

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Data-driven solution-based synthesis design for inorganic materials

Permalink

<https://escholarship.org/uc/item/1gf0q1mf>

Author

Wang, Zheren

Publication Date

2022

Peer reviewed|Thesis/dissertation

Data-driven solution-based synthesis design for inorganic materials

by

Zheren Wang

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 Kristin Persson

Professor Kwabena Bediako

Fall 2022

Data-driven solution-based synthesis design for inorganic materials

Copyright 2022
by
Zheren Wang

Abstract

Data-driven solution-based synthesis design for inorganic materials

by

Zheren Wang

Doctor of Philosophy in Engineering - Materials Science and Engineering

University of California, Berkeley

Professor Gerbrand Ceder, Chair

The development of a materials synthesis route is usually based on heuristics and experience. This thesis proposes to apply data-driven approaches to learn the patterns of synthesis from past experience and use them to predict the syntheses of novel materials. However, this route is impeded by the lack of a large-scale database of synthesis formulations. Scientific publications represent the largest repository of knowledge about material synthesis and can be used as a reliable source of data. However, human-written descriptions of syntheses require additional levels of interpretation for conversion into a codified, machine-operable format. Therefore, this thesis aims to achieve two objectives: 1) constructing a text-mining pipeline that extracts synthesis datasets from scientific publications, and 2) validating a novel synthesis hypothesis, minimum thermodynamic competition, by the text-mined dataset and systematic synthesis experiments.

To fulfill the first objective, we need to build a text-mining pipeline to extract essential information from scientific publications. Extraction of synthesis information is challenging, especially for extracting synthesis actions, because of the lack of a comprehensive labeled dataset using a solid, robust, and well-established ontology for describing synthesis procedures. In order to extract synthesis actions (Chapter 2), we propose the first *unified language of synthesis actions* (ULSA) for describing inorganic synthesis procedures. We created a dataset of 3,040 synthesis

procedures annotated by domain experts according to the proposed ULSA scheme, and then built a neural network-based model to map arbitrary inorganic synthesis paragraphs into ULSA and used it to construct synthesis flowcharts for synthesis procedures.

We constructed the first large dataset of solution-based inorganic materials synthesis procedures by designing an advanced text-mining pipeline (Chapter 3), including the ULSA synthesis action extraction framework alone with other natural language processing(NLP) and deep learning techniques. The dataset consists of 35,675 solution-based synthesis procedures. Each procedure contains essential synthesis information, including the precursors and target materials, their quantities, and the synthesis actions and corresponding attributes. Every procedure is also augmented with the reaction formula.

Digitizing and systemizing the large synthesis corpus of existing materials science publications provides a foundation not only to build machine learning models, but also empirically validate the fundamental physical theory. Thermodynamics has strong predictive power for materials synthesis by identifying the stability regions of target phases, which can guide synthesis planning for computationally-designed materials. However, a stability domain does not give explicit information about the relative competitiveness of undesired byproduct phases, nor does it identify a precise synthesis condition for optimized kinetics to produce the target phase.

To fulfill the second objective, in Chapter 4, we define thermodynamic competition as the difference in driving force between one phase and its competing phases, and we hypothesize that one approach to optimizing the kinetics of phase-pure synthesis is to minimize the thermodynamic competition between the desired target phase and its competing phases. We systematically validate this hypothesis with two approaches: (1) we analyze large-scale solution synthesis procedures as text-mined from the literature and show that experimentally-optimized synthesis conditions are near our predicted thermodynamic optimum point, and (2) direct experimental evaluation of synthesis in $\text{LiIn}(\text{IO}_3)_4$ and LiFePO_4 ; where we show phase-pure synthesis occurs only when thermodynamic competition is minimized. Our work demonstrates that thermodynamic competition is an effective descriptor for synthesis optimization and a promising tool for optimizing aqueous solution-based experimental synthesis conditions.

Finally, Chapter 5 summarizes the main findings of the dissertation and provides an outlook for the future directions of data-driven approaches synthesis design.

This thesis is dedicated to Jincong Wang
for her love and support.

Contents

Contents	ii
List of Figures	v
List of Tables	vi
1 Introduction	1
1.1 Data-driven approaches to solution-based synthesis design	2
1.2 Text-mining and NLP for material synthesis	3
1.3 Physical modeling synthesis	4
1.4 Structure of this thesis	5
2 ULSA: Unified Language of Synthesis Action	6
2.1 Unified Language of Synthesis Actions and annotation scheme . .	9
2.1.1 Dataset annotation	11
2.1.2 Annotation decisions and ambiguous cases	12
2.2 Synthesis terms mapping	14
2.2.1 Baseline model	14
2.2.2 Word embeddings	14
2.2.3 Neural network model	15
2.3 Data analysis	15
2.3.1 Reassignment of mixing terms	15
2.3.2 Assigning synthesis actions attributes	16
2.3.3 Constructing synthesis flowchart for paragraphs	16
2.4 Dataset statistics	16
2.5 Mapping model	18
2.6 Analysis of action embeddings	23

2.7	Analysis of graphs clustering	25
2.8	Discussion and Conclusions	27
2.9	Code and data availability	29
3	Text-mined dataset of solution-based synthesis	31
3.1	Content acquisition	32
3.2	Paragraph classification	33
3.3	Synthesis procedure extraction	34
3.3.1	Materials entity recognition (MER)	34
3.3.2	Extraction of synthesis actions and attributes	35
3.3.3	Extraction of material quantities	35
3.3.4	Building reaction formulas	35
3.4	Solution-based synthesis dataset	36
3.5	Technical Validation	40
3.5.1	Extraction completeness and accuracy	40
3.5.2	Exploratory data analysis	41
3.6	Conclusion	46
3.7	Code and data availability	47
4	Minimum thermodynamic competition for synthesis design	48
4.1	Thermodynamic competition in aqueous solution-based systems .	49
4.2	Large-scale analysis combining text-mining with first principles calculations	52
4.3	Systematic synthesis of $\text{LiIn}(\text{IO}_3)_4$ and LiFePO_4	56
4.3.1	$\text{LiIn}(\text{IO}_3)_4$ synthesis analysis	56
4.3.2	LiFePO_4 synthesis analysis	58
4.4	Conclusion	62
4.5	Methods	63
4.5.1	Thermodynamic potential in aqueous solution system . . .	63
4.5.2	Thermodynamic competition calculation and optimization	64
4.5.3	Text-mined dataset analysis	64
4.5.4	Synthesis of $\text{LiIn}(\text{IO}_3)_4$	65
4.5.5	Synthesis of LiFePO_4	66
4.5.6	Structure Characterization	66

5	Conclusions and outlooks	67
5.1	Conclusions of this thesis	68
5.2	Outlook	69
5.2.1	Improving text-mined synthesis dataset	69
5.2.2	Generalization of minimum thermodynamic competition hypothesis	70
5.2.3	Better physical modeling synthesis	71
	Bibliography	72

List of Figures

2.1	Schematic workflow of data annotation, extraction and analysis.	9
2.2	Quantitative characteristics of the annotated dataset	18
2.3	2D projection of word embeddings vectors	24
2.4	Visualization of the first two principal components for the adjacency matrices of synthesis action graphs	26
3.1	Extraction pipeline and example	33
3.2	The chemical space covered by the dataset	43
3.3	Correspondence between choice of synthesis route and selected types of targets	44
4.1	Schematic representation of approach for predicting aqueous solution- based synthesis conditions	50
4.2	Thermodynamic competition analysis on text-mined dataset	55
4.3	$\text{LiIn}(\text{IO}_3)_4$ thermodynamic competition analysis and the correspon- dence synthesis results.	59
4.4	Schematic Pourbaix diagram of Li-Fe-P- H_2O system and the synthesis results.	61

List of Tables

2.1	Quantitative characteristics of the dataset chosen for annotation with ULSA schema.	17
2.2	Fleiss' kappa score for inter-annotator agreement using ULSA scheme.	19
2.3	Performance of baseline and bi-LSTM models for mapping synthesis sentences into ULSA terms	20
2.4	Examples of processing a solid-state and hydrothermal synthesis paragraph by baseline models and the bi-LSTM model using ULSA scheme.	22
3.1	Format of each data record: description, key label, data type.	38
3.2	Performance of data extraction for dataset entries.	41
3.3	Ten most common targets in the dataset with their corresponding precursors.	42
4.1	Synthesis conditions with thermodynamic competition analysis and experimental results for $\text{LiIn}(\text{IO}_3)_4$ synthesis.	57
4.2	Synthesis conditions with thermodynamic competition analysis and experimental results for LiFePO_4 synthesis.	60

Acknowledgments

I am grateful to my advisor, Professor Gerbrand Ceder, for your support, guidance, and encouragement. You offered not only the freedom to explore research fields along with my interest, but also patience and advice when I encountered obstacles. I am constantly inspired by your vision and taste as a great scientist. You are my role model not only in academia but also in problem-solving, high standards, leadership, etc. Almost every perspective in life, so I couldn't list all of them here.

I would like to thank my mentor, Professor Wenhao Sun. You offered lots of insights to understand synthesis from thermodynamics and kinetics. Also, your work on the inverse phase stability approach is the foundation of my “minimum thermodynamic competition” hypothesis. My Ph.D. life as a junior student would have been much harder without your tremendous help.

Thank you to my thesis committee, Professor Kristin Persson and Professor Kwabena Bediako, and my qualifying exam committee, Professor Mark Asta and Professor Jie Yao, for advice, discussion, and guidance.

I would like to thank my lab mates from the Ceder group: Kevin Cruse, my colleague who taught me not only science but also guidance in professional English writing with infinite patience; Yingzhi Sun, who conducted experiments to demonstrate my theory, Tanjin He, who might be the kindest person I met in my life and taught me a lot of NLP knowledge, Yuxing Fei, who often explored interesting research ideas with me, Yan Zeng for teaching me fundamental synthesis knowledge, Haoyan Huo for tons of help in machine learning and NLP. I would also like to thank past members of our ML subgroup: Olga Kononova, Ziqin Rong, Vahe Tshitoyan, Tiago Botari, Chris Bartel, and Amalie Trewartha. It has been a pleasure working with you. Special mention to KyuJung Jun – my buddy from the thermodynamics class exam preparation, then the prelim preparation, and to the qualifying exam preparation. Also, I will miss the time when we were complaining about how hard Ph.D. life is together. In addition, I want to thank all other members in the Ceder Group for all the wonderful memories.

My friends, Junyi Shangguan, Mao Cheng, Zexuan Liu, Yiwen Jin, Minjia Li, Wenbin Guo, Yuxin Xu, Kang Yu, Yuan Tian, Changpeng Lu, and Xi Chen, thank you for the chat, discussion, and help during my Ph.D. study; Jingwen Ouyang, Chuchu Zhao, Wuming Zhang, Jie Yang, Boyu Liu, Haoye Cai, and Chunyan Bai, for exploring food in the Bay Area together frequently. Without you, my life during my Ph.D. would not have been so much fun.

I am here to thank my undergraduate research advisor, Prof. Bernard Kirtman at UCSB again. You led me into the world of science, and I would not have had the opportunity to be admitted to UC Berkeley Ph.D. program without your support.

Infinite gratitude to my parents, Jianbo Wang and Jiangbo Xu, and other family members. Thank you for always supporting me in any decisions I made in my life and enabling me to pursue my interests without worry.

Finally, to Jincong Wang. Our story started at high school, then we had a long-distance relationship across hundreds of miles between Wuhan and Shanghai in China for four years, and later thousands of miles between the Bay Area and St. Louis in the United States for one year and a half. Thank you for the substantial efforts spent making us reunite in the Bay Area. You complete not only my life but also my soul. I promise you that we will make a brighter and brighter future together.

Chapter 1

Introduction

Part of this chapter has been published in or included in:

Wang, Z., Kononova, O., Cruse, K., He, T., Huo, H., Fei, Y., Zeng, Y, Sun, Y., Cai, Z., Sun, W., & Ceder, G.(2022). Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. Scientific Data 9, 231. [[1](#)]

Wang, Z., Sun, Y., Cruse, K., Zeng, Y., Fei, Y., Liu, Z., Shangguan, J., Byeon, Y., Jun, K., He, T., Sun, W., & Ceder, G.. Optimal thermodynamic conditions to minimize kinetic byproducts in aqueous materials synthesis. Working paper which is pending publication at the time of filing this thesis.

1.1 Data-driven approaches to solution-based synthesis design

Big-data-driven approaches have helped to establish a new paradigm of scientific research[2–4]. In materials science specifically, the Materials Genome Initiative (MGI) effort has significantly facilitated and accelerated materials discovery and design by deploying large-scale *ab initio* computation and building computed databases of structure–property relationships [5–7]. Unlike computational data, the experimentally determined properties and structures of inorganic materials are mainly available in manually curated databases[8–12]. Such well-curated experimental databases have led to early machine learning models that address difficult problems in materials research, such as structure prediction[4]. The ability to efficiently design and predict the structure of novel advanced materials with the assistance of computed and experimental databases has shifted the materials innovation challenge toward understanding and determining the synthesis routes for novel materials[13]. In principle, this challenge could also be mastered using data-driven approaches since only limited predictive theories are available for materials synthesis[14–24].

In this thesis, we propose and demonstrate a data-driven approach to understanding and designing solution-based synthesis. Indeed, in organic chemistry, data-driven synthesis planning [25, 26] has already been successfully implemented in certain cases, such as in predicting retrosynthesis [27] and in complex natural product synthesis design [28]. Although datasets for the synthesis of organic materials are widely available [29, 30], there is not yet a large-scale database of inorganic synthesis routes, which is needed to build data-driven approaches to design inorganic materials synthesis. Therefore, preparing a large-scale and reliable synthesis dataset becomes the first step toward developing data-driven approaches for understanding inorganic materials synthesis and synthesizability.

1.2 Text-mining and NLP for material synthesis

Scientific publications represent the largest repository of knowledge about material synthesis and can be used as a reliable source of data. However, human-written descriptions of syntheses require additional levels of interpretation for conversion into a codified, machine-operable format. Additionally, manual extraction of synthesis information is laborious, even for a very limited number of papers[31–34]. Given these obstacles, an automated information extraction pipeline can accelerate data collection and assist in building structured synthesis procedures from scientific text. Natural language processing (NLP) approaches have been widely developed in the past decade, and various advanced tools for high-quality information extraction from unstructured text are available to researchers.

In materials science, NLP has been used to extract and analyze materials properties [35–37], applications [38, 39], and synthesis conditions for some limited cases [40]. Various NLP tools, including ChemDataExtractor [41], OSCAR[42], ChemicalTagger[43], and others [31, 44, 45], have been developed to extract information from chemical text. Recently, advanced models based on deep convolutional and recurrent neural networks [46–49] have been proposed to improve the accuracy of chemical data extraction.

Text-mining approaches in materials science have also been used to construct automated pipelines for collecting information about materials synthesis from publications and to build large-scale publicly available datasets from such collected data, including datasets of synthesis formulations for metal oxides[40, 50, 51], germanium-containing zeolites[52], and perovskites[53]. In recent work, Olga et al. have developed a text-mining pipeline to construct the first large-scale dataset of solid-state ceramics synthesis “recipes”, which includes not only the starting materials and final products but also the synthesis actions, their attributes, and balanced chemical-reaction equations [54], and an interpretable machine learning synthesis predictor has been trained on this text-mined dataset[55]. However, there is no large-scale codified datasets on solution-based synthesis procedures for inorganic materials were available. Constructing such a dataset is one of the primary goals of this thesis. The detailed text-mining techniques and the text-mined dataset will be discussed in Chapter 2 and 3.

1.3 Physical modeling synthesis

Digitizing and systemizing the large synthesis corpus of existing materials science publications provides a foundation not only for building machine learning models[55], but also empirically validating the fundamental physical theory. Thermodynamics has strong predictive power for materials synthesis [16, 20–22, 56] — for example, phase diagrams are routinely used to guide scientists to synthesize the target phase by indicating regions of thermodynamic stability corresponding to that phase [57–59]. However, a stability domain given by the phase diagram does not give explicit information about why certain conditions are more favorable than others for phase-pure synthesis. This is because traditional phase diagrams do not visualize the free energy axis[60], and so they do not provide the relevant information on thermodynamic competition between competing phases.

A major problem in synthesis arises from the formation of impurities or undesired byproduct phases[61], which compete kinetically with other phases in the synthesis space. When designing synthesis procedures, experimentalists carefully choose synthesis conditions that diminish the variety of competing phases which can appear as kinetic byproducts[62]. This process of tuning synthesis conditions is usually based on heuristics, and there is no universally accepted metric to evaluate competitiveness during synthesis design.

The difference in free energy between precursors and products is the fundamental underlying thermodynamic driving force that propels a reaction toward thermodynamic equilibrium. Thermodynamic driving force not only determines the direction towards which a reaction moves but also affects kinetics, such as the rate of nucleation and growth[63] and diffusion flux[64]. Therefore, a theoretical framework that embeds thermodynamic and kinetic mechanisms needs to be formulated and, more importantly, validated, which is another goal of this thesis. Our novel synthesis hypothesis with its validation will be discussed in Chapter 4.

1.4 Structure of this thesis

The work presented in this thesis aims at solving two problems: 1) developing an NLP and deep learning infrastructure that text-mines synthesis information from journal articles, and 2) using the text-mined dataset and systematic synthesis experiments to demonstrate the synthesis hypothesis, minimum thermodynamic competition, proposed by us.

Chapter 2 describes the first *unified language of synthesis actions* (ULSA) for describing inorganic synthesis procedures. This synthesis ontology helps create an annotation dataset and build a neural network-based model to extract synthesis actions, which is an important step in the text-mining pipeline to extract synthesis procedures from the scientific literature. The synthesis actions extraction approach was used in Chapter 3, where we show how we constructed the first large dataset of solution-based inorganic materials synthesis procedures by designing an advanced text-mining pipeline. The codified synthesis dataset contains essential information required to complete a synthesis route, which made it possible to build data-driven approaches for synthesis design. The text-mined synthesis dataset was used in Chapter 4 to demonstrate the thermodynamic synthesis hypothesis, minimum thermodynamic competition, we proposed. Furthermore, we designed systematic synthesis experiments as a complementary approach for large-scale analysis on the text-mined dataset to test the hypothesis. Finally, we summarize the work in this thesis and outline future works in Chapter 5.

Chapter 2

ULSA: Unified Language of Synthesis Action

Part of this chapter has been published in:

Wang, Z., Cruse, K., Fei, Y., Chia, A., Zeng, Y., Huo, H., He, T., Kononova, O., & Ceder, G. (2022). ULSA: Unified Language of Synthesis Actions for the Representation of Inorganic Synthesis Protocols. Digital Discovery 1, 313–324.[\[65\]](#)

Scientific text mining has received its recognition in the past few years [31, 40, 44, 66], providing the materials science community with datasets on a variety of materials and their properties [36, 37, 67] as well as synthesis protocols [50, 51, 54, 68]. Nonetheless, a majority of these text mining studies have been focused on extracting chemical entities such as material names, formulas, properties, and other characteristics [42, 45, 48, 69, 70]. There have only been a few attempts to extract information about chemical synthesis and reactions and compile them into a flowchart of synthesis actions [49, 71–78]. Hawizy et al. [71] were early developers for such extraction, using a combination of rule-based regular expressions (regex) [79] and syntax tree parsing to identify and classify action phrases in their tool, ChemicalTagger. This approach shows very good performance on organic synthesis procedures. Vaucher et al. [72] used a combination of rule-based approaches and machine learning models trained on over 2 million procedural sentences to extract synthesis actions from the organic chemistry patents texts and map them into well-defined language schemas. We found this work to be one of the most robust and accurate in describing organic synthesis procedures. Mehr et al. [73] developed a semi-automated workflow that uses NLP-based approaches to translate human-written text into an internal Chemical Description Language (so-called XDL) and then map it into robotic operations. To the best of our knowledge, this is the only work that applied the developed synthesis ontology to robotic synthesis for organic molecules. Mysore et al. [74] paved the way for synthesis action graph extraction from the inorganic synthesis text. For this, they applied several neural network-based models and used dependency tree parsing to combine the extracted information into synthesis graphs. Similarly, Kuniyoshi et al. used bi-LSTM combined with BERT word embeddings to construct synthesis graphs for solid-state batteries fabrication [49], which showed excellent results on the extraction of operations using the science literature-specific SciBERT pretrained language model.

As is apparent from the above survey, the automation of synthesis procedures for organic molecules has made significant progress. This is mainly due to the facts that a) organic synthesis is more deterministic and hence more common in materials science and biochemical domains, and b) there exist large-scale databases and repositories of organic reactions [29, 30] and annotated sets [80, 81] that help to speed up development of the machine-learning approaches for interpretation and

prediction of organic synthesis. Even with such data availability, to the best of our knowledge, there have been only a few attempts to create a publicly available annotated corpus containing materials synthesis protocols extracted from the text [68, 72, 75]. The dataset created by Mysore et al. [68] contains 230 labeled synthesis paragraphs with labels assigned to material entities, synthesis actions, and other synthesis attributes for inorganic synthesis, and is freely available to users. The dataset used by Vaucher et al. [72] was obtained by augmenting the existing Pistachio dataset [82] of organic synthesis procedures, and is available upon request. Kuniyoshi et al. [75] annotated an in-house dataset of inorganic materials synthesis entities that is publicly available.

A major obstacle in annotating synthesis actions in the text corpora is the lack of a solid, robust, and well-established ontology for describing synthesis procedures in materials science [83]. Indeed, researchers prefer to vaguely sketch “methods” sections of the manuscript in general human-readable language rather than follow a specific protocol. This significantly impacts reproducibility of the results, not to mention ambiguity in understanding even when read by a human expert [83].

In this chapter, we discuss a potential approach to the problem of inorganic synthesis ontology based on creating a *unified language of synthesis actions* (ULSA). We demonstrate an application of this approach in describing solid-state, sol-gel, precipitation, and solvo-/hydrothermal synthesis procedures, which cover the majority of inorganic synthesis procedures [84, 85]. Specifically, we built and created a dataset of 3,040 synthesis sentences labeled according to the ULSA schema and trained a neural network-based model that identifies a sequence of synthesis actions in a paragraph, maps them into the ULSA, and builds a graph of the synthesis procedure (Figure 2.1). We applied this model to thousands of synthesis paragraphs and analysed the resulting synthesis graphs. The obtained results show that our ULSA vocabulary is comprehensive enough to obtain high-accuracy extraction of synthesis actions as well as to identify the important features of each of the aforementioned synthesis types. Additionally, the ULSA as it is encoded in the labeled dataset can be easily customized and augmented to account for other inorganic synthesis methods. The dataset and the scripts for building such a synthesis flowchart are publicly available. We anticipate these results will be widely used by the researchers interested in scientific text mining and will help (a) to achieve a

breakthrough in predictive and AI-guided autonomous materials synthesis and (b) build a robust materials synthesis ontology.

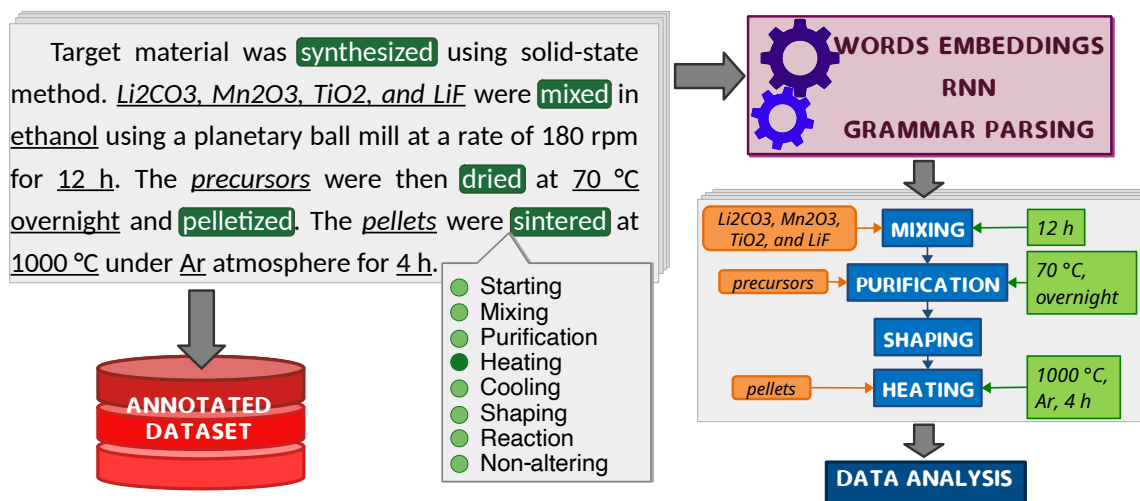


Figure 2.1: Schematic workflow of data annotation, extraction and analysis. First, the set of paragraphs were annotated using an Amazon Mechanical Turk engine. Highlighted in green are the action tokens that were annotated and then extracted using a neural network model. Other highlighted tokens and phrases (i.e. synthesis action attributes and subjects) were obtained using rule-based sentence parsing solely for the purpose of data analysis and are not presented in the annotated dataset. The obtained labeled dataset is stored as a single JSON file and is also used for training a neural network model to identify synthesis actions in the text. Obtained synthesis actions, attributes, and subjects were converted into synthesis flowcharts that were further used for data analysis.

2.1 Unified Language of Synthesis Actions and annotation scheme

To unify terminology used to describe a synthesis procedure, we defined 8 *action terms* that unambiguously identify a type of synthesis action. Every action word (or multi-word phrase) in the dataset is mapped to the corresponding action term according to the following rule: the word (or multi-word phrase) is recognized as an action if it (a) results in modification of the state of the material or mixture during the synthesis or (b) carries a piece of information affecting the outcome of

the synthesis procedure. The action terms used within the unified language are explained below. In each example, the text underlined is the word or phrase that is annotated.

- **Starting:** A word or a multi-word phrase that marks the beginning of a synthesis procedure. Specifically, this often indicates which materials will be produced. For example: “PMN-PT was synthesized by the columbite precursor method”, “Solid-state synthesis was used to prepare the target material”, “The powder was obtained after the aforementioned procedure”.
- **Mixing:** A word or a multi-word phrase that marks the combination of different materials (in a solid or liquid phase) to form one substance or mass. For example: “Precursors were weighted and ball -milled”, “Precursors were mixed in appropriate amounts”, “ Sb_2O_3 is added to the solution”, “The solution was neutralized”, “The mixture was stabilized by the addition of sodium citrate”.
- **Purification:** A word or a multi-word phrase that marks the separation of the sample phases. This also includes drying of a material. For example: “Samples were exfoliated from substrates”, “The liquid was discarded and the remaining product was filtered off and washed several times with distilled water”, “The precursors were heated in order to remove the moisture”, “The precipitate was collected by washing the solution in distilled water”.
- **Heating:** A word or a multi-word phrase that marks increasing or maintaining high temperature for the purpose of obtaining a specific sample phase or promoting a reaction rather than drying a sample. For example: “The powder sample was annealed to obtain a crystalline phase”, “The mixture was subjected to heating at 240 °C for 24 h”.
- **Cooling:** A word or a multi-word phrase that marks rapid, regular, or slow cooling of a sample. For example: “The product was cooled down to room temperature in the furnace”, “The sample was quenched rapidly in the solid CO_2 ”, “The product was left to cool down to room temperature”.
- **Shaping:** A word or a multi-word phrase that marks the compression of powder or forming the sample to a specific shape. For example: “The

powder was pressed into circular pellets”, “The powder was then pelletized with a uniaxial press”.

- **Reaction:** A word or a multi-word phrase that marks a transformation without any external action. For example: “*The sample was left to react for 6 hrs*”, “*The temperature was kept at 1000 K*”, “*The solution was maintained at 200 K for 12 hrs*”.
- **Non-altering:** A word or a multi-word phrase that marks an action done on a sample that either does not induce any transformation of the sample or does not belong to any of the above classes. “*The pellets were placed in a sealed alumina crucible*”, “*The reaction vessel was wrapped with aluminum foil*”, “*The sample was sealed in a tube*”, “*The gel was transferred to an oven*”.

2.1.1 Dataset annotation

To annotate synthesis paragraphs with the unified language of synthesis actions (ULSA), we selected 535 synthesis paragraphs from the database of 420K full-text publications acquired previously [54]. The paragraphs were chosen to proportionally represent four major types of inorganic synthesis: solid-state, sol-gel, solvo-/hydrothermal, and precipitation. The details of the content acquisition and synthesis type classification have been described in previous papers [54, 85].

The 535 paragraphs consisted of 3,781 tokenized sentences [69]. First, each sentence was classified as either related to synthesis or not related to synthesis. The latter case usually contains sentences about product characterization and other details. Next, we isolated 3,040 synthesis sentences and assigned labels to each word or multi-word phrase in the sentence on the basis of the ULSA protocol with annotation schema described in Section 2.1. Only words and phrases describing synthesis actions were annotated. The final dataset consists of these 3,040 labeled synthesis sentences. All annotations were performed using a custom Amazon Mechanical Turk-based server.

2.1.2 Annotation decisions and ambiguous cases

The ULSA was developed based on the authors' own experiences with the extraction of information from materials synthesis paragraphs [54] and extensive communication with experimentalists actively involved in various types of materials synthesis research. The annotation schema and the choice of action terms were designed to provide maximum flexibility to future users and allow them to adjust the schema according to their preferences and tasks. For example, the annotated multi-word phrases such as "left to react" and "heated to evaporate" were handled as one entity. This way, they can be split into individual terms or modified later with a simple set of rules to make a customized labeled dataset.

It is important to keep in mind that we mapped words into the terms of synthesis action per sentence, meaning that we used only information in the context of a given sentence to make a decision about the annotation of a word, rather than the whole paragraph. The reason for this choice is the multiple and diverse possibilities to combine and augment sentences leading to different meanings of the terms. The interpretation of the whole text or paragraph is an entirely separate field of research that is outside the scope of this work.

We chose to annotate those words that are characteristic of a synthesis procedure or result in the transformation of a substance. For example, in the sentence "the precursors were weighed and mixed," the term "weighed" is not a synthesis action since it is to be expected in synthesis, while "mixing" is a synthesis action because it may have a specific condition and transform the sample, or can be preceded by calcination of the precursors in other syntheses. The exclusion from this rule is the **Starting** action. Even though terms belonging to this action do not bring any special information or explicit action to the synthesis, we chose to distinguish "starting" actions because in a substantial number of cases they can serve as flags to separate multiple synthesis procedures from one another. An illustration of this situation is when precursors are prepared prior to synthesizing a target material, as in sol-gel synthesis.

For the annotation of **Mixing** synthesis actions, we did not differentiate between powder mixing, ball milling (grinding), addition of droplets, or dissolving

of substances. In many situations, this precise definition depends on the solubility of reactants and mixing environment, as well as on other details of the procedure that are never explicitly mentioned in the text. We leave it up to the user to create their own application-based definitions of these mixing categories. Nonetheless, in the application below we provide a rule-based example of how these types of synthesis actions can be identified in the text.

The **Non-altering** action term was introduced to make room for those synthesis actions that are not typical or do not fall into any other category but nevertheless appear as a synthesis action within our definitions. While **Non-altering** action terms can be easily confused with **Reaction** actions or non-actions, the decision depends on the sentence context and can be arbitrarily extended or removed by a user. Comparing “the sample was kept in the crucible” and “the sample was kept overnight,” the former is not a synthesis action while the latter should be considered an important synthesis step.

Ambiguous situations as in the ones mentioned above are ubiquitous in descriptions of syntheses. A substantial amount of these situations occur when authors try to be wordy or use flowery language when writing the synthesis methods. Unfortunately, this often presents a challenge for accurate machine interpretation of the text. We accounted for some of these cases when annotating the data as described below.

First, implicit mentions of synthesis actions (i.e. when a past participle form of a verb is used as a descriptive adjective referring to an already processed material) are the most frequent source of confusion. We chose to annotate these as synthesis actions. For example: “the calcined powder was pressed and annealed.” In this sentence, the descriptive adjective *calcined* could be either a restatement of the fact that there was a calcination step or it could be additional information which had not been mentioned previously. These situations can be later resolved with a rule-based approach, hence we leave it as a task for users of the data.

The situation when a method is specified along with the synthesis action is also common. In a phrase of the form “transformed by a specific procedure,” we consider only the key action (the transformation) as a synthesis action. For example: “the precipitates were separated by centrifugation.” When required, the

method can be retrieved with a set of simple rules.

Redundant action phrases are also abundant in many descriptions of the procedures. In a phrase of the form “subjected to a process”, we considered only the processing verb as a synthesis action. For example: “the samples were subjected to an initial calcination process.”

Finally, phrases that attempt to reason the purpose of the action, such as “left to react”, “brought to a boil”, “heated to evaporate,” are considered as one synthesis action. This is done for the purpose of providing flexibility to a user and to let them make a decision on how to treat these cases.

2.2 Synthesis terms mapping

We used lookup table (baseline) and neural network models to map synthesis sentences into the ULSA.

2.2.1 Baseline model

Two baseline models were implemented, both based on a lookup table. For the lookup table, we chose the most frequent words used to describe synthesis steps in the “methods” section of the papers. The first baseline model matches every token against the lookup table and assigns the corresponding action term if any appear. The second baseline model uses information about the part of speech of a given word (assigned by SpaCy [86]) and matches only verbs against the lookup table.

2.2.2 Word embeddings

Word embeddings were used as a vectorized representation of the word tokens for the neural network model. To create an embedding, we trained a Word2Vec model [87] implemented in the Gensim library [88]. We used ~420K paragraphs describing four synthesis types: solid-state, sol-gel, solvo-/hydrothermal and precipitation synthesis. The paragraphs were obtained as described in our previous work [54]. Prior to training, the text was normalized and tokenized using ChemDataEx-

tractor [69]. Conjunctive adverbs describing consequences, such as “therefore”, “whereas”, and “next”, were removed from the text. All quantity tokens were replaced with the keyword <NUM> and all chemical formulas were replaced with the keyword <CHEM>. All words that occur less than 5 times in the text corpus were replaced with the keyword <UNK>. We found that skip-gram with negative sampling loss ($n = 10$) performed best, and the final embedding dimension was set to 100.

2.2.3 Neural network model

We used a bi-directional long short-term memory (bi-LSTM) neural network model to map synthesis tokens into the aforementioned action terms. The model was implemented using the Keras library (<https://keras.io/>) with latent dimensionality 32 and dropout probability 0.2. Word embeddings were used as model input. The categorical cross-entropy was calculated as the loss function. The labeled dataset was split into training, validation, and test sets using a 70:20:10 split, respectively. Early stopping was used to obtain the best performance.

2.3 Data analysis

2.3.1 Reassignment of mixing terms

For the purposes of data analysis and to demonstrate potential directions for customization of the labeled dataset, we additionally reassigned **Mixing** synthesis action terms as **Dispersion Mixing**, **Solution Mixing**, and **Ball-Milling** whenever there was enough information to distinguish between the three, otherwise they were left as **Mixing** actions. Here, **Dispersion Mixing** is identified either by explicit “dispersion” action words or by words such as “disperse” or “suspend” plus any liquid environment. **Solution Mixing** is identified by a list of specific action words such as “dissolve”, “dropwise added”, and others. **Ball-Milling** is identified more specifically through terms related to “ball-milling”. This was achieved by constructing and traversing the dependency trees of the sentences using the SpaCy library [86] and by using dictionaries of common solution and mixing terms.

2.3.2 Assigning synthesis actions attributes

Synthesis actions identified as **Mixing**, **Heating**, and **Cooling**, as well as the actions referring to drying processes (identified by the stem “dry” and “evaporate”), were assigned attributes such as temperature, time, and environment. This was done by analysing dependency sub-trees associated with each action token [86] and by applying rule-based regular expression matching [79]. It is important to notice that this approach fails when the action and its attributes are not mentioned in the same context or the dependency tree is built incorrectly.

2.3.3 Constructing synthesis flowchart for paragraphs

For every paragraph in the set, we then applied the bi-LSTM mapping model (Section 2.2) to extract the sequence of action terms from every sentence. Next, we merged all the synthesis actions obtained from all sentences within the paragraph into a synthesis flowchart. This was performed with a rule-based approach by traversing grammar trees and analysing the surrounding words of each action term and comparing them to the words and action terms of the previous sentence. Finally, the flowchart of synthesis actions for a given paragraph was converted into an adjacency matrix. For this, synthesis action terms were ordered and assigned to rows and columns of the matrix and initialized with zeros, resulting in a 10 by 10 matrix for every paragraph (8 action terms from the vocabulary of ULSA plus three additional terms for **Mixing** and **Non-altering** terms removed). Whenever there was a step from action i to action j , the corresponding value in the matrix was incremented by 1. The matrices for all paragraphs were flattened and merged together for further principal component analysis.

2.4 Dataset statistics

The quantitative characteristics of the set are provided in Table 2.1 and displayed in Figure 2.2. Briefly, 535 synthesis paragraphs resulted in 3,781 sentences of which 3,040 describe actual synthesis procedures. While we tried to maintain an even distribution of the action terms in the labeled set, it is still highly skewed toward **Mixing** and **Purification** actions. This is not surprising, since mixing of precursors occupies any synthesis procedure and purification is required in almost

any non-solid-state method for inorganic synthesis. Heating is the next most prevalent synthesis action since it is also one of the basic operations in inorganic synthesis.

	Amount
Paragraphs used for annotation	535
Per synthesis type:	
– solid-state synthesis	199
– sol-gel synthesis	51
– solvo-/ hydrothermal synthesis	148
– precipitation	137
Total sentences	3781
Synthesis sentences	3040
Action tokens	5547
Per action category:	
– starting	619
– mixing	1853
– purification	1080
– heating	973
– cooling	259
– shaping	225
– reaction	232
– non-altering	306

Table 2.1: Quantitative characteristics of the dataset chosen for annotation with ULSA schema.

To probe the robustness of ULSA and our annotation schema, we asked 6 human experts to annotate the same paragraphs in our dataset and used Fleiss’ kappa score to estimate the inter-annotator agreement between the annotations [89]. In general, the Fleiss’ kappa score evaluates the degree from -1 to 1 to which different annotators agree with one another above the agreement expected by pure chance. A positive Fleiss’ kappa indicates good agreement, scores close to zero indicate near randomness in categorization, and negative scores indicate conflicting annotations. This is a generalized reliability metric and is useful for agreement between three or more annotators across three or more categories. Table 2.2

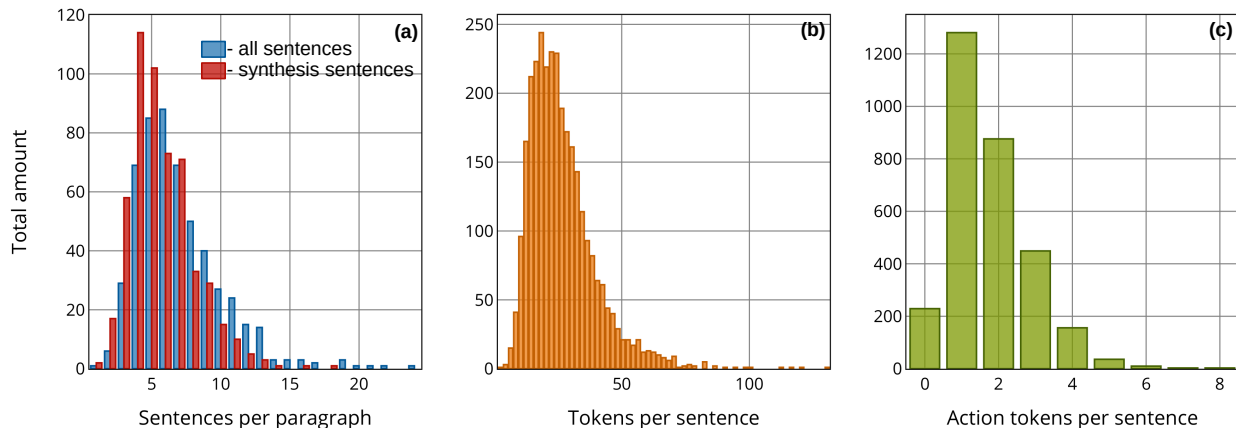


Figure 2.2: Quantitative characteristics of the annotated dataset. (a): Number of sentences per paragraph (blue), including sentences related to synthesis procedures (red). (b): Number of all tokens per sentence in the annotated set. (c): Number of tokens denoting a synthesis action per sentence in the annotated set.

lists the Fleiss’ kappa scores for agreement between human experts annotating the sentences according to the schema described in Section 2.1. The table shows good agreement on distinguishing synthesis sentences from non-synthesis sentences, as well as for all and for each individual synthesis action, including non-actions. The agreement across all action terms is 0.83. Among those, the action terms with lower scores are *Shaping* and *Non-altering*. The low score for *Non-altering* is expected since a wide range of actions which do not induce a transformation in the sample could be mapped into this category. The *Shaping* action term can also be associated with many synthesis operations. For instance, granulating procedures that break a sample into smaller chunks could be considered a *Shaping* action; at the same time, a bench chemist could consider “granulation” to be *Mixing* action term since it requires performing a grinding operation to obtain the new shape. Less ambiguous actions terms, such as *Heating* and *Mixing*, showed higher agreement.

2.5 Mapping model

As a first approach for mapping of synthesis paragraphs into ULSA, we used dictionary lookup constructed as described in Section 2.2.1. We use the labeled

	Score
Identification of synthesis sentences	0.69
Action terms tagging	0.83
Per action terms:	
– starting	0.82
– mixing	0.86
– purification	0.79
– heating	0.84
– cooling	0.88
– shaping	0.59
– reaction	0.66
– non-altering	0.45
– no action	0.87

Table 2.2: Fleiss’ kappa score for inter-annotator agreement using ULSA scheme.

dataset of 3,040 sentences to assess the performance of the model. We considered two options: mapping of all sentence words and mapping the verbs only. In both cases, the overall accuracy of the prediction (i.e. F1 score) is ~60-70% (Table 2.3). Nonetheless, mapping of all words shows relatively good recall and poor precision, while mapping of only verbs improves the precision but diminishes recall.

These results moved us toward considering a recurrent neural network model for mapping paragraphs into ULSA. It is generally accepted that recurrent neural networks (RNNs), and specifically bi-LSTMs, can effectively process sequential data and keep track of past events [90]. Indeed, bi-LSTM is simple enough and does not require exhaustive training and fine-tuning, as is common for BERT[91] and GPT models[92–94]. The bi-LSTM model combined with word embeddings (Section 2.2.3) was trained on the labeled dataset of 3,040 sentences. The bi-LSTM model significantly improves mapping accuracy, yielding >90% F1 score (Table 2.3). It is important to notice here that all the metrics for baseline and neural network models were computed per sentence, i.e. we evaluated the whole sentence being mapped correctly rather than individual terms.

Model	Precision	Recall	F1 score
Baseline 1	0.54	0.61	0.57
– solid-state	0.53	0.72	0.61
– sol-gel	0.57	0.75	0.65
– hydrothermal	0.54	0.53	0.54
– precipitation	0.55	0.50	0.53
Baseline 2	0.84	0.50	0.63
– solid-state	0.84	0.54	0.66
– sol-gel	0.79	0.62	0.69
– hydrothermal	0.84	0.47	0.61
– precipitation	0.84	0.44	0.54
bi-LSTM	0.90	0.88	0.89
– solid-state	0.90	0.90	0.90
– sol-gel	0.88	0.86	0.87
– hydrothermal	0.90	0.86	0.88
– precipitation	0.90	0.91	0.91

Table 2.3: Performance of baseline and bi-LSTM models for mapping synthesis sentences into ULSA terms. In Baseline 1, all words in the sentence are matched against a lookup table. In Baseline 2, only verbs tagged by SpaCy are matched against the lookup table. The quantities are computed per sentence, i.e. the number of sentences with all the action tokens identified and assigned correctly.

The output of the baseline models and the bi-LSTM model for exemplary solid-state and hydrothermal synthesis paragraphs are shown in Table 2.4, which shows significant improvement in the bi-LSTM model performance compared to the baseline models. There are a few reasons why the bi-LSTM model outperforms plain dictionary lookup. First, researchers use diverse vocabulary to describe synthesis procedures, hence there are unlimited possibilities in constructing a lookup table. For instance, “heating” can be referred to as “calcining”, “sintering”, “firing”, “burning”, “heat treatment”, and so on. In this case, a word embedding model helps to significantly improve the score even for those terms that have never appeared in the training set (e.g. “degas”, “triturate”). Second, a given verb is defined as a synthesis action term largely based on the context. Prominent examples are “heating rate”, “mixing environment”, “ground powder”, etc. That is well captured

by the recurrent neural network architecture. Lastly, synthesis actions are not only denoted by verb tokens, but also by nouns, adjectives, and gerunds. This can be also learnt by the neural network better than by a set of rules.

Paragraph	Baseline 1	Baseline 2	Bi-LSTM
Target was synthesized by solid state reaction. Precursors were milled together with a mortar and pestle for 10 min. The mixture was then placed into a alumina crucible, heated to 1200 °C at a heating rate of 5 °C/min in ambient air, held at 1200 °C for 2 h, and then cooled to room temperature at 20 °C/min. The resulting powder was then ground for 5 min to break up agglomerates.	Starting(synthesized) Mixing(milled) Mixing(mixture) Heating(heated) Heating(heating) Cooling(cooled)	Starting(synthesized) Mixing(milled) Cooling(cooled)	Starting(synthesized) Mixing(milled) Heating(heated) Reaction(held) Cooling(cooled) Mixing(ground)
Titanium isopropoxide and isopropanol were mixed and stirred for 30 min. Then, the solution was added to 40 mL nitric acid solution. The mixture was heated at 80 °C for 10 h under stirring, the resulting solution was transferred into Teflon-lined autoclave and kept at 200 °C for 24 h. The precipitates were washed thoroughly with distilled water and ethanol, and dried overnight.	Mixing(mixed) Mixing(stirred) Mixing(added) Mixing(mixture) Heating(heated) Mixing(stirring) Reaction(kept) Purification(dried)	Mixing(stirred) Mixing(added) Heating(heated) Reaction(kept) Purification(dried)	Mixing(mixed) Mixing(stirred) Mixing(added) Heating(heated) Mixing(stirring) Reaction(kept) Purification(washed) Purification(dried)

Table 2.4: Examples of processing a solid-state and hydrothermal synthesis paragraph by baseline models and the bi-LSTM model using ULSA scheme.

In summary, we designed a neural network-based model that maps any synthesis paragraph into ULSA with high accuracy and significantly outperforms a plain dictionary lookup approach.

2.6 Analysis of action embeddings

To analyse how well the ULSA represents the space of synthesis operations commonly used when describing inorganic synthesis processes, we plotted a 2D projection of the word embeddings calculated with a t-SNE approach. The results are shown in Figure 2.3. To achieve a clear representation, we only analysed those verbs that appear more than 10 times. We then mapped these paragraphs into ULSA by using the bi-LSTM model. Those verbs that were assigned with a ULSA label are color-coded in the figure correspondingly, the other non-synthesis action terms are colored in grey.

First, we observe that the verbs mapped into ULSA and hence representing synthesis actions are all grouped in the top-left corner of the projection. Indeed, analysis of the individual words in the rest of the space showed that those are the words that generally appear in synthesis paragraphs but do not carry any information about the synthesis procedure. For instance, these are verbs denoting characterization of a material (“detect”, “quantify”, “examine”, “measure”), naming of a sample (“denoted”, “referred”, “named”, “labeled”) or referring to a table or figure. The blob of dots in the middle of the plot are all words that were either mis-tokenized during text segmentation or mistakenly recognized as verbs by the SpaCy algorithm. In the embeddings mapping, these words are replaced with the <UNK> token.

A second interesting observation is that the embeddings of firing (blue dots), pelletizing (purple dots) and grinding into powder (orange dots) are all located next to each other. This agrees well with the fact that those actions together describe solid-state synthesis processes. Oppositely, the verbs describing solution mixing (orange dots) are in close proximity with the verbs referring to purification or drying (green dots). Similarly, verbs indicating cooling processes (magenta dots) and the verbs referring to reaction processes (red dots) are clustered together. This

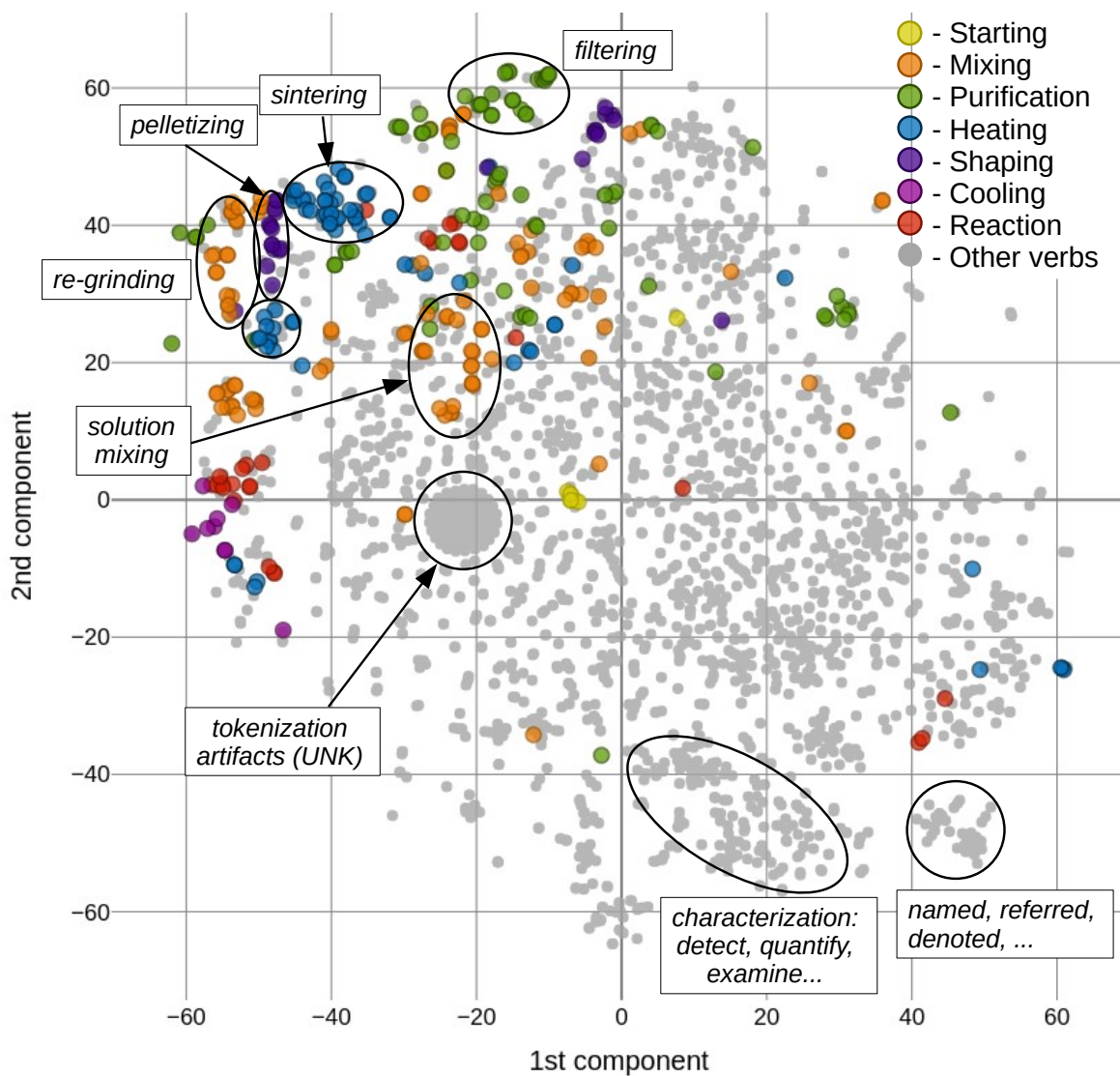


Figure 2.3: 2D projection of word embeddings vectors. Shown are the most frequent verb tokens encountered in the set of 420K paragraphs describing a synthesis procedure. Highlighted in different colors are the vectors that correspond to the common verbs from the categories of synthesis actions used for annotation. Other prominent clusters of vectors are denoted with circles and labeled by a common term. Dimensionality reduction was performed using a t-SNE approach.

agrees with the often encountered constructions of “left to react” or “kept and then cooled” describing the final steps of a given synthesis.

Taken together, these results demonstrate that (a) the embeddings model we

created reflects well the similarity of the verbs used for synthesis descriptions and (b) the vocabulary of ULSA covers all common synthesis actions used in inorganic synthesis.

2.7 Analysis of graphs clustering

As we showed above, ULSA can capture well the vocabulary commonly used for the description of synthesis and, further, we were able to design a high-accuracy model that maps arbitrary synthesis descriptions into ULSA. Here, we would like to demonstrate how the dataset can be modified and augmented with the additional user-defined data to apply it to a specific task. To show that unification of synthesis actions still allows for distinguishing between inorganic synthesis types, we constructed synthesis flowcharts for 4,000 paragraphs (1,000 per each synthesis type) randomly pulled from the set of 420K inorganic synthesis paragraphs (see Section 2.3.3 for procedure description). To construct the flowchart of synthesis (represented by an adjacency matrix), we used the synthesis action terms assigned to each sentence in a paragraph. Additionally, we augmented `Mixing` actions with three categories, `Dispersion Mixing`, `Solution Mixing`, and `Ball-Milling`, by using heuristics and dictionary lookup (Section 2.3.1). It is important to note here that we assume a linear order of synthesis actions, i.e. that the sequence of sentences and synthesis actions in a paragraph corresponds to the true sequence of synthesis steps done during experiment. According to our estimation, this assumption is violated only in 2% of paragraphs in the 420K paragraphs set. All the adjacency matrices were flattened and concatenated, resulting in a matrix of size 100×4000 , i.e. 10×10 matrix per each of 4,000 paragraphs, where 10 is the size of the ULSA vocabulary minus `Non-altering` and with three additional mixing actions. Next, principal component analysis was used to perform dimensionality reduction of the matrix.

Figure 2.4 displays the projection of the 1st and 2nd principal components. Each data point here corresponds to one synthesis paragraph, i.e. one synthesis flowchart. Different colors highlight different types of synthesis. A few observations can be made from the plot. First, the clusters of synthesis procedures are well separated and aggregated according to the synthesis types. Specifically, the data

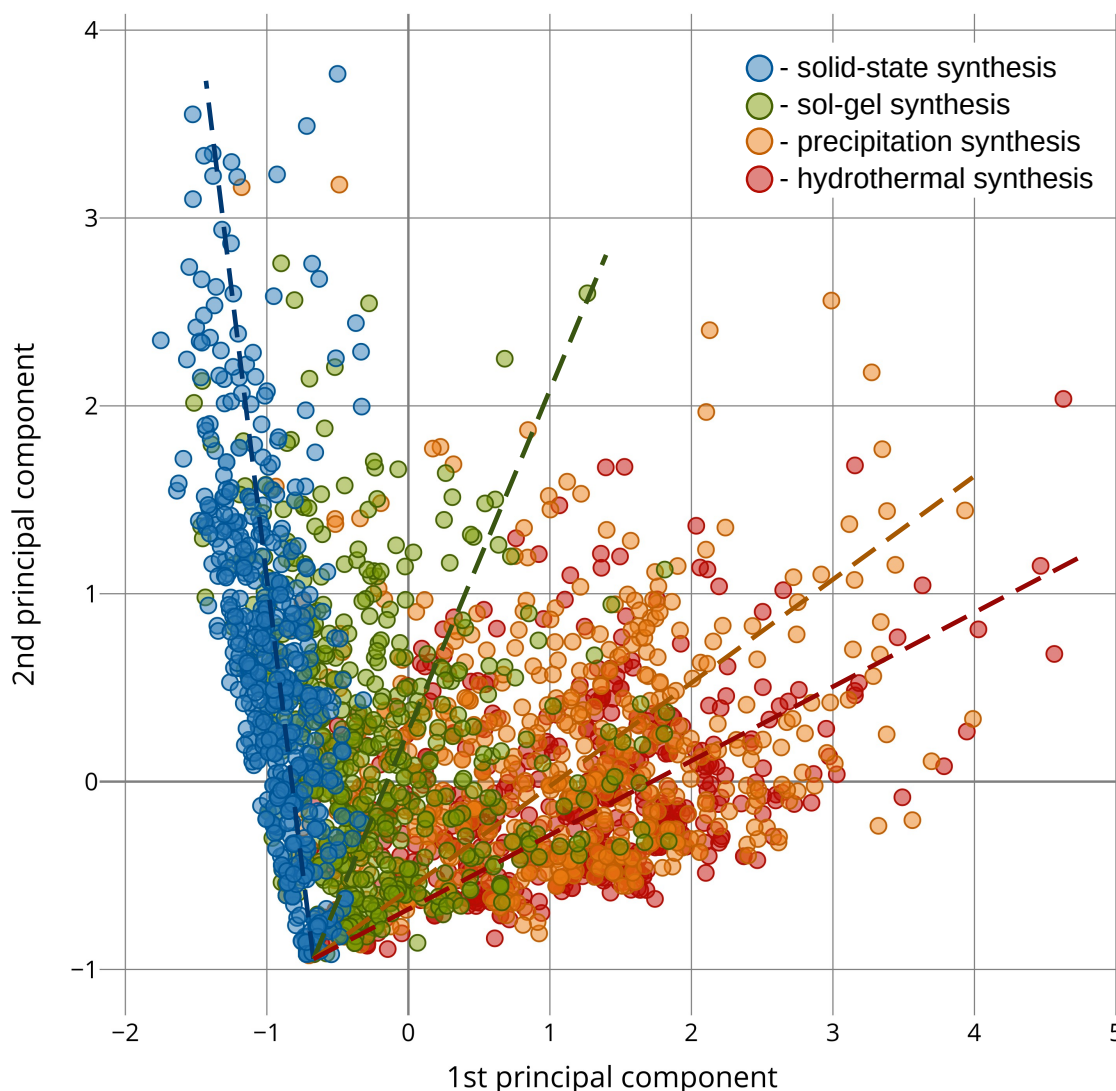


Figure 2.4: Visualization of the first two principal components for the adjacency matrices of synthesis action graphs. Each dot on the plot represents a synthesis graph colored according to its type. Dashed lines display linear fitting of each data subset and show the overall direction for clustering of each synthesis graph. Note that the lines were shifted to have a common origin for representation purposes while preserving the slope.

points corresponding to solid-state synthesis are narrowly clustered along a line with negative slope while the other synthesis types are spread more widely and the slope of their linear fit is positive. Second, the clusters of data points for precipitation and hydrothermal synthesis almost completely overlap and partially overlap

with sol-gel synthesis, while the overlap with solid-state synthesis is negligible.

These two observations agree well with the standard procedures associated with each of the four synthesis types. Indeed, solid-state syntheses usually operate with mixing powder precursors, firing the mixture, and obtaining final products; sol-gel synthesis is considered as a solid-state synthesis with solution-assisted mixing of precursors; hydrothermal and precipitation syntheses usually involve preparation of the sample in solution, then filtering (purification) to separate the liquid and obtain the final product instead of including a firing step.

To get further insights, we manually sampled and compared synthesis procedures corresponding to the data points along each of the fitted lines. The results show that the 1st principal component correlates with the involvement of solution mixing for precursors in synthesis procedures. In other words, the larger a coordinate the data point has along the 1st principal component, the more steps of dissolving and mixing precursors in solution as well as purification that data point involves. This agrees well with the fact that solid-state synthesis mostly operates with powders while hydrothermal and precipitation procedures are solution-based procedures, and sol-gel syntheses exist in between.

The 2nd principal component corresponds to the level of complexity of the synthesis procedure. The larger and more positive the data point along the 2nd principal component, the more steps are involved in the synthesis process. Interestingly, all four synthesis types exhibit simple synthesis procedures (fewer steps) and complex synthesis procedures (many steps). Nonetheless, solid-state synthesis has the largest deviation along the 2nd principal component compared to hydrothermal and precipitation synthesis since solid-state procedures can involve multiple heating and re-grinding steps for the sample to obtain the desired phase while in solution synthesis this can often be achieved in one or two steps.

2.8 Discussion and Conclusions

In this Chapter, we aim to fill the gap in automated synthesis information extraction from scientific publications by proposing a unified language for synthesis actions

(ULSA). We used the ULSA on an annotated set of 3,040 sentences about inorganic synthesis including solid-state, sol-gel, precipitation and solvo-/hydrothermal syntheses. The dataset is publicly available and can be easily customized by researchers accordingly to fit their application. As an example of such application, we used a recurrent neural network and grammar parsing to build a mapping model that converts written synthesis procedures into a ULSA-based synthesis flowchart. Analysis of the results demonstrates that the ULSA vocabulary spans the essential set of words used by researchers to describe synthesis procedures in scientific literature and that the flowchart representation of synthesis constructed using ULSA can capture important synthesis features and distinguish between solid-state, sol-gel, precipitation and solvo-/hydrothermal synthesis methods.

Despite these promising results, the ULSA scheme is not considered a complete language and can be significantly improved in the future with contributions from other researchers. First, we only demonstrated that it works for specific inorganic synthesis methods, and introduction of synthesis techniques such as deposition, crystal growth, and others will require extending the ULSA vocabulary or reconsidering the definitions of some terms. Second, the scheme and methodology will benefit from a robust approach to distinguish between various mixing procedures since this is one of the defining items in understanding synthesis protocols. This includes separation between, for example, dissolving precursors and dispersive mixing in a liquid environment, using ball-milling to homogenize the sample and using high-energy ball-milling to actually achieve the final product, adding reagents to promote reaction and adding precursors to compensate for loss due to volatility, and other cases. We have demonstrated that the details of mixing are important for distinguishing between inorganic synthesis methods using simple heuristics, however, the scheme will benefit from a high-fidelity approach. Nonetheless, we anticipate that our results and the ULSA schema will help researchers to develop a data-oriented methodology to predict synthesis routes of novel materials.

Efficient and controllable materials synthesis is a bottleneck in technological breakthroughs. While predicting materials with advanced properties and functionality has been brought to a state-of-the-art level with the development of computational and data-driven approaches, the design and optimization of synthesis routes for those materials is still a tedious experimental task. The progress in inor-

ganic materials synthesis is mainly impeded due to (a) a lack of publicly available large-scale repositories with high-quality synthesis data and (b) a lack of ontology and standardization for communication on synthesis protocols. Indeed, the first matter arises from the fact that the vast majority of experimental data gets buried in lab notebooks and is never published anywhere. As a result, researchers are liable to perform redundant and wasteful experimental screenings through those parameters of synthesis that have already been performed by someone, but are not reported. Even published experimental procedures face the problem of ambiguity of the language used by researchers. This creates a major challenge in acquiring synthesis data from publications by automated approaches including text mining.

The advantage of the paradigm we establish in this work is that it brings us closer to addressing important questions in materials synthesis: *“How should we think about the synthesis process?”*, *“What is the minimum information required to unambiguously identify a synthesis procedure?”*, and *“Can synthesis be thought of as a combination of fixed action blocks augmented with attributes such as temperature, time, and environment, or are there other important aspects that have to be taken into account?”*. These questions will become crucial when transitioning toward AI-driven synthesis.

2.9 Code and data availability

The dataset of 3,040 annotated synthesis sentences as well as the processing scripts are available

at CederGroupHub/synthesis-action-retriever at <https://doi.org/10.5281/zenodo.6383380>. In the dataset, each record contains the raw sentence tokens concatenated with a space between each token and a list of objects, each containing a token and the tag assigned to that token. For example:

```
{
  "annotations" :
  [
    {
      "tag" : token_tag,
      "token" : token
    }
  ]
}
```

```
        }  
    ],  
    "sentence" : sentence  
}
```

The repository also contains a script for training a bi-LSTM model that can be used to map words into action terms. Users are not limited to using only the provided dataset, but can augment their usage with other labeled data as long as they satisfy the data format described above. Finally, we also share scripts used for the inference of synthesis actions terms and for building synthesis flowcharts for a list of paragraphs. Examples of model application are available as well.

Chapter 3

Text-mined dataset of solution-based synthesis

Part of this chapter has been published in:

Wang, Z., Kononova, O., Cruse, K., He, T., Huo, H., Fei, Y., Zeng, Y, Sun, Y., Cai, Z., Sun, W., & Ceder, G. (2022). Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Scientific Data* 9, 231.

[\[1\]](#)

In this Chapter, we describe an advanced extraction pipeline (Figure. 3.1), which uses various advanced machine learning and natural language processing techniques to extract precise data for solution-based inorganic materials synthesis procedures from the scientific literature. Solution procedures are considerably more complex than solid state synthesis and require the precise extraction of not only the chemicals involved but also their respective amounts (since they determine concentration in solution). In addition, more complex organic and mixed organic-inorganic compounds are used to solubilize ions or to control solution conditions. By applying the extraction pipeline, we codified 35,675 solution-based inorganic materials synthesis procedures from over 4 million papers. Extracted information includes target material and precursors, their quantities, and the synthesis operations and their attributes. Information about the targets and precursors is then used to build a reaction formula for every synthesis procedure. This dataset is the first large-scale dataset of solution-based synthesis procedures, and provides a foundation to test and verify existing empirical synthesis rules, improve prediction accuracy, and even data-mine new rules to guide synthesis.

3.1 Content acquisition

The journal articles used in this work were downloaded with publisher consent from Wiley, Elsevier, the Royal Society of Chemistry, the Electrochemical Society, the American Chemical Society, the American Physical Society, the American Institute of Physics, and Nature Publishing Group. A customized web-scraper, Borges (see Codes Availability section below), was used to automatically download a broad selection of materials-relevant papers published after the year 2000 from publishers' websites in HTML/XML format. We selected 2000 as the cutoff year as parsing of materials science papers stored as image PDFs (as for most papers published before 2000) introduces a significant number of errors due to the limitations of currently available optical character recognition models on chemistry-containing text[95, 96].

To convert the articles from HTML/XML into raw-text files, we developed the LimeSoup toolkit (see Codes Availability section below), which takes into account the specific format standards of various publishers and journals. The full-text and

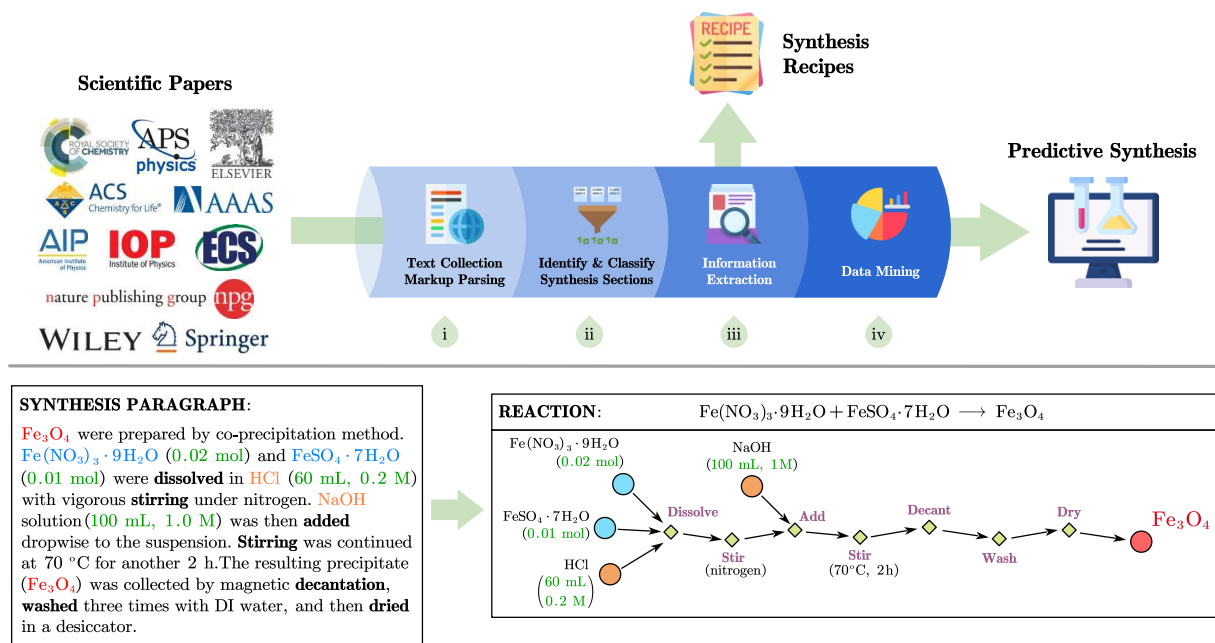


Figure 3.1: Extraction pipeline and example. *Top panel:* Schematic representation of the standard text mining pipeline: (i) scrape papers in markup format from the major publishers; (ii) identify and classify synthesis sections; (iii) extract key information including materials, amounts, sequenced operations, and conditions; (iv) store synthesis procedures into the database for future data mining. *Bottom panel:* Example of a codified procedure extracted from a synthesis paragraph.

metadata of the articles such as the journal name, article title, abstract, author names, etc., are stored in a MongoDB (www.mongodb.com) database collection. To date, we have accumulated 4.06 million articles, which are used for further processing down the pipeline (Figure 3.1).

3.2 Paragraph classification

Paragraphs containing information about solution synthesis (referred to as “synthesis paragraphs” throughout this paper) were identified using a Bidirectional Encoder Representations from Transformers (BERT) model [91]. The model was pre-trained on full-text paragraphs of 2 million papers randomly drawn from our database in a self-supervised way, i.e., by predicting masked words based on their surrounding context. After training the BERT model, we fine-tuned the paragraph classifier using 7,292 paragraphs labeled as either “solid-state synthesis”, “sol-gel

precursor synthesis”, “hydrothermal synthesis”, “precipitation synthesis”, or “none of the above”. The resulting F1 score of the paragraph classification is 99.5%, an improvement over the F1 score of 94.6% in our previous work [85], when evaluated using the same labeled training dataset.

3.3 Synthesis procedure extraction

A solution-based synthesis procedure includes the precursors and target materials, their quantities, and the synthesis actions and their attributes, properly sequenced. This is the minimum essential information required to complete a synthesis route. A schematic representation of the procedure is shown in the bottom panel in Figure 3.1. In the sections below, we provide a brief overview of the methods used for each step of the procedure extraction.

3.3.1 Materials entity recognition (MER)

Materials entities in synthesis paragraphs are identified and classified as *target*, *precursor*, or *other* via a two-step sequence-to-sequence model as introduced in our previous work [47]. In the current work, we replaced the original Word2Vec embedding model used previously [97] with a BERT model trained on papers from the materials science domain (see Section “Paragraph classification” above). First, each word token was transformed into a digitized BERT embedding vector. A bi-directional long-short-term memory neural network with a conditional random-field top layer (BiLSTM-CRF) was used to determine whether the token was a materials entity or a regular word, and each materials entity was replaced with the keyword <MAT> before being classified as either a *target*, *precursor*, or *other* material using a second BERT-based BiLSTM-CRF network. In addition to the 834 annotated solid-state synthesis paragraphs from 750 papers used in our previous work[47], we manually annotated 447 solution-based synthesis paragraphs from 405 papers by labeling each word token as *material*, *target*, *precursor*, or *outside*. The annotated dataset was split into training, validation, and test sets with a paper-wise ratio of 700:150:305 to train the aforementioned two neural networks.

3.3.2 Extraction of synthesis actions and attributes

We implemented an algorithm which combines a neural network and sentence dependency tree analysis to identify synthesis actions in the text. First, the Word2Vec model from the Gensim library [98] was re-trained on ~400,000 synthesis paragraphs of four synthesis types (see Section “Paragraph classification” above). These word embeddings were used as the input for a recurrent neural network that takes a sentence word-by-word and assigns labels to the verb tokens: *not-operation*, *mixing*, *heating*, *cooling*, *shaping*, *drying*, or *purifying*. For each obtained synthesis action, we parsed a dependency sub-tree using the SpaCy library [86] to obtain information about the corresponding temperature, time, and environment. To extract the corresponding values of these attributes, we used a rule-based regular expression approach [79].

3.3.3 Extraction of material quantities

To extract the numerical values of material quantities and assign them to the corresponding materials obtained using the MER model (see Section “MER” above), we applied a rule-based approach to search along the syntax tree [79]. The NLTK library [99] was used to build the syntax trees for each sentence in a paragraph. The words in given sentences are leaf nodes of syntax trees. We then applied an algorithm to cut the syntax tree of each sentence into the largest sub-trees for every material, with each sub-tree having only one material entity: 1. we first identified the materials on leaf nodes; 2. starting from each material, we identified the largest sub-trees, i.e., we traversed the syntax tree upwards until there was more than one material leaf node descending from the same node; 3. the largest sub-tree for a given material was defined as the sub-tree formed by the node and its descendants identified in step 2. Next, we searched for the quantities in each sub-tree given as molarity, concentration, or volume. Finally, we assigned the quantities found to the unique material entity in the sub-tree.

3.3.4 Building reaction formulas

For every synthesis procedure described in a paragraph, we built a chemical formula. Every material entity was converted from a text-string representation into a

chemical-data structure using an in-house material parser toolkit (see Codes Availability section below). The data structure included information about the material formula, composition, and ions. We then paired the target with precursors containing at least one element in the target except for hydrogen and oxygen and defined those precursors as “precursor candidates”. Next, we computed the oxidation state change of elements from each “precursor candidate” to the target and determined whether the precursor was oxidized or reduced. If precursors were reduced or oxidized, we also included the corresponding redox agents in the reaction formula. The agents can either be another “precursor candidate” or a commonly used oxidizing or reducing agent from the remaining material entities marked as *other* or *precursor* by the MER algorithm (see Section “MER” above).

3.4 Solution-based synthesis dataset

The dataset generation followed the protocol displayed in Figure 3.1. We downloaded a total of 4,061,814 papers using web scraping and identified the experimental sections by keyword matching in section headings, with keywords including “experiment”, “synthesis”, “preparation”, and their morphological derivations. ChemDataExtractor [41] was used to split the plain-text paragraphs into sentences and words. After classification (see Section “Paragraph classification” above), 364,076 paragraphs describing solid-state, hydrothermal, sol-gel, and precipitation syntheses were obtained. Among them, 189,553 paragraphs described hydrothermal or precipitation syntheses, which we categorize as solution-based synthesis methods. These paragraphs were further processed to extract the precursors, targets, quantities, and operations with corresponding conditions and to build the reaction formula (Figure 3.1).

The solution-based synthesis dataset is provided as a single JSON file, available at <https://doi.org/10.6084/m9.figshare.16583387.v4>[100]. There are 20,037 hydrothermal synthesis reactions and 15,638 precipitation synthesis reactions. Each record corresponds to a synthesis procedure extracted from a paragraph and is represented as an individual JSON object. If a paragraph reported the synthesis of several materials, the corresponding reactions were split into separate data records. In addition to the chemical formula, the metadata for each reaction

returns the data structure used in our previous work[54], which includes: DOI of the paper, a snippet of the corresponding synthesis paragraph (50 first and 50 last characters to facilitate its lookup), chemical information about the target and precursor materials used in the reaction, and operations with their corresponding attributes. We also included the materials with their corresponding quantities in the metadata. The details of the data format are given in Table 3.1

Data description	Data Key Label	Data Type
DOI of the original paper	doi	string
Snippet of the raw text	paragraph_string	string
Chemical formula	reaction	Object (dict): - left_side: list of strings - right_side: list of strings
Chemical formula in string format	reaction_string	string
Target material data	target	Object (dict): - material_string: string, - material_formula: string, - composition: list of Objects, ¹ - additives: list of strings - elements_vars: {var: list of strings} - amounts_vars: {var: list of Objects ² } - 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)
List of solvent formulas	solvents_string	list of strings
Sequence of synthesis steps and corresponding conditions	operations	list of Objects (dict): - token: string, - type: string - conditions: Object -- temperature: list of Objects ³ -- time: list of Objects ³ -- atmosphere: list of strings -- mixing_device: list of strings -- mixing_media: list of strings
Materials with corresponding quantities	quantities	list of Objects (dict): - material: string, - quantity: list of Objects ⁴
Synthesis type	type	string

¹ {formula: string, elements: {element: amount of element}, amount: string}² {max_value: float, min_value: float, values: list of floats}³ {max_value: float, min_value: float, values: list of floats, units: string }⁴ {number: float, unit: string }

Table 3.1: Format of each data record: description, key label, data type.

The chemical formula for the reaction is stored as a string (`reaction_string`) as well as in a dictionary containing lists of precursors (`left_side`) and target materials (`right_side`) in `reaction`.

The metadata for target materials and precursors used to construct the chemical formula are represented by the following data structure:

- **material_string**: string of material as given in the original paragraph before being parsed into a chemical composition.
- **material_formula**: chemical formula associated with the material (given originally or constructed empirically by parser).
- **composition**: chemical composition of the material derived from its formula. Aside from single-compound materials, we found that a large portion of the materials (predominantly target materials) are composites, mixtures, solid solutions, or alloys written as a sequence of compound-fraction pairs. Therefore, a chemical-composition entity is represented by a list of dictionary entries, where each item is associated with a compound found in the materials formula. The fraction of each compound in the material is given in **amount**, and its chemical composition (i.e., the elements and stoichiometry) is given in **elements**. If a material is one compound, the list has only one item and **amount**=1.0. If a material is a hydrate, water is added to the **composition** list with **amount** corresponding to the amount of water molecules (if specified).
- **additives**: list of additive elements (i.e., elements used for doping, stabilization, or substitution) resolved from the material string.
- **elements_vars**: lists all variable elements and their corresponding values found in the materials.
- **amounts_vars**: lists all variable element ratios and their corresponding values found in the material formula. The values of each variable are given as a structure with **values** listing the values of each specific variable and **max_value/min_value** values if a range is given in the paragraph.
- **oxygen_deficiency**: yes/no attribute that reflects whether a material was synthesized with unspecified oxygen stoichiometry.
- **mp_id**: ID of the lowest-energy polymorph entry in the Materials Project database (materialsproject.org) if the material is found there.

To facilitate querying of the dataset, the **targets_string** field contains the target material formulas, and the **solvents** field contains all solvent(s) from

matching material entities marked as *other* by the MER model with a table of common solvents adopted from Common Solvents Used in Organic Chemistry (organicchemistrydata.org/solvents).

3.5 Technical Validation

3.5.1 Extraction completeness and accuracy

To ensure high accuracy of the dataset, we included only those data that produced complete reaction formulas at the final step of the pipeline. This strategy reduced potential errors in the dataset that may have been caused by composition-parsing failure, incomplete extraction, or incomplete information provided by the text. We applied the extraction pipeline to 189,553 solution-based synthesis paragraphs, 28,749 of which generated a reaction formula, giving an extraction yield of ~15%. To evaluate the source of the loss, we randomly selected and manually checked 100 solution-based synthesis paragraphs that did not produce any reactions. Among those 100 paragraphs, 36 were written with an incomplete list of precursors or targets in the text, such that human experts would not be able to reconstruct the reaction based solely on the information provided in the paragraph. For the remaining 64 paragraphs, the loss was due to: 1. the use of organic precursors with complex groups or complicated notation (e.g., acronyms) that could not be parsed into a chemical composition by our parser or 2. MER misidentification resulting in an incomplete or incorrect list of precursors and (or) target entities such that the reactions could not be built.

To evaluate the quality of the dataset, we had a human expert test 100 randomly pulled entries. The human expert manually extracted the information presented in the procedure, and the results were compared with those extracted by the pipeline. Table 3.2 presents the accuracy statistics, which include the precision, recall, and F1 scores calculated from the tested entries. For the fields that included reaction, targets, precursors, operations, operation temperatures, time, and atmosphere, the F1 scores were over 90%. The relatively low recall, and hence F1 score, for the extraction of materials quantities can be mainly explained by the MER algorithm missing the corresponding material entity and, thus, the quantities not

Data attribute	Precision	Recall	F1 score
Reaction formulas	0.94	/	/
- targets	0.97	/	/
- precursors	0.98	0.99	0.98
Operations	0.96	0.85	0.90
Conditions			
- temperature	0.97	0.92	0.94
- time	0.98	0.89	0.93
- atmosphere	0.97	0.92	0.94
Quantities	0.90	0.85	0.87

Table 3.2: Performance of data extraction for dataset entries.

being assigned. The accuracy of the obtained dataset is comparable to that in our previous work[54] and of other text-mined datasets[50].

3.5.2 Exploratory data analysis

To test the diversity of the dataset and its coverage of the materials space, we analyzed unique materials (targets and precursors) and reactions. The dataset contains 11,603 unique reactions that include 2,870 unique precursors and 5,416 unique targets. The ten most frequent targets in the dataset and their corresponding precursors are listed in Table 3.3. The target list captures materials that have drawn substantial attention in the past two decades: catalysts (ZnO, Fe₂O₃, TiO₂, Fe₃O₄, SnO₂, ZrO₂, CuO), adsorbents (SiO₂), various materials for sensors (ZnO, Fe₂O₃, WO₃), quantum dots (CdS), and semiconductors (ZnO, TiO₂, SnO₂, CdS). Unsurprisingly, these most frequent target materials usually appear in multiple applications, as they possess desirable physical and chemical properties in many scientific and engineering fields.

Targets	Common Precursors
ZnO	Zn(NO ₃) ₂ , Zn(Ac) ₂ , ZnCl ₂
TiO ₂	Ti(OCH(CH ₃)) ₄ , Ti(OC ₄ H ₉) ₄ , TiCl ₄
Fe ₃ O ₄	FeCl ₃ , FeCl ₂
Fe ₂ O ₃	FeCl ₃ , Fe(NO ₃) ₃
SnO ₂	SnCl ₄
ZrO ₂	ZrOCl ₂ , ZrO(NO ₃) ₂
CuO	Cu(NO ₃) ₂ , Cu(Ac) ₂ , CuCl ₂ , CuSO ₄
SiO ₂	Si(OC ₂ H ₅) ₄
WO ₃	Na ₂ WO ₄ , WCl ₆ , H ₂ WO ₄
CdS	Na ₂ S, CH ₄ N ₂ S, CdCl ₂ , Cd(NO ₃) ₂

Table 3.3: Ten most common targets in the dataset with their corresponding precursors.

We use the periodic table representation (Figure 3.2) to visualize the chemical space covered by the dataset. For each element, the fraction of synthesis procedures containing this element in the target formula is shown with the yellow-to-navy blue gradient framed at the top of each element box. The most data-rich elements are transition metals in the third period, such as Zn, Fe, Ti, Ni, and Co, in accordance with the compounds listed in Table 3.3. The next-most prevalent targets are materials with Bi, Sn, Al, W, Mo, Cu, Zr, or Li. The least common elements are rare elements such as Ru, Rh, Hf, Ta, Re, and Ir. The elements Fr, Ra, Tc, and Pm are not present as target materials in the dataset, likely due to their radioactivity. Additionally, we calculated the frequency of co-occurrence of chemical elements and common ions in precursor materials to understand how different ions are brought into solution. In Figure 3.2, the frequencies for each ion are displayed as colored bars. The length of the bar is the fraction of one specific ion paired with the element normalized over all precursors for this element.

The commonly used precursors are mainly those that are widely available from companies such as Sigma-Aldrich and Fisher Scientific. For example, Li₂CO₃ or LiOH for Li and sulfate or chloride for Fe. Inorganic salts, such as nitrates, sulfates, and chlorides, are often used because of their high solubility [101]. Halides are not used for Lanthanide metals because Lanthanide halides are highly hygroscopic,

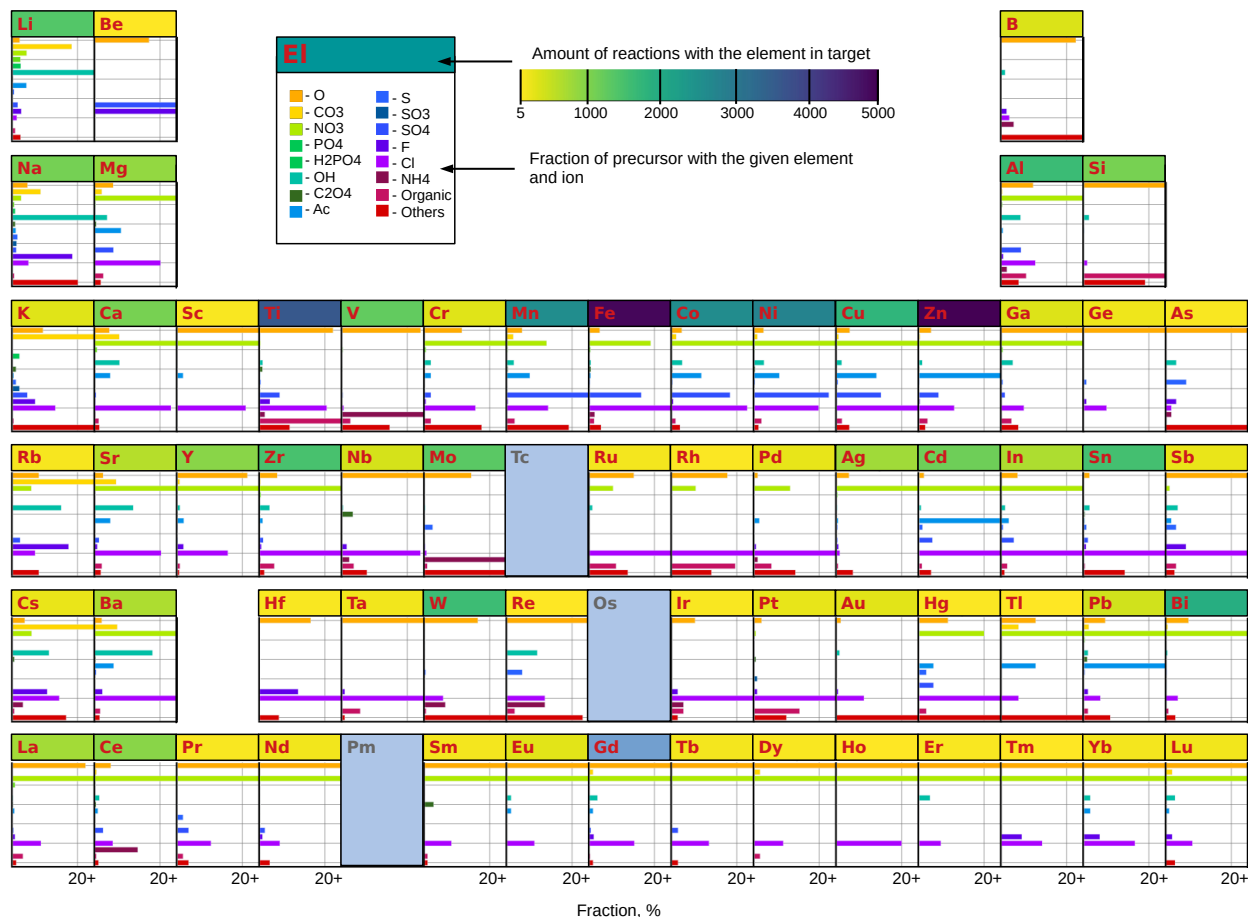


Figure 3.2: The chemical space covered by the dataset. For each element, the box containing the element name is colored in a yellow-to-navy blue gradient representing the total amount of reactions that produce a target compound containing the element. The bar graph below each element shows the list of ions paired with the element in precursor compounds. The fractions of the precursors (i.e. element+ion) used are shown by the length of the bars. Boxes with no bar graph represent elements occurring in five and fewer targets. “Ac” stands for acetate radical CH_3COO^- in the compound formula.

and thus their molar weight is not well defined. We observed that precursors with neighboring elements in the periodic table tend to have similar ions paired. For instance, nitrates, sulfates, and chlorides are commonly used anions for 3rd-period transition metals, whereas the precursors for lanthanides are mostly oxides and nitrates.

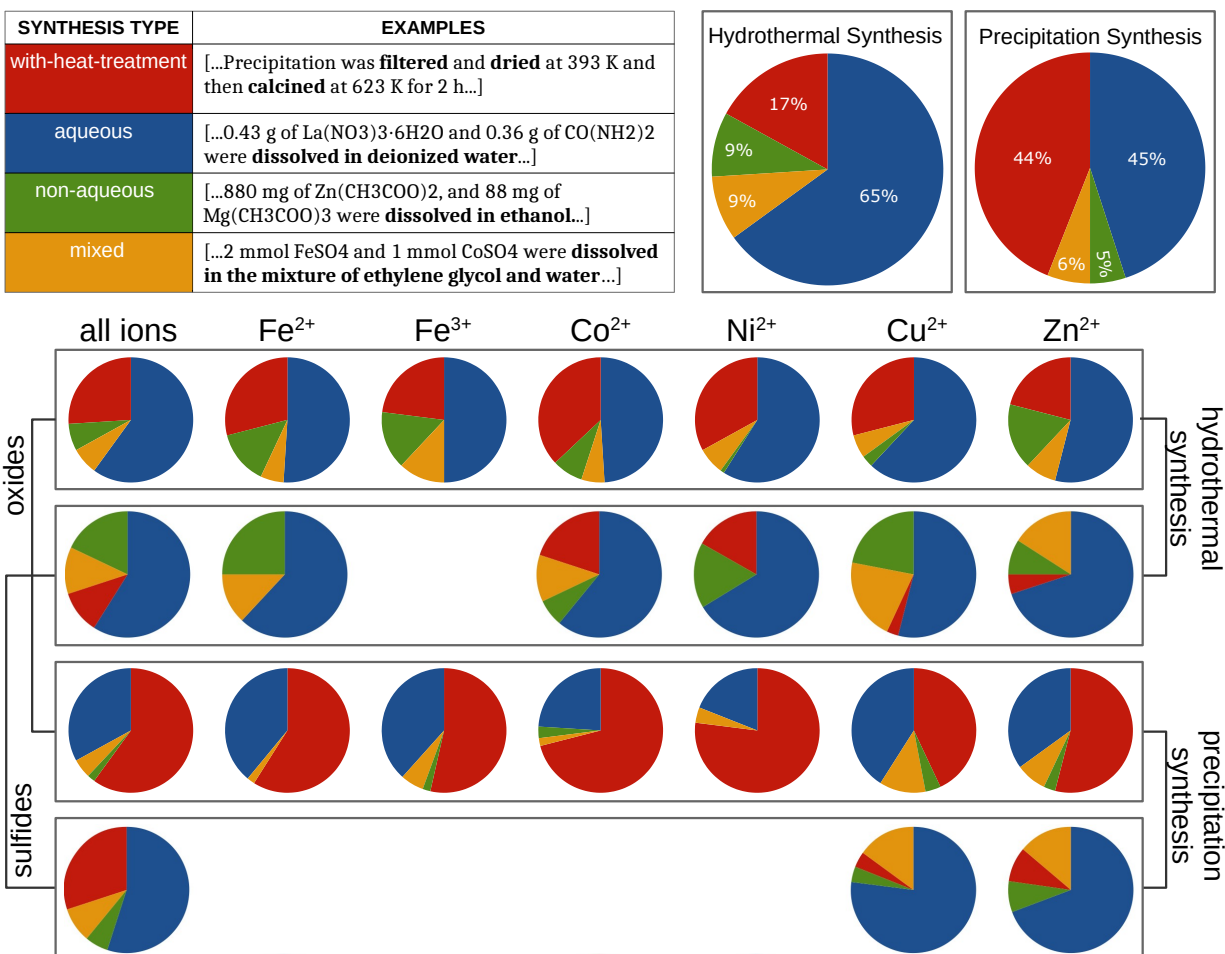


Figure 3.3: Correspondence between choice of synthesis route and selected types of targets. The top table gives an example of the four synthesis categories defined: with heat treatment step, aqueous, non-aqueous, and mixed. The two pie-charts on the top-right show the fractions of synthesis routes in the hydrothermal and precipitation datasets separately. The four rows of pie charts in the lower half of the figure represent the fractions of the four synthesis routes (given in the table) for all oxides, all sulfides, and individual oxides and sulfides with different oxidation states of data-rich transition metals separately. The first and second rows are results from the hydrothermal dataset. The third and fourth rows are results from the precipitation dataset. Each blank space means that there is not enough data to form a statistic for the corresponding type of target.

We used information about the extracted materials and sequences of synthesis actions to classify the solution-based synthesis procedures into four categories of synthesis protocols (table in Figure 3.3) according to the following definitions:

- *solution-mixing with heat treatment step* has a final heat treatment step after

the precipitate is obtained from the solution;

- *aqueous solution synthesis* has no final heat treatment step after precipitating the compound from the solution and the solvent is water;
- *non-aqueous solution synthesis* has no final heat treatment step after precipitating the compound from solution and the solvent(s) is (are) organic;
- *aqueous–non-aqueous mixed solution synthesis* has no final heat treatment step after precipitating the compound from solution and the solvents are a mixture of water and organic solvent(s).

The resulting distributions of synthesis protocols over the aforementioned categories are shown in the two pie charts in the top-right corner of Figure 3.3. Note that as solution-based synthesis includes both hydrothermal and precipitation synthesis according to our definition (see Section “Paragraph classification”), we analyzed these synthesis types separately. As observed in the pie charts, only 20% of the procedures in the hydrothermal synthesis subset have a heat treatment step after solution mixing. Among those that do not have heat treatment step, 63% use only water as a solvent, 8% use only organic solvents, and 9% use both water and organic solvents. In contrast, the fractions in the precipitation synthesis subset are 43%, 46%, 5%, and 6%, respectively.

A heat treatment step after solution mixing can be used to dehydrate the targets, decompose the intermediates to produce the final products, change the oxidation state, change the morphology, or improve crystallization[102–104]. To explore this in more detail, we split the targets according to their anion type (oxide, sulfide, etc.) and different oxidation states of several data-rich transition-metal elements. We then computed the distribution of synthesis categories for each of the split subsets. Figure 3.3 presents the results for the most prevalent subsets of oxides, sulfides, and elements Fe^{2+} , Fe^{3+} , Co^{3+} , Ni^{2+} , Cu^{2+} , and Zn^{2+} . The fraction of procedures with a heat treatment step in precipitation synthesis is larger than that in hydrothermal synthesis. This observation holds for all targets, all oxides, all sulfides, and individual oxides and sulfides with queried oxidation states of transition metals. This finding can be also interpreted as hydrothermal synthesis often being used to obtain final products in a “one-shot” process, without subsequent heat treatment after

solution mixing, likely because many compounds can be crystallized as anhydrous powders with controlled size and morphology directly from hydrothermal synthesis. In a standard hydrothermal synthesis procedure, the reaction is performed in an autoclave with autogenic pressure so that it can operate in a wider temperature window, including temperatures above the atmospheric boiling point of the solvent. In contrast, precipitation synthesis is performed under normal pressure. The higher temperature possible in hydrothermal synthesis is associated with enhanced kinetics in chemical transport, nucleation, and crystal growth and thus with a more effective dissolution–recrystallization process, which can help remove defects and improve crystallinity. Furthermore, the physico-chemical properties, such as the viscosity and dielectric constant of water or other solvents, change pronouncedly under conditions of hydrothermal synthesis, affecting the solubility and mobility of species in the solution and eventually facilitating crystallization [105]. Therefore, hydrothermal synthesis does not need a post-synthesis heat treatment as often as precipitation synthesis.

3.6 Conclusion

In this Chapter, we describe an advanced text-mining pipeline to generate a codified dataset of 35,675 solution-based inorganic materials synthesis procedures from over 4 million papers. Each procedure contains essential synthesis information, including the precursors and target materials, their quantities, and the synthesis actions and corresponding attributes. Tested by human experts, the precision of reaction formulas reconstruction and F1 scores of synthesis actions and attributes extraction were over 90%, demonstrating the state-of-the-art performance of our text-mining pipeline.

Because the dataset contains detailed information about chemical formulas as well as the compositions of the target materials and precursors for each procedure, it can be easily used to conduct a literature review by querying desired precursors and (or) targets in different chemical spaces. For example, selecting all TiO_2 synthesized from TiCl_4 allows an exploration of how other synthesis formulations, such as synthesis actions, attributes, and quantities, affect the results. Furthermore, the materials entries in the dataset are supplied with the Materials Project [6] iden-

tifiers, thus facilitating the integration of the procedures with the thermochemical data available in the Materials Project[106, 107]. In addition, this solution-based synthesis dataset keeps the same data structure as that in the solid-state dataset generated in our previous work [54]. Therefore, it is easy to analyze the procedures from the two datasets. More importantly, the codified text-mined dataset can be used to data-mine new synthesis rules to guide future synthesis[108]. In the next Chapter, we will show how to use this text-mined solution-based synthesis dataset to test the new synthesis hypothesis.

3.7 Code and data availability

The scripts used to classify paragraphs and extract procedures as well as to perform the data analysis are home-written codes which are publicly available at the GitHub repository: https://github.com/CederGroupHub/text-mined-solution-synthesis_public.

The underlying libraries used in this project are all open-source:

Tensorflow (www.tensorflow.org)

Keras (keras.io)

SpaCy (spacy.io) [86]

NLTK (<https://www.nltk.org/>) [99]

gensim (radimrehurek.com) [98]

scikit-learn (scikit-learn.org) [109]

ChemDataExtractor (chemdataextractor.org) [41]

Material Parser (github.com/CederGroupHub/MaterialParser)

Borges (github.com/CederGroupHub/Borges)

LimeSoup (github.com/CederGroupHub/LimeSoup).

Chapter 4

Minimum thermodynamic competition for synthesis design

Part of this chapter has been included in:

Wang, Z., Sun, Y., Cruse, K., Zeng, Y., Fei, Y., Liu, Z., Shangguan, J., Byeon, Y., Jun, K., He, T., Sun, W., & Ceder, G.. Optimal thermodynamic conditions to minimize kinetic byproducts in aqueous materials synthesis. Working paper which is pending publication at the time of filing this thesis.

We hypothesize that one avenue to determine favored synthesis conditions is to maximize the difference in driving force with competing phases. We define thermodynamic competition as the difference between the free energy of the target phase and the minimum free energy of all competing phases[110], and name the hypothesis “minimum thermodynamic competition” (MTC). Importantly, this identifies a specific *point* in thermodynamic space for optimal materials synthesis—in contrast with a stability *region* from the thermodynamic phase diagram. We note here that a similar concept based on “maximum driving force” was recently proposed by Walters and Rondinelli for the determination of which oxide phases form to protect against metal corrosion [110].

In this Chapter, we develop a synthesis optimization strategy that relies on the inherent concave geometry of the free-energy landscape for an aqueous electrochemical system (see Methods section), which leads to an efficient computational algorithm that is readily scalable to high-dimensional (multi-component) optimization spaces. We provide explicit empirical confirmation of our MTC hypothesis using two complementary approaches, as shown in Figure 4.1. First, we performed a large-scale analysis on a text-mined dataset of solution synthesis procedures[1], and found that within the enormous available parameter space of synthesis conditions, the empirically-reported syntheses lie near our predicted thermodynamic optimal conditions (i.e., the condition with minimum thermodynamic competition). Second, we performed a detailed thermodynamic competition analysis of two systems, $\text{LiIn}(\text{IO}_3)_4$ and LiFePO_4 , and found that the desired target materials can be synthesized when the thermodynamic competition is very negative. Our work illustrates the how a comprehensive understanding of the free-energy axis of the phase diagram, considering not only the stable phase but also its competing phases, enables the computational design and guidance of rational materials synthesis conditions.

4.1 Thermodynamic competition in aqueous solution-based systems

Here, we define a metric that captures both thermodynamic and kinetic aspects to guide synthesis design precisely. Our approach aims to maximize the thermody-

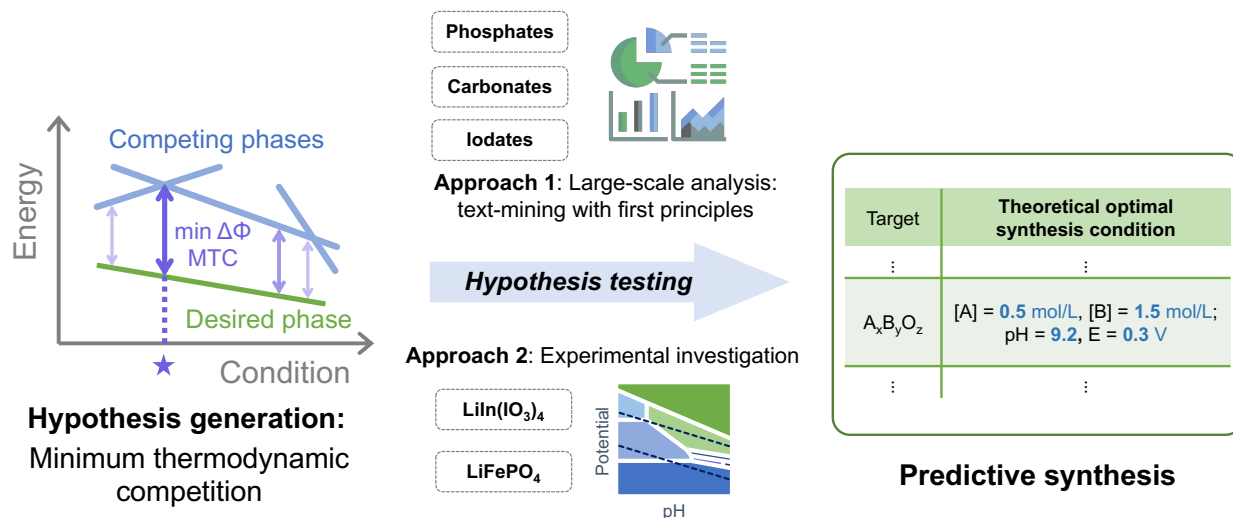


Figure 4.1: Schematic representation of approach for predicting aqueous solution-based synthesis conditions: 1) Generation of the hypothesis: conditions where target phase experiences minimum thermodynamic competition from other phases are favored for synthesis. 2) Hypothesis testing: Large-scale analysis based on the text-mined dataset and detailed investigation of $\text{LiIn}(\text{IO}_3)_4$ and LiFePO_4 . 3) Predictive synthesis: minimize the thermodynamic competition to predict synthesis condition.

dynamic driving force to the desired target and minimize that to competing phases, thus leading to favored synthesis conditions that maximize the chance for high phase purity while avoiding the formation and persistence of competing kinetic products[111]. When considering the stability of a target material, the thermodynamic boundary conditions should be framed to examine the stability response of the target material with respect to an open external chemical reservoir. In such case, the natural variables of the thermodynamic system are intensive, and the resulting free energy surfaces are concave. This transforms the conditions of heterogeneous equilibrium from a convex hull—which applies to extensive variables such as composition—to the concave envelope, which is solved by half-space intersection[112]. From this geometric description of the energy landscape, we can define the competition that the desired target phase experiences with other phases in the thermodynamic system with respect to the intensive natural variables, which enables us to determine a condition where thermodynamic competition is minimized.

The thermodynamic competition that the desired target phase experiences with

other phases is defined as the difference between the free energy of the desired target phase and the minimum free energy of all competing phases, whose magnitude corresponds to the length of the gap between the energy surface of desired target phase and this the inner envelope of the intersecting free-energy surfaces of competing phases. A schematic plot of thermodynamic competition is shown in the leftmost panel of Figure 4.1. Mathematically, we denote phase k as our desired target phase, Y as intensive variables, and I_c is the index set of other competing phases. The thermodynamic competition that the phase k experiences from other phases, k experiences from other phases, $\Delta\Phi(Y)$, can be written as:

$$\Delta\Phi(Y) = \Phi_k(Y) - \min_{i \in I_c} \Phi_i(Y) \quad (4.1)$$

Here, $\Phi_k(Y)$ is the free energy of the desired target phase, and $\min_{i \in I_c} \Phi_i(Y)$ is the minimum free energy of all competing phases. As $\Phi_k(Y)$ decreases, the thermodynamic competition that the desired target phase experiences with other phases becomes weaker. Specifically, $\Delta\Phi(Y) \leq 0$ means the desired target phase is the stable phase as the free energy of the desired target phase is smaller than the minimal free energy of competing phases. Thus, the thermodynamic optimal condition where the thermodynamic competition that phase k experiences is minimized is written as:

$$Y^* = \arg \min_Y \Delta\Phi(Y) = \arg \min_Y \left[\Phi_k(Y) - \min_{i \in I_c} \Phi_i(Y) \right] \quad (4.2)$$

In an aqueous solution-based system, the Pourbaix potential, Ψ , is used to depict the free energy of a phase with respect to pH and redox potential, E [107].

$$\bar{\Psi} = \frac{1}{N_M} \left((G - N_O \mu_{H_2O}) - RT \cdot \ln(10) \cdot (2N_O - N_H) pH - (2N_O - N_H + Q) E \right) \quad (4.3)$$

Here, N_M , N_O , N_H , and Q are the number of metal, oxygen, and hydrogen in the composition and the charge of phase, respectively. The molar Gibbs free energy of a phase, G , is its chemical potential, $\mu_i = \mu_i^0 + RT \ln[a_i]$, where μ_i^0 is given by the standard-state Gibbs formation free energy. An ideal solid with no defects has an activity of one, and so the $RT \ln[a_i]$ term is zero, but the chemical potential of aqueous ions ideally scales with the natural log of the metal ion concentration in solution. These concentration parameters produce additional synthesis parameters which need to be optimized along with pH and E for synthesis; which increases the

overall dimensionality of the synthesis problem. The Pourbaix potential, given by Eq (4.3), can be used to compute the thermodynamic competition that the desired target phase experiences from competing phases directly in the aqueous solution system with respect to intensive variables.

4.2 Large-scale analysis combining text-mining with first principles calculations

Recently, we used a variety of natural language processing algorithms to text-mine 35,675 solution synthesis procedures from the literature[1]. Here, we combine this dataset with first-principles thermochemical data from the Materials Project to interrogate our MTC hypothesis. When optimizing a solution synthesis procedure, there are many degrees of freedom available to experimental chemists—including the precursors, precursor molar concentration, pH , and redox potential. Often-times, experimentalists invest substantial trial-and-error efforts to optimize these synthesis parameters to produce the desired target phase. Here, we will show that the reported text-mined solution synthesis conditions are near the conditions predicted to be optimal by the MTC, which provides a post-hoc validation of the MTC hypothesis.

First, we confirm that the majority of materials in our text-mined dataset are stable phases on the Pourbaix diagram under some thermodynamic condition, and therefore exhibit negative thermodynamic competition ($\Delta\Phi(Y) \leq 0$). On Figure 4.2(a), we plot the distributions of thermodynamic competition, $\Delta\Phi(Y)$, for both the MTC-predicted thermodynamic optimal conditions (left violin plot) versus the reported text-mined synthesis conditions (right violin plot), using kernel density estimates (KDE) of the thermodynamic competition distributions for different ionic compounds, including phosphates, carbonates, and iodates (see Methods).

Given the fact that these materials have been synthesized, we can directly quantify the magnitude of the MTC that is necessary to synthesize materials in aqueous solutions. We find that phosphates have a 90th percentile $\Delta\Phi(Y) = -47$ meV/atom, carbonates at $\Delta\Phi(Y) = -122$ meV/atom, and iodates at $\Delta\Phi(Y) = +19$ meV/atom. This quantifies the MTC that was required to synthesize many known

ionic compounds, and serves as a guideline for the expected MTC magnitude when evaluating the synthesizability of novel computationally-predicted materials.

While Figure 4.2(a) shows overall distributions between the thermodynamic optimal conditions and the text-mined condition distributions, Figure 4.2(b) shows for specific compounds the 1-to-1 difference in energy between their text-mined thermodynamic conditions and the corresponding MTC-predicted optimal conditions. Going forward, we refer to this difference in experimental versus theoretically optimal energy difference as $\Delta(\Delta\Phi)$ or “energy above optimum”.

Figure 4.2(b) shows that text-mined reported procedures are always slightly non-optimal, which is evidenced by the a peak in probability distribution at $\Delta(\Delta\Phi)$ slightly above the exact MTC-predicted optimal conditions, where $\Delta(\Delta\Phi) = 0$. This suggests that either the experimentalists did not find the actual optimal synthesis condition, or that kinetic considerations shift the actual optimal synthesis condition slightly off from the thermodynamic optimal conditions. Nevertheless, the magnitude of $\Delta(\Delta\Phi)$ (medians ranging from 20 to 140 meV/atom) is much smaller than the magnitude of thermodynamic competition (medians ranging from -96 to -377 meV/atom) shown in Figure 4.2(a), which suggests that the experimentally-optimized conditions are near the thermodynamic optimal conditions. Most importantly, this provides a large-scale empirical validation that the MTC provides synthesis conditions that are likely to be experimentally viable, which will reduce the relevant trial-and-error processes.

A final intriguing observation from Figure 4.2(a) is that many synthesized materials have text-mined $\Delta\Phi(Y)$ values that are positive, which illustrates the opportunity that some compounds can be synthesized under conditions where they are not the lowest free-energy phase. For example, the 90th percentile boundary for phosphates is $\Delta\Phi(Y) = +83$ meV/atom for phosphates and +34 meV/atom for iodates. This compares closely to the 90th percentile of inorganic crystalline metastability, which we previously reported to be 67 meV/atom energy above the convex hull in [20]. It is well understood that the relatively low temperatures of solution synthesis often allow *chemie douce* kinetics[113] to dominate structure-selection during nucleation[114, 115], which may facilitate the formation of metastable compounds. This text-mined dataset here provides explicit quantita-

tive validation of this anecdotal intuition for the realization of metastable materials.

Overall, our text-mined dataset provides explicit post-hoc empirical validation of the MTC hypothesis, and furthermore highlights the value of such text-mined datasets in not only providing data to train machine-learning models[55], but also in empirically-validating new mechanistic theories.

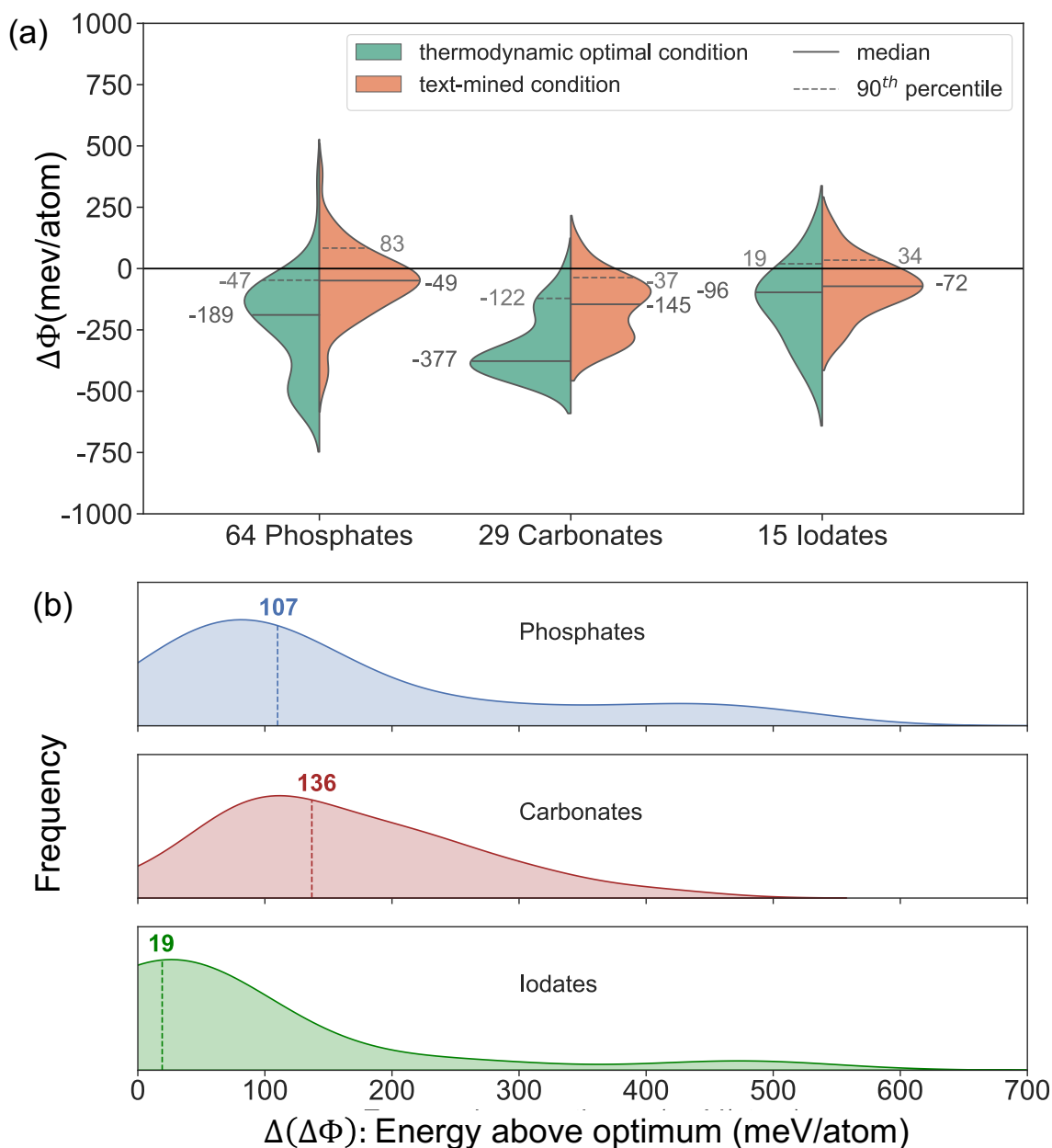


Figure 4.2: Thermodynamic competition analysis on text-mined dataset. *Upper panel:* Gaussian kernel density estimates of thermodynamic competition distributions of different multicomponent functional oxides. The green and orange parts are the distributions of different compounds in their corresponding thermal optimal conditions and text-mined conditions, respectively. The numbers on the plot are medians and 90th percentiles for different distributions. *Lower panel:* Gaussian kernel density estimates of the thermodynamic competition differences of different compounds between the text-mined conditions and their corresponding thermal optimal conditions. The top number is the median for all entries.

4.3 Systematic synthesis of $\text{LiIn}(\text{IO}_3)_4$ and LiFePO_4

In addition to our text-mined dataset validation of the MTC hypothesis, we also directly validate the MTC hypothesis by using it to guide the successful experimental synthesis of multicomponent functional oxides. Here, we apply the MTC to search for optimal aqueous electrochemical conditions to synthesize $\text{LiIn}(\text{IO}_3)_4$, a promising superionic conductor[116] and LiFePO_4 , a commercialized cathode material[117]. This represents two five-dimensional thermodynamic spaces in which to find optimized conditions, which are $[\text{Li}]-[\text{In}]-[\text{I}]-p\text{H}-E$ for $\text{LiIn}(\text{IO}_3)_4$ and $[\text{Li}]-[\text{Fe}]-[\text{P}]-p\text{H}-E$ for LiFePO_4 . We successfully show that only conditions with $\Delta\Phi(Y) \leq -55$ meV/atom and $\Delta\Phi(Y) \leq -29$ meV/atom result in the phase-pure $\text{LiIn}(\text{IO}_3)_4$ and LiFePO_4 , respectively, with MTC analysis. Notably, this information is inaccessible from traditional Pourbaix diagrams, which only show the stability region; they do not give the thermodynamic optimum.

4.3.1 $\text{LiIn}(\text{IO}_3)_4$ synthesis analysis

First, we focused on synthesizing phase-pure $\text{LiIn}(\text{IO}_3)_4$, which was predicted to exhibit one of the highest room temperature lithium ionic conductivity reaching up to 18 mS/cm in the pristine state[116]. Notably, when we attempted to reproduce the synthesis of this material using the published procedure from Yang et al. [118], we were unable to reproduce the target compound. However, using the MTC, we were able to identify robust synthesis conditions that led to the successful synthesis of $\text{LiIn}(\text{IO}_3)_4$.

MTC indicates that the most competitive phase (i.e., the phase with the lowest energy among all competing phases) to be $\text{In}(\text{IO}_3)_3$. The MTC analysis further suggests that to promote $\text{LiIn}(\text{IO}_3)_4$, we should maximize the aqueous concentration of $[\text{Li}]$, also drive the solution towards a highly oxidizing and acidic environment. To interpret this prediction conceptually, we can rationalize this prediction because increasing the concentration of Li^+ increases the supersaturation of $\text{LiIn}(\text{IO}_3)_4$, while it has no effect on the supersaturation of the competitive non-Li-bearing phase $\text{In}(\text{IO}_3)_3$. The oxidizing environment is required because the polyanion IO_3^- can only exist at a high redox potential, as the standard half-cell potential for decomposition to I_2 is at $(E^\ominus(\text{IO}_3^-/\text{I}_2(s)) = 1.18\text{V})$. Simultaneously, an acidic

environment is required to avoid the hydrolysis of In^{3+} to phases like InOOH . Following these MTC recommendations, we designed a series of synthesis experiments that spanned a range of $\Delta\Phi(Y)$ to evaluate the predictive power of our MTC analysis.

To create a highly oxidizing and acidic environment, we used an H_5IO_6 ($E^\ominus(\text{IO}_6^{5-}/\text{IO}_3^-) = 1.6\text{V}$) precursor as the iodine source. The experimental synthesis conditions and corresponding thermodynamic competition and final products are given in Table 4.1. Condition A1 is the optimal condition predicted by the MTC using LiCl as Li source. Condition A3 replicates the condition reported by Yang et al.[118], in which $\text{Li}(\text{CH}_3\text{COO})$ is used as Li source. A buffer solution could be formed with the existence of CH_3COO^- and it can influence the $p\text{H}$ change during reaction process. To keep this factor consistent, condition A2 is designed by changing the equal amount of LiCl in the precursor of A1 to $\text{Li}(\text{CH}_3\text{COO})$. To further investigate the influence of $[\text{Li}]$ and $p\text{H}$, conditions A4, A5, and A6 are designed with intermediate $[\text{Li}]$, using LiCl , $\text{Li}(\text{CH}_3\text{COO})/\text{LiCl}$ mixture, and $\text{Li}(\text{CH}_3\text{COO})$ as Li source, respectively. The redox potential data is taken from Pourbaix's Atlas[119]. The reduction potential of H_5IO_6 is much higher than the upper limit of the electrochemical stability window of H_2O ($E^\ominus(\text{O}_2/\text{H}_2\text{O}) = 1.23\text{V}$). During the reaction, IO_6^{5-} oxidizes H_2O and generates IO_3^- together with O_2 .

tag	[Li]	[In]	[I]	pH	E(V)	$\Delta\Phi(Y)(\text{meV/atom})$	XRD results
A1	1.5	0.1	0.8	0.66	1.61	-56	$\text{LiIn}(\text{IO}_3)_4$
A2	1.5	0.1	0.8	0.77	1.61	-55	$\text{LiIn}(\text{IO}_3)_4$
A3	0.15	0.1	0.8	0.91	1.60	-8	$\text{In}(\text{IO}_3)_3$
A4	0.8	0.1	0.8	0.71	1.61	-13	$\text{In}(\text{IO}_3)_3$
A5	0.8	0.1	0.8	0.87	1.60	-15	$\text{In}(\text{IO}_3)_3$
A6	0.8	0.1	0.8	2.6	1.50	-32	$\text{In}_2\text{O}_3 + \text{InOOH}$

Table 4.1: Synthesis conditions with thermodynamic competition analysis and experimental results for $\text{LiIn}(\text{IO}_3)_4$ synthesis.

$\text{LiIn}(\text{IO}_3)_4$ was successfully synthesized under conditions A1 and A2, whereas $\text{In}(\text{IO}_3)_3$ was the product phase for conditions A3-A5. The product phase for condition A6 is $\text{In}_2\text{O}_3 + \text{InOOH}$, which are the hydrolysis products of In^{3+} at high

pH. Figure 4.3 presents the XRD results (right panel) and the Pourbaix potentials of several competitive phases with respect to the $\text{LiIn}(\text{IO}_3)_4$ on conditions A1-A6 (left panel).

In line with the MTC theory, our experiments validate that the more negative the thermodynamic competition, the higher the kinetic propensity to synthesize a phase-pure material. Importantly, we also show that even within the Pourbaix stability region of $\text{LiIn}(\text{IO}_3)_4$, where $\Delta\Phi < 0$, impurity phases can form when $\Delta\Phi$ is not negative enough. Under conditions A1 and A2, the $\Delta\Phi$ are -56 meV/atom and -55 meV/atom, which means the thermodynamic competition $\text{LiIn}(\text{IO}_3)_4$ experiences from other phases is very weak, so $\text{LiIn}(\text{IO}_3)_4$ can be readily synthesized. Indeed, A1 was our predicted optimal condition to synthesize $\text{LiIn}(\text{IO}_3)_4$. Under conditions A3-A6, the thermodynamic competition was negative in value (-8 to -32 meV/atom), indicating that $\text{LiIn}(\text{IO}_3)_4$ is still the stable phase, but the thermodynamic competition is not as negative as that under conditions A1 and A2. Thus, kinetic considerations can still promote the formation of undesired byproducts under conditions A3-A6 than under conditions A1 and A2. Specifically, $\text{In}(\text{IO}_3)_3$ was the final product of conditions A3-A5, which complemented the finding from thermodynamic competition analysis that $\text{In}(\text{IO}_3)_3$ is the most competitive phase under conditions A3-A5 shown on the right panel of Figure 4.3. The excellent agreement between theoretical analysis and experimental results strongly supports the MTC hypothesis.

Our successful synthesis of $\text{LiIn}(\text{IO}_3)_4$ under conditions A1 and A2 indicate that the irreproducibility of the experiment from Yang et al. could be due to human error (i.e., reporting $[\text{Li}] = 0.15$ instead of an actual $[\text{Li}] = 1.5$). MTC not only predicts robust synthesis conditions to synthesize the $\text{LiIn}(\text{IO}_3)_4$, but can also provide a sanity check to verify literature synthesis recipes—overall providing a valuable tool to guide synthesis design.

4.3.2 LiFePO_4 synthesis analysis

In our second example, we demonstrate the advantage of MTC over traditional Pourbaix diagrams for providing precise condition for successful phase-pure syn-

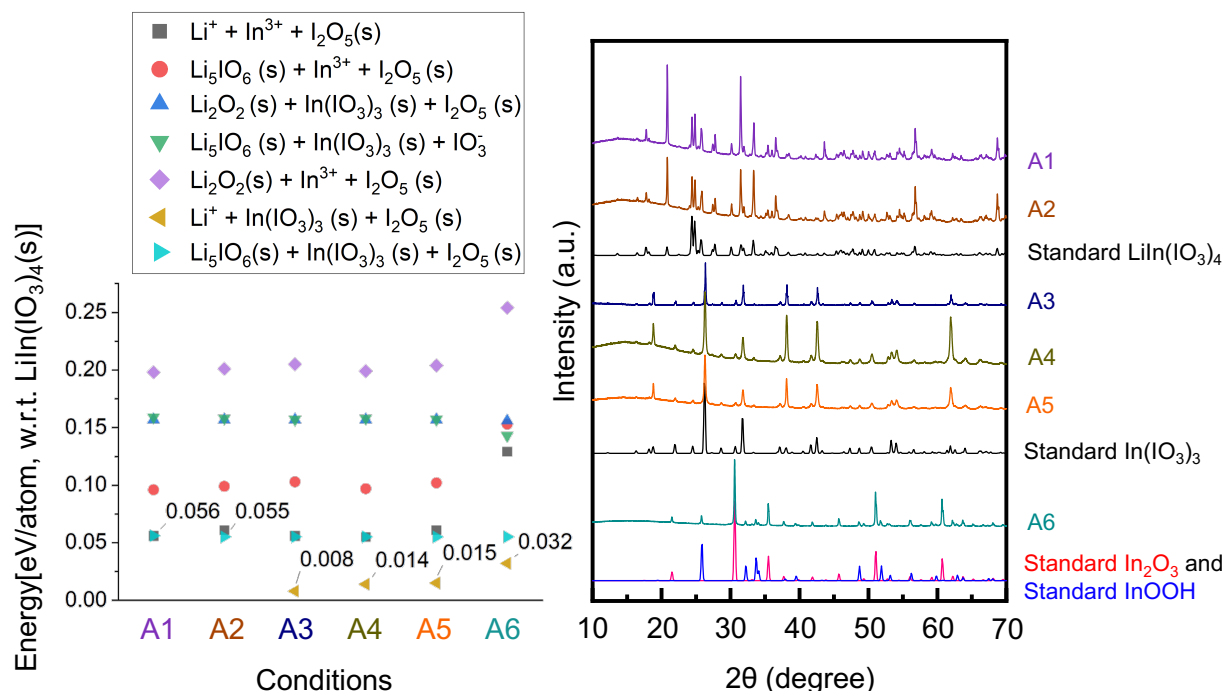


Figure 4.3: $\text{LiIn}(\text{IO}_3)_4$ thermodynamic competition analysis and the correspondence synthesis results. *Left panel:* the Pourbaix potentials of common competing phases with respect to the energy of $\text{LiIn}(\text{IO}_3)_4$ under condition A1-A6. *Right panel:* XRD experimental results for syntheses under condition A1-A6. A1-A6 are associated with varying [Li] concentration, pH, redox potential.

thesis of LiFePO_4 . The schematic Pourbaix diagram shown in Figure 4.4 presents the stable phases in the Li-Fe-P- H_2O aqueous electrochemical system. LiFePO_4 is stable only in the near pH-neutral region. However, the results reported by Liu et al. [120] indicate that LiFePO_4 was not successfully synthesized even within this thermodynamically stable region. Here we experimentally validate that the MTC identifies the optimal conditions within the LiFePO_4 stability window that successfully produces phase-pure LiFePO_4 .

Table 4.2 summarizes the concentrations of [Li], [Fe], [P], pH, redox potential, and final products under different synthesis conditions, which span a similar range of experimental conditions by Liu et al. [120]. Various amounts of H_3PO_4 were used to control the pH and the concentration of [P]. The pH values were 12.35, 9.27, 8.29, 7.21, 6.00, and 5.13, respectively (tagged as B1 - B6). We

estimated the redox potential based on glucose($E^\ominus(CO_2/C_6H_{12}O_6) = -0.016V$), which comes from the hydrolysis of sucrose, acting as a reducing agent in this system. Figure 4.4 presents the XRD results (right panel) for syntheses under conditions B1-B6. $LiFePO_4$ was successfully synthesized under conditions B2-B5. By contrast, $LiFePO_4$ does not form under basic condition B1 or acidic condition B6 — even though the pH for B6 is within the Pourbaix stability region for $LiFePO_4$.

tag	[Li]	[Fe]	[P]	pH	E(V)	$\Delta\Phi(Y)(\text{meV/atom})$	Product
B1	0.75	0.25	0.15	12.35	-0.74	1	$LiFe_5O_8 + Li_3PO_4$
B2	0.75	0.25	0.27	9.27	-0.56	-71	$LiFePO_4$
B3	0.75	0.25	0.28	8.29	-0.50	-58	$LiFePO_4$
B4	0.75	0.25	0.29	7.21	-0.44	-43	$LiFePO_4$
B5	0.75	0.25	0.44	6.00	-0.37	-29	$LiFePO_4$
B6	0.75	0.25	0.69	5.13	-0.31	-19	$Fe_5H_{10}P_4O_{20}$

Table 4.2: Synthesis conditions with thermodynamic competition analysis and experimental results for $LiFePO_4$ synthesis.

We calculated the $\Delta\Phi$ for the six conditions, and found that stability ($\Delta\Phi \leq 0$) is not a sufficient predictor for the formation of pure phase $LiFePO_4$. Similar to $LiIn(IO_3)_4$, $LiFePO_4$ can only be synthesized when thermodynamic competition is very negative ($\Delta\Phi(Y) \leq -29 \text{ meV/atom}$), indicating minimal thermodynamic competition from potential impurity phases, which maximizes the kinetic preference for the target $LiFePO_4$ phase. As the pH changes from basic to acidic, the thermodynamic competition first decreases and then increases, with resulting product phases agreeing well with the thermodynamic predictions. Under condition B1, when the pH is 12.35, the $\Delta\Phi$ is +1 meV / atom in value, so the thermodynamic competition $LiFePO_4$ experiences from other phases is very strong and $LiFePO_4$ is not the stable phase, which explains why $LiFePO_4$ was not the final product under this condition.

As the pH decreases to neutral and weak acidic regions (see conditions B2-B5), the $\Delta\Phi$ ranges from -71 to -29 meV / atom, so $LiFePO_4$ is not only the stable equilibrium phase, it also experiences weak thermodynamic competition compared to condition B1. $LiFePO_4$ was successfully synthesized as the final product under

these conditions. However, as pH further decreases to 5.13 (see condition B6), the $\Delta\Phi$ reduces in magnitude to -19 meV / atom, so the thermodynamic competition $LiFePO_4$ experiences becomes stronger than that in B2-B5. Accordingly, we see that the final product formed in this experiment is $Fe_5H_{10}P_4O_{20}$. This finding opposes information from the equilibrium Pourbaix diagram, which indicates stable $LiFePO_4$ at this pH and electrochemical potential combination.

Our ability to rationalize the kinetic unsuitability of certain conditions that are even within the Pourbaix stability region demonstrates the advantage of our MTC theory over traditional phase diagram methods.

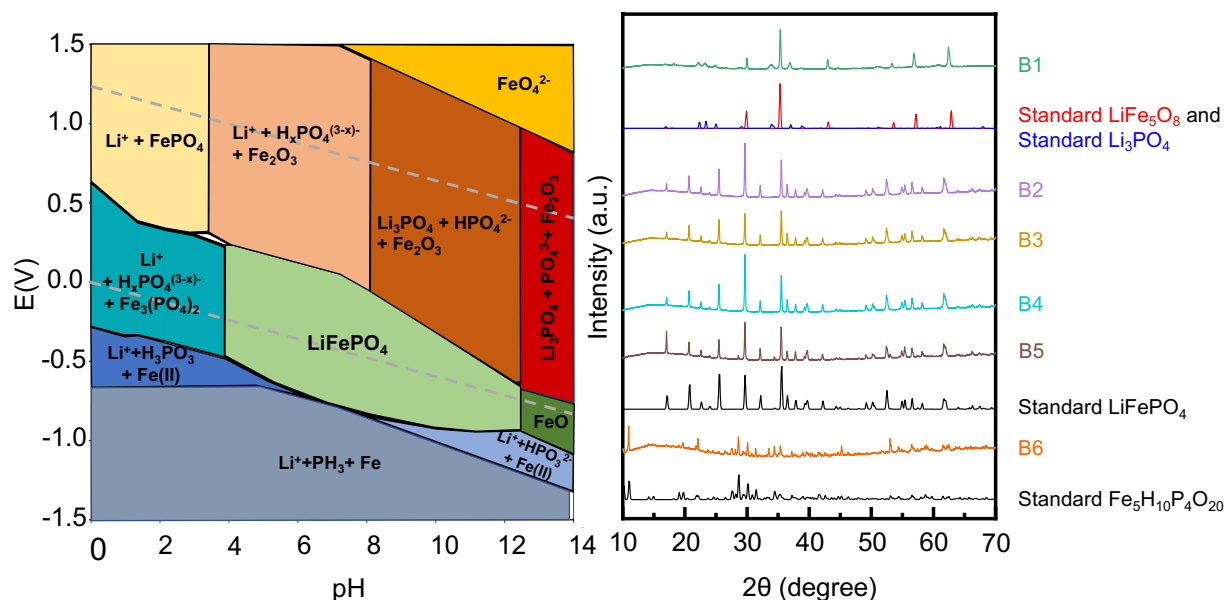


Figure 4.4: Schematic Pourbaix diagram of Li-Fe-P-H₂O system and the synthesis results. *Left panel:* Li-Fe-P-H₂O schematic Pourbaix diagram with $[Li]=0.75$ mol/L, $[Fe]=0.25$ mol/L, $[P]=0.3$ mol/L. The stability regions of different types of stable phases are marked in different colors. *Right panel:* XRD experimental results for syntheses under condition B1-B6. B1-B6 are associated with varying $[P]$ concentration, pH , redox potential.

4.4 Conclusion

In this Chapter, we present a computational approach to explicitly navigate multidimensional thermodynamic space and identify favored conditions to synthesize desired target phases, which breaks through some limitations of traditional phase diagrams and stability region analysis. As shown in our systematic investigation of $\text{LiIn}(\text{IO}_3)_4$ synthesis and LiFePO_4 system, thermodynamic stability may not be a sufficient to predict phase-pure synthesis conditions, and vice versa, that some compounds may be synthesizable by aqueous solution-based synthesis even under solution conditions where they are metastable[60]. Nevertheless, our large-scale statistical analysis and experiments reveal that target phases are more likely to be synthesized when they experience minimized $\Delta\Phi(Y)$ — in other words, weak thermodynamic competition. This is the start-of-the-art theoretical framework that embeds thermodynamic and kinetic mechanisms, and also illustrates the value of computational materials science in predictive synthesis.

The magnitude of the thermodynamic competition scale (tens of meV/atom) in aqueous solution-based synthesis is smaller than the thermodynamic limit for the synthesis of metastable inorganic materials (more than one hundred meV/atom) [22], and might be comparable with DFT errors[121–123]. However, we note that the DFT errors cannot be directly applied as error bars in our analysis because (1) it is reduced by correction based on the experimental aqueous states implemented in pymatgen[106] and (2) solid-state errors in DFT tend to be systematic and can be largely canceled between phases within the same chemistry[121]. Also, we would emphasize the DFT errors only affect the molar Gibbs free energy of a phase, given in Eq.(4.3). However, all the other terms in the Pourbaix potential are extensive quantities based on the composition (N_M , N_O , N_H , and charge Q). Those terms affect the “slopes” of the free energy planes, and therefore provide actionable directions in E , pH and concentration space to maximize the difference in driving force with competing phase to optimize synthesis conditions.

4.5 Methods

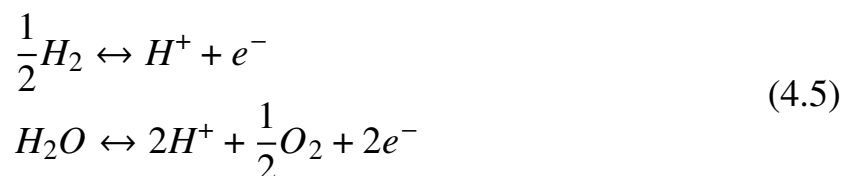
4.5.1 Thermodynamic potential in aqueous solution system

In order to apply the MTC hypothesis to the aqueous solution system, choosing a suitable thermodynamic grand potential is required. In the aqueous electrochemical system, the Pourbaix potential is used to depict the free energy of phases. The Pourbaix potential can be constructed from the Gibbs free energy by Legendre transformation:

$$\Phi(Y) = G - \frac{\partial G}{\partial N_H} N_H - \frac{\partial G}{\partial N_O} N_O - \frac{\partial G}{\partial Q} Q \quad (4.4)$$

where N_M , N_O , N_H is the number of metal, oxygen, hydrogen atoms in the composition, respectively, and Q is the charge of the phase. The partial derivative of G with respect to N_H is the chemical potential of hydrogen μ_H . The partial derivative with respect to N_O is the chemical potential of oxygen μ_O . The partial derivative with respect to Q is the redox potential E .

Considering the two constraints by chemical equilibrium in water:



which yields:

$$\begin{aligned} \mu_H &= \mu_{H^+} - E \\ \mu_O &= \mu_{H_2O} - 2 \cdot \mu_{H^+} + 2E \end{aligned} \quad (4.6)$$

After substituting the chemical equilibrium equations from Eq.(4.6) and the relationship between pH and μ_H ($\mu_{H^+} = -RT \cdot \ln(10) \cdot \text{pH}$) into the Legendre transformation Eq.(4.4) and normalizing the Pourbaix potential by the number of metal atoms, the Pourbaix potential can be obtained:

$$\bar{\Psi} = \frac{1}{N_M} ((G - N_O \mu_{H_2O}) - RT \cdot \ln(10) \cdot (2N_O - N_H) \text{pH} - (2N_O - N_H + Q) E) \quad (4.7)$$

The molar Gibbs free energy of a phase, G , is its chemical potential, $\mu_i = \mu_i^0 + RT\ln[a_i]$, where μ_i^0 is given by the standard-state Gibbs formation free energy. An ideal solid with no defects has an activity of one, and so the $RT\ln[a_i]$ term is zero, but the chemical potential of metal ions ideally scales with the natural log of the metal ion concentration in solution.

4.5.2 Thermodynamic competition calculation and optimization

The thermodynamic data in this paper were retrieved from the Materials Project[7] using the Python Materials (MP) Genomic (pymatgen) package [6]. All competing phases and corresponding Pourbaix potentials were generated and calculated through the Pourbaix Diagram module[106]. The linearity of a phase’s thermodynamic potential in an aqueous electrochemical system [107] turns the thermodynamic competition analysis formulated in Eq.(4.1) into a convex problem in its natural variables, as stated by the pointwise minimum theorem[124]. Therefore, optimization algorithms, such as gradient-based methods, can be used to optimize natural variables (i.e., pH, redox potential (E), and log-scaled concentrations ($\ln[X]$) in aqueous solution system) to obtain the minimized thermodynamic competition in the pre-defined searching synthesis space. In this work, the searching synthesis space defined as $0 \leq pH \leq 14$, $-1V \leq E \leq 2V$, and $0 \text{ mol/L} \leq [X] \leq 2 \text{ mol/L}$.

4.5.3 Text-mined dataset analysis

We used the dataset of text-mined solution-based synthesis procedures [1] to perform the thermodynamic competition analysis. We firstly filtered out syntheses that have a final heat treatment step after the precipitate is obtained from the solution. Then, we identified aqueous solution syntheses based on the solvent (i.e., solvent is water)[1]. 20,062 out of 35,675 procedures are aqueous solution-based without post heat treatment. In this subset, 4,431 unique targets were synthesized, and 1,370 of them can be found in the Materials Project database. 251 among those 1,370 unique targets are salts. Specifically, we selected procedures containing one of three common types of salts, phosphates, carbonates, and iodates, as synthesized targets,

and kept only those targets with an mp-id, meaning their thermodynamic data are available in MP database. The dataset is unbalanced, so in order to avoid statistical distributions being skewed by compounds with higher synthesis frequency, such as LiFePO_4 and CaCO_3 , we randomly sampled up to three procedures for every compound. To specify the synthesis conditions, we manually curated the synthesis procedures: 1. the quantities of materials were converted into concentrations of different ions; 2. pH was calculated based on the concentrations and pKa of different ions using pHcalc (<https://github.com/rnelsonchem/pHcalc>); 3. the redox potentials at the calculated pH were estimated. We only kept synthesis procedures that can provide enough information to construct the Pourbaix system. In the final dataset that is used for statistical analysis as Figure 4.2 shows, there are unique 40 phosphates, 15 carbonates, and 13 iodates. Pourbaix potentials for all target phases and their corresponding possible competing phases were then calculated based on the estimated synthesis conditions using the Pourbaix diagram module in pymatgen. Finally, we calculated the thermodynamic competition the target phases experience according to Eq.(4.1). We then performed the optimization to minimize the thermodynamic competition the target phases experience from competing phases for comparison purposes.

4.5.4 Synthesis of $\text{LiIn}(\text{IO}_3)_4$

$\text{Li}(\text{CH}_3\text{COO})$ (anhydrous, Sigma-Aldrich, 99.9%) and LiCl (Sigma-Aldrich, 99%) were used as lithium sources for the synthesis of $\text{LiIn}(\text{IO}_3)_4$. Different amounts of lithium salts (A1: 15 mmol LiCl , A2: 1.5 mmol $\text{Li}(\text{CH}_3\text{COO})$ + 13.5 mmol LiCl , A3: 1.5 mmol $\text{Li}(\text{CH}_3\text{COO})$, A4: 8 mmol LiCl , A5: 1.5 mmol $\text{Li}(\text{CH}_3\text{COO})$ + 6.5 mmol LiCl , and A6: 8 mmol $\text{Li}(\text{CH}_3\text{COO})$) were mixed with 0.5 mmol In_2O_3 and 8 mmol H_5IO_6 . The mixture was dissolved in 10 mL water. The obtained solution was sealed in an autoclave equipped with a Teflon liner (49 mL) and heated at 200 °C for 132 hours, followed by natural cooling to room temperature. The products were centrifuged, washed with water 3 times, and dried in a 70 °C vacuum oven for 12 hours.

4.5.5 Synthesis of LiFePO_4

LiFePO_4 was synthesized through hydrothermal method under nitrogen environment. 5 mmol $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ (Sigma-Aldrich, 99%), 15 mmol $\text{LiOH} \cdot \text{H}_2\text{O}$ (Sigma-Aldrich, 99.95%), 0.04 g sucrose and various amounts of 85 wt% H_3PO_4 (Sigma-Aldrich) were dissolved in 20 mL water and sealed in an autoclave equipped with a Teflon liner (49 mL). The pH of the mixture solution was controlled by the amount of 85 wt% H_3PO_4 . The autoclaves were heated at 170 °C for 12 hours, followed by natural cooling to room temperature. The products were centrifuged, washed by water and ethanol for 3 times, and dried in a 70 °C vacuum oven for 12 hours.

4.5.6 Structure Characterization

The X-ray diffraction (XRD) measurement was carried out using a Rigaku MiniFlex 600 diffractometer with Cu $K\alpha$ radiation ($\lambda = 1.5406 \text{ \AA}$) with a step size of 1°/min from 10° to 70°.

Chapter 5

Conclusions and outlooks

Part of this chapter has been published in or included in:

Wang, Z., Kononova, O., Cruse, K., He, T., Huo, H., Fei, Y., Zeng, Y., Sun, Y., Cai, Z., Sun, W., & Ceder, G.(2022). Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. Scientific Data 9, 231. [[1](#)]

Wang, Z., Sun, Y., Cruse, K., Zeng, Y., Fei, Y., Liu, Z., Shangguan, J., Byeon, Y., Jun, K., He, T., Sun, W., & Ceder, G.. Optimal thermodynamic conditions to minimize kinetic byproducts in aqueous materials synthesis. Working paper which is pending publication at the time of filing this thesis.

5.1 Conclusions of this thesis

This thesis demonstrated a data-driven approach to understanding aqueous solution-based synthesis by achieving two milestones: 1) developing an NLP and deep learning infrastructure that text-mines synthesis information from journal articles to create the first large-scale dataset of solution synthesis procedures, and 2) modeling synthesis and empirically validating the synthesis model by using the text-mined dataset and systematic synthesis experiments.

In chapter2, we proposed the first *unified language of synthesis actions* (ULSA) for describing inorganic synthesis procedures. We created a labeled dataset and built a neural network-based model extract and map synthesis actions into ULSA from inorganic synthesis paragraphs. ULSA serves as an important step in building the text-mining pipeline. In Chapter3, we constructed the first large dataset of solution-based inorganic materials synthesis by designing an advanced text-mining pipeline, including content acquisition (scraping, parsing), text classification (BERT-based), name entity recognition (Bi-LSTM), and relation extraction (heuristic) to extract essential information from the scientific literature. In Chapter4, we have formulated a thermodynamic framework that embeds thermodynamic and kinetic mechanisms to predict optimal synthesis conditions. We hypothesized that the condition where the desired target phase experiences minimal competition with other phases is favored for synthesis. We evaluated the hypothesis of MTC in the aqueous electrochemical conditions and surveyed the magnitude of the thermodynamic competition scale with respect to synthesizability for three families of multicomponent functional oxides through statistical analysis on a text-mined solution-based synthesis dataset. Our systematic investigation of $\text{LiIn}(\text{IO}_3)_4$ synthesis and LiFePO_4 system further demonstrated the predictive power of thermodynamic competition. The thermodynamic competition analysis paves the way to optimize synthesis conditions.

In this thesis, we utilized the text-mined dataset to empirically validate the fundamental physical hypothesis instead of building machine learning synthesis predictors. The main reason is that the accuracies of the machine-learned models trained from the biased and incomplete datasets are limited[55, 125, 126]. However, hypothesis testing can still uncover the hidden mechanisms from the imperfect

dataset[127, 128].

Finally, in this Chapter, we outline potential solutions to improve the synthesis dataset and some interesting related topics that could help understand synthesis mechanisms.

5.2 Outlook

5.2.1 Improving text-mined synthesis dataset

Solution-based synthesis is an important area of materials synthesis[85] and our text-mined dataset can help with advancing the science and model building for solution synthesis. Nevertheless, challenges remain in the mining of scientific literature and the construction of robust and accurate large-scale datasets.

Better materials science text parser. The organic precursors with complex radicals commonly used in solution-based synthesis pose a challenge for parsing and extracting chemical information. Constructing reaction formulas becomes problematic when the precursor information is lost. Therefore, these entries are mostly dropped out later in the pipeline. To address this issue, a universal parser that can parse chemical tokens needs to be developed.

Determining actual synthesis results. our data was extracted from the experimental section in the main body of each paper and does not include any information about the actual synthesis results, e.g., whether the material was synthesized using the reported procedure or which structure was obtained. This problem could be overcome by introducing a model that can parse characterization data (e.g., X-ray diffraction patterns or electron microscopy images) and relate them to the corresponding synthesis conditions, something which, to the best of our knowledge, has not yet been performed.

Balanced dataset. Even though the actual results of a synthesis can be extracted from a paper, there remains the challenge of data interpretation and usage, as the authors usually report only successful and “cherry-picked” experimental re-

sults. This introduces a significant anthropogenic bias toward “positive” data with little “negative” content in the dataset, thus limiting the tasks for future machine-learning applications [125, 126]. A promising approach to solve this issue is to incorporate results obtained by autonomous robotic synthesis platforms that can provide a vast amount of “negative” data in a reasonable time frame [108, 129]. Here, we would like to point out that implementation of abstract synthesis representations to an autonomous robotic platform will require well-defined and robust mapping onto a fixed set of manipulations and devices available to the robot. The unified language we propose in this thesis can become a solid foundation for the future development in this direction.

Morphology extraction. solution-based synthesis is advantageous when the control of specimen morphology is required, e.g., when synthesizing noble-metal nanoparticles. However, this dataset does not provide information about the morphology of the synthesized materials, though such information is often contained in characterization or results paragraphs instead of the experimental section. The extraction of morphology and other solution synthesis outcomes is another text-mining challenge in materials science research that requires the development of advanced algorithms and models [31]. Recently, Cruse et al. have constructed a dataset of codified gold nanoparticle synthesis protocols[130], providing morphologies and size entities information, and can be used to help understand the formation of specific morphology or size.

5.2.2 Generalization of minimum thermodynamic competition hypothesis

In this thesis, though we only provide a demonstration of MTC in an aqueous electrochemical system, we note that our schema is generalizable to any experimental synthesis method. For other experimental synthesis, appropriate thermodynamic grand potentials based on different natural handles are required. Fortunately, Materials Genome Initiative efforts have not only led to large-scale *ab initio* computed databases[5, 7] that provide reliable sources of thermodynamic data, but also deliver materials analysis platforms[6] that can automate competing phase generation and grand potential calculations. Therefore, different thermodynamic grand po-

tentials could be easily constructed based on the thermodynamic data using the Legendre transformation by considering different boundary conditions[106, 131]. For example, the surface energy term should be considered in nanoparticle synthesis analysis[107, 132, 133]; mechanical work should be included in ball-milling synthesis analysis[134–136]; an electromagnetic susceptibility term is important to take into account when electric and magnetic fields are applied[137, 138]. Thus, the MTC hypothesis can be extended to different synthesis systems, though the applicability of MTC in different synthesis requires further comprehensive and systematical evaluation.

5.2.3 Better physical modeling synthesis

The Pourbaix potential, the free energy we chose in our work, can be improved in the future. Pourbaix potential, which is the grand potential of an open system with boundary conditions pH, redox potential, and concentration, was used to analyze aqueous solution synthesis. However, during the chemical reaction in the aqueous solution system, the precursors are consumed. Thus, a new thermodynamic potential with closed boundary conditions would need to be constructed to better model the aqueous solution based synthesis.

Bibliography

1. Wang, Z. *et al.* Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Scientific Data* **9**, 231 (2022).
2. Pankratius, V. *et al.* Computer-Aided Discovery: Toward Scientific Insight Generation with Machine Support. *IEEE Intelligent Systems* **31**, 3–10 (2016).
3. Tolle, K., Tansley, D. & Hey, A. The Fourth Paradigm: Data-Intensive Scientific Discovery. *Proceedings of the IEEE* **99**, 1334–1337 (2011).
4. Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nature Materials* **5**, 641–646 (2006).
5. Holden, J. *Materials Genome Initiative for global competitiveness* tech. rep. (National Science and Technology Council, 2011).
6. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319. ISSN: 0927-0256 (2013).
7. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
8. Bergerhoff, G. & Brown, I. D. *Crystallographic Databases* (F.H. Allen *et al.* (Hrsg.) Chester, International Union of Crystallography, 1987).
9. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. B* **58**, 364–369 (2002).

10. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* (eds Linstrom, P. & Mallard, W.) (National Institute of Standards and Technology, Gaithersburg MD, 20899, 2019).
11. Blokhin, E. & Villars, P. in, 1–26 (Springer, Cham, 2018).
12. Villars, P. & Cenzual, K. *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on DVD)* ASM International®, Materials Park, Ohio, USA. 2018.
13. Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. A bridge for accelerating materials by design. *npj Computational Materials* **1**, 15008 (2015).
14. Bianchini, M. *et al.* The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature Materials* **19**, 1088–1095 (2020).
15. Bartel, C. J. *et al.* Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nature Communications* **9**, 4168 (2018).
16. McDermott, M. J., Dwaraknath, S. S. & Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nature Communications* **12**, 3097 (2021).
17. Kovnir, K. Predictive Synthesis. *Chemistry of Materials* **33**, 4835–4841 (2021).
18. Kohlmann, H. Looking into the Black Box of Solid-State Synthesis. *European Journal of Inorganic Chemistry* **2019**, 4174–4180 (2019).
19. Chamorro, J. R. & McQueen, T. M. Progress toward Solid State Synthesis by Design. *Accounts of Chemical Research* **51**, 2918–2925 (Nov. 2018).
20. Sun, W. *et al.* The thermodynamic scale of inorganic crystalline metastability. *Science Advances* **2**, e1600225 (2016).
21. Sun, W. *et al.* A map of the inorganic ternary metal nitrides. *Nature Materials* **18**, 732–739 (2019).

22. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Science Advances* **4**, eaaq0148 (2018).
23. Aykol, M. *et al.* Network analysis of synthesizable materials discovery. *Nature Communications* **10**, 2018 (2019).
24. 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).
25. Duvenaud, D. K. *et al.* in *Advances in Neural Information Processing Systems* 28 (eds Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 2224–2232 (Currant Associates, Inc., 2015).
26. Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic Synthesis: March of the Machines. *Angew. Chem. Int. Edit.* **54**, 3449–3464 (2015).
27. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
28. Mikulak-Klucznik, B. *et al.* Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
29. Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **49**, 2897–2898 (12 2009).
30. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2018).
31. Kononova, O. *et al.* Opportunities and challenges of text mining in Materials research. *iScience* **24** (2021).
32. Gaultois, M. *et al.* Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations. *Chem. Mater.* **25**, 2911–2920 (2013).
33. Ghadbeigi, L., Harada, J. K., Lettiere, B. R. & Sparks, T. D. Performance and resource considerations of Li-ion battery electrode materials. *Energy Environ. Sci.* **8**, 1640–1650 (6 2015).
34. Oh, E. *et al.* Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nature Nanotech.* **11**, 479 (2016).

35. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5**, 180111 (2018).
36. 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).
37. Huang, S. & Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. *Scientific Data* **7**, 260 (2020).
38. Onishi, T., Kadohira, T. & Watanabe, I. Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity. *Sci. Technol. Adv. Mater.* **19**, 649–659 (2018).
39. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
40. Kim, E. *et al.* Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater* **29**, 9436–9444 (2017).
41. Swain, M. C. & Cole, J. M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
42. Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L. & Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminformatics* **3**, 41 (2011).
43. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminformatics* **3**, 17 (2011).
44. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **117**, 7673–7761 (2017).
45. Eltyeb, S. & Salim, N. Chemical Named Entities Recognition: A Review on Approaches and Applications. *J. Cheminformatics* **6**, 17 (2014).

46. Korvigo, I., Holmatov, M., Zaikovskii, A. & Skoblov, M. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. *J. Cheminform.* **10**, 28 (2018).
47. He, T. *et al.* Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chemistry of Materials* **32**, 7861–7873 (2020).
48. Weston, L. *et al.* Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *J. Chem. Inf. Model.* **59**, 3692–3702 (2019).
49. Kuniyoshi, F., Makino, K., Ozawa, J. & Miwa, M. *Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature in Proceedings of The 12th Language Resources and Evaluation Conference* (European Language Resources Association, Marseille, France, May 2020), 1941–1950.
50. Kim, E. *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
51. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater* **3**, 53 (2017).
52. Jensen, Z. *et al.* A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).
53. Kim, E. *et al.* Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
54. Kononova, O. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data* **6**, 203 (2019).
55. Huo, H. *et al.* Machine-Learning Rationalization and Prediction of Solid-State Synthesis Conditions. *Chemistry of Materials* **34**, 7323–7336 (Aug. 2022).
56. Sun, W. *et al.* Thermodynamic Routes to Novel Metastable Nitrogen-Rich Nitrides. *Chem. Mater* **29**, 6936–6946 (2017).

57. Zhao, Y. *et al.* Pourbaix diagram for HP-13Cr stainless steel in the aggressive oilfield environment characterized by high temperature, high CO₂ partial pressure and high salinity. *Electrochimica Acta* **293**, 116–127. ISSN: 0013-4686 (2019).
58. Huang, L. -., Hutchison, M. J., Santucci, R. J., Scully, J. R. & Rondinelli, J. M. Improved Electrochemical Phase Diagrams from Theory and Experiment: The Ni–Water System and Its Complex Compounds. *The Journal of Physical Chemistry C* **121**, 9782–9789 (May 2017).
59. Huang, L.-F., Ha, H. M., Lutton Cwalina, K., Scully, J. R. & Rondinelli, J. M. Understanding Electrochemical Stabilities of Ni-Based Nanofilms from a Comparative Theory–Experiment Approach. *The Journal of Physical Chemistry C* **123**, 28925–28940 (Nov. 2019).
60. Singh, A. K. *et al.* Electrochemical stability of metastable materials. *Chemistry of Materials* **29**, 10159–10167 (2017).
61. Walters, L. N., Zhang, C., Dravid, V. P., Poeppelmeier, K. R. & Rondinelli, J. M. First-Principles Hydrothermal Synthesis Design to Optimize Conditions and Increase the Yield of Quaternary Heteroanionic Oxychalcogenides. *Chemistry of Materials* **33**, 2726–2741 (Apr. 2021).
62. Wang, Y., He, J., Liu, C., Chong, W. H. & Chen, H. Thermodynamics versus Kinetics in Nanosynthesis. *Angewandte Chemie International Edition* **54**, 2022–2051 (2015).
63. Hoffman, J. D. Thermodynamic Driving Force in Nucleation and Growth Processes. *The Journal of Chemical Physics* **29**, 1192–1193 (1958).
64. Whitman, J. R., Aranovich, G. L. & Donohue, M. D. Thermodynamic driving force for diffusion: Comparison between theory and simulation. *The Journal of Chemical Physics* **134**, 094303 (2011).
65. Wang, Z. *et al.* ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discovery* **1**, 313–324 (2022).
66. Olivetti, E. *et al.* Data-driven Materials Research Enabled by Natural Language Processing. *Appl. Phys. Rev.* **7**, 041317 (2020).

67. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data* **5**, 180111 (2018).
68. Mysore, S. *et al.* The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. *LAW 2019 - 13th Linguistic Annotation Workshop, Proceedings of the Workshop*, 56–64. arXiv: [1905.06939](https://arxiv.org/abs/1905.06939) (2019).
69. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
70. Hiszpanski, A. *et al.* Nanomaterials Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge. *J. Chem. Inf. Model.* **60**, 2876–2887 (2020).
71. 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).
72. Vaucher, A. *et al.* Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).
73. Mehr, S. H. M., Craven, M., Leonov, A. I., Keenan, G. & Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101–108 (2020).
74. Mysore, S. *et al.* Automatically Extracting Action Graphs from Materials Science Synthesis Procedures. <https://arxiv.org/abs/1711.06872> (2017).
75. Kuniyoshi, F., Makino, K., Ozawa, J. & Miwa, M. Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature. <https://arxiv.org/abs/2002.07339> (2020).
76. Dieb, T., Yoshioka, M., Hara, S. & Newton, M. Framework for automatic information extraction from research papers on nanocrystal devices. *Beilstein J. Nanotechnol.* **6**, 1872–1882 (2015).

77. Kulkarni, C., Xu, W., Ritter, A. & Machiraju, R. *An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2018), 97–106.
78. Friedrich, A. *et al.* *The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2020), 1255–1268.
79. 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).
80. Kim, J.-D., Ohta, T., Tateisi, Y. & Tsujii, J. GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics* **19**, i180–i182 (2003).
81. 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).
82. Mayfield, J., Lagerstedt, I. & Sayle, R. *Pistachio in NIH Virtual Workshop on Reaction Informatics* (2021).
83. Kim, E., Huang, K., Kononova, O., Ceder, G. & Olivetti, E. Distilling a Materials Synthesis Ontology. *Matter* **1**, 8–12 (2019).
84. Xu, R.-R. in *Modern Inorganic Synthetic Chemistry (Second Edition)* (eds Xu, R. & Xu, Y.) Second Edition, 1–7 (Elsevier, Amsterdam, 2017). ISBN: 978-0-444-63591-4.
85. Huo, H. *et al.* Semi-Supervised Machine-Learning Classification of Materials Synthesis Procedures. *npj Computational Materials* **5**, 62 (2019).
86. 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, Sept. 2015), 1373–1378.

87. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. *Distributed Representations of Words and Phrases and Their Compositionality* in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Curran Associates Inc., Lake Tahoe, Nevada, 2013), 3111–3119.
88. Řehůřek, R. & Sojka, P. *Software framework for topic modelling with large corpora* English. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, May 2010), 45–50.
89. Fleiss, J. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382 (5 1971).
90. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
91. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 4171–4186.
92. Radford, A. & Narasimhan, K. *Improving Language Understanding by Generative Pre-Training* in (2018).
93. Radford, A. *et al.* *Language Models are Unsupervised Multitask Learners* in (2019).
94. Brown, T. *et al.* *Language Models are Few-Shot Learners* in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) **33** (Curran Associates, Inc., 2020), 1877–1901.
95. Mouchère, H., Zanibbi, R., Garain, U. & Viard-Gaudin, C. Advancing the state of the art for handwritten math recognition: the CROHME competitions, 2011–2014. *IJDAR* **19**, 173–189 (2016).

96. Mahdavi, M., Zanibbi, R., Mouchère, H., Viard-Gaudin, C. & Garain, U. *ICDAR 2019 CROHME+ TFD: Competition on recognition of handwritten mathematical expressions and typeset formula detection in 2019 International Conference on Document Analysis and Recognition (ICDAR)* (2019), 1533–1538.
97. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. in *Advances in Neural Information Processing Systems 26* (eds Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 3111–3119 (Curran Associates, Inc., 2013).
98. Řehůřek, R. & Sojka, P. *Software framework for topic modelling with large corpora* English. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, May 2010), 45–50.
99. Bird, S., Edward, L. & Ewan, K. *Natural Language Processing with Python*. (O'Reilly Media Inc, 2009).
100. Wang, Z. *et al.* *Dataset of Solution-based Inorganic Materials Synthesis Procedures Extracted from the Scientific Literature* figshare <https://doi.org/10.6084/m9.figshare.16583387.v4>. 2021.
101. Wang, D., Xie, T. & Li, Y. Nanocrystals: Solution-based synthesis and applications as nanocatalysts. *Nano Research* **2**, 30–46 (2009).
102. Rakhi, R. B., Chen, W., Cha, D. & Alshareef, H. N. Influence of calcination temperature on the morphology and energy storage properties of cobalt oxide nanostructures directly grown over carbon cloth substrates. *Materials for Renewable and Sustainable Energy* **2**, 17 (2013).
103. Augustin, M. *et al.* Manganese oxide phases and morphologies: A study on calcination temperature and atmospheric dependence. *Beilstein journal of nanotechnology* **6**, 47–59 (2015).
104. Ellingham, H. Transactions and Communications. *Journal of the Society of Chemical Industry* **63**, 125–160 (1944).
105. Byrappa, K. & Yoshimura, M. *Handbook of Hydrothermal Technology* ISBN: 978-0-12-375090-7 (William Andrew Publishing, Oxford, 2013).

106. Persson, K. A., Waldwick, B., Lazic, P. & Ceder, G. Prediction of solid-aqueous equilibria: Scheme to combine first-principles calculations of solids with experimental aqueous states. *Phys. Rev. B* **85**, 235438 (23 June 2012).
107. Sun, W., Kitchaev, D. A., Kramer, D. & Ceder, G. Non-equilibrium crystallization pathways of manganese oxides in aqueous solution. *Nature Communications* **10**, 573 (2019).
108. Szymanski, N. J. *et al.* Toward autonomous design and synthesis of novel inorganic materials. *Mater. Horiz.* **8**, 2169–2198 (8 2021).
109. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
110. Walters, L. N., Wang, E. L. & Rondinelli, J. M. Thermodynamic Descriptors to Predict Oxide Formation in Aqueous Solutions. *The Journal of Physical Chemistry Letters* **13**, 6236–6243 (July 2022).
111. Demopoulos, G. Aqueous precipitation and crystallization for the production of particulate solids with desired properties. *Hydrometallurgy* **96**, 199–214 (2009).
112. Patel, A. M., Nørskov, J. K., Persson, K. A. & Montoya, J. H. Efficient Pourbaix diagrams of many-element compounds. *Phys. Chem. Chem. Phys.* **21**, 25323–25327 (45 2019).
113. Gopalakrishnan, J. Chimie douce approaches to the synthesis of metastable oxide materials. *Chemistry of Materials* **7**, 1265–1275 (1995).
114. Martin, S. T. *Precipitation and Dissolution of Iron and Manganese Oxides* in (2003).
115. Murray, J. W., Dillard, J. G., Giovanoli, R., Moers, H. & Stumm, W. Oxidation of Mn(II): Initial mineralogy, oxidation state and ageing. *Geochimica et Cosmochimica Acta* **49**, 463–470. ISSN: 0016-7037 (1985).
116. Jun, K. *et al.* Lithium superionic conductors with corner-sharing frameworks. *Nature Materials* **21**, 924–931 (2022).
117. Manthiram, A. & Goodenough, J. Lithium insertion into Fe₂(SO₄)₃ frameworks. *Journal of Power Sources* **26**. 4th International Meetings on Lithium Batteries, 403–408 (1989).

118. Yang, B.-P., Sun, C.-F., Hu, C.-L. & Mao, J.-G. A series of new alkali metal indium iodates with isolated or extended anions. *Dalton Trans.* **40**, 1055–1060 (5 2011).
119. POURBAIX, M. Atlas of Electrochemical Equilibria in Aqueous Solution. *NACE* **307** (1974).
120. Liu, J., Jiang, R., Wang, X., Huang, T. & Yu, A. The defect chemistry of LiFePO₄ prepared by hydrothermal method at different pH values. *Journal of Power Sources* **194**. XIth Polish Conference on Fast Ionic Conductors 2008, 536–540. ISSN: 0378-7753 (2009).
121. Bartel, C. J. *et al.* A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials* **6**, 97 (2020).
122. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA + U framework. *Phys. Rev. B* **73**, 195107 (19 May 2006).
123. Gong, S. *et al.* Calibrating DFT Formation Enthalpy Calculations by Multifidelity Machine Learning. *JACS Au* **2**, 1964–1977 (Sept. 2022).
124. Boyd, S. & Vandenberghe, L. *Convex optimization* (Cambridge university press, 2004).
125. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
126. Jia, X. *et al.* Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).
127. Li, J. J. & Tong, X. Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines. *Patterns* **1**, 100115. ISSN: 2666-3899 (2020).
128. Lopez-Paz, D. & Oquab, M. *Revisiting Classifier Two-Sample Tests* 2017. <https://arxiv.org/abs/1610.06545>.
129. Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
130. Cruse, K. *et al.* Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. *Scientific Data* **9**, 234 (2022).

131. Ong, S. P., Wang, L., Kang, B. & Ceder, G. Li-Fe-P-O₂ Phase Diagram from First Principles Calculations. *Chemistry of Materials* **20**, 1798–1807 (Mar. 2008).
132. Navrotsky, A., Mazeina, L. & Majzlan, J. Size-Driven Structural and Thermodynamic Complexity in Iron Oxides. *Science* **319**, 1635–1638 (2008).
133. Ranade, M. R. *et al.* Energetics of nanocrystalline TiO₂. *Proceedings of the National Academy of Sciences* **99**, 6476–6481 (2002).
134. Yang, L., Moores, A., Frišćić, T. & Provatas, N. Thermodynamics Model for Mechanochemical Synthesis of Gold Nanoparticles: Implications for Solvent-Free Nanoparticle Production. *ACS Applied Nano Materials* **4**, 1886–1897 (Feb. 2021).
135. Rana, R. *et al.* Insights into the Mechanism of the Mechanochemical Formation of Metastable Phases. *ACS Applied Materials & Interfaces* **13**, 6785–6794 (Feb. 2021).
136. Tricker, A. W., Samaras, G., Hebisch, K. L., Realff, M. J. & Sievers, C. Hot spot generation, reactivity, and decay in mechanochemical reactors. *Chemical Engineering Journal* **382**, 122954 (2020).
137. Aber, J. E., Arnold, S., Garetz, B. A. & Myerson, A. S. Strong dc Electric Field Applied to Supersaturated Aqueous Glycine Solution Induces Nucleation of the γ Polymorph. *Phys. Rev. Lett.* **94**, 145503 (14 Apr. 2005).
138. Potticary, J. *et al.* An unforeseen polymorph of coronene by the application of magnetic fields during crystal growth. *Nature Communications* **7**, 11555 (2016).