately, these decisions regarding dependencies of precursors are usually empirical and hard to standardize.

Different from assessment on substituting precursors for an alternative recipe in Chapter 3, precursor selection for novel target materials is more challenging because the chemical space of targets is much larger than that of precursors. In a set of 33,343 solid-state synthesis reactions, only hundreds of precursors are frequently used (in at least 10 reactions) and the number tends to converge, but the number of targets almost increases linearly with the number of reactions (Figure 4.3). As a result, it is impractical to enumerate all pairs of targets and evaluate the similarity between them, which is the proposed solution to the

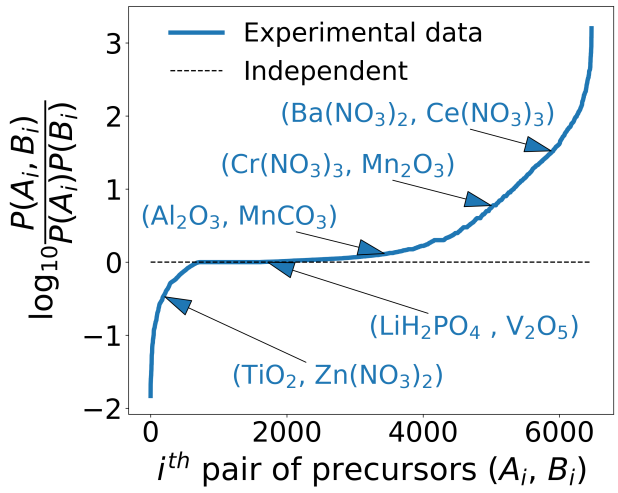

Figure 4.2: Pairwise dependency of precursors A_i and B_i characterized by $\frac{P(A_i,B_i)}{P(A_i)P(B_i)}$. Probability is estimated from the frequency of occurrence in the solid-state synthesis dataset. The value of $\log_{10} \frac{P(A_i,B_i)}{P(A_i)P(B_i)}$ is zero when A_i and B_i are independent, positive when A_i and B_i tends to be used in the same experiment more frequently than $P(A_i)P(B_i)$, negative otherwise.

similarity of precursors in Chapter 3. More importantly, the similarity of targets should be capable of estimating the distance of an unseen target to other known targets because the synthesis of novel materials is the most interesting.

In the following sections, we show our efforts of using machine learning as a possible solution to ingest the heuristics that underlie precursor selections for new target materials.

4.2 Materials encoding for precursor selection

Our precursor recommendation model for the synthesis of a novel target will mimic the human approach of trying to identify similar target materials for which successful synthesis reactions are known. To find similar materials, digital processing requires an encoding model that transforms any arbitrary inorganic material into a numerical vector. For organic synthesis, structural fingerprinting such as Morgan2Feat [120] is a good choice [25] because it is natural to track the conservation and change of functional groups in organic reactions, but the concept of functional groups is not applicable to inorganic synthesis. Chemical formulas of inorganic solids have been represented using a variety of approaches (e.g., Magpie [121, 122], Roost [123], CrabNet [124]). However, these representations are typically used as inputs to predict thermodynamic or electronic properties of materials. Here, we attempt to directly incorporate synthesis information into the representation of a material with arbi-

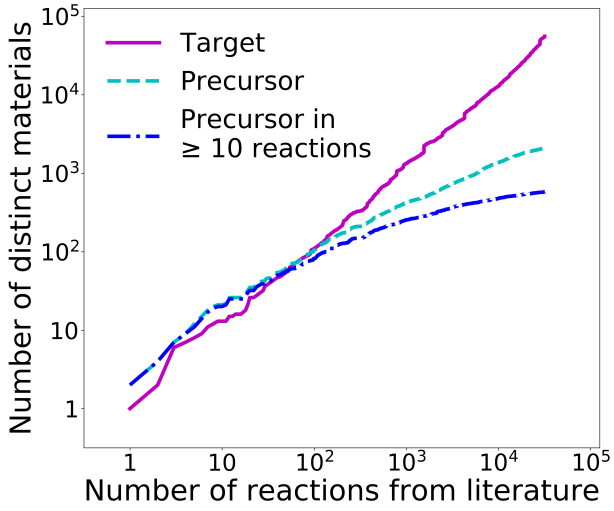

Figure 4.3: Count of target and precursor materials in the text-mined solid-state synthesis dataset.

trary composition. Local text-based encodings such as Word2Vec [54, 125] and FastText [24] are able to capture contextual information from the materials science literature, of which synthesis information is a part; however, they are not applicable to unseen materials when the materials text (sub)strings are not in the vocabulary or when the materials are not in the predefined composition space. For example, Pei et al. [125] computed the similarity of high-entropy alloys as the average similarity of element strings by assuming the elements are present in equal proportions in the material (e.g., CoCrFeNiV). However, this approach is not applicable to unseen materials different from such composition template, and consequently would not be practical in our work on synthesis of diverse inorganic materials. Substitution modeling can evaluate similarity of precursors by assessing the viability of substituting one precursor with another while retaining the same target, but it cannot be used to identify analogues for new target materials [29]. In this work, we propose a synthesis context-based encoding model utilizing the idea that target materials produced with similar synthesis variables are similar.

Analogous to how language models [38, 43, 44] pre-train word representations by predicting context for each word, we use a self-supervised representation learning model to encode arbitrary materials by predicting precursors for each target material, which we refer to as PrecursorSelector encoding (Figure 4.4A). The upstream part is an encoder where properties of the target material are projected into a latent space as the encoded vector representation. In principle, any intrinsic materials property could be included at this step. Here, we use only composition for simplification. The downstream part consists of multiple tasks where the encoded vector is used as the input to predict different variables related to precursor selection.

Here, we use a masked precursor completion (MPC) task (Figure 4.4B) to capture (i) the correlation between the target and precursors and (ii) the dependency between different precursors in the same experiment. For each target material and corresponding precursors in the training set, we randomly mask part of the precursors and use the remaining precursors as a condition to predict the complete precursor set. We also add a task of reconstructing the chemical composition to conserve the compositional information of the target material. The downstream task part is designed to be extensible; other synthesis variables such as operations and conditions can be incorporated by adding corresponding prediction tasks in a similar fashion. By training the entire neural network, the encoded vectors for target materials with similar precursors are automatically dragged closer to each other in the latent space because that reduces the overall prediction error. This PrecursorSelector encoding thus takes the correlation induced by precursor selection and serves as a useful metric to measure similarity of target materials in syntheses.

To demonstrate that the neural network is able to learn precursor information, we present the results of the MPC task (Figure 4.4B) for $LaAlO_3$ as an example (Table 4.1). $LaAlO_3$ is a ternary material that normally requires two precursors (one to deliver each cation, La and Al). In this test, we masked one precursor and asked the model to predict the complete precursor set. For the same target conditioned with different partial precursors, the predicted probabilities of precursors strongly depend on the given precursor and agree with some rules of thumb for precursor selection. When the partial precursors are oxides such as La_2O_3 or Al_2O_3, the most probable precursors are predicted to be oxides for the other element, i.e., Al_2O_3 for La_2O_3 and La_2O_3 for Al_2O_3 [126]. When the partial precursors are nitrates such as $La(NO_3)_3$ or $Al(NO_3)_3$, nitrates for the other element are prompted with higher probabilities, i.e., $Al(NO_3)_3$ for $La(NO_3)_3$ and $La(NO_3)_3$ for $Al(NO_3)_3$ [127]. If both precursors are masked, oxides rank first in the prediction because the common precursors for elements La and Al are La_2O_3 and Al_2O_3, respectively. The simple successful prediction shows our PrecursorSelector encoding model is able to learn the correlation between the target and precursors in different contexts of synthesis without explicit input of chemical rules about synthesis. In addition, the use of different precursors suggests various synthetic routes may lead to the same target material. When a practical preference for a particular route exists, the framework we introduce in this work can be extended to include more constraints, such as synthesis type, temperature, morphology, particle size, and cost of precursors, by learning from pertinent datasets [119, 128, 129].

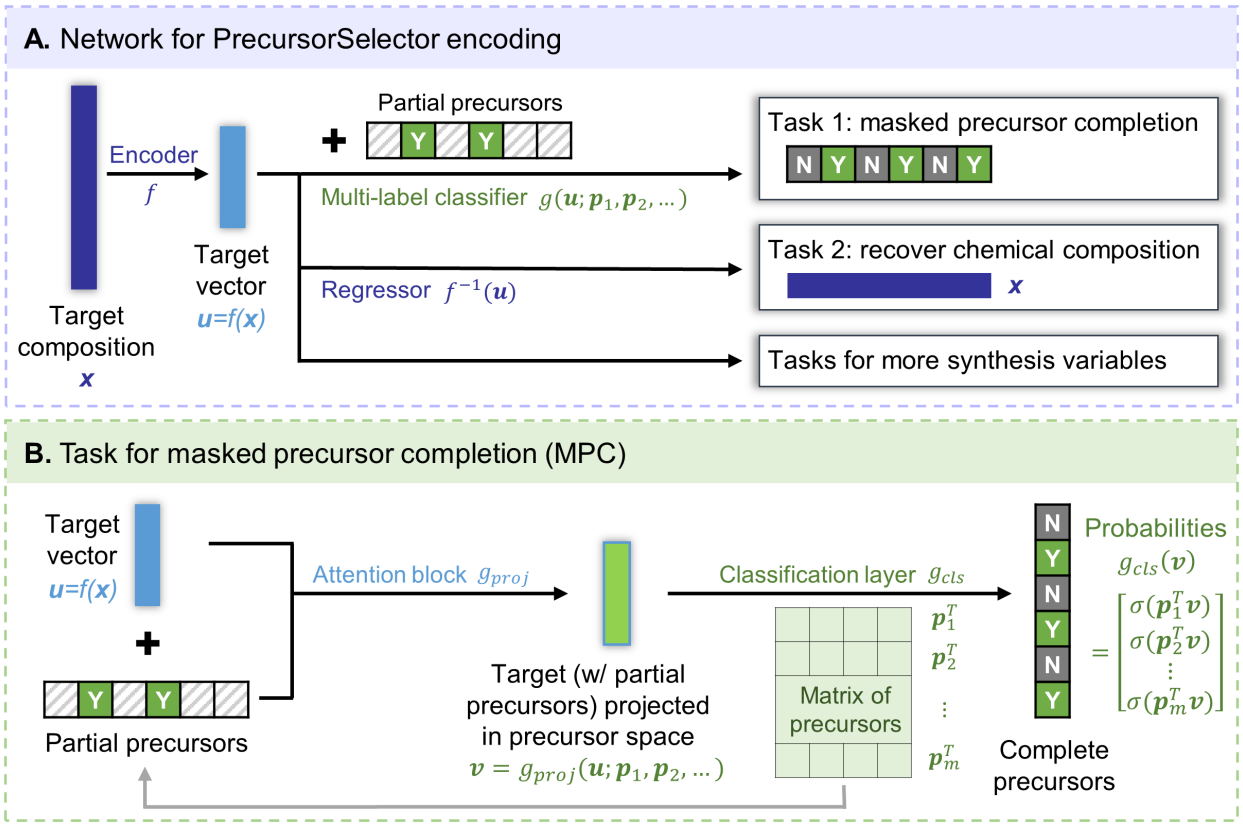

Figure 4.4: Representation learning to encode precursor information for target materials. **(A)** Multi-task network structure to encode the target material in the upstream and to predict the complete precursor set, chemical composition, and more synthesis variables in the downstream. $\boldsymbol{x}$ and $\boldsymbol{u}$ represent the composition and encoded vector of the target material, respectively. $\boldsymbol{p}_i$ represents the i^{th} precursor in a predefined ordered precursor list. Dense layers are used in each layer unless specified differently. **(B)** Submodel of multi-label classification for the masked precursor completion (MPC) task. Part of the precursors are randomly masked; the remaining precursors (marked as "Y") are used as a condition to predict the probabilities of other precursors for the target material. The probabilities corresponding to the complete precursors (marked as "Y") are expected to be higher than that of unused precursors (marked as "N"). The attention block g_{proj} [42] is used to aggregate the target vector and conditional precursors. The final classification layer g_{cls} and the embedding matrix for conditional precursors share the same weights. σ represents the sigmoid function.

Table 4.1: MPC conditioned on different partial precursors for the same target material $LaAlO_3$. The predicted complete precursors are the ones with the highest probabilities (bold).

Partial precursors (condition)	Probability to use different precursors (output)					
	La_2O_3	Al_2O_3	$La(NO_3)_3$	$Al(NO_3)_3$	$La_2(CO_3)_3$	$Al(OH)_3$
La_2O_3	**0.75**	**0.71**	0.58	0.57	0.57	0.57
Al_2O_3	**0.72**	**0.73**	0.58	0.57	0.58	0.56
$La(NO_3)_3$	0.60	0.59	**0.64**	**0.63**	0.61	0.61
$Al(NO_3)_3$	0.62	0.58	**0.65**	**0.65**	0.62	0.60
N/A	**0.70**	**0.69**	0.59	0.58	0.59	0.59

4.3 Similarity of target materials

Similarity establishes a link between a novel material to synthesize and the known materials in the knowledge base because it is reasonable to assume similar target materials share similar synthesis variables in experiments. Although the understanding of similarity is generally based on heuristics, the PrecursorSelector encoding introduced in Section 4.2 provides a meaningful representation for quantified similarity analysis. Dedicated to precursor prediction in this study, we define the similarity of two target materials as the similarity of the precursors used in their respective syntheses. Although precursors for a new target material are not known in advance, the PrecursorSelector encoding serves as a proxy reflecting the potential precursors to use. In that latent space, we can take the cosine similarity [43, 44, 54] of the PrecursorSelector encoding as a measure of the similarity (Sim) of two target materials $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:

$$\mathrm{Sim}(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \cos(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)), \tag{4.1}$$

where f is the encoder part of the PrecursorSelector model transforming the composition of the target material $\boldsymbol{x}$ into the encoded target vector (Figure 4.4A).

To demonstrate that the similarity estimated from PrecursorSelector encoding is reasonable, we show typical materials with different levels of similarity to an example target material $NaZr_2(PO_4)_3$ (Table 4.2). The most similar materials are the ones with the same elements such as Zr-containing phosphates and other NASICON materials. The similarity decreases slightly as additional elements are introduced (e.g., $Na_3Zr_{1.9}Ti_{0.1}Si_2PO_{12}$) or when one element is substituted (e.g., $LiZr_2(PO_4)_3$). When the phosphate groups are replaced with another anion, the similarity decreases further, with oxides having generally mild similarity to the phosphate $NaZr_2(PO_4)_3$. The similarity decreases even further for compounds with

Table 4.2: Different levels of similarity between $NaZr_2(PO_4)_3$ and materials in the knowledge base.

Target	Similarity	Target	Similarity
$Zr_3(PO_4)_4$	0.946	$Li_{1.8}ZrO_3$	0.701
$Na_3Zr_2Si_2PO_{12}$	0.929	$NaNbO_3$	0.600
$Na_3Zr_{1.8}Ge_{0.2}Si_2PO_{12}$	0.921	$Li_2Mg_2(MoO_4)_3$	0.500
$Na_3Ca_{0.1}Zr_{1.9}Si_2PO_{11.9}$	0.908	$Sr_2Ce_2Ti_5O_{16}$	0.400
$Na_3Zr_{1.9}Ti_{0.1}Si_2PO_{12}$	0.900	$Ga_{0.75}Al_{0.25}FeO_3$	0.300
$LiZr_2(PO_4)_3$	0.896	Cu_2Te	0.200
$NaLa(PO_3)_4$	0.874	$Ni_{60}Fe_{30}Mn_{10}$	0.100
$Sr_{0.125}Ca_{0.375}Zr_2(PO_4)_3$	0.852	$AgCrSe_2$	0.000
$Na_5Cu_2(PO_4)_3$	0.830	$Zn_{0.1}Cd_{0.9}Cr_2S_4$	-0.099
$LiGe_2(PO_4)_3$	0.796	Cr_2AlC	-0.202

no anion (e.g., intermetallics) and for non-oxygen anions (e.g., chalcogenides). This finding agrees with our experimental experience that when seeking a reference material, researchers will usually refer to compositions in the same chemical system or to cases where some elements are substituted. It is also worth noting that our quantitative similarity is purely a data-driven abstraction from the literature and uses no externally chemical knowledge.

To better understand the similarity, we conducted a relationship analysis [43, 44, 54] by visualizing four groups of target materials synthesized using one shared precursor and one distinct precursor (Figure 4.5). For example, the syntheses of $YCuO_2$, $Ba_3Y_4O_9$, and $Ti_3Y_2O_9$ share Y_2O_3 as a precursor and separately use CuO, $BaCO_3$, and TiO_2. The three other groups share the precursors In_2O_3, Al_2O_3, and Fe_2O_3, respectively. To separate the effect of the precursor variation, we align the original points of the target vectors by first projecting each target vector to the same vector space as the precursors and then subtracting the vector of the shared precursor, providing a difference vector showing the relationship between the target material and the shared precursor (more details in Section 4.7). Next, we plot the top two principal components [130] of these difference vectors in a two-dimensional plane. The difference vectors are automatically separated into three clusters according to the precursor variate, representing three types of relationships, "react with $BaCO_3$", "react with CuO", and "react with TiO_2", respectively. For example, $Ba_3Y_4O_9$ is to Y_2O_3 as $BaAl_2O_4$ is to Al_2O_3 (i.e., $Ba_3Y_4O_9 - Y_2O_3 \approx BaAl_2O_4 - Al_2O_3$) because both syntheses use $BaCO_3$. The consistency between this automatic clustering and the chemical intuition again affirms the

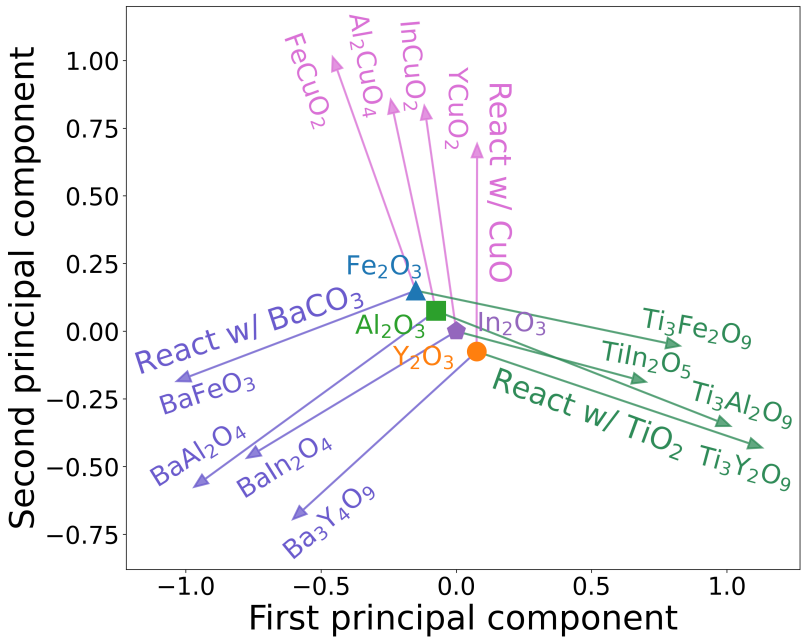

Figure 4.5: Relationships between targets and their shared precursors. Four groups of target materials are synthesized each using one shared precursor shown as the original point (Y_2O_3, In_2O_3, Al_2O_3, or Fe_2O_3) and one distinct precursor shown as the edge ($BaCO_3$, CuO, or TiO_2). The relationship of "react with another precursor" is visualized as the first two principal components of the difference vector between the target and the shared precursor $g_{proj}(f(\boldsymbol{x})) - \boldsymbol{p}_i$. The original points corresponding to different precursors $\boldsymbol{p}_i$'s are jittered for clarity.

efficacy of using PrecursorSelector encoding as a similarity metric.

4.4 Recommendation of precursor materials

With the capability of measuring similarity, a natural solution to precursor selection is to replicate the literature-based approach used by experimental researchers. Given a novel material to synthesize, we initialize our recommendation by first proposing a recipe consisting of common precursors for each metal/metalloid element in the target material because this might be the first attempt in a lab. Then, we encode the novel target material and known target materials in the knowledge base using PrecursorSelector encoding model from Section 4.2 and calculate the similarity between the novel target and each known material with Eq. 4.1. We rank known materials based on their similarity to the target such that a reference material can be identified that is the most similar to the novel target. When the precursors

used in the synthesis of the reference material cannot cover all elements of the target, we use MPC in Figure 4.4B to predict the missing precursors. For example, for Y_2FeSbO_7 (Figure 4.1B), the most similar material in the knowledge base is $FeSbO_4$. It is reasonable to assume that the precursors Fe_2O_3 and Sb_2O_5 used in the synthesis of $FeSbO_4$ [131] can also be used to synthesize Y_2FeSbO_7. Because the Y source is missing, MPC finds Y_2O_3 is likely to fit with Fe_2O_3 and Sb_2O_5 for the synthesis of Y_2FeSbO_7, ending up as a complete precursor set (Fe_2O_3, Sb_2O_5, and Y_2O_3) [132]. Multiple attempts of recommendation are feasible by moving down the list of known materials ranked to be most similar to the novel target.

To evaluate our recommendation pipeline, we conduct a validation (Figure 4.6) using the 33,343 synthesis recipes text-mined from the scientific literature. Using the knowledge base of 24,034 materials reported by the year 2014, we predict precursors for 2,654 test target materials newly reported from 2017 to 2020 (more details in Section 4.7). Because multiple precursors exist for each element, the number of precursor combinations increases combinatorially with the number of elements present in the target material. A good precursor prediction algorithm is anticipated to select from hundreds of possible precursor combinations those that have a higher probability of success. For each test material, we attempt to propose five different precursor sets. For each attempt, we calculate the percentage of test materials being successfully synthesized, where success means at least one set of proposed precursors has been observed in previous experiments. The similarity-based reference already increases the success rate to 73% at the second attempt. The first guess is set to default to the most common precursors which leads to 36% success rate. Within five attempts, the success rate of our recommendation pipeline using PrecursorSelector encoding is 82%, comparable to the performance of recommendations for organic synthesis [25]. We note that as defined here, "success" will be underestimated since some suggested precursor sets may actually lead to successful target synthesis even though they may not have been tried (and therefore do not appear in the data).

We also establish a baseline model ("Most frequent" in Figure 4.6) that ranks precursor sets based on the product of frequencies with which different precursors are used in the literature (more details in Section 4.7). This baseline simulates the typical early stage of the trial-and-error process where researchers grid-search different combinations of precursors matching elements present in the target material without the knowledge of dependency of precursors (Figure 4.2). The success rate of this baseline is 58% within five attempts. Our recommendation pipeline performs better because the dependency of precursors is more easily captured when the combination of precursors is sourced from a previously used successful recipe for a similar target. Through in-situ diffraction of synthesis [5–7], it is now better understood that some precursor sets do not lead to the target material because they form intermediate phases which have consumed much of the overall reaction energy, thereby leaving a low driving force to form the target. It is likely that our literature informed precursor prediction approach implicitly captures some of this reactivity and pathway information, resulting in a higher prediction power than random selection or selection based on how common a precursor is.

In addition, we compare with three other baseline models ("Magpie encoding", "FastText encoding", and "Raw composition" in Figure 4.6) using the same recommendation strategy but different encoding methods (more details in Section 4.7). Magpie encoding [121, 122] is a set of attributes computed using the fraction of elements in a material, including stoichiometric attributes, elemental property statistics, electronic structure attributes, and ionic compound attributes. Precursor recommendation with Magpie encoding achieves a success rate of 68% within five attempts; it performs reasonably well because these properties reflect the material composition and generally materials with close compositions tend to be similar. Similarly, precursor recommendation directly with the raw material composition achieves a success rate of 66% within five attempts. FastText encoding [24] utilizes the FastText model [133] to capture information about the co-occurrences of context words around material formulas/names in the literature. However, only 1,985 test materials can be digitized with FastText encoding due to the conflict between the limited vocabulary of n-grams and the variety of float numbers in material formulas. The success rate using FastText encoding is 56% within five attempts. Overall, the recommendation with PrecursorSelector encoding performs substantially better because Magpie and FastText encodings are more generic but not dedicated to predictive synthesis. The PrecursorSelector encoding and MPC capture the correlation between synthesis variables and known target materials, which better extends to novel materials.

4.5 Discussion

Because of its heuristic nature, it is challenging to capture the decades of synthesis knowledge established in the literature. By establishing a materials similarity measure that is a natural handle of chemical knowledge and leveraging a large-scale dataset of precedent synthesis recipes, our similarity-based recommendation strategy mimics human synthesis design and succeeds in precursor selection. The incorporation of precursor information into materials representations (Figure 4.4) leads to a quantitative similarity metric that successfully reproduces a known precursor set 82% of the time in five attempts or less (Figure 4.6). We discuss the strengths and weaknesses of this recommendation algorithm and its generalizability to broader synthesis prediction problems.

In this work, materials similarity is learned through an automatic feature extraction process mapping a target material to the combination of precursors. While learning the usage of precursors, useful chemical knowledge for synthesis practice is accordingly embedded in PrecursorSelector encoding. The first level of knowledge about materials similarity is based on composition. For example, to synthesize $Li_7La_3Nb_2O_{13}$, PrecursorSelector encoding finds $Li_5La_3Nb_2O_{12}$ as a reference target material (Table 4.3) because their difference in composition is only one Li_2O unit. PrecursorSelector encoding also reflects the consideration of valence in synthesis. Although it is not necessary to keep the valence in the precursor the same as that in the target, a precursor with similar valence states

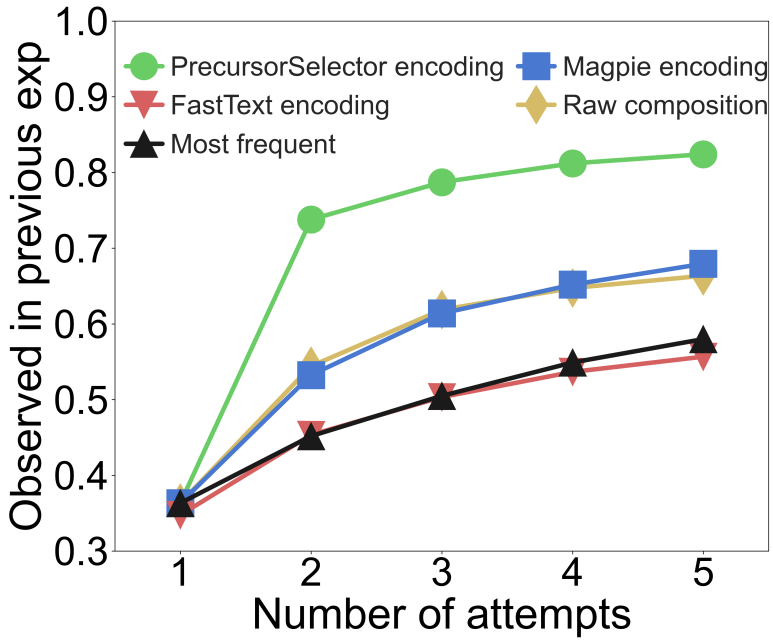

Figure 4.6: Performance of various precursor prediction algorithms. For each of the 2,654 test target materials, the algorithm attempts to propose n ($1 \leq n \leq 5$ as the x-axis) precursor sets. The y-axis shows the success rate that at least one out of the n proposed precursor set is observed in previous experimental records. PrecursorSelector encoding: this work. Magpie encoding/FastText encoding/Raw composition: similar recommendation pipeline to this work but using Magpie representation[121, 122]/FastText representation[24]/the raw material composition. Most frequent: select precursors by frequency.

to the target is frequently used in practical synthesis [29]. For example, to synthesize $NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ [134], $MnCO_3$ was used as the Mn source because the valence state of Mn is 2+ in both the target and precursor. PrecursorSelector encoding finds $Mn_{0.24}Zn_{1.76}SiO_4$ similar to $NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ because the valence state of Mn is also 2+ in $Mn_{0.24}Zn_{1.76}SiO_4$, despite $NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ containing large fractions of Na and Ga while $Mn_{0.24}Zn_{1.76}SiO_4$ does not. Our algorithm also captures the similarity of syntheses between compounds which have one element substituted. For example, PrecursorSelector encoding refers to CaZnSO for synthesizing SrZnSO because the elements Ca and Sr are regarded as similar. While such knowledge may appear obvious to the trained chemist, our approach enables it to be automatically extracted and convoluted as a vectorized representation (Figure 4.4), making it thereby available in a mathematical form, convenient to be used in recommendation engines or automated labs [135].

Table 4.3: Representative successful and failed examples for precursor prediction using the similarity-based recommendation pipeline in this study.

Target	Reference Target(s)	Expected Precursors	Error in Recommendation
Successful			
$Li_7La_3Nb_2O_{13}$ [136]	$Li_5La_3Nb_2O_{12}$ [137]	LiOH, La_2O_3, Nb_2O_5	N/A
$NaGa_{4.6}Mn_{0.01}Zn_{1.69}Si_{5.5}O_{20.1}$ [134]	$Mn_{0.24}Zn_{1.76}SiO_4$ [138]	$MnCO_3$, Na_2CO_3, Ga_2O_3, SiO_2, ZnO	N/A
SrZnSO [139]	CaZnSO [140]	$SrCO_3$, ZnS	N/A
$Na_3TiV(PO_4)_3$ [141]	$Na_3V_2(PO_4)_3$ [142]	NaH_2PO_4, NH_4VO_3, TiO_2	N/A
$GdLu(MoO_4)_3$ [143]	$Gd_2(MoO_4)_3$ [144]	$(NH_4)_6Mo_7O_{24}$, Lu_2O_3, Gd_2O_3	N/A
$BaYSi_2O_5N$ [145]	$YSiO_2N$ [146]	Si_3N_4, SiO_2, $BaCO_3$, Y_2O_3	N/A
$Cu_3Yb(SeO_3)_2O_2Cl$ [147]	$Cu_4Se_5O_{12}Cl_2$ [148]	CuO, $CuCl_2$, SeO_2, Yb_2O_3	N/A
$LiMn_{0.5}Fe_{0.5}PO_4$ [149, 150]	$LiMn_{0.8}Fe_{0.2}PO_4$ [151], $LiMn_{0.9}Fe_{0.1}PO_4$ [152]	$MnCO_3$, FeC_2O_4, LiH_2PO_4; $Mn(CH_3COO)_2$, FeC_2O_4, LiH_2PO_4	N/A
Failed			
Li_3CoTeO_6 [153]	$LiCoO_2$ [154]	Co, Te, Li_2CO_3	Co_3O_4, TeO_2, LiOH
$Sr_4Al_6SO_{16}$ [155]	$SrAl_2O_4$ [156]	$SrCO_3$, $SrSO_4$, $Al(OH)_3$	$SrCO_3$, H_2SO_4, $Al(OH)_3$
$Ca_{7.5}Ba_{1.5}Bi(VO_4)_7$ [157]	$Bi_3Ca_9V_{11}O_{41}$ [158]	$BaCO_3$, NH_4VO_3, $CaCO_3$, Bi_2O_3	BaO, NH_4VO_3, $CaCO_3$, Bi_2O_3

Because of this customized synthesis similarity of materials and our precursor recommendation pipeline, we are able to not only recommend trivial solutions for target synthesis such as the use of common precursors, but also deal with more challenging situations. One typical scenario is the adoption of uncommon precursors. For example, Lalère et al. [141] used NaH_2PO_4 as the source of Na and P to synthesize $Na_3TiV(PO_4)_3$, while the common precursors for Na and P are Na_2CO_3 and $NH_4H_2PO_4$, respectively. It is not apparent to conclude from the composition of $Na_3TiV(PO_4)_3$ that the uncommon precursor NaH_2PO_4 is needed. However, the similarity-based recommendation pipeline successfully predicts the use of NaH_2PO_4 by referring to a similar material $Na_3V_2(PO_4)_3$ [142]. A plausible reason for the choice of NaH_2PO_4 for $Na_3TiV(PO_4)_3$ can also be inferred from the synthesis of $Na_3V_2(PO_4)_3$. Feng et al. [142] reported that NaH_2PO_4 was used to implement a one-pot solid-state synthesis of $Na_3V_2(PO_4)_3$, while Fang el al. [159] reported that a reductive agent and additional complex operations are needed when using Na_2CO_3 and $NH_4H_2PO_4$. Similar outcomes may also apply to the synthesis of $Na_3TiV(PO_4)_3$. A second example is the successful precursor recommendation for target compound $GdLu(MoO_4)_3$. Instead of the common precursor MoO_3, an uncommon precursor $(NH_4)_6Mo_7O_{24}$ was adopted as the Mo source [143]. The use of $(NH_4)_6Mo_7O_{24}$ may facilitate the mixing of different ions in the synthesis of $GdLu(MoO_4)_3$. The adoption of uncommon precursors also provides clues in underexplored chemical spaces such as mixed-anion compounds [160]. Taking the pentanary oxynitride material $BaYSi_2O_5N$ [145] as an example, the five component system with multiple anions means many options that could potentially yield the target phase, including oxides, nitrides, carbonates, etc. Our recommendation pipeline correctly identifies that a combination of SiO_2 and Si_3N_4 facilitates the formation of $BaYSi_2O_5N$ by referring to a quaternary oxynitride material, $YSiO_2N$ [146]. Another challenging situation is that multiple precursors may be used for the same element. Usually, only one precursor is used for each metal/metalloid element in the target material, but exceptions do exist. For example, CuO and $CuCl_2$ were used as the Cu source in the synthesis of $Cu_3Yb(SeO_3)_2O_2Cl$ [147]. Through analogy to $Cu_4Se_5O_{12}Cl_2$ [148], the recommended precursor set includes both CuO and $CuCl_2$. Moreover, it is possible to predict multiple correct precursor sets by referring to multiple similar target materials. For example, two different sets of precursors for $LiMn_{0.5}Fe_{0.5}PO_4$ were reported by Zhuang et al. [149] and Wang et al. [150]. The recommendation pipeline predicts both by repurposing the precursor sets for $LiMn_{0.8}Fe_{0.2}PO_4$ [151] and $LiMn_{0.9}Fe_{0.1}PO_4$ [152].

The recommendation of precursors presented here is still imperfect. The engine we present is inherently limited by the knowledge base it is trained on, thereby biasing recommendations toward what has been done previously and lacking creativity for unprecedented combinations of precursors. For example, metals Co and Te were used in the synthesis of Li_3CoTeO_6 [153], but no similar materials in the knowledge base use the combination of Co and Te as precursors. Another example is that $SrCO_3$ and $SrSO_4$ were used in the synthesis of $Sr_4Al_6SO_{16}$ [155]. Although the recommendation pipeline is, in principle, able to predict multiple precursors for the same element, a similar case using both $SrCO_3$ and $SrSO_4$ as the

Sr source is not found in the knowledge base. Both examples end up being mispredictions. This situation could be improved when more data from text mining and high-throughput experiments [135] are added to the knowledge base. Furthermore, the success rate of the recommendation strategy may be underestimated in some cases. For example, BaO is predicted as the Ba source for synthesizing $Ca_{7.5}Ba_{1.5}Bi(VO_4)_7$, while $BaCO_3$ is used in the reported synthesis [157]. Given the slight difference between BaO and $BaCO_3$, BaO may actually be suitable.

Besides the prediction of precursors, the similarity-based recommendation framework is a potential step toward general synthesis prediction. The same strategy can be extended to the recommendation of more synthesis variables, such as operations, device setups, and experimental conditions, by adding corresponding prediction tasks to the downstream part of the multi-task network (Figure 4.4) for similarity measurement. For example, we may infer that reduced atmosphere is necessary for synthesizing $Na_3TiV(PO_4)_3$ [141] because it is used in the synthesis of a similar material $Na_3V_2(PO_4)_3$ [142]. Moreover, synthesis constraints such as the type of synthesis method, temperature, morphology of the target material, particle size, and cost can be added as conditions of synthesis prediction. For example, we may integrate our effort of synthesis temperature prediction [94] to prioritize the predicted precursors within expected temperature regime. Our automated algorithm, mimicking human design process for the synthesis of a new target, provides a practical solution to query decades of heuristic synthesis data in recommendation engines and autonomous laboratories.

4.6 Conclusion

We presented a similarity-based recommendation strategy for predictive solid-state synthesis of novel inorganic materials. Through representation learning based on synthesis information from a large knowledge base of 29,900 synthesis procedures, the similarity of target materials is quantified. Applying such similarity and a recommendation pipeline in the prediction of precursors, the observed precursor sets are among the top five proposed ones for 82% of 2,654 test target materials. Our quantitative recommendation pipeline can serve as a predictive tool to help experimental researchers rapidly plan materials synthesis for new compounds. It also provides meaningful initial solutions in the active learning and decision-making process for autonomous synthesis of inorganic materials.

4.7 Additional details of methods

Representation learning for similarity of materials

The neural network consists of an encoder part for encoding target materials and a task part for predicting variables related to precursor selection. The encoder part f is a three-layer fully connected submodel transforming the composition of the target material $\boldsymbol{x}$ into

a 32-dimensional target vector $\boldsymbol{u} = f(\boldsymbol{x})$. The input composition is an array with 83 units showing the fraction of each element. The reduced dimension of the encoded target vector is inspired by the bottleneck architecture of autoencoders [161]. By limiting the dimension of the encoded vector, the network is forced to learn a more compact and efficient representation of the input data, which is more appropriate for the precursor selection-related downstream tasks [162]. The task part uses different network architectures for different tasks of prediction, including precursor completion and composition recovery in this work. The masked precursor completion (MPC) task replaces part of the precursors with `[MASK]` at random and uses the remaining precursors as a condition to predict the complete precursor set for the target material, which is formulated as a multi-label classification problem [163]. An attention block g_{proj} [42] is used to aggregate the target vector and the vectors for conditional precursors as a projected vector $\boldsymbol{v} = g_{proj}(\boldsymbol{u}; \boldsymbol{p}_1, \boldsymbol{p}_2, \dots)$ with dimensionality of 32. Then, $\boldsymbol{v}$ is passed to the precursor classification layer represented by a 417×32 matrix $\boldsymbol{P}$, of which each row is the 32-dimensional vector representation of a potentially used precursor $\boldsymbol{p}_i$. To avoid having too many neural network weights to learn, the precursor completion task only considers 417 precursors used in at least five reactions in the knowledge base. The probability to use each precursor is indicated by $\text{sigmoid}(\boldsymbol{p}_i^\top \boldsymbol{v})$, allowing non-exclusive prediction of multiple precursors [163]. Here, $\boldsymbol{v}$ acts as a probe corresponding to the target material projected in the precursor space and is used to search for $\boldsymbol{p}_i$'s with similar vector representations via a dot product. The conditional precursors input to g_{proj} share the same trainable vector representations as $\boldsymbol{p}_i$'s. Circle loss [164] is used because of its benefits in capturing the dependency between different labels in multi-label classification and deep feature learning. The composition recovery task is a two-layer fully connected submodel decoding back to the chemical composition $\boldsymbol{x}$ from the target vector $\boldsymbol{u}$, similar to the mechanism of autoencoders [161, 165]. Mean squared error loss is used because it is the most popular for regression. More tasks predicting other synthesis variables such as operations and conditions can be appended in a similar fashion. To combine the loss functions in this multi-task neural network, an adaptive loss [166] is used to automatically weigh different loss by considering the homoscedastic uncertainty of each task.

Baseline models

"Most frequent". This baseline model ranks precursor sets based on an empirical joint probability without considering the dependency of precursors (Figure 4.2). Assuming that the choices of precursors are independent from each other, the joint probability of selecting a specific set of precursors can be estimated as the product of their marginal probabilities. For each metal/metalloid element, different precursors can be used as the source. The marginal probability to use a precursor is estimated as the relative frequency of using that precursor over all precursors contributing the same metal/metalloid element. For example, the precursor set ranked in first place is always the combination of common precursors for each metal/metalloid element in the target material, which is also typically the first attempt in the lab.

"Magpie encoding". This baseline model uses the same recommendation strategy as Figure 4.1, except that the similarity is calculated using Magpie encoding [121, 122]. The composition of each target material is converted into a vector consisting of 132 statistical quantities such as the average and standard deviation of various elemental properties. The cosine similarity is used, as shown in Eq. 4.1. When the precursors from the reference target material cannot cover all elements of the novel target, the common precursors for the missing elements are supplemented because MPC (Figure 4.4B) is only trained for PrecursorSelector encoding.

"FastText encoding". Similar to the baseline of "Magpie encoding", this baseline model uses the same recommendation strategy as Figure 4.1, except that the similarity is calculated using FastText encoding [24]. The formula of each target material is converted into a 100-dimensional vector using the FastText model trained with materials science papers [24]. The total number of target materials tested in this baseline model is 1,985 instead of 2,654 because some n-grams such as certain float numbers corresponding to the amount of elements are not in the vocabulary.

"Raw composition". Similar to the baseline of "Magpie encoding", this baseline model uses the same recommendation strategy as Figure 4.1, except that the similarity is calculated using the cosine similarity of raw material composition. The formula of each target material is converted into an 83-dimensional vector corresponding to the fraction of each element.

Data preparation

In total, 33,343 inorganic solid-state synthesis recipes extracted from 24,304 materials science papers [30] were used in this work. Because some material strings (e.g., $Ba_{1-x}\,Sr_x\,TiO_3$) extracted from the literature contain variables corresponding to different amounts of elements, we substituted these variables with their values from the text to ensure that a material in any reaction only corresponds to one composition, resulting in 49,924 expanded reactions and 28,598 target materials. An ideal test for generalizability and applicability of this method would be to synthesize many entirely new materials using recommended precursors. In the absence of performing extensive new synthesis experiments, we designed a robust test to simulate precursor recommendation for target materials that are new to the trained model. We split the data based on the year of publication, i.e., training set (or knowledge base) for reactions published by 2014, validation set for reactions in 2015 and 2016, and test set for reactions from 2017 to 2020. In addition, to avoid data leakage where the synthesis of the same material can be reported again in a more recent year, we placed reactions for target materials with the same prototype formula in the same data set as the earliest record. The prototype formula was defined as the formula corresponding to a family of materials including (1) the formula itself, (2) formulas derived from a small amount (< 0.3) of substitution (e.g., $Ca_{0.2}La_{0.8}MnO_3$ for prototype formula $LaMnO_3$), and (3) formulas able to be coarse-grained

by rounding the amount of elements to one decimal place (e.g., $Ba_{1.001}La_{0.004}TiO_3$ for the prototype formula $BaTiO_3$). In the end, the number of reactions in the training/validation/test set was 44,736/2,254/2,934 from 29,900/1,451/1,992 original recipes. The number of target materials in the training/validation/test set was 24,304/1,910/2,654, respectively.

Model training and validation

To train the PrecursorSelector encoding model, 44,736/2,254/2,934 synthesis reactions were used as the training/validation/test set as discussed in Section 4.7. Each reaction consists of a target material and multiple precursor materials extracted from the literature. For the purpose of training and validation, a random subset of precursor materials is selected to be replaced with a placeholder `[MASK]` [38] in each reaction, referred to as the masked reaction. Because a combinatorial number of masked reactions can be generated from the same reaction, the sampling space of masked reactions is much larger than that of original reactions. To sample as many different masked reactions during the training phase, we employ a dynamic masking strategy [167] that randomly samples a batch of reactions and re-generates the masking pattern in every training step. Different from the training samples, the validation samples are generated using static masking during data pre-processing because keeping the validation set unchanged is necessary for model selection afterwards. In this work, we trained with a batch size of 8 masked reactions for 500,000 steps, or 50 epochs with 10,000 steps per epoch. The optimizer used was Adam [168] with learning rate of 5×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Starting from the 2,254 original reactions in the validation set, we applied the masking procedure and randomly sampled 3,320 masked reactions for validation. The optimal model selected was the one with minimal loss on the validation samples to minimize overfitting [61].

Two tasks are implemented in the representation learning model: (1) the masked precursor completion (MPC) task that predicts the complete precursor set based on the target material and the synthesis context provided by the unmasked precursors, and (2) the composition recovery task that predicts the chemical composition of the target material from the encoded target vector. The loss function in the MPC task, denoted as L_1, is the circle loss [164] to maximize the within-class similarity and minimize the between-class similarity in multi-label classification. Here, the within-class similarity corresponds to the similarity of precursor materials present in the same reaction, while the between-class similarity corresponds to the similarity between used and unused precursor materials. The loss function in the composition recovery task, denoted as L_2, is the mean squared error (MSE) loss to compare the difference between predicted composition and the real composition of the target material. The total loss, denoted as L_{multi}, is an adaptive multi-task loss [166] to automatically weigh L_1 and L_2 with

$$L_{multi} = \frac{1}{\sigma_1^2} L_1 + \frac{1}{\sigma_2^2} L_2 + \log \sigma_1^2 + \log \sigma_2^2, \tag{4.2}$$

where σ_1 and σ_2 are the model's observation noise parameters which are learned alongside

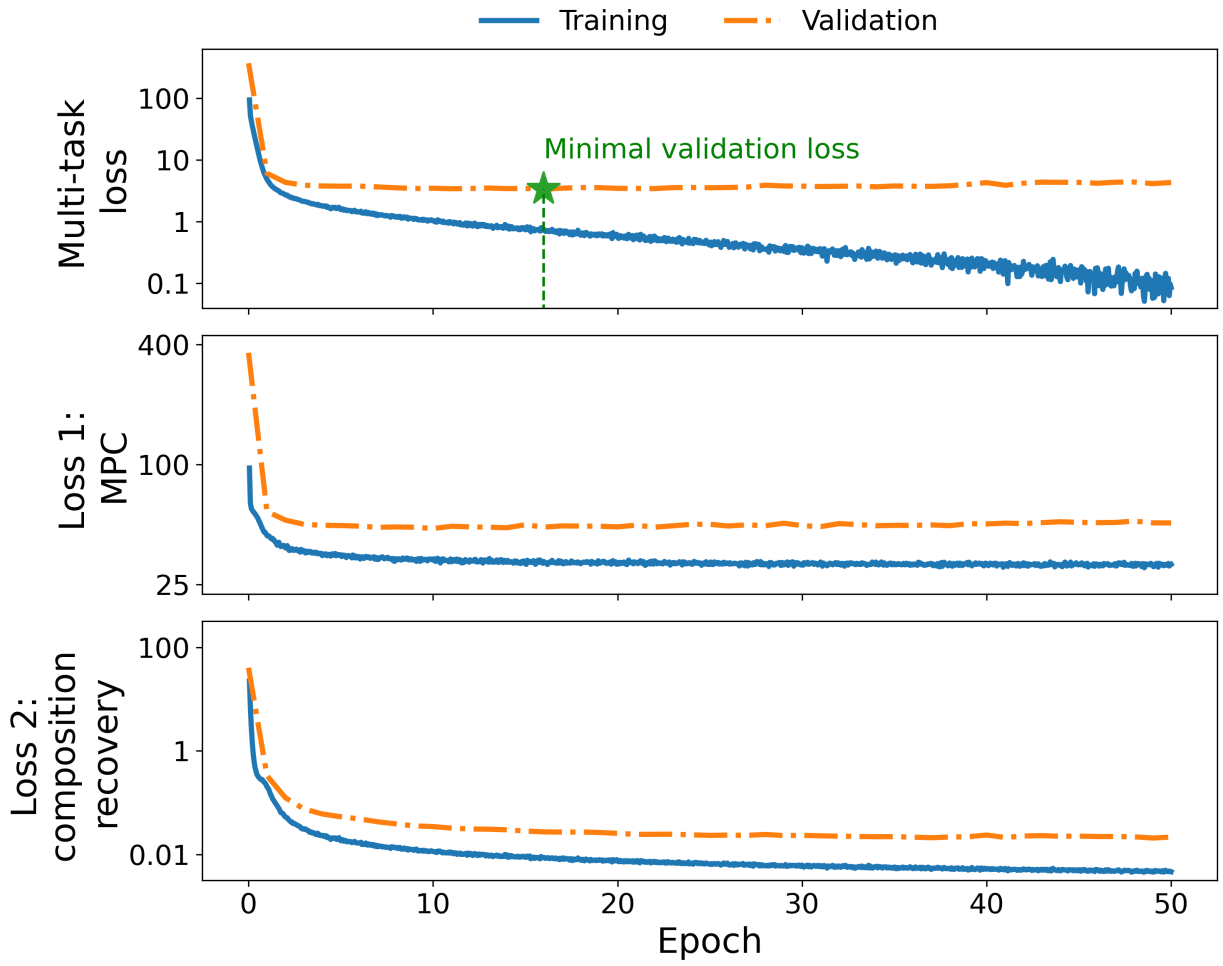

Figure 4.7: Evolution of training and validation loss while training the PrecursorSelector encoding model. Top: total multi-task loss, L_{multi}. Middle: loss for the MPC task, L_1. Bottom: loss for the composition recovery task, L_2.

other model parameters. The training and validation losses for each task and the total are shown in Figure 4.7. The training loss is averaged every 500 training steps to estimate performance on a substantial number of training samples, which in this study is 4,000. The validation loss is evaluated before training and at the end of each training epoch. As the training loss continues to decrease, the validation loss initially decreases and then increases. The minimal total validation loss is achieved at the end of 15^{th} epoch, leading to the optimal model. The final performance of the optimal model is tested by predicting precursors for 2,654 unseen target materials in the test set. Our similarity-based recommendation strategy successfully reproduces a known precursor set 82% of the time in five attempts or less.

Computation time for similarity evaluation

In this work, all 24,034 materials in the knowledge base are converted to 32-dimensional vectors in advance, forming a $24,034 \times 32$ matrix. For the 2,654 test materials, we monitored the time required to vectorize them one by one and calculate their cosine similarity to the target vectors in the pre-stored matrix. The similarity evaluation took merely 26 seconds for all the test materials (i.e. 0.01 seconds/material) because of the fast matrix multiplication.

Chapter 5

Conclusions and outlook

5.1 Conclusions

In conclusion, this dissertation employs interdisciplinary methods that combine materials science, NLP, and machine learning to tackle the challenge of predictive solid-state synthesis. Two major accomplishments include: (1) we have developed core NLP algorithms and an automated text-mining pipeline to extract large volumes of structured inorganic synthesis data from material science literature, and (2) we have developed successful precursor recommendation algorithms for predictive synthesis by machine learning materials similarity from the text-mined synthesis dataset. This work represents a step forward in the prediction of solid-state synthesis.

In Chapter 2, we successfully applied text mining to the field of inorganic solid-state synthesis. We developed a two-step SMR model based on Bi-LSTM to extract the precursor and target materials from textual descriptions of synthesis experiments. The F_1 scores for the extraction of precursors and targets are 90.0% and 84.5%, respectively. Through comparison with baseline models, we demonstrated the challenges and solutions for building information retrieval models in the context of inorganic synthesis. By integrating both the SMR model and other in-house NLP tools, we developed a fully automated text-mining pipeline for inorganic materials synthesis science. Starting from 4,973,165 materials science papers, we applied our text-mining pipeline and successfully extracted 33,343 solid-state synthesis recipes. The quality of the text-mined synthesis dataset is validated by the high accuracy of 93% at the chemistry level (correct precursors, targets, and reactions). The usage of the text-mined synthesis dataset was exemplified by exploratory data analysis with various aspects, including coverage of chemical space, synthesis temperature, synthesis routes, synthesis time, reaction energy, and the application to a specific chemical system. This dataset for inorganic solid-state synthesis is currently the largest of its kind and provides new opportunities for the development of data-driven approaches toward rational synthesis design.

In Chapter 3, we conducted a meta-analysis on the similarity of precursors for the creation of alternative recipes. Using the solid-state synthesis data extracted from materials science literature, we created a substitution model to quantify the probability of substituting one precursor with another while the target remains unchanged. By establishing distance metrics from the substitution model and the distribution of synthesis temperature, we proposed a multi-feature distance metric to characterize the similarity of precursors. Through hierarchical clustering based on the similarity of precursors, we demonstrated that "chemical similarity" in solid-state synthesis can be captured from text mining, without the need to include any explicit domain knowledge. The similarity could help guide the selection of precursors when researchers alter existing recipes by replacing precursors. The predictive power of this precursor substitution is validated by the higher true positive rate (TPR) than the false positive rate (FPR) in a large-scale cross-validation. This quantitative similarity metric offers a reference to rank precursor candidates and provides a foundation for developing a predictive synthesis model.

In Chapter 4, we studied similarity of targets to enable precursor selection for new materials. The similarity of precursors is limited to creating alternative recipes for existing target materials in Chapter 3. We stepped forward by developing a precursor recommendation strategy based similarity of target materials. This recommendation strategy mimics the human approach of referring the synthesis of a novel material to similar target materials for which successful synthesis reactions are known. Through representation learning based on synthesis information from our text-mined synthesis procedures, the similarity of target materials is quantified. Applying such similarity and our recommendation pipeline in the prediction of precursors, the observed precursor sets are among the top five proposed ones for 82% of 2,654 test target materials. Our quantitative recommendation pipeline can serve as a predictive tool to help experimental researchers rapidly plan materials synthesis for new compounds. It also provides meaningful initial solutions in the active learning and decision-making process for autonomous synthesis of inorganic materials.

5.2 Outlook

This work exemplifies a successful practice of applying text mining to inorganic synthesis science. However, because of time and energy constraints, many problems remain unsolved and many new research directions need to be explored. Here, we list a few of them.

First, the potential of text mining has not been fully exploited. In this work, we used research articles as the only data source. In fact, text mining can be applied to various data sources. Patents may serve as another rich and high-quality data source because (a) R&D institutions and companies are motivated to publish details of synthesis experiments to protect their intellectual property, and (b) patents are usually strictly examined by national patent offices before granted. By extending text mining to patents, the size of the extracted

dataset can be enlarged quickly. Furthermore, we only tried to extract data from textual descriptions. In addition to plain text, a research article also contains numerous valuable pieces of information in the form of tables and figures. Typically, tables and figures are well-formatted and present information more densely. By combining NLP and computer vision methods, it is possible to also extract data from tables and figures.

Next, large language models (LLMs) could ease the development of text-mining algorithms. In Chapter 2, we introduced an early attempt at fine-tuning BERT [38] for the extraction of precursor and target materials. Although the performance is not significantly better than our SMR model, we still value the potential of LLMs. When increasing the number of parameters of the neural network, emergent abilities [169] that are not present in small models appear in LLMs. Recently, several LLMs that are thousands of times larger than BERT, such as GPT-4, [49] have been announced. It is promising that the use of the most recent LLMs can further improve the accuracy of the text-mining pipeline because of the emergent abilities.

Finally, the predictive synthesis models are far from perfect. This work is focused on the problem of precursor selection, while many other synthesis variables also influence the synthesis outcome. Many a time, the solid-state synthesis may involve regrinding and reheating rather than a one-shot synthesis of "Shake 'n bake". There are also specific conditions for mixing, heating, cooling, etc. More importantly, these synthesis variables may be correlated with each other. For example, the selection of a precursor set may ask for a specific heating temperature. More work needs to be done to predict these synthesis variables for more thorough synthesis suggestions. In addition, the machine learning approaches used in this work try to suggest synthesis procedures without revealing the mechanisms underlying solid-state synthesis. Physics-based machine learning models are wanted to integrate the possible synthesis mechanisms.

Bibliography

1. Holden, J. *Materials Genome Initiative for global competitiveness* tech. rep. (National Science and Technology Council, 2011). `https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf`.
2. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **1,** 011002 (2013).
3. Hemminger, J. C., Sarrao, J., Crabtree, G., Flemming, G. & Ratner, M. *Challenges at the frontiers of matter and energy: Transformative opportunities for discovery science* tech. rep. (USDOE Office of Science (SC)(United States), 2015).
4. Grover, V., Mandal, B. P. & Tyagi, A. Solid State Synthesis of Materials. *Handbook on Synthesis Strategies for Advanced Materials: Volume-I: Techniques and Fundamentals,* 1–49 (2021).
5. Miura, A., Bartel, C. J., Goto, Y., Mizuguchi, Y., Moriyoshi, C., Kuroiwa, Y., Wang, Y., Yaguchi, T., Shirai, M., Nagao, M., *et al.* Observing and Modeling the Sequential Pairwise Reactions that Drive Solid-State Ceramic Synthesis. *Advanced Materials* **33,** 2100312 (2021).
6. Bianchini, M., Wang, J., Clément, R. J., Ouyang, B., Xiao, P., Kitchaev, D., Shi, T., Zhang, Y., Wang, Y., Kim, H., *et al.* The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature materials* **19,** 1088–1095 (2020).
7. Jiang, Z., Ramanathan, A. & Shoemaker, D. P. In situ identification of kinetic factors that expedite inorganic crystal formation and discovery. *Journal of Materials Chemistry C* **5,** 5709–5717 (2017).
8. Corey, E. J. *The logic of chemical synthesis* (John Wiley, 1991).
9. Corey, E. Robert Robinson lecture. Retrosynthetic thinking—essentials and examples. *Chemical society reviews* **17,** 111–133 (1988).
10. Stein, A., Keller, S. W. & Mallouk, T. E. Turning down the heat: Design and mechanism in solid-state synthesis. *Science* **259,** 1558–1564 (1993).

11. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555,** 604–610 (2018).

12. Chamorro, J. R. & McQueen, T. M. Progress toward solid state synthesis by design. *Accounts of chemical research* **51,** 2918–2925 (2018).

13. Kohlmann, H. Looking into the Black Box of Solid-State Synthesis. *European Journal of Inorganic Chemistry* **2019,** 4174–4180 (2019).

14. Schäfer, H. Preparative solid state chemistry: the present position. *Angewandte Chemie International Edition in English* **10,** 43–50 (1971).

15. Martinolich, A. J. & Neilson, J. R. Toward reaction-by-design: achieving kinetic control of solid state chemistry with metathesis. *Chemistry of Materials* **29,** 479–489 (2017).

16. Sun, W., Dacek, S. T., Ong, S. P., Hautier, G., Jain, A., Richards, W. D., Gamst, A. C., Persson, K. A. & Ceder, G. The thermodynamic scale of inorganic crystalline metastability. *Science advances* **2,** e1600225 (2016).

17. Sun, W., Jayaraman, S., Chen, W., Persson, K. A. & Ceder, G. Nucleation of metastable aragonite CaCO3 in seawater. *Proceedings of the National Academy of Sciences* **112,** 3199–3204 (2015).

18. Chen, B.-R., Sun, W., Kitchaev, D. A., Mangum, J. S., Thampy, V., Garten, L. M., Ginley, D. S., Gorman, B. P., Stone, K. H., Ceder, G., *et al.* Understanding crystallization pathways leading to manganese oxide polymorph formation. *Nature communications* **9,** 2553 (2018).

19. Sun, W., Holder, A., Orvañanos, B., Arca, E., Zakutayev, A., Lany, S. & Ceder, G. Thermodynamic routes to novel metastable nitrogen-rich nitrides. *Chemistry of Materials* **29,** 6936–6946 (2017).

20. Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *Journal of the American Chemical Society* **143,** 9244–9259 (2021).

21. McDermott, c. J., Dwaraknath, S. S. & Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nature communications* **12,** 1–12 (2021).

22. Raccuglia, P., Elbert, K. C., Adler, P. D., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J. & Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533,** 73–76 (2016).

23. Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G. & Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials* **29,** 9436–9444 (2017).

24. Kim, E., Jensen, Z., van Grootel, A., Huang, K., Staib, M., Mysore, S., Chang, H.-S., Strubell, E., McCallum, A., Jegelka, S., *et al.* Inorganic materials synthesis planning with literature-trained neural networks. *Journal of chemical information and modeling* **60,** 1194–1201 (2020).

25. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **3,** 1237–1245 (2017).

26. Feng, F., Lai, L. & Pei, J. Computational chemical synthesis analysis and pathway design. *Frontiers in chemistry* **6,** 199 (2018).

27. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS central science* **2,** 725–732 (2016).

28. *Reaxys* `https://www.reaxys.com/`. Accessed: 2022-07-16.

29. He, T., Sun, W., Huo, H., Kononova, O., Rong, Z., Tshitoyan, V., Botari, T. & Ceder, G. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chemistry of Materials* **32,** 7861–7873 (2020).

30. Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V. & Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data* **6,** 1–11 (2019).

31. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications* **82,** 3713–3744 (2023).

32. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* **13,** 55–75 (2018).

33. Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chemical reviews* **117,** 7673–7761 (2017).

34. Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y.-J. & Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews* **7,** 041317 (2020).

35. Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A. & Ceder, G. Opportunities and challenges of text mining in materials research. *Iscience* **24,** 102155 (2021).

36. Bengio, Y., Ducharme, R. & Vincent, P. A neural probabilistic language model. *Advances in neural information processing systems* **13** (2000).

37. Collobert, R. & Weston, J. *A unified architecture for natural language processing: Deep neural networks with multitask learning* in *Proceedings of the 25th international conference on Machine learning* (2008), 160–167.

38. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
39. Lopez, M. M. & Kalita, J. Deep Learning applied to NLP. *arXiv preprint arXiv:1703.03091* (2017).
40. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9,** 1735–1780 (1997).
41. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
43. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **26** (2013).
44. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
45. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1,** 9 (2019).
46. Liu, Q., Kusner, M. J. & Blunsom, P. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278* (2020).
47. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv: 210713586 (2021).
48. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.* Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35,** 27730–27744 (2022).
49. OpenAI. GPT-4 Technical Report. *arXiv* (2023).
50. Young, S. R., Maksov, A., Ziatdinov, M., Cao, Y., Burch, M., Balachandran, J., Li, L., Somnath, S., Patton, R. M., Kalinin, S. V., *et al.* Data mining for better material synthesis: The case of pulsed laser deposition of complex oxides. *J. Appl. Phys.* **123,** 115303 (2018).
51. Court, C. & Cole, J. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. *npj Comput. Mater* **6,** 1–9 (2020).
52. Court, C. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5,** 180111 (2018).

53. Jensen, Z., Kim, E., Kwon, S., Gani, T., Roman-Leshkov, Y., Moliner, M., Corma, A. & Olivetti, E. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5,** 892–899 (2019).

54. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. & Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571,** 95–98 (2019).

55. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., *et al.* The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.* **7,** S2 (2015).

56. Imanaka, N., Kamikawa, M., Tamura, S. & Adachi, G. Carbon dioxide gas sensing with the combination of trivalent Sc3+ ion conducting Sc2 (WO4) 3 and O2- ion conducting stabilized zirconia solid electrolytes. *Solid State Ionics* **133,** 279–285 (2000).

57. Lafferty, J., McCallum, A. & Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

58. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

59. Narayanan, S. & Thangadurai, V. Effect of Y substitution for Nb in Li5La3Nb2O12 on Li ion conductivity of garnet-type solid electrolytes. *Journal of Power Sources* **196,** 8085–8090 (2011).

60. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling* **56,** 1894–1904 (2016).

61. Prechelt, L. in *Neural Networks: Tricks of the trade* 55–69 (Springer, 2002).

62. Bird, S., Loper, E. & Klein, E. *Natural Language Processing with Python* (O'Reilly Media Inc., 2009).

63. Honnibal, M. & Johnson, M. *An Improved Non-monotonic Transition System for Dependency Parsing* in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Lisbon, Portugal, 2015), 1373–1378.

64. Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L. & Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminform.* **3,** 41 (2011).

65. Leaman, R., Wei, C.-H. & Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminform.* **7,** S3 (2015).

66. Rocktäschel, T., Weidlich, M. & Leser, U. Chemspot: A hybrid system for chemical named entity recognition. *Bioinformatics* **28,** 1633–1640 (2012).

67. Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M. & Todirascu, A. Multiword expression processing: A survey. *Computational Linguistics* **43,** 837–892 (2017).

68. Taslimipoor, S. & Rohanian, O. Shoma at parseme shared task on automatic identification of vmwes: Neural multiword expression tagging with high generalisation. *arXiv preprint arXiv:1809.03056* (2018).

69. Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Mititelu, V. B., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., *et al. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions* in *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (2018), 222–240.

70. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

71. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

72. Bota, P., Silva, J., Folgado, D. & Gamboa, H. A semi-automatic annotation approach for human activity recognition. *Sensors* **19,** 501 (2019).

73. Li, P., Wang, Z., Yang, Z., Guo, Q. & Li, X. Emission features of LiBaBO3: Sm3+ red phosphor for white LED. *Materials Letters* **63,** 751–753 (2009).

74. Luo, L., Zhou, L., Zou, X., Zheng, Q. & Lin, D. Structure, piezoelectric and multiferroic properties of Bi (Ni 0.5 Mn 0.5) O 3-modified BiFeO 3–BaTiO 3 ceramics. *Journal of Materials Science: Materials in Electronics* **26,** 9451–9462 (2015).

75. Paik, J.-H., Kim, S.-K., Lee, M.-J., Choi, B.-H., Lim, E.-K. & Nahm, S. Ordering structure of barium magnesium niobate ceramic with A-site substitution. *Journal of the European Ceramic Society* **26,** 2885–2888 (2006).

76. Leng, S., Zheng, L., Li, G., Zeng, J., Yin, Q., Xu, Z. & Chu, R. Impedance spectroscopy analysis for high-Tc BaTiO3-(Bi1/2Na1/2) TiO3 lead-free PTCR ceramics. *physica status solidi (a)* **208,** 1099–1104 (2011).

77. Hashimoto, D., Han, D. & Uda, T. Dependence of lattice constant of Ba, Co-contained perovskite oxides on atmosphere, and measurements of water content. *Solid State Ionics* **262,** 687–690 (2014).

78. Zajac, M. A., Zakrzewski, A. G., Kowal, M. G. & Narayan, S. A novel method of caffeine synthesis from uracil. *Synthetic communications* **33,** 3291–3297 (2003).

79. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **47,** D1102–D1109 (2019).

80. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

81. *Borges* `https://github.com/CederGroupHub/Borges`. Accessed: 2020-08-12. 2020.

82. Developers, S. *Scrapy: An open source and collaborative web crawling framework for Python* `https://github.com/scrapy/scrapy`. Accessed: 2023-03-20. 2008.

83. MongoDB, Inc. *MongoDB Manual* 4.4. Accessed: 2023-03-20. MongoDB, Inc. (2021). `https://docs.mongodb.com/manual/`.

84. *LimeSoup* `https://github.com/CederGroupHub/LimeSoup`. Accessed: 2020-08-12. 2020.

85. Jurafsky, D. & Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* ISBN: 9780131873216 (Pearson Prentice Hall, 2009).

86. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform.* **3,** 1–13 (2011).

87. Huo, H., Rong, Z., Kononova, O., Sun, W., Botari, T., He, T., Tshitoyan, V. & Ceder, G. Semi-supervised machine-learning classification of materials synthesis procedures. *Npj Computational Materials* **5,** 62 (2019).

88. Of California, U. *Why UC split with publishing giant Elsevier* Accessed: 2023-03-20. Mar. 2019. `https://www.universityofcalifornia.edu/news/why-uc-split-publishing-giant-elsevier`.

89. *MaterialParser* `https://github.com/CederGroupHub/MaterialParser`. Accessed: 2020-08-12. 2020.

90. Wang, Z., Cruse, K., Fei, Y., Chia, A., Zeng, Y., Huo, H., He, T., Deng, B., Kononova, O. & Ceder, G. ULSA: Unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discovery* **1,** 313–324 (2022).

91. Bartel, C. J., Millican, S. L., Deml, A. M., Rumptz, J. R., Tumas, W., Weimer, A. W., Lany, S., Stevanović, V., Musgrave, C. B. & Holder, A. M. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nature communications* **9,** 4168 (2018).

92. *FREED-Thermodynamic Database* `https://www.thermart.net/freed-thermodynamic-database/`. Accessed: 2022-07-16.

93. *NIST Chemistry WebBook* `https://webbook.NIST.gov/chemistry/`. Accessed: 2022-07-16.

94. Huo, H., Bartel, C. J., He, T., Trewartha, A., Dunn, A., Ouyang, B., Jain, A. & Ceder, G. Machine-learning rationalization and prediction of solid-state synthesis conditions. *Chemistry of Materials* **34,** 7323–7336 (2022).

95. Jin, E. M., Jin, B., Jeon, Y.-S., Park, K.-H. & Gu, H.-B. Electrochemical properties of LiMnO2 for lithium polymer battery. *Journal of Power Sources* **189,** 620–623 (2009).

96. Jang, Y.-I., Moorehead, W. D. & Chiang, Y.-M. Synthesis of the monoclinic and orthorhombic phases of LiMnO2 in oxidizing atmosphere. *Solid State Ionics* **149,** 201–207 (2002).

97. Idemoto, Y., Mochizuki, T., Ui, K. & Koura, N. Properties, crystal structure, and performance of o-LiMnO2 as cathode material for Li secondary batteries. *Journal of The Electrochemical Society* **153,** A418 (2006).

98. Bessekhouad, Y., Gabes, Y., Bouguelia, A. & Trari, M. The physical and photo electrochemical characterization of the crednerite CuMnO 2. *Journal of materials science* **42,** 6469–6476 (2007).

99. He, T., Huo, H., Bartel, C. J., Wang, Z., Cruse, K. & Ceder, G. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science Advances.* In press (2023).

100. Hautier, G., Fischer, C., Ehrlacher, V., Jain, A. & Ceder, G. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry* **50,** 656–663 (2011).

101. Yang, L. & Ceder, G. Data-mined similarity function between material compositions. *Physical Review B* **88,** 224107 (2013).

102. West, A. R. *Solid state chemistry and its applications* (John Wiley & Sons, 2022).

103. Zhao, H., Li, F., Liu, X., Xiong, W., Chen, B., Shao, H., Que, D., Zhang, Z. & Wu, Y. A simple, low-cost and eco-friendly approach to synthesize single-crystalline LiMn2O4 nanorods with high electrochemical performance for lithium-ion batteries. *Electrochimica Acta* **166,** 124–133 (2015).

104. Liu, S. Z., Wang, T. X. & Yang, L. Y. Low temperature preparation of nanocrystalline SrTiO3 and BaTiO3 from alkaline earth nitrates and TiO2 nanocrystals. *Powder technology* **212,** 378–381 (2011).

105. Mercury, J. R., De Aza, A., Turrillas, X. & Pena, P. The synthesis mechanism of Ca3Al2O6 from soft mechanochemically activated precursors studied by time-resolved neutron diffraction up to 1000 C. *Journal of Solid State Chemistry* **177,** 866–874 (2004).

106. Varshney, D., Mansuri, I., Kaurav, N., Lung, W. & Kuo, Y. Influence of Ce doping on electrical and thermal properties of La0. 7- xCexCa0. 3MnO3 ($0.0 \leq x \leq 0.7$) manganites. *Journal of magnetism and magnetic materials* **324,** 3276–3285 (2012).

107. Bhattacharya, S., Pal, S., Mukherjee, R., Chaudhuri, B., Neeleshwar, S., Chen, Y., Mollah, S. & Yang, H. Development of pulsed magnetic field and study of magnetotransport properties of K-doped La1–xCax–yKyMnO3 CMR materials. *Journal of magnetism and magnetic materials* **269,** 359–371 (2004).

108. Fajans, K. & Barber, S. W. Properties and Structures of Vitreous and Crystalline Boron Oxide1. *Journal of the American Chemical Society* **74,** 2761–2768 (1952).

109. Ferlat, G., Seitsonen, A. P., Lazzeri, M. & Mauri, F. Hidden polymorphs drive vitrification in B2O3. *Nature materials* **11,** 925–929 (2012).

110. Sakthipandi, K. & Rajendran, V. Metal insulator transition of bulk and nanocrystalline La1- xCaxMnO3 perovskite manganite materials through in-situ ultrasonic measurements. *Materials characterization* **77,** 70–80 (2013).

111. Selvadurai, A. P. B., Pazhanivelu, V. & Murugaraj, R. Strain correlated effect on structural, magnetic, and dielectric properties in Ti4+ substituted Bi0. 8Ba0. 2Fe1-xTixO3. *Solid State Sciences* **46,** 71–79 (2015).

112. Liu, Y. & Chen, D. Protective coatings for Cr2O3-forming interconnects of solid oxide fuel cells. *international journal of hydrogen energy* **34,** 9220–9226 (2009).

113. Garskaite, E., Gibson, K., Leleckaite, A., Glaser, J., Niznansky, D., Kareiva, A. & Meyer, H.-J. On the synthesis and characterization of iron-containing garnets (Y3Fe5 O12, YIG and Fe3Al5O12, IAG). *Chemical physics* **323,** 204–210 (2006).

114. Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A. & Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319 (2013).

115. Sun, W., Bartel, C. J., Arca, E., Bauers, S. R., Matthews, B., Orvañanos, B., Chen, B.-R., Toney, M. F., Schelhas, L. T., Tumas, W., *et al.* A map of the inorganic ternary metal nitrides. *Nature materials* **18,** 732–739 (2019).

116. Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics,* 857–871 (1971).

117. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58,** 236–244 (1963).

118. *Strem Chemicals* Accessed: 2019-09-18. `https://www.strem.com/`.

119. Jia, X., Lynch, A., Huang, Y., Danielson, M., Lang'at, I., Milder, A., Ruby, A. E., Wang, H., Friedler, S. A., Norquist, A. J., *et al.* Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573,** 251–255 (2019).

120. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50,** 742–754 (2010).

121. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2,** 1–7 (2016).

122. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152,** 60–69 (2018).

123. Goodall, R. E. & Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature communications* **11,** 1–9 (2020).

124. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials* **7,** 1–10 (2021).

125. Pei, Z., Yin, J., Liaw, P. K. & Raabe, D. Toward the design of ultrahigh-entropy alloys via mining six million texts. *Nature Communications* **14,** 54 (2023).

126. Mao, Z.-y., Zhu, Y.-c., Fei, Q.-n. & Wang, D.-j. Investigation of 515 nm green-light emission for full color emission LaAlO3 phosphor with varied valence Eu. *Journal of luminescence* **131,** 1048–1051 (2011).

127. Mendoza-Mendoza, E., Padmasree, K. P., Montemayor, S. M. & Fuentes, A. F. Molten salts synthesis and electrical properties of Sr-and/or Mg-doped perovskite-type LaAlO3 powders. *Journal of Materials Science* **47,** 6076–6085 (2012).

128. Wang, Z., Kononova, O., Cruse, K., He, T., Huo, H., Fei, Y., Zeng, Y., Sun, Y., Cai, Z., Sun, W., *et al.* Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Scientific Data* **9,** 1–11 (2022).

129. Cruse, K., Trewartha, A., Lee, S., Wang, Z., Huo, H., He, T., Kononova, O., Jain, A. & Ceder, G. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. *Scientific Data* **9,** 234 (2022).

130. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2,** 559–572 (1901).

131. Zvereva, E., Savelieva, O., Titov, Y. D., Evstigneeva, M., Nalbandyan, V., Kao, C., Lin, J.-Y., Presniakov, I., Sobolev, A., Ibragimov, S., *et al.* A new layered triangular antiferromagnet Li 4 FeSbO 6: Spin order, field-induced transitions and anomalous critical behavior. *Dalton Transactions* **42,** 1550–1566 (2013).

132. Luan, J., Zhang, L., Ma, K., Li, Y. & Zou, Z. Preparation and property characterization of new Y2FeSbO7 and In2FeSbO7 photocatalysts. *Solid state sciences* **13,** 185–194 (2011).

133. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* **5,** 135–146 (2017).

134. Lv, S., Shanmugavelu, B., Wang, Y., Mao, Q., Zhao, Y., Yu, Y., Hao, J., Zhang, Q., Qiu, J. & Zhou, S. Transition metal doped smart glass with pressure and temperature sensitive luminescence. *Advanced Optical Materials* **6,** 1800881 (2018).

135. Szymanski, N. J., Zeng, Y., Huo, H., Bartel, C. J., Kim, H. & Ceder, G. Toward autonomous design and synthesis of novel inorganic materials. *Materials Horizons* **8,** 2169–2198 (2021).

136. Peng, H., Luan, X., Li, L., Zhang, Y. & Zou, Y. Synthesis and ion conductivity of Li7La3Nb2O13 ceramics with cubic garnet-type structure. *Journal of The Electrochemical Society* **164,** A1192 (2017).

137. Van Wüllen, L., Echelmeyer, T., Meyer, H.-W. & Wilmer, D. The mechanism of Li-ion transport in the garnet Li 5 La 3 Nb 2 O 12. *Physical Chemistry Chemical Physics* **9,** 3298–3303 (2007).

138. Park, K., Lim, H., Park, S., Deressa, G. & Kim, J. Strong blue absorption of green Zn2SiO4: Mn2+ phosphor by doping heavy Mn2+ concentrations. *Chemical Physics Letters* **636,** 141–145 (2015).

139. Chen, C., Zhuang, Y., Tu, D., Wang, X., Pan, C. & Xie, R.-J. Creating visible-to-near-infrared mechanoluminescence in mixed-anion compounds SrZn2S2O and SrZnSO. *Nano Energy* **68,** 104329 (2020).

140. Duan, C., Delsing, A. & Hintzen, H. Photoluminescence properties of novel red-emitting Mn2+-activated MZnOS (M= Ca, Ba) phosphors. *Chemistry of Materials* **21,** 1010–1016 (2009).

141. Lalère, F., Seznec, V., Courty, M., Chotard, J. & Masquelier, C. Coupled X-ray diffraction and electrochemical studies of the mixed Ti/V-containing NASICON: Na 2 TiV (PO 4) 3. *Journal of Materials Chemistry A* **6,** 6654–6659 (2018).

142. Feng, P., Wang, W., Wang, K., Cheng, S. & Jiang, K. Na 3 V 2 (PO 4) 3/C synthesized by a facile solid-phase method assisted with agarose as a high-performance cathode for sodium-ion batteries. *Journal of Materials Chemistry A* **5,** 10261–10268 (2017).

143. Wang, B., Li, X., Zeng, Q., Yang, G., Luo, J., He, X. & Chen, Y. Efficiently enhanced photoluminescence in Eu3+-doped Lu2 (MoO4) 3 by Gd3+ substituting. *Materials Research Bulletin* **100,** 97–101 (2018).

144. Thirumalai, J., Krishnan, R., Shameem Banu, I. & Chandramohan, R. Controlled synthesis, formation mechanism and lumincence properties of novel 3-dimensional Gd 2 (MoO 4) 3: Eu 3+ nanostructures. *Journal of Materials Science: Materials in Electronics* **24,** 253–259 (2013).

145. Yasunaga, T., Kobayashi, M., Hongo, K., Fujii, K., Yamamoto, S., Maezono, R., Yashima, M., Mitsuishi, M., Kato, H. & Kakihana, M. Synthesis of Ba1- xSrxYSi2O5N and discussion based on structure analysis and DFT calculation. *Journal of Solid State Chemistry* **276,** 266–271 (2019).

146. Kitagawa, Y., Ueda, J., Brik, M. G. & Tanabe, S. Intense hypersensitive luminescence of Eu3+-doped YSiO2N oxynitride with near-UV excitation. *Optical Materials* **83,** 111–117 (2018).

147. Markina, M., Zakharov, K., Ovchenkov, E., Berdonosov, P., Dolgikh, V., Kuznetsova, E., Olenev, A., Klimin, S., Kashchenko, M., Budkin, I., *et al.* Interplay of rare-earth and transition-metal subsystems in C u 3 Yb (Se O 3) 2 O 2 Cl. *Physical Review B* **96,** 134422 (2017).

148. Zhang, D., Berger, H., Kremer, R. K., Wulferding, D., Lemmens, P. & Johnsson, M. Synthesis, crystal structure, and magnetic properties of the copper selenite chloride Cu5 (SeO3) 4Cl2. *Inorganic Chemistry* **49,** 9683–9688 (2010).

149. Zhuang, H., Bao, Y., Nie, Y., Qian, Y., Deng, Y. & Chen, G. Synergistic effect of composite carbon source and simple pre-calcining process on significantly enhanced electrochemical performance of porous LiFe0. 5Mn0. 5PO4/C agglomerations. *Electrochimica Acta* **314,** 102–114 (2019).

150. Wang, L., Li, Y., Wu, J., Liang, F., Zhang, K., Xu, R., Wan, H., Dai, Y. & Yao, Y. Synthesis mechanism and characterization of LiMn0. 5Fe0. 5PO4/C composite cathode material for lithium-ion batteries. *Journal of Alloys and Compounds* **839,** 155653 (2020).

151. Zou, Q.-Q., Zhu, G.-N. & Xia, Y.-Y. Preparation of carbon-coated LiFe0. 2Mn0. 8PO4 cathode material and its application in a novel battery with Li4Ti5O12 anode. *Journal of Power Sources* **206,** 222–229 (2012).

152. Yi, H., Hu, C., Fang, H., Yang, B., Yao, Y., Ma, W. & Dai, Y. Optimized electrochemical performance of LiMn0. 9Fe0. 1- xMgxPO4/C for lithium ion batteries. *Electrochimica Acta* **56,** 4052–4057 (2011).

153. Heymann, G., Selb, E., Kogler, M., Götsch, T., Köck, E.-M., Penner, S., Tribus, M. & Janka, O. Li 3 Co 1.06 (1) TeO 6: synthesis, single-crystal structure and physical properties of a new tellurate compound with Co II/Co III mixed valence and orthogonally oriented Li-ion channels. *Dalton Transactions* **46,** 12663–12674 (2017).

154. Alcantara, R., Jumas, J., Lavela, P., Olivier-Fourcade, J., Pérez-Vicente, C. & Tirado, J. X-ray diffraction, 57Fe Mössbauer and step potential electrochemical spectroscopy study of LiFeyCo1- yO2 compounds. *Journal of power sources* **81,** 547–553 (1999).

155. Ndzila, J. S., Liu, S., Jing, G., Wu, J., Saruchera, L., Wang, S. & Ye, Z. Regulation of Fe3+-doped Sr4Al6SO16 crystalline structure. *Journal of Solid State Chemistry* **288,** 121415 (2020).

156. Zhu, Y., Zeng, J., Li, W., Xu, L., Guan, Q. & Liu, Y. Encapsulation of strontium aluminate phosphors to enhance water resistance and luminescence. *Applied surface science* **255,** 7580–7585 (2009).

157. Dorbakov, N. G., Titkov, V. V., Stefanovich, S. Y., Baryshnikova, O. V., Morozov, V. A., Belik, A. A. & Lazoryak, B. I. Barium-induced effects on structure and properties of β-Ca3 (PO4) 2-type Ca9Bi (VO4) 7. *Journal of Alloys and Compounds* **793,** 56–64 (2019).

158. Radosavljevic, I., Howard, J. A., Sleight, A. W. & Evans, J. S. Synthesis and structure of Bi3Ca9V11O41. *Journal of Materials Chemistry* **10,** 2091–2095 (2000).

159. Fang, Y., Xiao, L., Ai, X., Cao, Y. & Yang, H. Hierarchical carbon framework wrapped Na3V2 (PO4) 3 as a superior high-rate and extended lifespan cathode for sodium-ion batteries. *Advanced materials* **27,** 5895–5900 (2015).

160. Kageyama, H., Hayashi, K., Maeda, K., Attfield, J. P., Hiroi, Z., Rondinelli, J. M. & Poeppelmeier, K. R. Expanding frontiers in materials chemistry and physics with multiple anions. *Nature communications* **9,** 772 (2018).

161. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35,** 1798–1828 (2013).

162. Tschannen, M., Bachem, O. & Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069* (2018).

163. Herrera, F., Charte, F., Rivera, A. J. & Jesus, M. J. d. in *Multilabel Classification* 17–31 (Springer, 2016).

164. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z. & Wei, Y. Circle loss: A unified perspective of pair similarity optimization. *arXiv preprint arXiv:2002.10857* (2020).

165. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **313,** 504–507 (2006).

166. Kendall, A., Gal, Y. & Cipolla, R. *Multi-task learning using uncertainty to weigh losses for scene geometry and semantics* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 7482–7491.

167. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

168. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

169. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., *et al.* Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).