Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2022, **1**, 313

# ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols

Zheren Wang, †[ab] Kevin Cruse, †[ab] Yuxing Fei,[ab] Ann Chia,‡[a] Yan Zeng,[b] Haoyan Huo, [ab] Tanjin He,[ab] Bowen Deng,[ab] Olga Kononova§*[a] and Gerbrand Ceder *[ab]

Applying AI power to predict syntheses of novel materials requires high-quality, large-scale datasets. Extraction of synthesis information from scientific publications is still 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 this work, we propose the first unified language of synthesis actions (ULSA) for describing inorganic synthesis procedures. We created a dataset of 3040 synthesis procedures annotated by domain experts according to the proposed ULSA scheme. To demonstrate the capabilities of ULSA, we built a neural network-based model to map arbitrary inorganic synthesis paragraphs into ULSA and used it to construct synthesis flowcharts for synthesis procedures. Analysis of the flowcharts showed that (a) ULSA covers essential vocabulary used by researchers when describing synthesis procedures and (b) it can capture important features of synthesis protocols. The present work focuses on the synthesis protocols for solid-state, sol–gel, and solution-based inorganic synthesis, but the language could be extended in the future to include other synthesis methods. This work is an important step towards creating a synthesis ontology and a solid foundation for autonomous robotic synthesis.

Received 4th November 2021
Accepted 18th April 2022

DOI: 10.1039/d1dd00034a

rsc.li/digitaldiscovery



## 1 Introduction

In the past decade, we have witnessed the growing success of data-driven and artificial intelligence (AI)-based methodologies promoting breakthroughs in predicting materials structure, properties, and functionality.[1–3] Still, adapting the power of AI to predict and control materials synthesis and fabrication is challenging and requires substantial effort in gathering high-quality large-scale datasets. One approach to gather such datasets of synthesis parameters and conditions would be running high-throughput experiments. This requires a costly setup and substantial human labor and expertise, and is typically limited to a small part of chemical space. Another way to acquire the data or augment existing datasets is to extract information about materials synthesis from the wealth of scientific publications (*e.g.* papers, archives, patents) available online.

Scientific text mining has received its recognition in the past few years,[4–7] providing the materials science community with datasets on a variety of materials and their properties[8–10] as well as synthesis protocols.[11–14] Nonetheless, a majority of these text mining studies have been focused on extracting chemical entities such as material names, formulas, properties, and other characteristics.[15–19] There have only been a few attempts to extract information about chemical synthesis and reactions and compile them into a flowchart of synthesis actions.[20–28] Hawizy *et al.*[20] were early developers for such extraction, using a combination of rule-based regular expressions (regex)[29] 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.*[21] 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.*[22] 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.*[23] paved the way for synthesis action graph extraction from the inorganic synthesis text. For this, they applied several neural network-

[a]*Department of Materials Science & Engineering, University of California, Berkeley, CA 94720, USA. E-mail: olga_kononova@berkeley.edu; gceder@berkeley.edu*

[b]*Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

† Equal contribution.

‡ Present address: Nanyang Technological University, Republic of Singapore, 639798.

§ Present address: Roivant Sciences, New York, NY 10036, USA.

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,[24] 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[30,31] and annotated sets[32,33] 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.[13,21,25] The dataset created by Mysore *et al.*[13] 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.*[21] was obtained by augmenting the existing Pistachio dataset[34] of organic synthesis procedures, and is available upon request. Kuniyoshi *et al.*[25] 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.[35] 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.[35] While such ambiguity is inconvenient for human readers, the growing interest in automated AI-guided materials synthesis demands a robust and unified language for describing synthesis protocols in order to make them applicable to autonomous robotic platforms.[22,36,37]

The previous works describing inorganic synthesis action extraction from the text[23,24] have laid the groundwork for extending such methods to this materials field, and have made their datasets available for interested researchers; however, neither provide an ontology for the actions that their models extract. Although development of synthesis action extraction from the text in organic chemistry has significantly accelerated and some groups have developed specific ontologies[21,22] for such vocabulary, we found that the existing models do not transfer well to the inorganic synthesis space due to the disparate natures of these two approaches. For example, we found that vocabulary unique to inorganic synthesis like sintering and calcining would be frequently misclassified. Additionally, existing models with developed ontologies do not include tags for important inorganic synthesis actions like shaping of samples into pellets.

In this work, 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.[38,39] Specifically, we built and created a dataset of 3040 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 (Fig. 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

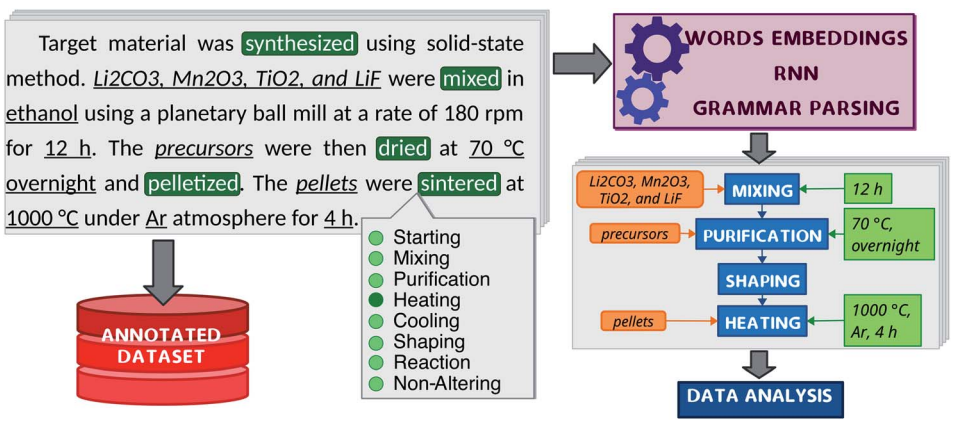


Fig. 1 Schematic workflow of data annotation, extraction and analysis. First, the set of paragraphs were annotated using an Amazon Mechanical Turk engine. Highlighted in green are the action tokens that were annotated and then extracted using a neural network model. Other highlighted tokens and phrases (*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.

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.

# 2 Methodology

## 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 h", "the temperature was **kept** at 1000 K", "the solution was **maintained** at 200 K for 12 h".

• *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".

## 2.2 Dataset annotation

To annotate synthesis paragraphs with the unified language of synthesis actions (ULSA), we selected 535 synthesis paragraphs from the database of ~420 000 full-text publications acquired previously.[12] The paragraphs where 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.[12,39]

The 535 paragraphs consisted of 3781 tokenized sentences.[16] 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 3040 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 3040 labeled synthesis sentences. All annotations were performed using a custom Amazon Mechanical Turk-based server.

## 2.3 Annotation decisions and ambiguous cases

The ULSA was developed based on the authors' own experiences with the extraction of information from materials synthesis paragraphs[12] 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.

### 2.4 Synthesis terms mapping

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

**2.4.1 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[40]) and matches only verbs against the lookup table.

**2.4.2 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 Word2-Vec model[41] implemented in the Gensim library.[42] We used ~420 000 paragraphs describing four synthesis types: solid-state, sol–gel, solvo-/hydrothermal and precipitation synthesis. The paragraphs were obtained as described in our previous work.[12] Prior to training, the text was normalized and tokenized using ChemDataExtractor.[16] 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.

**2.4.3 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.

### 2.5 Data analysis

**2.5.1 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[40] and by using dictionaries of common solution and mixing terms.

**2.5.2 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[40] and by applying rule-based regular expression matching.[29] 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.

**2.5.3 Constructing synthesis flowchart for paragraphs.** For every paragraph in the set, we then applied the bi-LSTM mapping model (Section 2.4) 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.

## 3 Results

### 3.1 Code and data availability

The dataset of 3040 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.

### 3.2 Dataset statistics

The quantitative characteristics of the set are provided in Table 1 and displayed in Fig. 2. Briefly, 535 synthesis paragraphs resulted in 3781 sentences of which 3040 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.

To probe the robustness of ULSA and our annotation schema, we asked 6 human experts to annotate the same

Table 1 Quantitative characteristics of the dataset chosen for annotation with ULSA schema

| | 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 |

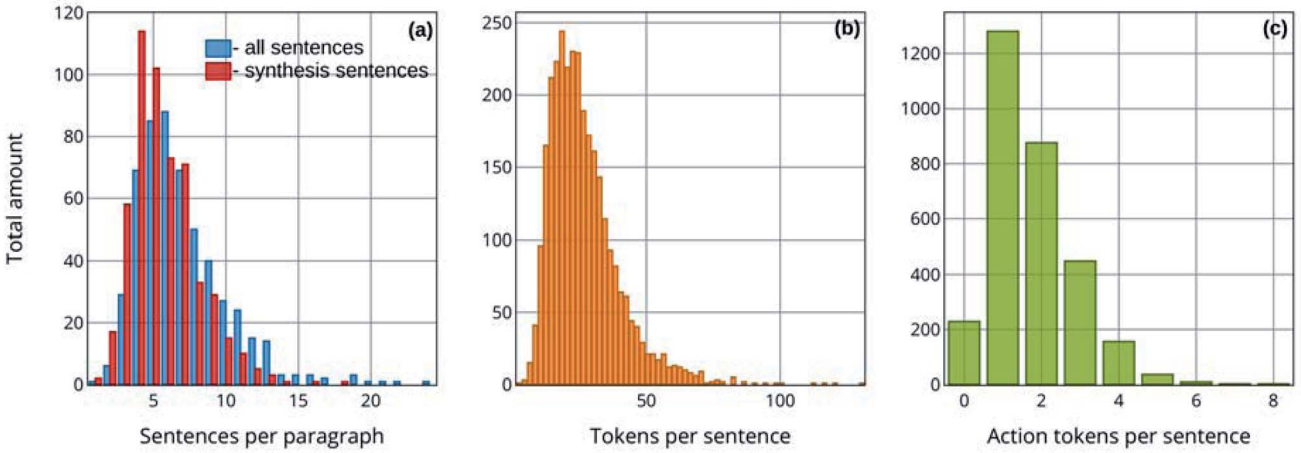


Fig. 2 Quantitative characteristics of the annotated dataset. (a) Number of sentences per paragraph (blue), including sentences related to synthesis procedures (red). (b) Number of all tokens per sentence in the annotated set. (c) Number of tokens denoting a synthesis action per sentence in the annotated set.

paragraphs in our dataset and used Fleiss' kappa score to estimate the inter-annotator agreement between the annotations.[43] 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 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.

Table 2 Fleiss' kappa score for inter-annotator agreement using ULSA scheme

| | 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 |

### 3.3 Mapping synthesis procedures into a unified language of synthesis actions

**3.3.1 Mapping model.** As a first approach for mapping of synthesis paragraphs into ULSA, we used dictionary lookup constructed as described in Section 2.4.1. We use the labeled dataset of 3040 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 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.[44] Indeed, bi-LSTM is simple enough and does not require exhaustive training and fine-tuning, as is common for BERT[45] and GPT models.[46–48] The bi-LSTM model combined with word embeddings (Section 2.4.3) was trained on the labeled dataset of 3040 sentences. The bi-LSTM model significantly improves mapping accuracy, yielding >90% F1 score (Table 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.

The output of the baseline models and the bi-LSTM model for exemplary solid-state and hydrothermal synthesis

Table 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

| 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 |

paragraphs are shown in Table 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.

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.

**3.3.2 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 Fig. 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

Table 4 Examples of processing a solid-state and hydrothermal synthesis paragraph by baseline models and the bi-LSTM model using ULSA scheme

| 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 $min^{-1}$ in ambient air, held at 1200 °C for 2 h, and then cooled to room temperature at 20 °C $min^{-1}$. The resulting powder was then ground for 5 min to break up agglomerates | *Starting* (synthesized)<br>*Mixing* (milled)<br>*Mixing* (mixture)<br>*Heating* (heated)<br>*Heating* (heating)<br>*Cooling* (cooled) | *Starting* (synthesized)<br>*Mixing* (milled)<br>*Cooling* (cooled) | *Starting* (synthesized)<br>*Mixing* (milled)<br>*Heating* (heated)<br>*Reaction* (held)<br>*Cooling* (cooled)<br>*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)<br>*Mixing* (stirred)<br>*Mixing* (added)<br>*Mixing* (mixture)<br>*Heating* (heated)<br>*Mixing* (stirring)<br>*Reaction* (kept)<br>*Purification* (dried) | *Mixing* (stirred)<br>*Mixing* (added)<br>*Heating* (heated)<br>*Reaction* (kept)<br>*Purification* (dried) | *Mixing* (mixed)<br>*Mixing* (stirred)<br>*Mixing* (added)<br>*Heating* (heated)<br>*Mixing* (stirring)<br>*Reaction* (kept)<br>*Purification* (washed)<br>*Purification* (dried) |

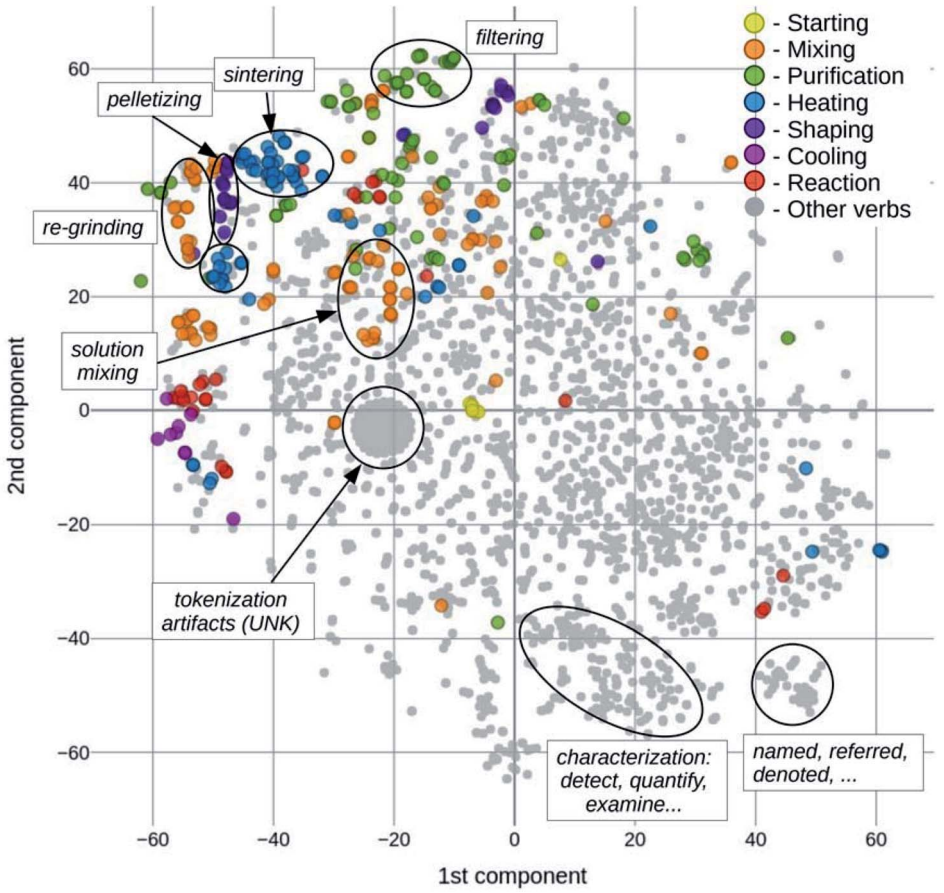


Fig. 3 2D projection of word embeddings vectors. Shown are the most frequent verb tokens encountered in the set of ~420 000 paragraphs describing a synthesis procedure. Highlighted in different colors are the vectors that correspond to the common verbs from the categories of synthesis actions used for 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.

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 related to sintering (blue dots), pelletizing (purple dots) and re-grinding (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, such as filtering or drying (green dots). Similarly, verbs indicating cooling processes (magenta dots) and the verbs referring to reaction processes (red dots) are clustered together. This 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.

**3.3.3 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 4000 paragraphs (1000 per each synthesis type) randomly pulled from the set of ~420 000 inorganic synthesis paragraphs (see Section 2.5.3 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 2.5.1). 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 ~420 000 paragraph set. All the adjacency matrices were flattened and concatenated, resulting in a matrix of size 100-by-4000, *i.e.* 10-by-10 matrix per each of 4000 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.

Fig. 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 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.

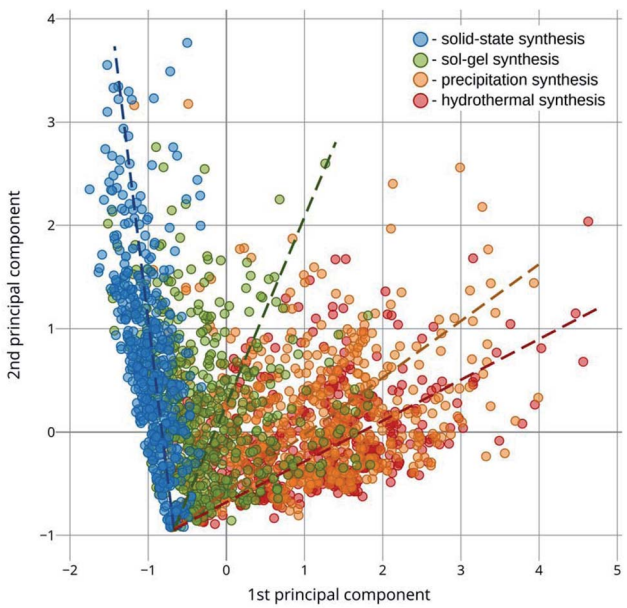


Fig. 4 Visualization of the first two principal components for the adjacency matrices of synthesis action graphs. Each dot on the plot represents a synthesis graph colored according to its type. Dashed lines display linear fitting of each data subset and show the overall direction for clustering of each synthesis graph. Note that the lines were shifted to have a common origin for representation purposes while preserving the slope.

## 4 Discussion and conclusions

In this work, 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 3040 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 inorganic 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.

Recent developments in autonomous robotic synthesis and the attempts to "close the feedback loop" in making decisions for the next synthesis step make the question of synthesis ontology and unification especially important.[36,37,49] Indeed, while theoretical decision-making and AI-guided systems can operate with abstract synthesis representations, implementation of this methodology 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 work can become a solid foundation for the future development in this direction.

## Author contributions

Z. W., K. C., O. K., and G. C. conceived the idea and drafted the manuscript. Z. W., K. C., A. C., and O. K. implemented the algorithms and analysed the data. Z. W., Y. F., and H. H. built the annotation tool. Z. W., K. C., Y. F., Y. Z., and O. K. defined the annotation schema. Z. W., K. C., Y. F., H. H., T. H., and B. D. prepared the annotation dataset. All authors discussed and revised the manuscript.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

The authors would like to thank the team of librarians from the University of California, Berkeley: Anna Sackmann (Data Services Librarian), Rachael Samberg (Scholarly Communication Officer) and Timothy Vollmer (Scholarly Communication & Copyright Librarian) for helping us to navigate through publishers copyright policies and related issues. We also thank Prof. Wenhao Sun (University of Michigan) for helpful discussions and thoughts about materials synthesis. This work was primarily supported by the National Science Foundation under Grant No. DMR-1922372. Dr Olga Kononova was in part supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 within the GENESIS EFRC program (DE-SC0019212). Expert validation of the extracted data was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (D2S2 program KCD2S2).

## References

1 K. Alberi, *et al.*, The 2019 materials by design roadmap, *J. Phys. D: Appl. Phys.*, 2018, **52**, 013001.
2 L. Himanen, A. Geurts, A. Foster and P. Rinke, Data-driven materials science: Status, challenges, and perspectives, *Adv. Sci.*, 2019, **6**(21), 1900808.
3 J. Schmidt, M. Marques, S. Botti and M. Marques, Recent advances and applications of machine learning in solid-state materials science, *npj Comput. Mater.*, 2019, **5**, 83.
4 O. Kononova, *et al.*, Opportunities and challenges of text mining in materials research, *iScience*, 2021, **24**, 102155.
5 E. Olivetti, *et al.*, Data-driven materials research enabled by natural language processing, *Appl. Phys. Rev.*, 2020, **7**, 041317.
6 M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal and A. Valencia, Information retrieval and text mining technologies for chemistry, *Chem. Rev.*, 2017, **117**, 7673–7761.

7 E. Kim, *et al.*, Materials synthesis insights from scientific literature *via* text extraction and machine learning, *Chem. Mater.*, 2017, **29**, 9436–9444.
8 S. Huang and J. M. Cole, A database of battery materials auto-generated using chemdataextractor, *Sci. Data*, 2020, **7**, 1–13.
9 C. Court and J. M. Cole, Auto-generated materials database of curie and néel temperatures *via* semi-supervised relationship extraction, *Sci. Data*, 2018, **5**, 180111.
10 C. Court and J. Cole, Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning, *npj Comput. Mater.*, 2020, **6**, 1–9.
11 E. Kim, *et al.*, Machine-learned and codified synthesis parameters of oxide materials, *Sci. Data*, 2017, **4**, 170127.
12 O. Kononova, *et al.*, Text-mined dataset of inorganic materials synthesis recipes, *Sci. Data*, 2019, **6**, 1–11.
13 S. Mysore *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* pp. 56–64, 2019). 1905.06939.
14 E. Kim, K. Huang, S. Jegelka and E. Olivetti, Virtual screening of inorganic materials synthesis parameters with deep learning, *npj Comput. Mater.*, 2017, **3**, 53.
15 S. Eltyeb and N. Salim, Chemical named entities recognition: A review on approaches and applications, *J. Cheminf.*, 2014, **6**, 1–12.
16 M. C. Swain and J. M. Cole, Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
17 D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy and P. Murray-Rust, Oscar4: a flexible architecture for chemical text-mining, *J. Cheminf.*, 2011, **3**, 41.
18 L. Weston, *et al.*, Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, *J. Chem. Inf. Model.*, 2019, **59**, 3692–3702.
19 A. Hiszpanski, *et al.*, Nanomaterials synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge, *J. Chem. Inf. Model.*, 2020, **60**, 2876–2887.
20 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, Chemicaltagger: A tool for semantic text-mining in chemistry, *J. Cheminf.*, 2011, **3**, 1–13.
21 A. Vaucher, *et al.*, Automated extraction of chemical synthesis actions from experimental procedures, *Nat. Commun.*, 2020, **11**, 3601.
22 S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan and L. Cronin, A universal system for digitization and automatic execution of the chemical synthesis literature, *Science*, 2020, **370**, 101–108.
23 S. Mysore *et al.*, *Automatically extracting action graphs from materials science synthesis procedures* (2017). 1711, p. 06872.
24 F. Kuniyoshi, K. Makino, J. Ozawa and M. Miwa, Annotating and extracting synthesis process of all-solid-state batteries from scientific, in *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, 2020, pp. 1941–1950.
25 F. Kuniyoshi, J. Ozawa and M. Miwa, *Analyzing research trends in inorganic materials literature using nlp*, 2021, vol. 2106, p. 14157.
26 T. Dieb, M. Yoshioka, S. Hara and M. Newton, Framework for automatic information extraction from research papers on nanocrystal devices, *Beilstein J. Nanotechnol.*, 2015, **6**, 1872–1882.
27 C. Kulkarni, W. Xu, A. Ritter & R. Machiraju 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)*, vol. 97–106, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018.
28 A. Friedrich *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, pp. 1255–1268.
29 D. Jurafsky & J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2nd edn, 2009.
30 J. Goodman, Computer software review: Reaxys, *J. Chem. Inf. Model.*, 2009, **49**, 2897–2898.
31 S. Kim, *et al.*, PubChem 2019 update: improved access to chemical data, *Nucleic Acids Res.*, 2018, **47**, D1102–D1109.
32 J.-D. Kim, T. Ohta, Y. Tateisi and J. Tsujii, Genia corpus – a semantically annotated corpus for bio-textmining, *Bioinformatics*, 2003, **19**, i180–i182.
33 M. Krallinger, *et al.*, The chemdner corpus of chemicals and drugs and its annotation principles, *J. Cheminf.*, 2015, **7**, S2.
34 J. Mayfield, I. Lagerstedt and R. Sayle, Pistachio, in *NIH Virtual Workshop on Reaction Informatics*, May 2021.
35 E. Kim, K. Huang, O. Kononova, G. Ceder and E. Olivetti, Distilling a materials synthesis ontology, *Matter*, 2019, **1**, 8–12.
36 N. J. Szymanski, *et al.*, Toward autonomous design and synthesis of novel inorganic materials, *Mater. Horiz.*, 2021, **8**, 2169–2198.
37 A. J. S. Hammer, A. I. Leonov, N. L. Bell and L. Cronin, Chemputation and the standardization of chemical informatics, *JACS Au*, 2021, **1**, 1572–1587.
38 R.-R. Xu, Chapter 1 - introduction, in *Modern Inorganic Synthetic Chemistry*, ed. R. Xu and Y. Xu, 2nd edn, Elsevier, Amsterdam, 2017, pp. 1–7.
39 H. Huo, *et al.*, Semi-supervised machine-learning classification of materials synthesis procedures, *npj Comput. Mater.*, 2019, **5**, 1–7.
40 M. Honnibal & M. Johnson An improved non-monotonic transition system for dependency parsing. in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1373–1378.

41 T. Mikolov, I. Sutskever, K. Chen, G. Corrado & J. Dean *Distributed representations of words and phrases and their compositionally,* 2013. vol. 1310, p. 4546.
42 R. Řehůřek & P. Sojka Software framework for topic modelling with large corpora. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, vol. 45–50, ELRA, Valletta, Malta, 2010.
43 J. Fleiss, Measuring nominal scale agreement among many raters, *Psychol. Bull.*, 1971, **76**, 378–382.
44 S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.*, 1997, **9**, 1735–1780.
45 J. Devlin, M.-W. Chang, K. Lee & K. Toutanova *BERT: Pre-training of deep bidirectional transformers for language understanding*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, URL **https://aclanthology.org/N19-1423**.
46 A. Radford & K. Narasimhan *Improving language understanding by generative pre-training* (2018).
47 A. Radford *et al.*, *Language models are unsupervised multitask learners* (2019).
48 T. Brown *et al.*, Language models are few-shot learners, *Advances in Neural Information Processing Systems*. ed. Larochelle H., Ranzato M., Hadsell R., Balcan M. F. & Lin H., vol. 33, 1877–1901 (Curran Associates, Inc., 2020).
49 B. Burger, *et al.*, A mobile robotic chemist, *Nature*, 2020, **583**, 237–241.