Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning

Citation: Kim, Edward et al. "Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning." Chemistry of Materials 29, 21 (October 2017): 9436–9444 © 2017 American Chemical Society

As Published: http://dx.doi.org/10.1021/acs.chemmater.7b03500

Publisher: American Chemical Society (ACS)

Persistent URL: https://hdl.handle.net/1721.1/129530

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

CHEMISTRY OF MATERIALS
Article
Cite This: Chem. Mater. XXXX, XXX, XXX-XXX
pubs.acs.org/cm

Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning

Edward Kim,[†] Kevin Huang,[†] Adam Saunders,[‡] Andrew McCallum,[‡] Gerbrand Ceder,[§] and Elsa Olivetti*[,†]

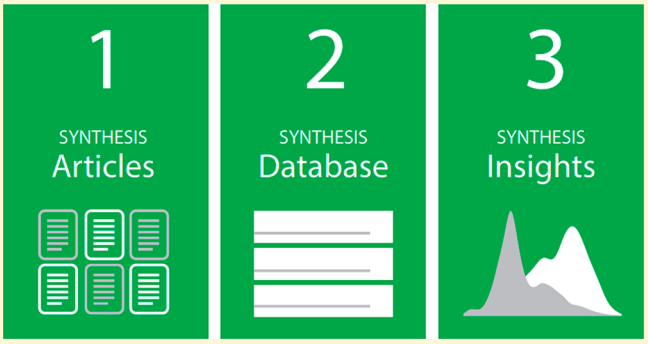
[†]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

[‡]Computer Science Department, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States

[§]Materials Science and Engineering, University of California Berkeley, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: In the past several years, Materials Genome Initiative (MGI) efforts have produced myriad examples of computationally designed materials in the fields of energy storage, catalysis, thermoelectrics, and hydrogen storage as well as large data resources that are used to screen for potentially transformative compounds. The bottleneck in high-throughput materials design has thus shifted to materials synthesis, which motivates our development of a methodology to automatically compile materials synthesis parameters across tens of thousands of scholarly publications using natural language processing techniques. To demonstrate our framework's capabilities, we examine the synthesis conditions for various metal oxides across more than 12 thousand manuscripts. We then apply machine learning methods to predict the critical parameters needed to synthesize titania nanotubes via hydrothermal methods and verify this result against known mechanisms. Finally, we demonstrate the capacity for transfer learning by using machine learning models to predict synthesis outcomes on materials systems not included in the training set and thereby outperform heuristic strategies.



INTRODUCTION

First-principles materials design, open access materials property databases,[1−3] and machine learning[4,5] have accelerated novel compound identification for a variety of applications, including energy storage, catalysis, thermoelectrics, and hydrogen storage.[6−14] To fully realize the vision of the Materials Genome Initiative of accelerating materials development,[15−18] we must, in a comprehensive and accessible way, link the compositions, structures, and morphologies of these computationally discovered materials to the synthesis conditions that can produce them. This work represents a small step in the direction toward this goal of systematically understanding the relationships between synthesized materials and reaction conditions by broadly data mining the literature.

The materials design community remains gated by the use of heuristic synthesis guidelines once a particular material of interest has been identified, either by direct first-principles computations or screening methods.[6,19,20] As a result, the synthesis of targeted novel compounds is rapidly becoming the slow step in computationally driven materials design. With direct modeling of the complex kinetic processes occurring during synthesis out of reach, a data-driven, machine learning approach that learns from the hundreds of thousands of published synthesis recipes may be more productive. As a step toward this objective, we use recent advances in full-text publisher application programming interfaces (APIs)[21] and natural language processing (NLP)[22−26] to develop a statistical learning approach to materials synthesis. While numerous studies have focused on text extraction from scientific literature,[22−24,27−29] we present here a framework focused on the problem of extracting and data-mining materials synthesis conditions.

Using a variety of machine learning and natural language processing techniques, our platform automatically retrieves articles and then extracts and codifies the materials synthesis conditions and parameters found in the text. By combining these text-mined synthesis parameters at large scale, this synthesis database can be mined to discover the underlying relationships between synthesis conditions and the materials they produce. This literature-based data mining strategy also complements and benefits from current combinatorial and in situ synthesis studies which produce libraries of materials with

Received: August 21, 2017
Revised: October 8, 2017

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

varied compositions to explore a materials parameter space.[30−32]

Here, we present a platform that leverages the large body of published synthesis recipes through natural language processing and uses these recipes to train machine learning models that aid in developing insights into the key parameters that drive the synthesis of specific, technologically relevant materials at a high level of automation.[12,33,34]

METHODS

The methods used for text extraction are briefly discussed in the following sections, and these methods are based on the techniques used by Kim et al.[35]

Article Retrieval. To construct a corpus of journal articles, the CrossRef Application Programming Interface (API)[21] is used to programmatically retrieve large lists of Digital Object Identifiers (DOIs), which serve as unique article identifiers, related to chosen search queries (e.g., "battery + electrode + synthesis"). Following this, a number of publisher APIs are used to download full-text journal articles, using a click-through service provided by CrossRef. We retrieve articles in both Portable Document Format (PDF) and plain text format, depending on availability.

Plain-Text Conversion and Classification. Using PDF text-processing tools (located at https://github.com/iesl/watr-works), we convert the collected PDF files to plain text files. The body-text contained within the articles is fed into a paragraph relevance classifier, both to reduce the data volume in later stages of data processing and to differentiate between similar sections of text, such as the experimental synthesis and materials characterization subsections.

To determine which paragraphs contain materials synthesis information, we manually applied binary labels to several hundred paragraphs from approximately 100 different journal articles, with positive samples representing materials synthesis paragraphs and negative samples representing all other paragraphs. We then use a logistic regression classifier to classify relevant articles, as implemented in the *scikit-learn* Python module.[36]

To reframe our journal article paragraphs as mathematical objects, a "word embedding" approach is used to transform the paragraphs into real-valued vectors, where each paragraph is represented by an average of context-sensitive word vectors.[25,37] This word-embedding approach has become standard in machine learning literature and was found to yield good performance on materials science vocabularies.[35]

The logistic regression classifier then applies binary labels to the paragraphs, with a label of 1 indicating a synthesis paragraph and 0 representing a paragraph unrelated to synthesis. Because there are many fewer synthesis paragraphs than nonsynthesis paragraphs, we use a class-weighting scheme to assign proportionally greater loss to the rarer category during the automated training of the algorithm. This logistic regression achieves an overall accuracy of 95% on unseen test data.

Parsing and Extraction. After identification of relevant synthesis paragraphs, the paragraphs are transformed into dependency parse trees using the ChemDataExtractor and *SpaCy* parsers.[24,25] As a part of the parsing process, word tokenization and part-of-speech tagging are performed. The former refers to splitting each sentence into a list of its constituent words (or tokens), and the latter is the process of applying grammatical labels to each word token, such as noun or verb. It is often the case that the synthesis verb of interest (which is automatically detected by a neural network) is placed at the root of the tree,[35] with the relevant synthesis parameters and materials appearing as children within the subtree of the root node.

Extraction of synthesis parameters is handled by a mixture of neural network word labeling and traversal of the dependency parse tree.[35] The parse tree of each sentence in a synthesis paragraph is scanned for the presence of key synthesis verbs (e.g., sinter, dissolve, mill), and the dependency parse trees are then iteratively traversed to find operating parameters (e.g., sintering temperature, stirring speed) by matching to specific character patterns.

The parse tree for each sentence is then scanned again for noun phrases (e.g., LiOH, ethanol, gold NPs, powder), and these phrases are matched against the *PubChem* database,[38] a character n-gram classifier (which achieves 82% accuracy for identifying materials), and the pretrained ChemDataExtractor model[24] to validate whether or not they are references to meaningful materials. These matches, along with word embedding vectors[37] trained on our corpus of papers, are fed into a neural network which predicts word categories (i.e., material, operation, amount, or condition). Training this neural network with ∼5000 human annotated words yields an overall accuracy of 86% for word categories, as measured against a set of approximately 100 human-annotated synthesis articles.[35]

Some of the errors made by the text extraction algorithms are corrected systematically by considering technological and practical limitations. As an example, the "degree" symbol is sometimes decoded from PDF documents as a numerical digit, which can adversely affect temperature parsing; this issue is fixed by pattern searches when temperatures are parsed above or below reasonable limits for related synthesis steps.

The authors note that the parsing and text extraction techniques presented here focus solely on information written in the main body text of scientific journal articles (and specifically the title, abstract, and methods sections). As a contrasting example, Swain et al. designed a system for data extraction from tables,[24] and the authors intend for future work to make use of both plain-text and tabular synthesis data, as such approaches would be informative for linking synthesis methods with resulting materials properties. Furthermore, text extraction from elsewhere within a manuscript, such as the results section, would provide information on the quality of the resulting material.

Verification of Annotated Data. In addition to the set of annotated data used to measure predictive accuracies for text extraction, another set of 30 articles was annotated in duplicate, independently, by two materials science researchers. These encoded articles were used to both confirm the consistency of the annotation procedure and provide a baseline for the upper-bounds of expected performance from any machine learning algorithms, by comparing the level of human agreement between the encoded articles. Inspection of these articles postannotation confirmed that annotation differences stemmed from details irrelevant to key synthesis details (e.g., annotating an intermediate material as a solution versus a sample was a common difference).

Data Mining and Machine Learning. Extracted synthesis parameters are encoded and compiled into a monolithic database, which can then be programmatically queried. We use this database to quantitatively analyze synthesis steps such as hydrothermal and calcination reactions reported in the literature. Additionally, the database is used to train machine learning models by providing examples of synthesis parameters and synthesis outcomes.

We note that our machine learning models do not yet robustly handle the separation of multiple synthesis routes described in a single paper, as it is nontrivial to detect natural language boundaries between synthesis routes. Therefore, we focus our data mining on specific reaction steps rather than entire synthesis routes to avoid conflating the end of one synthesis route with the beginning of another.

All machine learning models are implemented with *scikit-learn* and *tensorflow*.[36,39] The details of all machine learning model parameters are provided in the Supporting Information. The machine learning setup in Figure 2, involving a decision tree along with a linear classifier, was motivated by similar models used by Raccuglia et al. which were found to produce state-of-the-art results on their data. Raccuglia et al. also motivated the use of the machine learning approach shown in Figure 3.[33]

RESULTS

From a collection of over half a million journal articles in our database, we first apply topic and material-level text queries to select a set of articles in which metal oxides are synthesized. As an example of basic information that can be retrieved and examined in an exploratory manner, we present in Figure 1a the

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

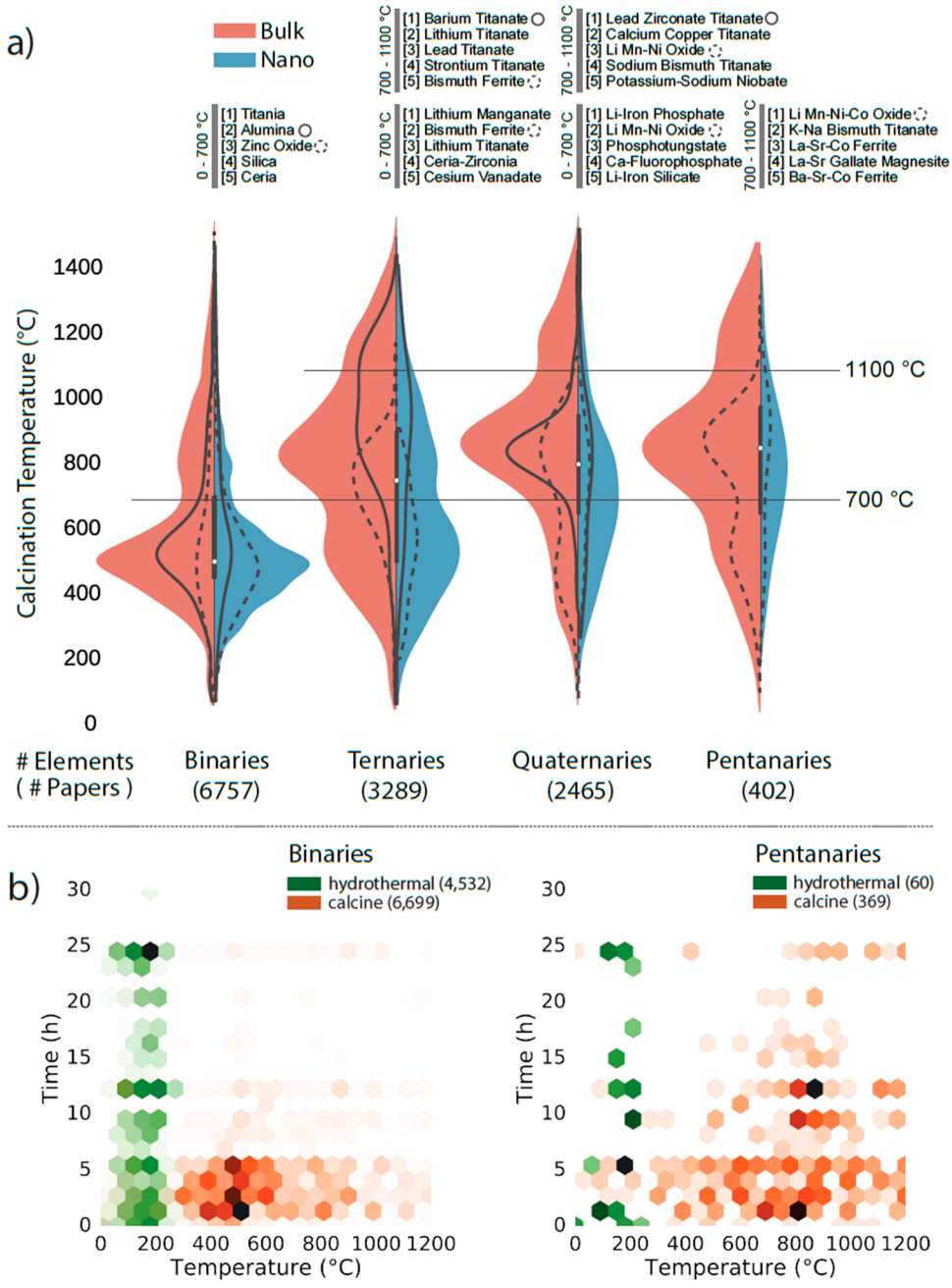


Figure 1. Synthesis parameter distributions across metal oxide systems. (a) Violin-histogram Gaussian kernel density estimate distributions of calcination temperatures for various oxides. Blue and red histogram areas are normalized to reflect relative counts between Bulk and Nano sections. The lists above each violin denote top-five occurring material systems in those temperature ranges. A few select material systems are included as solid and dashed curves within the violins. Each solid and dashed curve has been rescaled to emphasize differences in temperature peaks and relative counts of Bulk and Nano. (b) 2D hexagonally binned normalized histograms of hydrothermal reaction and calcination times and temperatures for binary and pentanary oxides. Number of papers is indicated in parentheses after each method label in the legend.

distribution of calcination temperatures used in 12 913 syntheses recipes of metal oxides, grouped by their number of constituent elements and whether or not the targets are nanostructured.

In each category we list the top 5 materials, by occurrence, and we delineate arbitrary temperature windows 0−700 and 700−1100 °C to make the peaks easier to see. We also show example curves for specific materials (with a solid or dashed line) within each distribution, and additional curves are provided in Supporting Information Figure S2. Each pair of nanostructured versus bulk distributions is scaled by paper count.

Several interesting observations can already be made from these plots. High calcination temperatures are found more frequently in the synthesis of bulk materials with greater elemental complexity. The difference in calcination temperature is particularly pronounced between the binaries and higher-component systems. Indeed, a binary oxide is often formed by straightforward substitution of the carbonate, hydroxyl, or similar anion group in the precursor by oxygen. In contrast, the phase-pure synthesis of multicomponent systems additionally

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

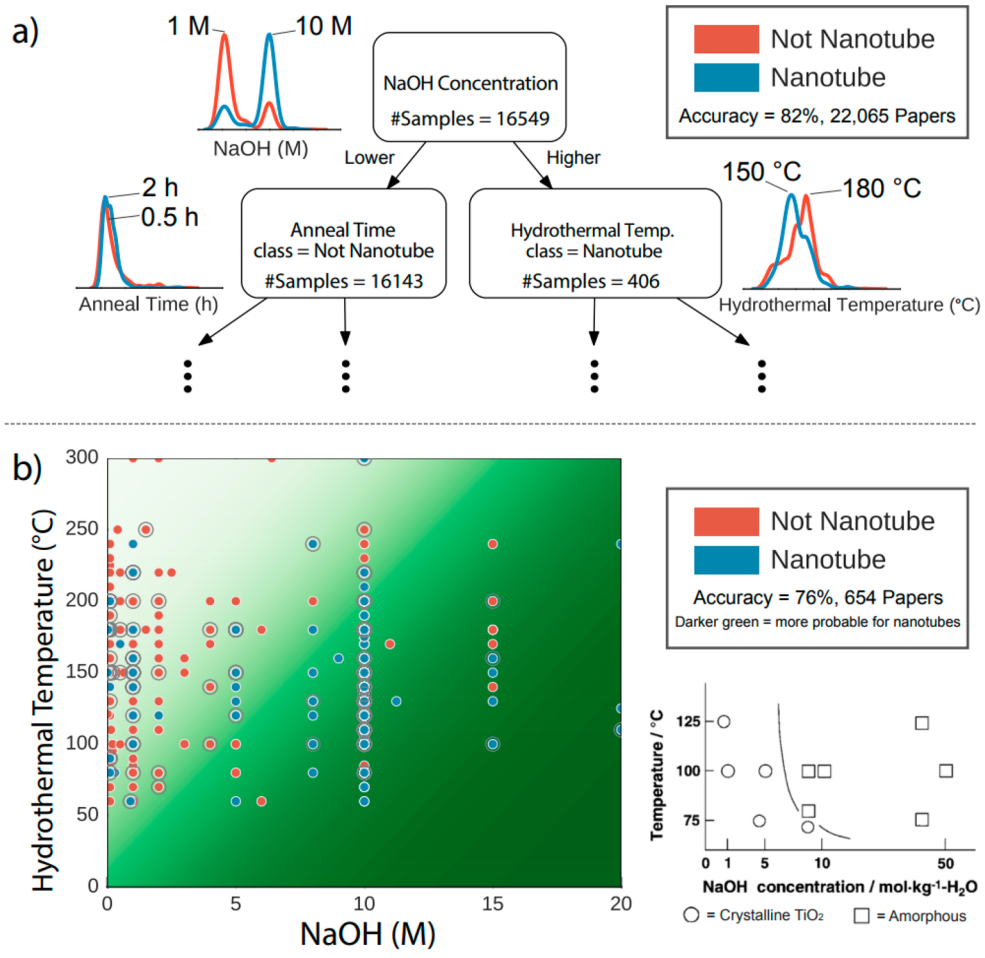


Figure 2. Autonomously learned decision boundaries for titania nanotubes. (a) Overview of decision tree model trained and tested on a total of 22 065 journal articles. Rounded boxes represent decision points in the tree with univariate normalized histograms plotted alongside the decision nodes. The dots represent further levels of the decision tree extending downward, which are included in the Supporting Information. (b) Probabilistic machine-learned decision space overlaid in the parameter-space of hydrothermal temperature and NaOH concentration. Circled data points denote testing points used to compute classification accuracy, and all other points are training points used to learn the decision space. The green gradient in the background denotes machine-learned probabilistic estimates for nanotube formation with darker green corresponding to higher likelihoods. Additionally, an experimentally determined plot of a boundary between hydrothermally produced crystalline titania and amorphous layered sodium titanates is adapted from Tomiha et al.[44]

requires the interdiffusion of multiple metals from the precursors, necessitating higher temperature. The increase in calcination temperature with number of components is clearest for the nanoexamples, where the temperature is kept as low as possible to prevent crystal growth, but mixing of multiple components demands a higher temperature.

Because each of the additive distributions in Figure 1a represents a compilation of many oxides, we comment on some of the key differences among select materials to provide a brief explanation for the location of each peak. Binary oxides, for instance, tend to have calcination temperatures between around 450 and 550 °C. Some of these materials, however, are predominantly synthesized in nanostructured forms, while others are not. For example, we see that alumina (solid line in the binary violin-histogram) has greater representation to the left side of the violin (bulk), while zinc oxide (dashed line) has greater representation to the right side (nano), consistent with the ample use of nanoarchitectures for zinc oxide in applications such as sensors, optoelectronics, and biomedical devices.[40]

For the ternary systems shown in Figure 1a, we see a more varied set of histograms. We contrast barium titanate (solid line) with bismuth ferrite (dashed line). In the ferrite system, the calcination step stabilizes a rhombohedrally distorted perovskite phase within a relatively narrow temperature range because of the tendency to form impurities such as $Bi_2Fe_4O_9$ at higher temperatures.[41] This results in a peak at ~750 °C for bulk materials, whereas particle size control demands lower temperature calcination closer to 600 °C for nanobismuth ferrite. Barium titanate, on the other hand, has a higher and broader calcination temperature range. The synthesis of this material occurs primarily by solid state reaction where precursors of barium carbonate, titania, and others are calcined between 900 and 1100 °C.[42] We compare this to the quaternary lead zirconate titanate (solid line), which must be calcined at temperatures lower than that for $BaTiO_3$ to prevent loss of lead.

Finally, we see distinct bimodal distributions for the multicomponent transition-metal oxides used in batteries, LiMnNi- and LiMnNiCo-oxides, shown in the quaternary and pentanary violins, respectively. The solid state synthesis

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

approach to making these oxides involves calcination at 800−900 °C to crystallize a new phase. Sol−gel and coprecipitation synthesis methods, however, also involve calcination between 400 and 500 °C to decompose organic constituents or to decompose the carbonate into an oxide.[43]

Beyond our analysis of the calcination temperatures used in the synthesis of various oxides, we also investigate calcination times and conditions for hydrothermal reactions. Most hydrothermal reactions, for instance, are carried out between 150 and 200 °C for 12 or 24 h. Such reactions are conducted above room temperature and in the presence of autogenous pressure to increase the solubility and reactivity of the precursors. An upper bound to practically employed reaction temperatures exists, however, due to the critical points of the solvents typically used for such reactions (e.g., water, ethanol). Accordingly, as illustrated in Figure 1b, the hydrothermal reactions used to synthesize both simple and complex oxides occur at similar and only modestly high temperatures but often with fairly long times.

The calcination temperature (occurring at much higher temperatures and shorter times) is typically material-specific and driven by the structural change being sought. For example, binary oxides, largely representing titania, alumina, and zinc oxide, are most often calcined at 400−500 °C for fewer than 5 h. This observed behavior is dominated by TiO_2, given the high frequency with which its synthesis is reported. Calcination is used, in this particular case, to obtain larger grained anatase phase product. More complex oxides must be crystallized at significantly higher temperatures (800−900 °C) and often for more than 5 h for the diffusion reasons explained previously.

To reveal further relations in our database, we use feature selection and classification techniques to identify the key factors that drive synthesis outcomes as well as highly probable values for these reaction parameters. Each synthesis route extracted from a journal article is composed of many attributes and parameters, including the temperatures and times of heating operations, the precursors used, and aspects of the morphologies of the synthesized products. One approach to feature selection is to inspect the probabilistic model learned by a decision tree and automatically select a strongly predictive reduced subset of synthesis parameters which drive the behavior of the model.[33] By training a decision tree across 22 065 journal articles, a hierarchy of single-variable divisions for titania nanotube formation is selected from a pool of 27 synthesis variables (e.g., annealing temperature, drying time). Figure 2a shows an excerpt of the learned decision tree with the nodes nearest to the root node representing the foremost rules learned by the model for maximally separating nanotube and not-nanotube results. The root node in the decision tree splits the data set by NaOH concentration, and the distributions of the data projected onto this univariate axis confirm that the majority of recipes use NaOH at concentrations of either 1 or 10 M. Hydrothermal temperature is also learned as a driving factor for nanotube synthesis, and examination of the temperature distributions shows some difference between two peaks at 150 and 180 °C, although there is a significant amount of overlap.

Examination of the annealing time, on the other hand, clearly shows that the two distributions are not easily separated, suggesting that this lone feature is not as strongly predictive when compared to NaOH concentration or hydrothermal temperature. For this reason, NaOH concentration and hydrothermal temperature are selected as the variables for further analysis based on the construction of the learned decision tree rules (in Figure 2b). We chose exactly two features for further analysis to facilitate a visualizable and easily interpretable two-dimensional diagram.

However, this does come at the necessary cost of some classification accuracy, which can be observed by comparing the accuracies of the full decision tree (82%) and the 2D logistic regression classifier (76%). Although these classification accuracies are not perfect, they are indeed sufficient for the machine learning models to learn an overall correlation between synthesis conditions and the resultant product morphologies.

Figure 2b plots a machine-learned phase diagram in this synthesis parameter-space, using the axes determined by the decision tree. In contrast to diagrams which may plot more direct chemical axes (e.g., free energies, ion activities), we instead restrict our diagram to experimentally accessible (and experimentally reported) axes to facilitate practical synthesis route planning. We note that this reduced 2D parameter space now considers only a particular set of syntheses which report hydrothermal temperatures and NaOH concentrations, and so Figure 2b does not reflect other viable ways of producing titania nanotubes, such as anodization.[45] Recipes using higher NaOH concentrations and lower hydrothermal temperatures, and thus falling in the darker green area, are more likely to produce titania nanotubes, and indeed these darker green regions contain a higher density of nanotube-producing (blue) data points.

This decision rule agrees with the currently accepted mechanism of titania nanotube formation: titania, with the addition of sodium ions, transforms into disordered, layered sodium titanate, and subsequent ion exchange (e.g., via acid washing) induces a rolling effect on the layers, producing titania nanotubes. Accordingly, it is reasonable to expect that a minimum concentration of sodium ions is required to induce sodium titanate (and subsequently nanotube) formation.[44,46−50] Bavykin et al. report that increasing the hydrothermal temperature changes the final product from nanotubes to solid (i.e., nonhollow) fibrous structures, which again agrees with our learned decision boundary.[51] The synthesis condition axes learned automatically by the decision tree also agree with literature findings: the subplot in Figure 2b, reproduced from Tomiha et al.,[44] shows a similar trend for the hydrothermal synthesis of amorphous sodium titanates. By comparison, our machine-learned diagram contains many more data points aggregated across a range of experimental studies and additionally extends the span of the data points along the temperature axis.

Figure 2b reveals a link between two related topics in experimental materials science. Our phase diagram is constructed entirely from titania synthesis journal articles, many of which directly produce titania nanotubes. Tomiha et al.'s phase diagram in the subplot of Figure 2b is adapted from a study which only discusses the synthesis of various alkali-titanates. It makes no mention of nanotube-like morphologies.[44] This type of data-driven analysis can therefore also be used to guide literature review by highlighting correlations and patterns which are only made apparent when many journal articles are examined simultaneously.

Finally, we show the potential for transfer learning by producing synthesis outcome predictions across diverse materials systems using this text-mined synthesis data set. While the previous example focused on hydrothermal reactions,

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

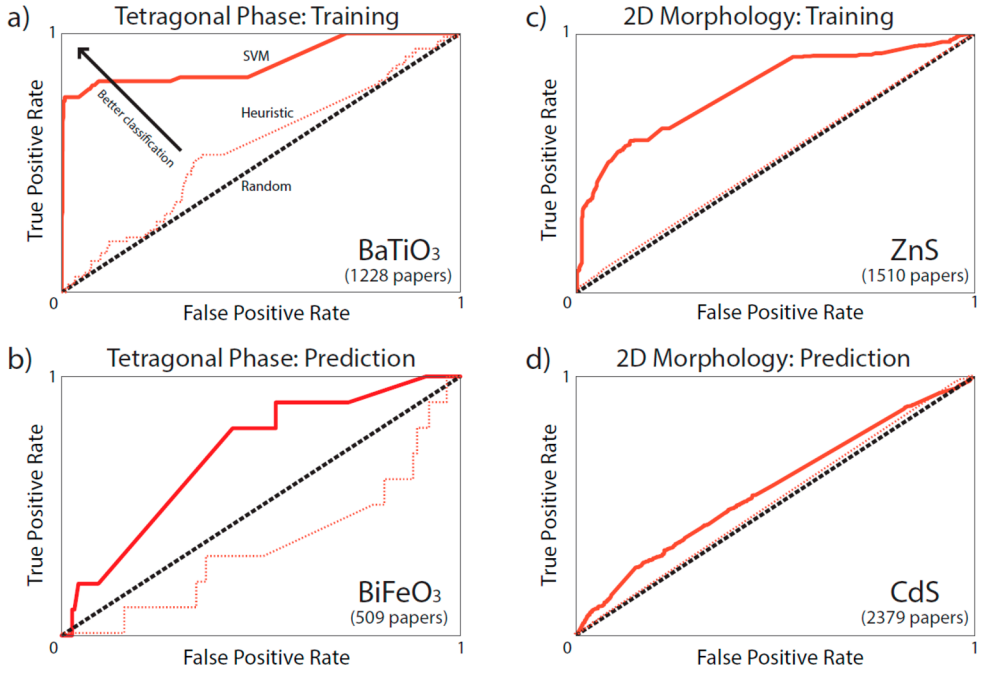


Figure 3. Machine-learned classifiers and predictions across materials systems. (a) Receiver operating characteristic (ROC) curves for tetragonal phase prediction in BaTiO3 on training data. In all four subplots, the three models shown are a Gaussian-kernel support vector machine (SVM) using 20 features (solid red line), a simplified linear heuristic classifier (dotted red line), and a random guessing strategy (black dashed line). Curves closer to the upper-left corner of the diagram represent more accurate classifiers, with the point exactly on the upper-left corner denoting a perfect classifier. (b) ROC curves for tetragonal phase prediction in BiFeO3 on unseen test data. (c) ROC curves for 2D morphology prediction in ZnS on training data. (d) ROC curves for 2D morphology prediction in CdS on unseen test data.

here we examine a set of syntheses which span across a few additional synthesis methods (e.g., hydrothermal, sol−gel, and solid state). We predict tetragonal phase formation (versus all other reported phases) in $BaTiO_3$ and in $BiFeO_3$, which is important because the ferroelectric properties of these materials are deeply linked to polarizations and consequently to the symmetry of the phases.[52,53] Additionally, 2D-like morphology (e.g., nanosheet) predictions are performed on ZnS and CdS (vs non-2D morphologies), as such morphologies have applications in areas ranging from catalysis to data storage.[54] Two-dimensional materials also allow for further tuning of key properties: as an example, confinement effects play a significant role in two-dimensional materials and alter the electronic environment such that band gaps may differ significantly between 3D and 2D morphologies.[55,56]

In each case, we seek a binary classifier which separates the desirable outcome (e.g., symmetry equals tetragonal, or morphology equals 2D) from the other outcomes (e.g., not 2D). In each subplot of Figure 3, the receiver operating characteristic (ROC) curves are shown for three different binary classifiers: a nonlinear Gaussian kernel support vector machine (SVM), a linear heuristic classifier, and a random guessing classifier. Each point on these curves represents the performance of a classifier at a particular decision threshold and the corresponding true and false positive rates. The curves are generated by continuously sweeping through decision threshold values from maximally conservative (never predicting positive) in the lower left corner, to minimally conservative (always predicting positive) in the upper right corner. The upper left corner of these subplots denotes a perfect classification strategy which maximizes true positive results while making no false positive errors. The three classifiers can be compared by examining which ROC curves lie closer to the upper left corner or, equivalently, by comparing the areas under the curves (AUC).

Figures 3a and b show tetragonal phase prediction in $BaTiO_3$ and $BiFeO_3$. Figure 3a shows how well the classifiers can reproduce the $BaTiO_3$ data on which it is trained. Figure 3b then shows the prediction quality of that classifier in another system of $BiFeO_3$. The linear heuristic assumes that tetragonal phase formation can be predicted from synthesis temperatures,[52,53] and so we use a linear classifier which considers only synthesis temperatures as a representative intuition-based rule. While this heuristic strategy outperforms random guessing at training time in $BaTiO_3$, it has no predictive capability on the unseen test data in $BiFeO_3$. Applying a more complete set of general synthesis features, including reaction times, solvent choices, and pH modifiers, while also using a nonlinear SVM to better capture complicated interactions between synthesis parameters yields a far superior result in terms of classification accuracy and consistency between training data and unseen test data. Note that because the classifier is trained only on $BaTiO_3$, it cannot capture the dependence in choice of synthesis parameters on chemistry, and hence, one would not expect it to give highly accurate results. The fact that there is clearly some predictive capability of this classifier in $BaFeO_3$ indicates that there may be some intrinsic aspects of the synthesis conditions that lead to tetragonal phase formation. Hence, more accurate classification results can be expected when training is performed over larger, chemically diverse sets.

In Figures 3c and d, the same comparison between robust nonlinear machine learning and an intuitive heuristic is shown for ZnS training data and CdS unseen test data, where the goal is to predict synthesized 2D morphologies. Two-dimensional materials synthesis is a rapidly developing field where new chemistries are frequently discovered and broadly applicable

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

heuristic synthesis rules are scarce beyond high-level experimental strategies such as etching.[57] Here, the heuristic strategy chosen is pH adjustment, represented by acid and base concentrations used, as a driving factor for predicting the formation of 2D morphologies.[58,59] Evidently, this linear heuristic is not particularly robust and indeed closely resembles random predictions, suggesting that this is a difficult prediction problem where generic synthesis parameters do not hold strong predictive power.

We then apply an SVM, using the same expanded feature set used in Figures 3a and b, and find that it improves performance on both training and unseen test data beyond the heuristic, though the prediction on CdS leaves room for future improvement.

These results indicate that materials systems where robust heuristics are not readily available, machine learning approaches, which jointly consider a broad collection of synthesis parameters, may outperform better-than-heuristic predictions, even those that consider a general set of synthesis descriptors. Further improvements in predictive performance for two-dimensional ZnS and CdS could be achieved by considering domain-specific synthesis parameters such as those related to etching and exfoliation techniques, although this would introduce a trade-off between accuracy and universality.

The examples outlined in Figure 3 demonstrate the potential applicability of this text-mining and machine learning framework across varied materials systems. Additionally, we find that our current predictive performance is achieved at a level comparable to that of recent algorithms used in predicting DFT-computed formation energies and similarly outperforms existing heuristic techniques.[60] The accuracies of the models and features used for the classifiers in Figure 3 are provided in Supplementary Table S1.

In this work, we presented a framework to automatically construct a large-scale materials synthesis database. This framework enables us to explore the materials synthesis conditions that produce selected materials properties at large scales. We use text analysis and extraction to populate a database of materials synthesis parameters drawn from tens of thousands of previously published journal articles. From this, we examine the correlations in synthesis conditions and materials properties across many papers at once. As it would have required researchers to read through papers and manually enter relevant information, such analysis was previously rather inefficient. Furthermore, by applying relevant machine learning tools, we are able to identify and analyze the specific synthesis recipe features that produce desired materials properties in both an automated fashion and without requiring prior or a priori knowledge of the system of interest.

The examples shown in this work represent a sample of what is possible with this approach, and continued exploration, expansion, and data set refinement are planned by the authors. Further work to improve this methodology will include consideration of newly published papers and disproportionate impact of highly cited papers to better complement existing synthesis planning techniques used by theorists and experimentalists. While this work has data-mined the influences of using certain solvents or temperatures (focused on a limited set of synthesis methods), future work will examine additional methods of synthesis, including thin film depositions, catalyst-driven reactions, and noncrystalline materials synthesis. Text extraction for these other synthesis approaches would likely change the focus of what subparts of extraction accuracy are more critical. For example, in the case of thin film deposition, likely equipment settings become essential to extract accurately.

We present our database of extracted synthesis parameters and data mined insights as a publicly accessible Web site, named the Synthesis Project, which complements existing computational materials property databases (www.synthesisproject.org).[2] By providing researchers with a toolkit for exploring and understanding text-mined and machine-learned synthesis data, the Synthesis Project will act as a further catalyst for data-driven materials screening and development for the community at large.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.chemmater.7b03500.

Machine learning model details and train/test data splits; decision tree parameters and layout; full features and classifier performances for models built in Figure 3; and individual histograms for materials shown in Figure 1 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: elsao@mit.edu.

ORCID

Edward Kim: 0000-0002-0781-5531

Elsa Olivetti: 0000-0002-8043-2385

Author Contributions

E.K. wrote the data processing and machine learning algorithms. K.H. led data mining and annotation efforts. A.S. wrote text conversion code. A.M. provided guidance in text extraction techniques. G.C. provided guidance in computational materials science perspectives. G.C. and E.O. developed the concept and structure of this work. E.O., E.K., G.C., and K.H. wrote the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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 N00014-16-1-2432, and the MIT Energy Initiative. Early work was collaborative under the Department of Energy's Basic Energy Science Program through the Materials Project under Grant EDCBEE. 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 who provided the substantial content required for our analysis, and research input from Yan Wang, Daniil Kitchaev, Wenhao Sun, Olga Kononova, Emma Strubell, Craig Greenberg, Rachel Osmundsen, Vicky Gong, Sara Matthews, and Alex Tomala.

■ REFERENCES

(1) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *Nat. Publ. Gr.* **2015**, *1*, 15010.
(2) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 11002.
(3) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sanchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.
(4) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory. *Chem. Mater.* **2010**, *22*, 3762−3767.
(5) Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory Using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304−1311.
(6) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191−201.
(7) Kalinin, S. V.; Sumpter, B. G.; Archibald, R. K. Big−deep−smart Data in Imaging for Guiding Materials Design. *Nat. Mater.* **2015**, *14*, 973−980.
(8) Hattrick-Simpers, J.; Wen, C.; Lauterbach, J. The Materials Super Highway: Integrating High-Throughput Experimentation into Mapping the Catalysis Materials Genome. Catal. *Catal. Lett.* **2015**, *145*, 290−298.
(9) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials Science with Large-Scale Data and Informatics: Unlocking New Opportunities. *MRS Bull.* **2016**, *41*, 399−409.
(10) Jain, A.; Shin, Y.; Persson, K. A. Computational Predictions of Energy Materials Using Density Functional Theory. *Nat. Rev. Mater.* **2016**, *1*, 15004.
(11) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
(12) Ghadbeigi, L.; Harada, J. K.; Lettiere, B. R.; Sparks, T. D. Performance and Resource Considerations of Li-Ion Battery Electrode Materials. *Energy Environ. Sci.* **2015**, *8*, 1640−1650.
(13) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324−7331.
(14) Pankajakshan, P.; Sanyal, S.; De Noord, O. E.; Bhattacharya, I.; Bhattacharyya, A.; Waghmare, U. Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chem. Mater.* **2017**, *29*, 4190−4201.
(15) White, A. The Materials Genome Initiative: One Year on. *MRS Bull.* **2012**, *37*, 715−716.
(16) Rajan, K. Materials Informatics: The Materials "Gene" and Big Data. *Annu. Rev. Mater. Res.* **2015**, *45*, 153−169.
(17) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A. A.; Chae, H. S. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I. I.; Baldo, M.; Adams, R. P. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120−1127.
(18) Holdren, J. P. *Materials Genome Initiative for Global Competitiveness*; Natl. Sci. Technol. Counc. OSTP: Washington, United States, 2011.
(19) Lin, L.-C.; Berger, A. H.; Martin, R. L.; Kim, J.; Swisher, J. A.; Jariwala, K.; Rycroft, C. H.; Bhown, A. S.; Deem, M. W.; Haranczyk, M.; Smit, B. In Silico Screening of Carbon-Capture Materials. *Nat. Mater.* **2012**, *11*, 633−641.
(20) Haldoupis, E.; Nair, S.; Sholl, D. S. Finding MOFs for Highly Selective CO_2/N_2 Adsorption Using Materials Screening Based on Efficient Assignment of Atomic Point Charges. *J. Am. Chem. Soc.* **2012**, *134*, 4313−4323.
(21) Lammey, R. CrossRef Text and Data Mining Services. *Insights* **2015**, *28*, 62−68.
(22) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *J. Cheminf.* **2011**, *3*, 1−13.
(23) Rocktäschel, T.; Weidlich, M.; Leser, U. ChemSpot: A Hybrid System for Chemical Named Entity Recognition. *Bioinformatics* **2012**, *28*, 1633−1640.
(24) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894−1904.
(25) Honnibal, M.; Johnson, M. An Improved Non-Monotonic Transition System for Dependency Parsing. *Emnlp* **2015**, 1373−1378.
(26) Jones, D. E.; Igo, S.; Hurdle, J.; Facelli, J. C. Automatic Extraction of Nanoparticle Properties Using Natural Language Processing: NanoSifter an Application to Acquire PAMAM Dendrimer Properties. *PLoS One* **2014**, *9*, e83932.
(27) Friedman, C.; Kra, P.; Yu, H.; Krauthammer, M.; Rzhetsky, A. GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics* **2001**, *17*, S74−82.
(28) Lewinski, N. A.; McInnes, B. T. Using Natural Language Processing Techniques to Inform Research on Nanotechnology. *Beilstein J. Nanotechnol.* **2015**, *6*, 1439−1449.
(29) Tchoua, R. B.; Chard, K.; Audus, D.; Qin, J.; De Pablo, J.; Foster, I. A Hybrid Human-Computer Approach to the Extraction of Scientific Facts from the Literature. *Procedia Comput. Sci.* **2016**, *80*, 386−397.
(30) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011**, *13*, 579−633.
(31) 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.* **2011**, *59*, 630−639.
(32) Jansen, M. The Energy Landscape Concept and Its Implications for Synthesis Planning. *Pure Appl. Chem.* **2014**, *86*, 883−898.
(33) Raccuglia, P.; Elbert, K. C.; Adler, P.; Falk, C.; Wenny, M.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73−76.
(34) Gaultois, M. W.; Sparks, T. D.; Borg, C. K. H.; Seshadri, R.; Bonificio, W. D.; Clarke, D. R. Data-Driven Review of Thermoelectric Materials: Performance and Resource Onsiderations. *Chem. Mater.* **2013**, *25*, 2911−2920.
(35) Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-Learned and Codified Synthesis Parameters of Oxide Materials. Sci. *Sci. Data* **2017**, *4*, 170127.
(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*, 2825−2830.
(37) Mikolov, T.; Corrado, G.; Chen, K.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *Proc. Int. Conf. Learn. Represent. (ICLR 2013)* **2013**, 1−12.
(38) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−1213.

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX

(39) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X.; Brain, G. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*; 2016; pp 265−284.
(40) Wang, Z. L. Zinc Oxide Nanostructures: Growth, Properties and Applications. *J. Phys.: Condens. Matter* **2004**, *16*, R829−R858.
(41) Lebeugle, D.; Colson, D.; Forget, A.; Viret, M.; Bonville, P.; Marucco, J. F.; Fusil, S. Room-Temperature Coexistence of Large Electric Polarization and Magnetic Order in $BiFeO_3$ Single Crystals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 1−8.
(42) Eckert, J. O.; Hung-Houston, C. C.; Gersten, B. L.; Lencka, M. M.; Riman, R. E. Kinetics and Mechanisms of Hydrothermal Synthesis of Barium Titanate. *J. Am. Ceram. Soc.* **1996**, *79*, 2929−2939.
(43) Ammundsen, B.; Paulsen, J. Novel Lithium-Ion Cathode Materials Based on Layered Manganese Oxides. *Adv. Mater.* **2001**, *13*, 943−956.
(44) Tomiha, M.; Masaki, N.; Uchida, S.; Sato, T. Hydrothermal Synthesis of Alkali Titanates from Nano Size Titania Powder. *J. Mater. Sci.* **2002**, *37*, 2341−2344.
(45) Mor, G. K.; Shankar, K.; Paulose, M.; Varghese, O. K.; Grimes, C. A. Enhanced Photocleavage of Water Using Titania Nanotube Arrays. *Nano Lett.* **2005**, *5*, 191−195.
(46) Roy, P.; Berger, S.; Schmuki, P. TiO2 Nanotubes: Synthesis and Applications. *Angew. Chem., Int. Ed.* **2011**, *50*, 2904−2939.
(47) Yu, J.; Su, Y.; Cheng, B.; Zhou, M. Effects of pH on the Microstructures and Photocatalytic Activity of Mesoporous Nanocrystalline Titania Powders Prepared via Hydrothermal Method. *J. Mol. Catal. A: Chem.* **2006**, *258*, 104−112.
(48) Zhao, B.; Lin, L.; He, D. Phase and Morphological Transitions of Titania/titanate Nanostructures from an Acid to an Alkali Hydrothermal Environment. *J. Mater. Chem. A* **2013**, *1*, 1659−1668.
(49) Yang, J.; Jin, Z.; Wang, X.; Li, W.; Zhang, J.; Zhang, S.; Guo, X.; Zhang, Z. Study on Composition, Structure and Formation Process of Nanotube $Na_2Ti_2O_4(OH)_2$. *Dalt. Trans.* **2003**, *4*, 3898.
(50) Tsai, C. C.; Teng, H. Structural Features of Nanotubes Synthesized from NaOH Treatment on TiO_2 with Different Post-Treatments. *Chem. Mater.* **2006**, *18* (2), 367−373.
(51) Bavykin, D. V.; Parmon, V. N.; Lapkin, A. A.; Walsh, F. C. The Effect of Hydrothermal Conditions on the Mesoporous Structure of TiO_2 Nanotubes. *J. Mater. Chem.* **2004**, *14*, 3370.
(52) Zhang, Q.; Cagin, T.; Goddard, W. a. The Ferroelectric and Cubic Phases in $BaTiO_3$ Ferroelectrics Are Also Antiferroelectric. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 14695−14700.
(53) Siemons, W.; Biegalski, M. D.; Nam, J. H.; Christen, H. M. Temperature-Driven Structural Phase Transition in Tetragonal-like $BiFeO_3$. *Appl. Phys. Express* **2011**, *4*, 1−4.
(54) Butler, S. Z.; Hollen, S. M.; Cao, L.; Cui, Y.; Gupta, J. a.; Gutierrez, H. R.; Heinz, T. F.; Hong, S. S.; Huang, J.; Ismach, A. F.; Johnston-Halperin, E.; Kuno, M.; Plashnitsa, V. V.; Robinson, R. D.; Ruoff, R. S.; Salahuddin, S.; Shan, J.; Shi, L.; Spencer, M. G.; Terrones, M.; Windl, W.; Goldberger, J. E. Progress, Challenges, and Opportunities in Two-Dimensional Materials Beyond Graphene. *ACS Nano* **2013**, *7*, 2898−2926.
(55) Chhowalla, M.; Shin, H. S.; Eda, G.; Li, L.-J.; Loh, K. P.; Zhang, H. The Chemistry of Two-Dimensional Layered Transition Metal Dichalcogenide Nanosheets. *Nat. Chem.* **2013**, *5*, 263−275.
(56) Miró, P.; Audiffred, M.; Heine, T. An Atlas of Two-Dimensional Materials. *Chem. Soc. Rev.* **2014**, *43*, 6537−6554.
(57) Anasori, B.; Lukatskaya, M. R.; Gogotsi, Y. 2D Metal Carbides and Nitrides (MXenes) for Energy Storage. *Nat. Rev. Mater.* **2017**, *2*, 16098.
(58) Xu, Y.; Zhao, W.; Xu, R.; Shi, Y.; Zhang, B. Synthesis of Ultrathin CdS Nanosheets as Efficient Visible-Light-Driven Water Splitting Photocatalysts for Hydrogen Evolution. *Chem. Commun. (Cambridge, U. K.)* **2013**, *49*, 9803−9805.
(59) Ben Nasr, T.; Kamoun, N.; Kanzari, M.; Bennaceur, R. Effect of pH on the Properties of ZnS Thin Films Grown by Chemical Bath Deposition. *Thin Solid Films* **2006**, *500*, 4−8.
(60) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 94104.

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. XXXX, XXX, XXX−XXX