

## MIT Open Access Articles

*Machine-learned and codified synthesis  
parameters of oxide materials*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Kim, Edward et al. "Machine-Learned and Codified Synthesis Parameters of Oxide Materials." Scientific Data 4 (September 2017): 170127

**As Published:** <http://dx.doi.org/10.1038/SDATA.2017.127>

**Publisher:** Nature Publishing Group

**Persistent URL:** <http://hdl.handle.net/1721.1/116340>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of use:** Attribution 4.0 International (CC BY 4.0)



# SCIENTIFIC DATA

## OPEN Data Descriptor: Machine-learned and codified synthesis parameters of oxide materials

Edward Kim<sup>1</sup>, Kevin Huang<sup>1</sup>, Alex Tomala<sup>1</sup>, Sara Matthews<sup>1</sup>, Emma Strubell<sup>2</sup>, Adam Saunders<sup>2</sup>, Andrew McCallum<sup>2</sup> & Elsa Olivetti<sup>1</sup>

Received: 10 April 2017

Accepted: 1 August 2017

Published: 12 September 2017

Predictive materials design has rapidly accelerated in recent years with the advent of large-scale resources, such as materials structure and property databases generated by *ab initio* computations. In the absence of analogous *ab initio* frameworks for materials synthesis, high-throughput and machine learning techniques have recently been harnessed to generate synthesis strategies for select materials of interest. Still, a community-accessible, autonomously-compiled synthesis planning resource which spans across materials systems has not yet been developed. In this work, we present a collection of aggregated synthesis parameters computed using the text contained within over 640,000 journal articles using state-of-the-art natural language processing and machine learning techniques. We provide a dataset of synthesis parameters, compiled autonomously across 30 different oxide systems, in a format optimized for planning novel syntheses of materials.

Design Type(s)	data integration objective • modeling and simulation objective
Measurement Type(s)	protocol
Technology Type(s)	data item extraction from journal article
Factor Type(s)	oxide
Sample Characteristic(s)	

<sup>1</sup>Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. <sup>2</sup>University of Massachusetts Amherst, Amherst, Massachusetts 01003, USA. Correspondence and requests for materials should be addressed to E.O. (email: elsao@mit.edu).

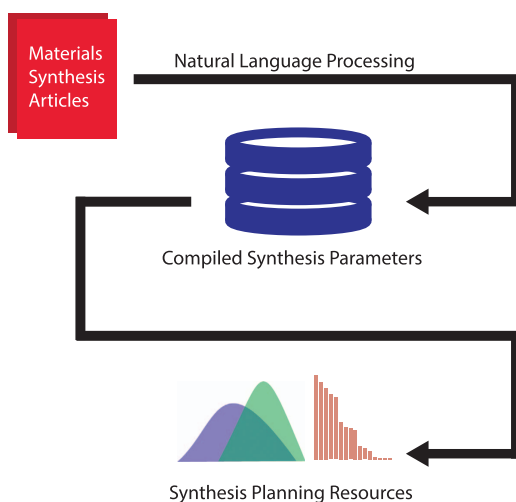
## Background & Summary

Materials Genome Initiative efforts have led to the proliferation of open-access materials properties databases, resulting in the rapid acceleration of materials discovery and design<sup>1–6</sup>. Given these advances in screening for novel compounds, the realization of comprehensive frameworks for predicting novel synthesis routes is now a primary bottleneck in materials design<sup>7</sup>. Recent high-throughput and data-driven explorations of materials syntheses have focused on optimizing a particular material system of interest<sup>8,9</sup>. Yet, the general landscape of synthesizable materials<sup>10,11</sup> spanning across material systems remains largely unexplored. To further encourage rapid and open synthesis discovery in the materials science community, we present here a dataset which collates key synthesis parameters aggregated by chemical composition (e.g., BiFeO<sub>3</sub>) across 30 commonly-reported oxide systems.

Many of the largest-volume databases consist primarily of data which are computed *ab initio* (e.g., using density functional theory)<sup>1,12</sup>. There are, however, ongoing efforts which make use of human-collected information extending beyond what can be computed from first principles: Ghadbeigi *et al.*<sup>4</sup> provide human-retrieved performance indicators for Li-ion battery electrode materials, extracted from ~200 articles, and Raccuglia *et al.*<sup>9</sup> apply a similar human-data-retrieval technique to lab notebooks to compile ~4,000 reaction conditions for training machine-learned syntheses of vanadium selenite crystals. Additionally, high-throughput experimental syntheses are capable of producing vast combinatorial materials ‘libraries’ for the purposes of materials screening<sup>8,13,14</sup>. These approaches lay the groundwork towards a broader approach using automated data collection techniques.

To accelerate the materials science community towards the goal of rapidly hypothesizing viable synthesis routes, a method for programmatically querying the body of existing syntheses is necessary. Such a resource may serve as a starting point for literature review, or an initial survey of ‘common’ and ‘outlier’ synthesis parameters, or as supplementary input data for other large-scale text mining studies on materials science literature. Indeed, approaches to high-throughput synthesis screening have seen recent success in organic chemistry<sup>15–22</sup>, since organic reaction data is well-tabulated in machine readable formats<sup>23</sup>.

In this work, we provide a set of tabulated and collated synthesis parameters across 30 oxide systems commonly reported in the literature. This data is retrieved by first training machine learning (ML) and natural language processing (NLP) algorithms using a broad collection of over 640,000 materials synthesis journal articles. These trained algorithms are then used to parse a subset of 76,000 articles discussing the syntheses of our selected oxide materials. Figure 1 provides a schematic overview of the methods used for transforming human-readable articles into machine-readable synthesis parameters and synthesis planning resources. No direct human intervention is necessary in this methodology: Our automated text processing approach downloads articles, extracts key synthesis information, codifies this information into a database, and then aggregates the data by material system.



**Figure 1. Schematic overview of text extraction and database construction.** Each colored object represents a high-level step in the automated workflow for retrieving journal articles and processing text into codified synthesis parameters. Materials synthesis articles are fed into a NLP pipeline, which computes a machine-readable database of synthesis parameters across numerous materials systems. These parameters can then be queried to produce synthesis planning resources, including, empirical distributions of real-valued parameters and ranked lists of keywords.

## Methods

### Article retrieval

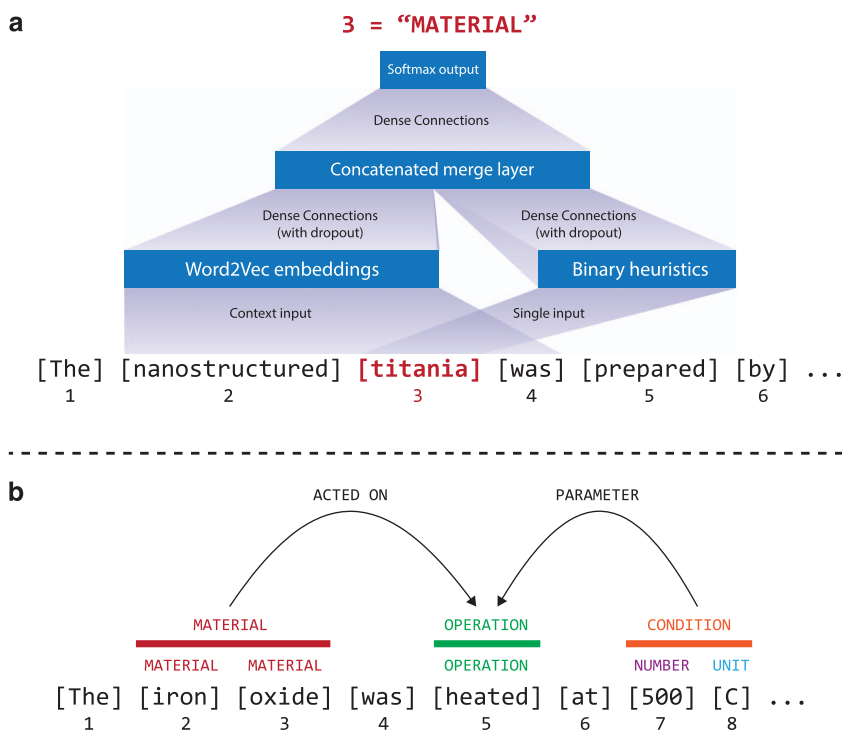
Using the CrossRef search Application Programming Interface (API)<sup>24</sup>, journal articles are programmatically queried and downloaded using additional API routes, approved individually by each publisher we access. These articles are downloaded in HTML and PDF formats, and we convert these articles to plain text for further text extraction and processing. PDF articles are converted using the open source *watr-works* Scala program.

### Article section relevance

In order to determine which paragraphs contain materials synthesis information, we have manually applied binary labels to thousands of paragraphs from approximately 100 different journal articles, with positive samples representing materials synthesis paragraphs and negative samples representing all other paragraphs. We use this data to train a binary logistic regression classifier, implemented in *scikit-learn*<sup>25</sup>.

Each paragraph in an article is represented by binary counts of frequently occurring words (commonly referred to as a ‘bag of words’ vector), and this is concatenated with a vector of simple binary heuristics (e.g., if the section title is ‘Experimental’ or ‘Methods’). A logistic regression classifier then applies categorical labels to the paragraphs, with a label of 1 indicating a synthesis paragraph and 0 representing a paragraph unrelated to synthesis. We find an overall F1 score of **96%** using this method, where the F1 score is computed from binary precision (true positives/all positive guesses) and recall (true positives/all positive samples). This score, which emphasizes the ability of a classifier to identify true positive samples, is used since most paragraphs in an article do not describe the synthesis of a material (and so synthesis paragraphs are the rarer category).

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$



**Figure 2. Neural-network and parse-based synthesis parameter extraction.** (a) A hierarchical neural network assigns labels (e.g., ‘MATERIAL’) to words one-at-a-time by converting words to embedding and heuristic vector representations, and outputting to a classifier. The embeddings of a five-word window are considered for each prediction. Each layer is densely connected, with the hidden layer concatenating each of the two input layers. The final layer is a *softmax* (classifier) computed over each possible word category. (b) A grammatical parse of a sentence is used to resolve word-level labels (below colored bars) into sequential word-chunk-level labels (above colored bars), followed by resolution into word-chunk relations (curved arcs).

## Text extraction

A schematic overview of the text extraction procedure is provided in Fig. 2. In Fig. 2a, word-level labels are applied using a neural network which predicts the category of each word (e.g., material, amount, number, irrelevant word), using a mixture of word embedding vectors from nearby words and domain-knowledge-driven heuristics. These heuristics include known word matches (e.g., ‘calcine’) and outputs of existing databases and models<sup>18,26,27</sup> which we have incorporated into our framework. A full description of the word labels is provided in Table 1.

In order to predict word-level labels, a *transfer learning* setup is used<sup>28</sup>, in which we first learn, in an unsupervised manner, a feature mapping function for words using unlabeled data. Then, we use these learned features to create context-sensitive word inputs during supervised training on a smaller set of data with high-accuracy labels applied by humans.

First, the Word2Vec algorithm is pre-trained on 640,000 unlabeled full-text materials synthesis articles in order to learn accurate vector representations for domain-specific words (e.g., anneal), which do not appear frequently in English-language documents<sup>29</sup>. This process yields a transformation function which accepts a plain-text word and outputs a dense, real-valued, fixed-length vector. These word vectors are used alongside the binary heuristic vectors to produce the fully-realized inputs for the neural network.

Following this pre-training, a *baseline* neural network is trained. A set of word labels is computed on this same set of 640,000 articles using only the binary heuristic rules (without the neural network, using simple ‘if...then’ logic). The *baseline* neural network is then trained to mimic the output of these heuristic rules using the embedding vectors as additional supporting data features. The neural network achieves a categorical accuracy of >99% and an F1 score of >99% on the task of replicating the heuristic rule word labels. This *baseline* version of the neural network achieves a categorical accuracy of 78% and an F1 score of 66% on a *test set* constructed from 10 human-labelled articles (~1,700 words). The *baseline* neural network thus serves as a lower-bound accuracy benchmark which represents the effectiveness of a database lookup strategy for categorizing words, using deterministic rules including matches to known chemical formulas (e.g., ‘Fe<sub>2</sub>O<sub>3</sub>’) or known verbs (e.g., ‘sinter’).

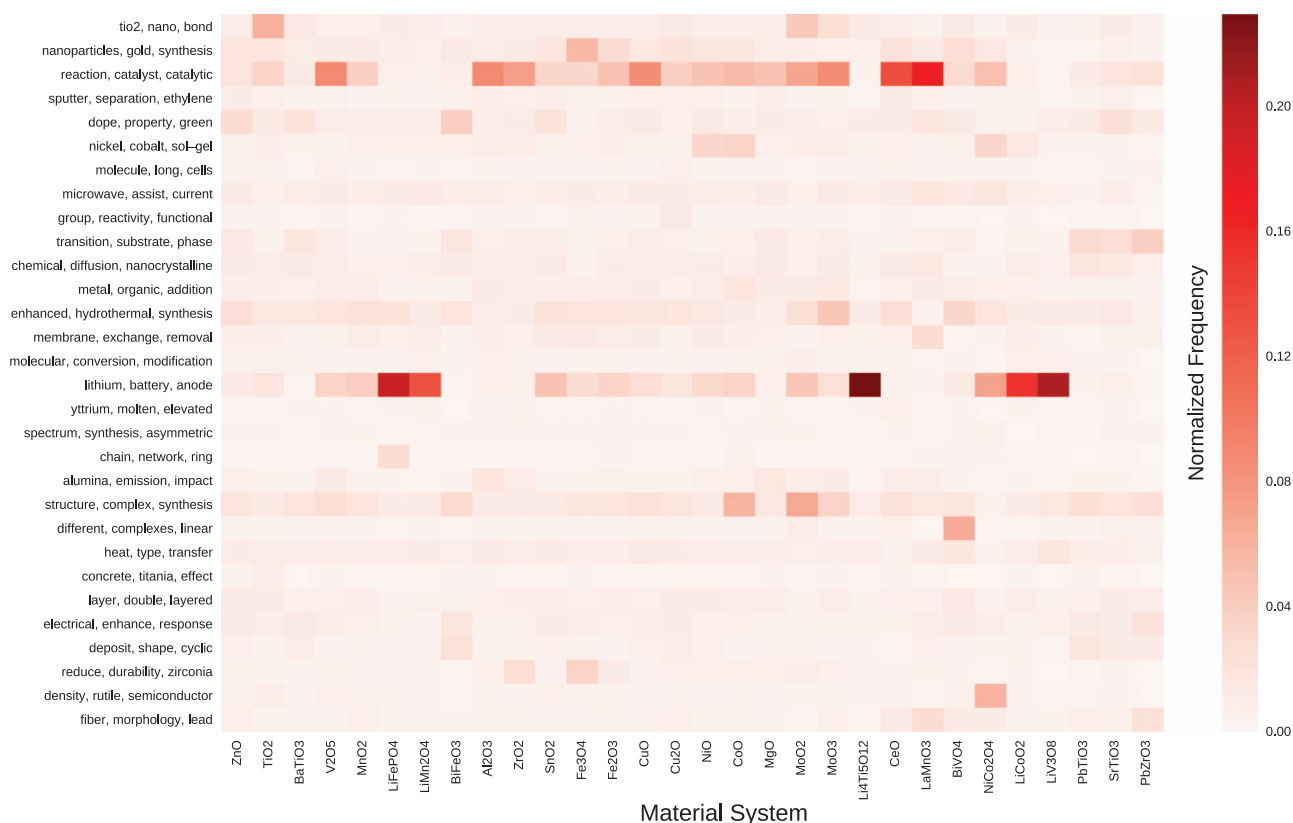
Separately, a *human-trained* neural network learns from annotated data labels which have been applied by materials scientists: A training set of 20 articles (~5,200 words) with human-applied word labels is used to learn the weights of the neural network. These human-applied word labels are marked while reading through the synthesis sections of journal articles, and thus emphasize scientific and linguistic context rather than a strict adherence to deterministic rules and database lookups. This *human-trained* network indeed achieves higher accuracies compared to the *baseline*: the categorical accuracy of this neural network classifier, as measured against the same *test set*, is 86% and its F1 score is 81%. This is comparable to the performance achieved by the recent *ChemDataExtractor* model on a similar task<sup>18</sup>, which is trained to extract relevant text from chemistry articles.

Word Label	Interpretation	Examples
TARGET	Final synthesized material	TiO <sub>2</sub> , BiFeO <sub>3</sub>
UNSPECIFIED	Generic references to materials	Solution, powder
MATERIAL	Non-target named materials	TiCl <sub>4</sub> , NaOH
OPERATION	Action on a material	Dissolve, sinter
AMOUNT MISC	Unspecified amount	Several, dropwise
AMOUNT UNIT	Amount-type unit	ml, mmol
CONDITION MISC	Unspecified parameter of an action	Slowly, ambient
CONDITION UNIT	Unit for parameter of an action	hours, °C
SYNTHESIS APPARATUS	Experimental equipment	Autoclave, furnace
CHARACTERIZATION APPARATUS	Experimental equipment	s.e.m., diffractometer
DESCRIPTOR	Qualitative material morphology	Layered, nanorods
PROPERTY MISC	Qualitative aspect of a material	Denser, brittle
PROPERTY UNIT	Quantitative aspect of a material	Mpa, cm <sup>2</sup>
PROPERTY TYPE	Type of aspect of a material	Strength, area
NUMBER	Numerical words	120, five
META	Synthesis route type	Solvothermal, sol-gel
BRAND	Commercial brand	Sigma, Fisher
REF	Citation or reference	[14], 2007
NULL	All other words	The, before

**Table 1. In-domain word categories and examples.** Word-level labels and examples of words belonging to each label. Each word in an article is assigned to exactly one of these labels.

Data Description	Data Key Label	Data Type
Name of material system	Name	String
Number of papers used to compute synthesis parameters	Num_papers	Integer
Top occurring synthesizing actions used for a material system	Associated_operations	Array of strings
Top co-occurring materials in synthesis sections for a material system	Associated_materials	Array of strings
Topic distribution for a material system	Topics	Object (dictionary) of topic strings and frequency floats
All temperatures reported in syntheses for a material system, aggregated as a kernel density estimate	Temperature_kde	Object (dictionary) of x and y floats
Hydrothermal temperatures and times reported in syntheses for a material system, aggregated as a kernel density estimate	Hydrothermal_kde	Object (dictionary) of x and y floats
Calcination temperatures and times reported in syntheses for a material system, aggregated as a kernel density estimate	Calcine_kde	Object (dictionary) of x and y floats

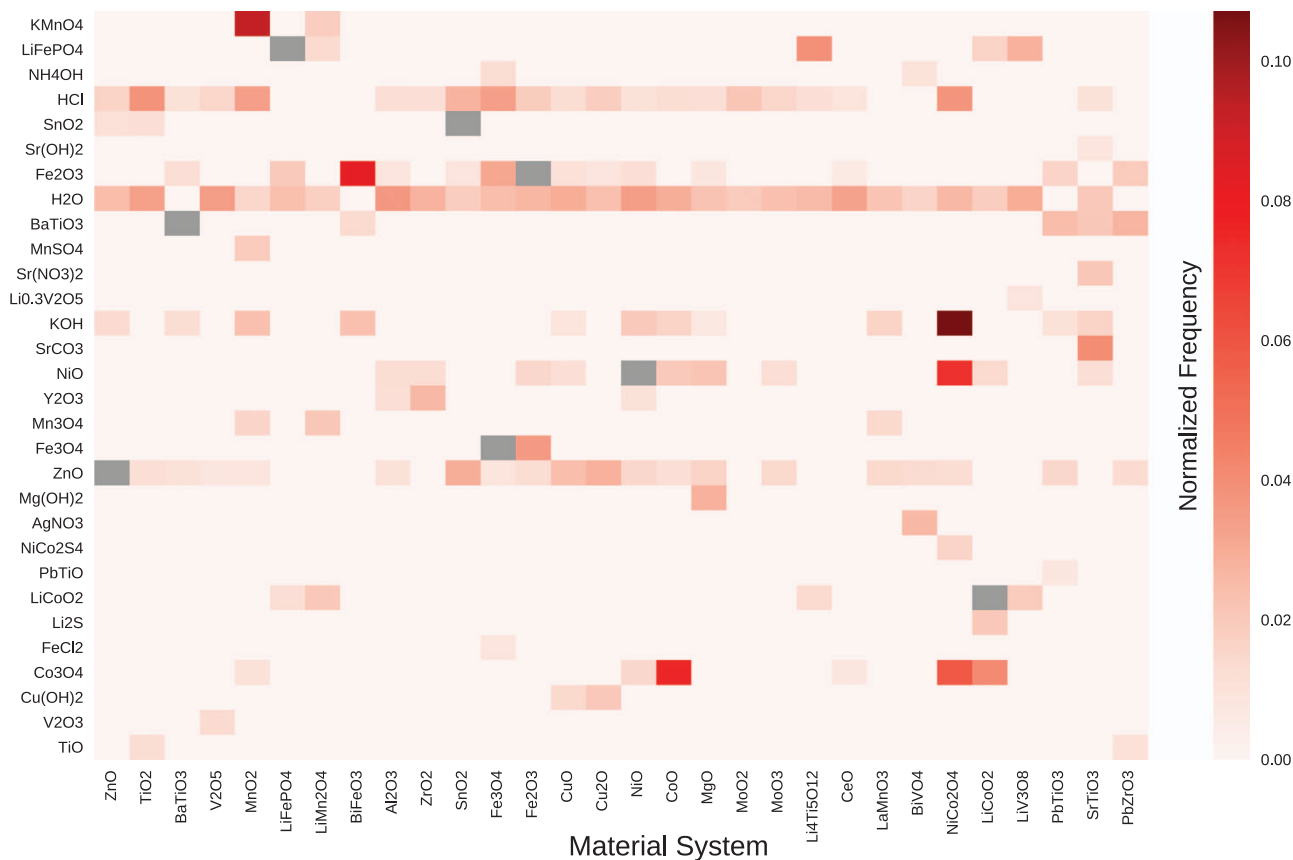
**Table 2. Schema of data records.** Overview of formatting for each data record, with each row representing a data record key. For each key, the key label name is provided, along with the data type.



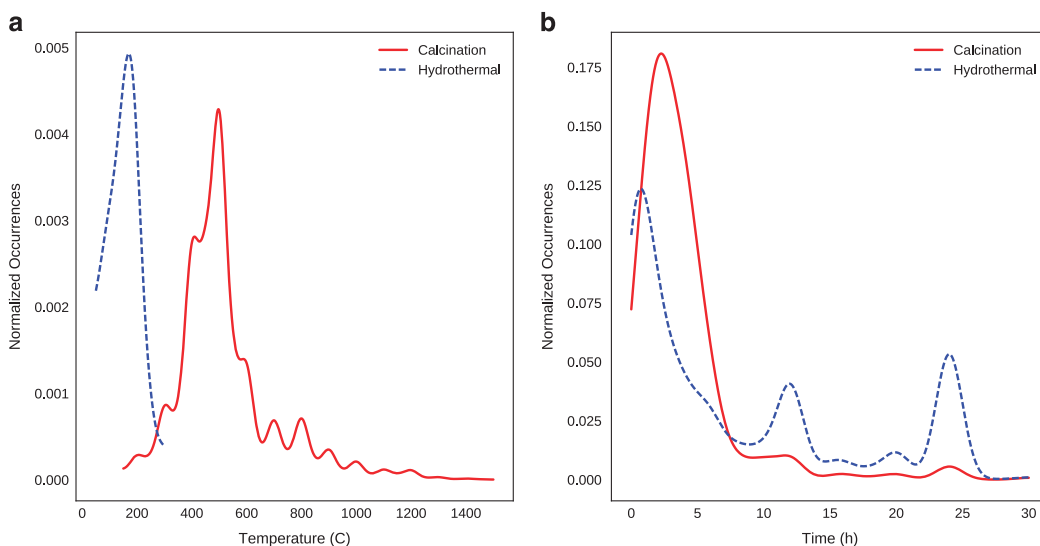
**Figure 3. Topic and synthesis target distributions within the database.** Heatmap showing a sample of topic distributions plotted against material systems of interest. Topics are computed from training a Latent Dirichlet Allocation model on 640,000 journal articles, and are labelled by their top-ranked keywords<sup>32</sup>. Values of the heatmap represent column-normalized counts across all articles within a material system.

Figure 2b shows the process of interpreting higher-level relations in the text, including resolving multiple sequential words into a single ‘chunked’ entity and relating these chunks to each other. This process is done by applying heuristic rules to the outputs of a grammatical parser<sup>27</sup>, where relations between chunks are computed from parse tree dependencies, with word-order proximity used as a fallback measure<sup>30</sup>. To illustrate via the example in Fig. 2b, ‘500 C’ is grammatically dependent on ‘heated,’ and so a relation is assigned between them.

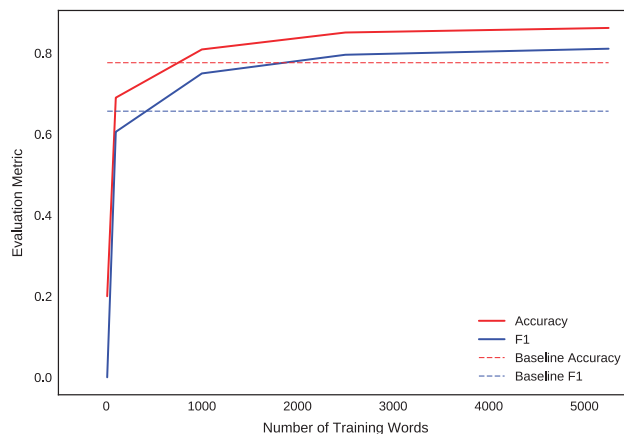
The extracted synthesis parameters from these articles are then filtered by material systems of interest, and aggregated for each material. Histograms and ranked lists of keywords are computed by counting occurrences within documents, and all histograms are normalized to integrate to unity. In this paper, we



**Figure 4. Top occurring material mentions per target material system.** Heatmap showing a sample of co-occurring mentions of materials within synthesis routes for material systems of interest. Values of the heatmap represent column-normalized counts across all articles within a material system. Counts of self-mentioning co-occurrences (e.g., ZnO mentioned in papers synthesizing ZnO) are fixed to zero prior to column normalization and plotted in grey.



**Figure 5. Temperature and time distributions for titania.** (a) Calcination and hydrothermal temperature kernel density estimate for titania, normalized to unit area. (b) Calcination and hydrothermal time kernel density estimate for titania, normalized to unit area. All density estimates are computed using Gaussian kernels computed from counts of temperatures and times extracted from synthesis sections of journal articles.



**Figure 6. Learning curve for neural-network word classifier.** The baseline accuracy and F1 score are plotted as horizontal lines, computed from the *baseline* neural network on the maximum number of training words. The solid curves are computed from the *human-trained* neural network, showing accuracy and F1 score as a function of training data volume.

present a dataset extracted from a sample of 76,000 articles which discuss the syntheses of 30 oxide systems of interest. Beyond this, we also provide a more extensive and continuously-updated dataset at [www.synthesisproject.org].

#### Code availability

The code used to compute and analyze the data is available at [www.github.com/olivettigroup/sdata-data-plots/]. Additionally, the compiled Word2Vec embedding vector model is available at [www.github.com/olivettigroup/materials-word-embeddings/]. The underlying machine learning libraries used in this project are all open-source: *Tensorflow*<sup>31</sup>, *SpaCy*<sup>27</sup>, and *scikit-learn*<sup>25</sup>.

#### Data Records

The data are provided as a single JSON file, available at [www.synthesisproject.org] and through figshare (Data Citation 1). Each record, corresponding to data for a single material system, is represented as a JSON object in a top-level list. The details of the data format are given in Table 2.

Metadata for each material system is provided in the JSON dataset, including topic distributions computed with Latent Dirichlet Allocation<sup>32</sup>. These topic distributions are visualized as a heatmap in Fig. 3, and demonstrate correlations between material chemistries and device applications, experimental apparatuses, and product morphologies.

Numerical synthesis parameters (e.g., calcination temperatures) for each material system are provided as kernel density estimates, computed across all journal articles discussing the synthesis of a given material. Such a format allows for rapid visualizations to aid high-level synthesis planning: for example, Fig. 4 shows co-occurrences between materials systems and mentions of other materials extracted from synthesis sections of articles.

#### Technical Validation

For scientific validation, we briefly compare the aggregated data in our provided dataset to known results. As an example, Fig. 5a displays frequent usages of temperatures near the anatase-rutile phase boundary for titania<sup>33</sup>. This data thus agrees with the intuitive reasoning that such temperatures are used to either crystallize an anatase-phase product, or convert to a rutile-phase product<sup>34</sup>. We also observe additional patterns which agree with intuitive expectations: Fig. 5a shows that hydrothermal reactions are confined to a narrow temperature range, peaking between 100–200 C, and Fig. 5b confirms that hydrothermal reactions more typically occur for long periods of time compared to calcination.

Besides validating the accuracy of the text extraction and word-labelling methods (as discussed in the Methods), we reiterate here the accuracy of the parsing algorithm used, which is reported as **91.85%** (ref. 27). We use this parsing algorithm as a dependency to resolve higher-level relations in our extracted text data (e.g., relating ‘500 C’ to ‘heated’ in Fig. 2b). We also report a training curve in Fig. 6, to demonstrate that several thousand labelled words is indeed a sufficient volume of data for training a neural network word classifier in the materials science domain.

#### Usage Notes

As this data is provided in the language-agnostic JSON format, no specific technical setup is required as a dependency. The authors have found it useful to load the data into the *Python* programming language,



especially for downstream integration with data from the Materials Project provided via their *pymatgen* library<sup>35</sup>.

A detailed, web-accessible Python tutorial for loading and analyzing the dataset is available at [<https://github.com/olivettigroup/sdata-data-plots/blob/master/SDATA-data-plots.ipynb>]. This web tutorial provides the exact Python code used to generate the figures in this article, along with commentary which explains the technical setup.

Empirical histograms provided in this dataset, along with ranked lists of frequent synthesis parameters, serve as useful starting points for literature review and synthesis planning: for example, selecting the most frequent synthesis parameters (e.g., most common reaction temperatures and precursors) would yield a starting point for a viable synthesis route.

Additionally, the topic labels provided in this dataset may prove useful in studies related to metadata and text mining in the materials science literature. As a motivating example, authorship and citation links have been analyzed in biomedical papers to reveal insights related to the impact of papers over time<sup>36</sup>; such analyses could potentially be extended to topic models in materials science.

While this paper details a static ‘snapshot’ of collated synthesis data, a continuously updated and rapidly-expanding dataset is also available via an API at [[www.synthesisproject.org](http://www.synthesisproject.org)].

## References

- Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater* **1**, 11002 (2013).
- Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
- Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
- Ghadbeigi, L., Harada, J. K., Lettiere, B. R. & Sparks, T. D. Performance and resource considerations of Li-ion battery electrode materials. *Energy Environ. Sci* **8**, 1640–1650 (2015).
- Saal, J. E., Kirklín, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
- Holdren, J. P. Materials Genome Initiative for Global Competitiveness. *National Science and Technology Council* [https://www.mgi.gov/sites/default/files/documents/materials\\_genome\\_initiative-final.pdf](https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf) (2011).
- Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. A bridge for accelerating materials by design. *npj Comput. Mater* **1**, 15008 (2015).
- Potyrailo, R. *et al.* Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS Comb. Sci* **13**, 579–633 (2011).
- Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
- Jansen, M. The energy landscape concept and its implications for synthesis planning. *Pure Appl. Chem.* **86**, 883–898 (2014).
- Jansen, M. Conceptual inorganic materials discovery—A road map. *Adv. Mater.* **27**, 3229–3242 (2015).
- Gunter, D. *et al.* Community Accessible Datastore of High-Throughput Calculations: Experiences from the Materials Project. *2012 SC Companion High Perform. Comput. Netw. Storage Anal* 1244–1251 (2012).
- Suh, C., Gorrie, C. W., Perkins, J. D., Graf, P. A. & Jones, W. B. Strategy for the maximum extraction of information generated from combinatorial experimentation of Co-doped ZnO thin films. *Acta Mater.* **59**, 630–639 (2011).
- Green, M. L., Takeuchi, I. & Hatrick-Simpers, J. R. Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials. *J. Appl. Phys.* **113** (2013).
- Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform* **3**, 17 (2011).
- Duvenaud, D. K. *et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst* **28**, 2215–2223 (2015).
- Rocktäschel, T., Weidlich, M. & Leser, U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**, 1633–1640 (2012).
- 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).
- Kano, Y. *et al.* U-Compare bio-event meta-service: compatible BioNLP event extraction services. *BMC Bioinformatics* **12**, 481 (2011).
- Szymkuć, S. *et al.* Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie - International Edition* **55**, 5904–5937 (2016).
- Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: March of the machines. *Angew. Chemie—Int. Ed* **54**, 3449–3464 (2015).
- Segler, M., Preuß, M. & Waller, M. P. Towards ‘AlphaChem’: Chemical Synthesis Planning with Tree Search and Deep Neural Network Policies. Preprint at <http://arxiv.org/abs/1702.00020> (2017).
- Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **49**, 2897–2898 (2009).
- Lammey, R. CrossRef text and data mining services. *Science Editing* **2**, 22–27 (2015).
- Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Mach. Learn.* **12**, 2825–2830 (2012).
- Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2015).
- Honnibal, M. & Johnson, M. An Improved Non-monotonic Transition System for Dependency Parsing. *Proc. 2015 Conf. Empir. Methods Nat. Lang. Process* 1373–1378 (2015).
- Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering.* **22**, 1345–1359 (2010).
- Mikolov, T., Corrado, G., Chen, K. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at <http://arxiv.org/abs/1301.3781> (2013).
- De Pablo, J. J., Jones, B., Kovacs, C. L., Ozolins, V. & Ramirez, A. P. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci* **18**, 99–117 (2014).
- Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *None* **1**, 19 (2015).
- Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2002).
- Ren, R., Yang, Z. Z. & Shaw, L. L. Polymorphic transformation and powder characteristics of TiO<sub>2</sub> during high energy milling. *J. Mater. Sci.* **35**, 6015–6026 (2000).

34. Primo, A., Corma, A. & García, H. Titania supported gold nanoparticles as photocatalyst. *Phys. Chem. Chem. Phys.* **13**, 886–910 (2011).
35. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
36. Catalini, C., Lacetera, N. & Oettl, A. The incidence and role of negative citations in science. *Proc. Natl. Acad. Sci* **112**, 13823–13826 (2015).

## Data Citation

1. Kim, E. *et al.* *figshare* <https://doi.org/10.6084/m9.figshare.5221351> (2017).

## Acknowledgements

The authors would like to acknowledge funding from the National Science Foundation Award #1534340, DMREF that provided support to make this work possible, support from the Office of Naval Research (ONR) under Contract No. N00014-16-1-2432, and the MIT Energy Initiative. E.K. was partially supported by NSERC. The authors would also like to acknowledge the tireless efforts of Ellen Finnie in the MIT libraries, support from seven major publishers, and research input from Yan Wang, Daniil Kitchaev, Wenhao Sun, Craig Greenberg, Rachel Osmundsen, and Vicky Gong.

## Author Contributions

E.K., A.T. and K.H. worked on the machine learning algorithms and data analysis. K.H., A.T., and S.M. produced annotated training and testing data. A.S., E.S., and A.M. provided natural language processing and data annotation insight. E.K., K.H. and E.O. wrote the manuscript.

## Additional Information

**Competing interests:** The authors declare no competing financial interests.

**How to cite this article:** Kim, E. *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* 4:170127 doi: 10.1038/sdata.2017.127 (2017).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2017