npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Semi-supervised machine-learning classification of materials synthesis procedures

Haoyan Huo[1,2], Ziqin Rong[1], Olga Kononova[1], Wenhao Sun[2], Tiago Botari[1,2], Tanjin He[1,2], Vahe Tshitoyan[2,3] and Gerbrand Ceder[1,2]

Digitizing large collections of scientific literature can enable new informatics approaches for scientific analysis and meta-analysis. However, most content in the scientific literature is locked-up in written natural language, which is difficult to parse into databases using explicitly hard-coded classification rules. In this work, we demonstrate a semi-supervised machine-learning method to classify inorganic materials synthesis procedures from written natural language. Without any human input, latent Dirichlet allocation can cluster keywords into topics corresponding to specific experimental materials synthesis steps, such as "grinding" and "heating", "dissolving" and "centrifuging", etc. Guided by a modest amount of annotation, a random forest classifier can then associate these steps with different categories of materials synthesis, such as solid-state or hydrothermal synthesis. Finally, we show that a Markov chain representation of the order of experimental steps accurately reconstructs a flowchart of possible synthesis procedures. Our machine-learning approach enables a scalable approach to unlock the large amount of inorganic materials synthesis information from the literature and to process it into a standardized, machine-readable database.

npj Computational Materials (2019)5:62 ; https://doi.org/10.1038/s41524-019-0204-1

INTRODUCTION

Over the last 30 years, advances in computational materials science have led to tremendous successes in materials design, with dozens of computationally designed novel compounds,[1,2] and on-demand availability of ab initio predicted properties.[3] However, the materials discovery pipeline remains bottlenecked by the challenges of experimental synthesis, which can require months of trial-and-error before a novel compound can be made. At present, it remains difficult to design how to synthesize predicted materials in a laboratory, or whether or not it is even possible.[4]

Current approaches toward understanding and predicting materials synthesis have involved in situ X-ray diffraction (XRD) investigations,[5,6] ab initio thermodynamic modeling,[7–9] classical thermodynamics perspectives,[4] and machine-learning guided synthesis parameters search.[10,11] Recently, exciting applications of machine-learning methods to retrosynthesis in organic chemistry are proving to be impactful,[12–14] inspiring the application of similar methods to predict inorganic materials synthesis. These machine-learning investigations of organic chemistry synthesis reactions have been enabled by organic chemistry reaction databases, such as Reaxys, which include >12 million single-step reactions. There is currently no analogous database that comprehensively catalogs the synthesis reactions of inorganic materials syntheses. However, even limited databases of materials synthesis reactions can yield valuable insights on the relationships between synthesis parameters and reaction products, as for example exemplified by Kim et al.[15–17] and others.[11,18]

To build a comprehensive inorganic materials synthesis database, synthesis procedures must be classified with high-resolution at multiple levels: at a high-level, the synthesis methodology; at an intermediate-level, individual experimental steps; and at a detailed-level, specific processing parameters. In principle, we could analyze sentence grammar and keywords to build a rule-based classification algorithm to identify different types of synthesis procedures. However, this is impractical, due to both the notorious ambiguity of natural language[19–21] and the complexity of solid-state chemistry terminology. Statistical classification algorithms, such as deep-learning neural networks[22,23] can achieve good text classification performances[24] with large amounts of training data.[25] However, no large annotated text data sets to train on exist in materials science or chemistry.

Recent advances in machine-learning research have demonstrated that semi-supervised learning methods can solve similar classification problems with much lessannotated data than supervised learning methods.[26–28] Here, we present a semi-supervised machine-learning approach (that uses a small amount of labeled data and a large amount of unlabeled data) for the accurate classification of synthesis procedures as described in written natural language. Using a body of 2,284,577 articles, we applied latent Dirichlet allocation (LDA)[29] to identify the experimental steps implied in sentences in an unsupervised manner. The "experimental steps" are grouped as topics and LDA provides a probabilistic topic distribution for each sentence. To this topic distribution, we apply the random decision forests (RF) algorithm[30]—a supervised machine-learning method—to classify different types of synthesis procedures: solid-state synthesis, hydrothermal synthesis, sol–gel precursor synthesis, or none of the above. We demonstrate that the RF models can achieve high classification performance with training data sets as small as a few hundred paragraphs, which can be readily prepared by manual annotation efforts. By combining these unsupervised and

[1]Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA; [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [3]Present address: Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
Correspondence: Gerbrand Ceder (gceder@berkeley.edu)

Received: 14 February 2019 Accepted: 31 May 2019
Published online: 08 July 2019

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

supervised approaches, our machine-learning algorithm accurately captures the features and subtleties of different synthesis procedures, with high classification performance, with results that can be presented in a way that is readily understood and interpretable by humans. Finally, we construct a machine-learned flowchart of synthesis procedures, which demonstrates that our method can build a "machine intuition" of materials synthesis procedures beyond classification.

RESULTS

Unsupervised learning of synthesis processes

Humans can categorize sentences into topics by recognizing familiar keywords. However, this objective can be difficult to train a computer to achieve, because it is impractical to code explicit rules for keywords of an English vocabulary that is both large (> 10,000) and open for new materials science/chemistry terms. Furthermore, in natural language various synonyms can often be used to represent the same topic, which introduces ambiguity and complexity into hard-coded rules. LDA[29,31,32] is an unsupervised topic modeling algorithm that observes common keywords over a large number of papers, then automatically clusters these synonymous keywords together into "topics". We applied LDA to identify topics of synthesis from the scientific literature, and we demonstrate that the topical grouping is closely related to conventional experimental classification of synthesis steps.

We first use LDA to identify topic–word distributions, which are a set of multinomial probability distributions over a cluster of keywords conditioned on certain topics. To demonstrate, in Table 1 we list two topics learned by LDA. (A complete list of all 200 topics can be found in Table S1.) We first show in Table 1 some representative sentences that we consider to discuss similar topics. From a collection of thousands of unlabeled sentences, LDA learns topic–word distributions using a Bayesian inference method. As shown in the second column of Table 1, the keywords (words of highest probability) of topics match the vocabulary often used by chemists to discuss each topic, making it possible for chemists to interpret the learned topics. For example, in Table 1, we interpret topic T_1 as "(ball-)milling", and topic T_2 as "high temperature sintering". We emphasize that the topic names, "(ball-)milling" and "sintering", are assigned by us for the sake of convenience, and the choice of names does not affect the topic–word distributions learned by LDA.

The distribution of topics in a sentence infers a "document–topic" distribution, which is quantified by the probability that each topic appears in a sentence. For example, in a sentence excerpted from our database, "the dried powders were calcined twice at 850 °C for 2 h and then ball milled again for 8 h.",[33] 39 and 60% of the words discuss the LDA-learned topics T_1 and T_2, respectively. LDA then interprets this sentence as having two topics, corresponding to the experimental steps "ball milling" and "sintering". More examples can be found in Table S2. Using document–topic distributions, a computer is able to quantitatively identify topics relevant to experimental steps in sentences, which are then used as input features for synthesis procedure classifiers.

Supervised classification of synthesis methodologies

LDA has now been used to automatically identify various topic–word distributions, which we labeled as specific experimental steps, for example, sintering, grinding, etc. These individual steps are subprocesses of an overall synthesis methodology, such as solid-state synthesis, hydrothermal, sol–gel precursor synthesis, etc. Based on the topic distributions learned by LDA, the machine is next trained to classify which of these three synthesis methodologies a synthesis paragraph corresponds to.

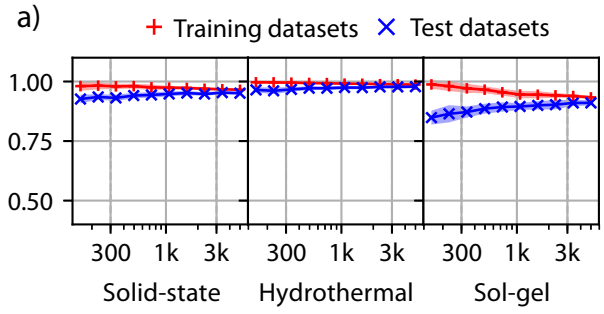
To build the classifier, we use the random forest (RF) algorithm,[30,34] which is a supervised machine-learning algorithm that uses an ensemble of decision-making trees to make classifications. We constructed a training set of synthesis paragraphs that was annotated by synthesis experts, which consists of 1000 training paragraphs for each of the three types of synthesis (solid-state, hydrothermal, and sol–gel precursor synthesis) as well as 3000 randomly sampled negative paragraphs from the database that do not contain any of the above three synthesis procedures. To provide input features for RF, we use the "topic n-gram",[35] which represents the sequence of LDA-derived topics in adjacent sentences within a paragraph. We used the *scikit-learn* Python package[36] to construct learning curves to understand how much training data is needed by the RF algorithm.

Figure 1a gives the learning curves of the RF algorithm, showing the F1 score versus the amount of training data. The RF algorithm reaches high F1 scores of ~90% when the training data set size is >3000, but surprisingly, the models can consistently converge to >80% F1 scores even when the training data set is as small as a few hundred paragraphs. These training data sets are small enough that they can be readily prepared by manual annotation efforts, indicating that LDA + RF methods are practicable machine-learning methods for classification problems of similar complexity. As summarized in Fig. 1b, the recall and precision scores are also >90%, signifying that our RF classification model is

Table 1. Two topics (topic–word distributions) selected from 200 topics learned by LDA using sentences in our database

Sample sentences	Words of highest probability	Topics
"As-received ZrB2 powder was mixed with 2 wt% B4C powder (4.5 vol%) and 1 wt% carbon (2.5 vol%) in acetone by ball milling for 24 h using WC media."[44] "The Al powder was first ball milled in an atmosphere of supra-pure hydrogen for removing the small amount of oxide film on the surface."[45]	P(ball) = 0.065 P(milling) = 0.051 P(h) = 0.042 P(milled) = 0.032 P(powder) = 0.031 P(mill) = 0.027 …	T_1 (ball-)milling
"The solid product obtained was filtered, dried at 110 °C and finally calcined in air at 550 °C for 6 h at a heating rate of 1 °C/min."[46] "Finally, the solid was calcined in air from RT to 500 °C at a heating rate of 2 °C min^{-1} and maintained for 4 h, which led to the formation of the MgO-Al2O3 support."[47]	P(°C) = 0.139 P(h) = 0.104 P(air) = 0.038 P(calcined) = 0.035 P(dried) = 0.028 P(K) = 0.016 …	T_2 sintering

Each topic is represented by a multinomial probability distribution over words. By interpreting the keywords (words of highest probability), we assign a human comprehensible label for each topic. Sample sentences from four articles[44–47] are used to demonstrate different topics



b)

Category	Precision	Recall	F1
Solid-state	96.9±1.0%	93.4±1.5%	95.1±0.6%
Hydrothermal	97.3±1.0%	98.3±1.0%	97.8±0.6%
Sol-gel precursor	95.5±1.4%	87.0±2.1%	91.0±1.1%

Fig. 1 **a** Learning curves of the RF model demonstrating F1 score improves with more training data. The red plus and blue cross symbols represent model F1 scores tested on training data sets and test data sets, respectively. The shaded areas denote the standard deviations of the curve. The performance converges to high F1 scores with training data sets as small as a few hundred paragraphs. **b** Precision/Recall/F1 scores of the RF model. The model was trained using 5000 training paragraphs and cross-validated using 1000 test paragraphs. Training paragraphs were randomly drawn from the annotated data set several times to calculate the standard deviation

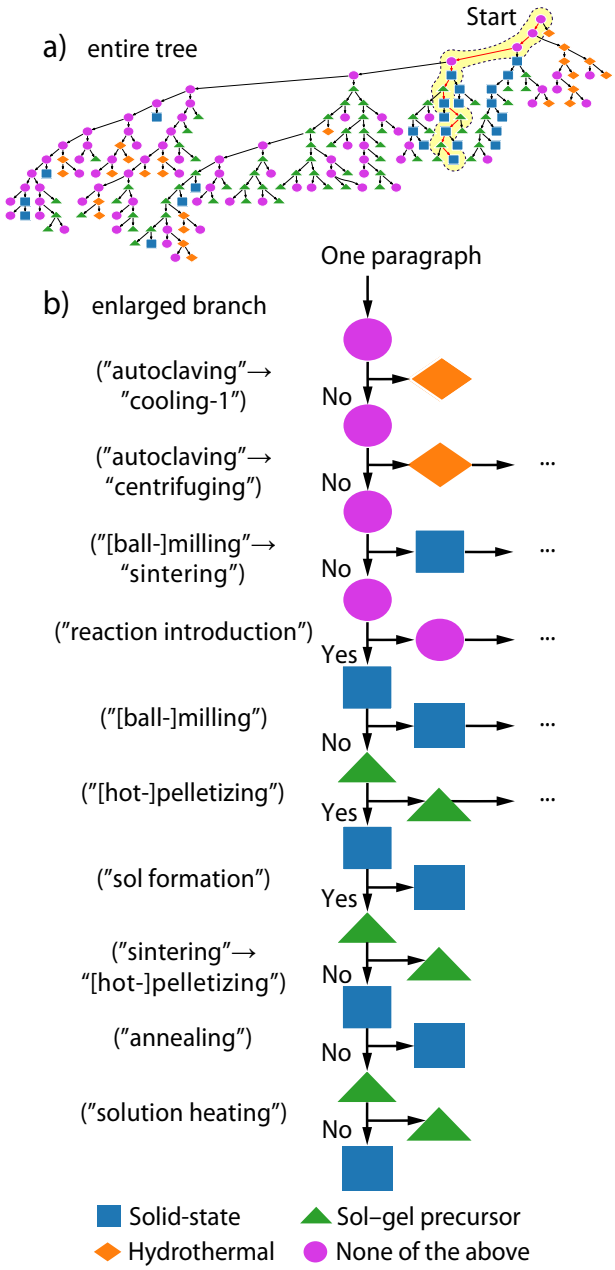


Fig. 2 **a** One of 20 decision trees learned by RF. **b** One particular branch. Starting from the topmost node, branch is made when certain topic pairs exist in a paragraph. When no branch can be made, a terminal node predicts the type of synthesis. A RF classifier consists of many trees and selects the majority of predictions

robust against false-positive and false-negative classification errors.

The RF algorithm consists of an ensemble of similar decision trees, which ultimately vote together on the final synthesis classification. Using hyperparameter optimization, we determined that 20 RF trees give the best model performance (See Methods section and Fig. S2). To visualize how our model classifies different types of synthesis procedures, we show in Fig. 2a one out of the 20 learned decision trees in our RF model. In Fig. 2a, the decision tree starts from the topmost node, and branches into one of two child nodes according to whether certain topic n-grams exist in a paragraph, as defined by the criterion of each node. We highlight a representative branch from Fig. 2a in yellow, and show the enlarged branch in Fig. 2b. For a paragraph that has topic "cooling-1" after topic "autoclaving" in two consecutive sentences, the decision tree changes its classification of the synthesis method from "none of the above" to the "hydrothermal" category. Because this "hydrothermal" node does not have any child nodes, no more decisions will be made and the decision tree predicts the paragraph as having a hydrothermal synthesis procedure.

In many ways, the RF algorithm classifies materials synthesis procedures similarly to how a solid-state chemist would—by looking for patterns of experimental procedures. For example, "shake-and-bake" is a common pattern for solid-state synthesis. If a paragraph is organized as "mix the precursors and then sinter the mixture", then one would likely classify it as solid-state synthesis. This same classification decision can be found in our computer-generated decision trees, where each node contains a pattern of experimental steps (represented by LDA topic results), such as ("[ball-]milling" → "sintering") in the third node of Fig. 2b. Moreover, our model represents patterns of synthesis as topic pairs, and we can study how words affect the detection of such patterns. As demonstrated in Fig. 2b, when a paragraph contains more keywords of topics "(ball-)milling", "(hot-)pelletizing", and "annealing" than keywords of topics "sol formation" and "solution heating", such as "milling", "pressed", and "annealed", chances are that our model predicts solid-state synthesis instead of sol–gel precursor synthesis. In general, the decision trees largely resemble the underlying procedures of materials synthesis methods, explaining why the RF algorithm can automatically pick out human-understandable features and weigh them accordingly.

Constructing a flowchart of synthesis procedures

In materials synthesis procedures, experimental steps do not appear randomly—they usually follow a certain procedural order, in patterns that are specific to different types of synthesis methodologies. Similarly, LDA-learned topics do not appear in

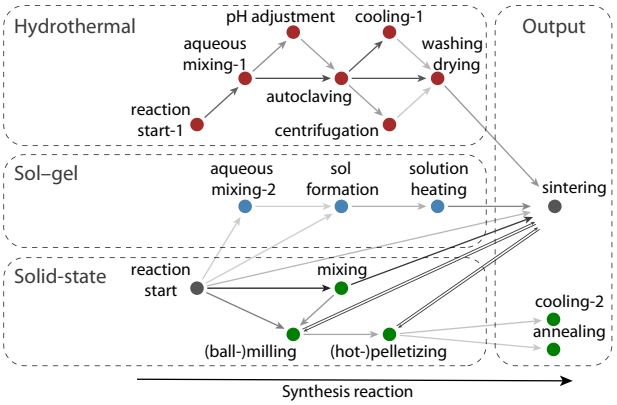


Fig. 3 Machine-learned flowchart showing the transition between experimental steps for different types of synthesis. The topics associated with the nodes can be found in Table 2 and Table S1. Edges represent transitions from one step to another, and the arrows show transition directions. Double-lined edges represent transitions in both directions. A darker edge indicates a more-probable transition

random sequences in the written synthesis paragraphs. By data-mining the transition probability from one LDA topic to another between adjacent sentences, we can construct a Markov chain representation of how various experimental steps proceed into others. We visualize these Markov chains as synthesis flowcharts, shown in Fig. 3, using a directed graph consisting of nodes and directed edges, where a node represents an experimental step, and an edge represents a transition from one experimental step to another one.

The computer-generated flowchart demonstrated in Fig. 3 largely summarizes three types of synthesis procedures. In Fig. 3, core experimental steps of syntheses are found, for example, the experimental steps "mixing", "(ball-)milling", "(hot-)pelletizing", and "sintering" (plus "cooling-2" and "annealing") are all found in the solid-state synthesis category, which matches a chemist's intuition of solid-state synthesis. The algorithm also learns important ordering information, for example, "(hot-)pelletizing" usually follows "(ball-)milling", but "(ball-)milling" never follows "(hot-)pelletizing". The edges between "sintering" and "(hot-)pelletizing" or "(ball-)milling" are found in both directions, indicating it is a common practice to regrind and pelletize sintered products in solid-state synthesis. In addition, the algorithm automatically captures subtleties regarding syntheses, for example, that "solution heating" is an intermediate step between "sol formation" and "sintering", which physically is because gel-like precursor states are formed when the particle density in the colloid is increased by evaporating liquid solvent; whereas that "pH adjustment" is an optional step between "aqueous mixing" and "autoclaving", as sometimes, but not always, the formation of the final product depends on specific pH values. Figure 3 reproduces common experimental processes from different synthesis procedures, because LDA allows computers to understand individual experimental steps, and the Markov chain construction enables general procedural orderings to be learned as they were recorded in synthesis paragraphs.

DISCUSSION

Much of the technical content in solid-state chemistry papers is locked-up in the ambiguities of written natural language. Topic modeling algorithms can teach computers to automatically elucidate structure and meaning from these complicated written texts. In this work, we combined unsupervised (LDA) and supervised (RF) machine-learning algorithms to accurately categorize different types of inorganic materials synthesis procedures by topic keywords. LDA can, without any human supervision, automatically learn keywords associated with specific experimental steps in materials synthesis procedures, which produces topic representations of sentences written in natural language. Using these topic representations, we used RF algorithms to classify different synthesis methods with high accuracy, using a relatively modest number of manually annotated synthesis paragraphs. Finally, a Markov chain representation of synthesis processes enables the construction of flowcharts, which capture many of the subtleties involved in inorganic materials synthesis. Because little annotation effort is required, our machine-learning classifier can be readily scaled up to categorize and interpret the millions of solid-state chemistry papers from the scientific literature, which can then be data-mined and analyzed using large-scale informatics tools.

LDA helps achieve high classification performance by reducing the ambiguity of natural language. Oftentimes in English, one meaning can be expressed using different synonyms. This ambiguity of English is also very common in the synthesis literature. For example, "grinding" and "milling" are often used interchangeably in experiment descriptions. LDA is designed to solve the ambiguity problem by identifying the same topic (for example, topic "(ball-)milling" in Table 2) in different ways of expression. A major advantage of LDA is that it can learn topic representations without human input. This is in contrast to other NLP methods, such as named-entity recognition (NER) or sentence dependency parsing used in similar works,[15,37] which are supervised classification models that require training on all different synonyms with the same meaning. This training is challenging owing to the limited availability of data sets in materials science with labeled text, meaning there are not enough cases for supervised learning. Another risk of neural networks trained to classify paragraphs is that the large number of parameters could lead to overfitting, and they would be unable to classify paragraphs that use synonyms for synthesis process that were not included in the training set.

One well-known limitation of LDA is that it has poor performance when modeling topics in short sentences or paragraphs.[38] We observed some incorrect classification results for short paragraphs, but these occurrences are rare, as it is nearly impossible to describe a full synthesis procedure in only a few words, and it is easy to filter all short paragraphs by the length of word sequences.

From the perspective of building an inorganic materials synthesis database, we argued that three levels of information are required: high-level classification of synthesis methodologies, intermediate-level experimental steps, and detailed-level processing parameters. We have shown that LDA is well-poised to learn the high-level synthesis methodologies and the intermediate-level experimental steps. However, LDA should be less capable of identifying the detailed-level processing parameters because it is designed to model topics (collections of common objects, ideas, facts[31]), whereas processing parameters appear as single words or phrases and need to be extracted using word-level algorithms, such as NER. Nevertheless, LDA is capable of constraining the problem domain by clustering[39] and smoothing[40] documents, and thus promoting performance of NER tasks.[41,42]

Good examples of mining materials synthesis parameters from journal articles have been previously shown by Kim et al.,[15,16] where they used NER to extract synthesis parameters and applied LDA as a post-processing analysis to cluster the chemistry of materials. These algorithms are trained and evaluated on materials synthesis paragraphs without a specific domain. However, online journal articles describe a large variety of synthesis methodologies, such as the solid-state, hydrothermal and sol–gel precursor syntheses studied in this work, where different domain knowledge is implicitly assumed, such as the vocabulary of describing

Table 2. List of topics relevant to solid-state, hydrothermal and sol–gel synthesis procedures

Assigned topic name	Cluster of keywords
Annealing	°C, h, min, air, annealed, samples, atmosphere, films, heat, treatment, annealing, furnace, treated, temperatures, temperature
Aqueous mixing-1	g, mL, water, solution, ml, dissolved, added, stirring, distilled, deionized, typical, M, mixed, ethanol, aqueous
Autoclaving	°C, autoclave, h, Teflon, lined, stainless, steel, transferred, microwave, heated, mixture, mL, solution, sealed, min
(ball-)milling	Ball, milling, h, milled, powder, mill, powders, balls, mixed, rpm, planetary, ratio, speed, zirconia, steel
Centrifuging	Water, washed, times, distilled, remove, ethanol, deionized, solution, dried, filtered, centrifugation, precipitate, three, collected, washing
Cooling-2	°C, min, temperature, rate, heating, h, heated, room, samples, cooling, furnace, cooled, K, Cmin-1, sample
(hot-)pelletizing	mm, pressed, diameter, powder, pressure, powders, pellets, pressing, hot, die, thickness, sintered, press, sintering, samples
Mixing	Materials, mixed, purity, starting, powders, stoichiometric, prepared, grade, mortar, amounts, raw, ratio, high, powder, composition
pH adjustment	pH, solution, M, NaOH, adjusted, solutions, buffer, HCl, acid, prepared, aqueous, sodium, phosphate, concentration, added
Reaction start	Prepared, method, solid, state, reaction, synthesized, x, samples, powders, conventional, gel, doped, sol, powder, synthesis
Sintering	°C, h, air, calcined, dried, K, powder, obtained, heated, powders, sintered, finally, samples, furnace, atmosphere
Sol formation	Acid, solution, ratio, added, glycol, water, citric, TEOS, molar, ethylene, prepared, agent, sol, ethanol, titanium
Solution heating	°C, h, mixture, stirred, reaction, heated, temperature, solution, min, stirring, bath, water, room, cooled, oil

By interpreting the keywords, we assigned a label of experimental steps to each topic. Topics labeled with "*−1/*−2" such as "aqueous mixing-1" and "cooling-2" are merely labeled with the same name but are learned as two independent topics. The complete list can be found in Table S1

experimental steps (Table 2) and the organization of these steps (Fig. 3). Proper consideration of the subtle domain knowledge is essential for machine learning to understand the synthesis literature in a higher resolution. Our semi-supervised approach allows paragraphs to be automatically clustered into small sub-domains of synthesis methodology, which provides a foundation for codifying domain knowledge and creating a more sophisticated analysis of synthesis information.

Our semi-supervised machine-learning algorithms benefit from high-classification performance while being trained on data sets small enough to be manually annotated by individual experts. Although this work has been a case study specifically for classifying materials synthesis paragraphs, the applicability of our method is general. For example, our method can also be used for extracting materials characterization information, which is a valuable text source for identifying the phases of synthesized materials. There are undoubtedly further opportunities to apply topic modeling methods to extract other important data and concepts from scientific articles published in materials science and other fields. We believe that this work gives a blueprint for how written information, contained in the large body of published literature, can be extracted and made machine-interpretable.

METHODS

Scientific articles used in this work are journal publications published by Springer, Wiley, Elsevier, the Royal Society of Chemistry, and the Electrochemical Society from which we received permissions to download large amounts of articles. For each publisher, we manually identified all materials science related journals available for download. A web scraping engine was built using scrapy (https://scrapy.org/). Only full-text articles published after 2000 were downloaded, including metadata such as journal name, article title, article abstract, authors, etc. All data were stored in a document-oriented database implemented using a MongoDB (https://www.mongodb.com/) database instance. Because downloaded articles are in HTML/XML format, which contains irrelevant markups and stylesheets, we developed a customized library for parsing article markup strings into text paragraphs while keeping the structures of paper and sections headings. The current snapshot of the database contains 2,284,577 papers, from which we used 3,210,525 paragraphs in the experimental sections of each paper to conduct this research. The experimental sections were identified by using case-insensitive keyword matching in section headings. (These keywords are "experiment", "synthesis", and their morphological derivations.)

Plain text paragraphs were segmented into sentences and tokenized into words using ChemDataExtractor tokenizer,[43] which is purposely trained on scientific corpus to handle abbreviations, chemical formulas, etc. Lemmatization preprocessing[35] was not practiced to keep the meanings of different word forms such as verb *fired* and noun *fire*. Common English stop-words serving as grammatical function words such as *the*, *be*, *on*, *that* were removed from each sentence.

We used the Mallet package[32] to train LDA topic models. Two parameters α and β, which control the Dirichlet prior distribution of the topic distributions and the words distributions, respectively, were set to $\alpha = 5/N$ and $\beta = 0.01$, where N is the number of topics. Inappropriate settings of the number of topics downgrade the quality of topics learned by LDA. By maximizing LDA model probability likelihood,[29] we found that setting the number of topics $N = 200$ produces the best performance of the LDA model without overfitting, as demonstrated by Fig. S1.

We used the RF module in the *scikit-learn* Python package[36] to train classification models. The "topic n-gram" feature is created as indicator variables for n-topic tuples in consecutive sentences (T_i, T_{i+1}, …, T_{i+n-1}). Each T_i is a topic in the i-th sentence with probability > 0.05. n denotes the length of the tuple, and we used $1 \le n \le 3$ in our study.

The training data set was annotated by synthesis experts in our research group and consists of 1000 training paragraphs for each of the three types of synthesis (solid-state, hydrothermal, and sol–gel precursor synthesis) as well as 3000 randomly sampled negative paragraphs from the database that do not contain any of the above three synthesis procedures. We annotated the data set according to a list of self-consistent definitions developed by us. These definitions can be found in the supplementary material. In total, 6000 annotated paragraphs were obtained. When developing this annotated data set, we found it important to use as few annotators as possible, as the use of a large number of annotators led to inconsistencies in annotation due to variations in interpretations on what each delineates each synthesis method. Part of this ambiguity of the annotation task is intrinsic. In solid chemistry, there are no formal definitions of different synthesis methodologies and hybrids of different methods are sometimes used. The issues with annotation are described in detail in the supplementary material. We used 10-fold cross-validations to test the robustness of our model. We ran cross-validation 20 times to estimate standard deviations of performance scores. In each run, the training data set contains 5000 samples, and the test data set contains 1000 samples. We did not use a development data set because we found that the model performance is nearly independent of the hyperparameters, once the number of trees ≥ 15 and the maximum depth of trees ≥ 15, as demonstrated by the grid search hyperparameter optimization in Fig. S2. Thus, we set the number of trees to 20 and the maximum depth of trees to 20 in all RF training.

To generate Fig. 3, we obtained sentence topics with probability > 0.05 in our annotated data set of paragraphs, and counted the topic pairs in

adjacent sentences, such as "mixing → sintering". By collecting all topic pairs, we can compute the probability that one topic pair follows another. This allows us to order a collection of topics into a Markov chain, which can be visualized using a directed graph, where each node is a topic and each edge is a topic pair. We weighted the edges by normalized frequencies of topic pairs observed in paragraphs. Edges with lower occurrence frequencies were plotted with a more transparent stroke in Fig. 3, and edges with occurrence frequencies lower than 0.3 were removed from the figure.

DATA AVAILABILITY

The trained LDA and RF models that support the findings of this study are available on request from the corresponding author Gerbrand Ceder (email: gceder@berkeley.edu). Copyright restrictions limit the distribution of extracted journal article paragraphs.

ACKNOWLEDGEMENTS

Funding to support this work was provided by the Energy & Biosciences Institute through the EBI-Shell program, Office of Naval Research (ONR) Award #N00014-14-1-0444, and the National Science Foundation under Grant No 5710003959. Computational study is conducted on the Savio computational cluster resource by the Berkeley Research Computing program at UC Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). We thank Anna Sackmann (Science Data and Engineering Librarian at UC Berkeley) for helping us to obtain Text and Data Mining agreements with the specified publishers. We also thank the valuable collaboration and discussions with Prof. Elsa Olivetti, Edward Kim, Alexander Van Grootel, Zach Jensen, Nicolas Mingione, Ram Balachandran, and Padmini Rajagopalan.

AUTHOR CONTRIBUTIONS

H.H. and G.C. conceived the project. H.H. implemented the algorithms and analyzed the data. Z.R. and V.T. downloaded the articles and developed the database. Z.R. and T.B. developed the HTML markup parser. H.H., Z.R., O.K. and T.H. prepared the annotation of the training data. H.H. and G.C. drafted the manuscript. All authors discussed and revised the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-019-0204-1).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 15004 (2016).
2. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191 (2013).
3. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
4. Sun, W. et al. The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225 (2016).
5. Jiang, Z., Ramanathan, A. & Shoemaker, D. P. In situ identification of kinetic factors that expedite inorganic crystal formation and discovery. *J. Mater. Chem. C.* **5**, 5709–5717 (2017).
6. Martinolich, A. J. & Neilson, J. R. Toward reaction-by-design: achieving kinetic control of solid state chemistry with metathesis. *Chem. Mater.* **29**, 479–489 (2017).
7. Sun, W., Jayaraman, S., Chen, W., Persson, K. A. & Ceder, G. Nucleation of metastable aragonite CaCO3 in seawater. *Proc. Natl. Acad. Sci.* **112**, 3199–3204 (2015).
8. Chen, B.-R. et al. Understanding crystallization pathways leading to manganese oxide polymorph formation. *Nat. Commun.* **9**, 2553 (2018).
9. Sun, W. et al. Thermodynamic routes to novel metastable nitrogen-rich nitrides. *Chem. Mater.* **29**, 6936–6946 (2017).
10. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73 (2016).
11. Xu, R. J. et al. Understanding structural adaptability: a reactant informatics approach to experiment design. *Mol. Syst. Des. Eng.* **3**, 473–484 (2018).
12. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604 (2018).
13. Feng, F., Lai, L. & Pei, J. Computational chemical synthesis analysis and pathway design. *Front. Chem.* **6** (2018).
14. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
15. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
16. Kim, E. et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
17. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational. Materials* **3**, 53 (2017).
18. Young, S. R. et al. Data mining for better material synthesis: the case of pulsed laser deposition of complex oxides. *J. Appl. Phys.* **123**, 115303 (2018).
19. Wasow, T., Perfors, A. & Beaver, D. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, 265–282 (2005).
20. Manning, C. D. & Schütze, H. *Foundations of statistical natural language processing*. (MIT press, 1999).
21. Nickel, M., Murphy, K., Tresp, V. & Gabrilovich, E. A review of relational machine learning for knowledge graphs. *Proc. IEEE* **104**, 11–33 (2016).
22. Maas, A. L. et al. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. 142–150 (Association for Computational Linguistics).
23. Pang, B., Lee, L. & Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. 79–86 (Association for Computational Linguistics).
24. Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* **117**, 7673–7761 (2017).
25. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
26. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
27. Goodfellow, I. et al. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
28. Chapelle, O., Scholkopf, B. & Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Trans. Neural Netw.* **20**, 542–542 (2009).
29. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
30. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
31. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).
32. McCallum, A. K. Mallet: a machine learning for language toolkit (2002).
33. Zhao, W., Zuo, R. & Fu, J. Temperature-insensitive large electrostrains and electric field induced intermediate phases in (0.7−x) Bi (Mg1/2Ti1/2) O3–xPb (Mg1/3Nb2/3) O3–0.3 PbTiO3 ceramics. *J. Eur. Ceram. Soc.* **34**, 4235–4245 (2014).
34. Denil, M., Matheson, D. & de Freitas, N. Narrowing the gap: random forests in theory and in practice. In *International conference on machine learning (ICML)*.
35. Jurafsky, D. & Martin, J. H. *Speech and language processing*. (Pearson, London, 2014).
36. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. Mysore, S. et al. Automatically extracting action graphs from materials science synthesis procedures. *arXiv preprint arXiv:1711.06872* (2017).
38. Cheng, X., Yan, X., Lan, Y. & Guo, J. Btm: topic modeling over short texts. *IEEE Transactions on Knowledge & Data Engineering* **26**, 2928–2941 (2014).
39. Xie, P. & Xing, E. P. Integrating document clustering and topic modeling. *arXiv preprint arXiv:1309.6874* (2013).
40. Yi, X. & Allan, J. A comparative study of utilizing topic models for information retrieval. In *European conference on information retrieval*. 29–41 (Springer).
41. Kim, H., Sun, Y., Hockenmaier, J. & Han, J. Etm: Entity topic models for mining documents associated with entities. In *2012 IEEE 12th International Conference on Data Mining*. 349–358 (IEEE).
42. Guo, H. et al. Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 281–289 (Association for Computational Linguistics).
43. 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).

44. Zhu, S., Fahrenholtz, W. G., Hilmas, G. E. & Zhang, S. C. Pressureless sintering of zirconium diboride using boron carbide and carbon additions. *J. Am. Ceram. Soc.* **90**, 3660–3663 (2007).
45. Xiao, X. et al. Influence of temperature and hydrogen pressure on the hydriding/dehydriding behavior of Ti-doped sodium aluminum hydride. *Int. J. Hydrog. Energy* **32**, 3954–3958 (2007).
46. Liang, C., Wei, M.-C., Tseng, H.-H. & Shu, E.-C. Synthesis and characterization of the acidic properties and pore texture of Al-SBA-15 supports for the canola oil transesterification. *Chem. Eng. J.* **223**, 785–794 (2013).
47. Li, G. et al. Highly selective hydrodecarbonylation of oleic acid into n-heptadecane over a supported Nickel/Zinc oxide–alumina catalyst. *ChemCatChem* **7**, 2646–2653 (2015).