ACS central science

Research Article

Cite This: ACS Cent. Sci. 2019, 5, 892−899

http://pubs.acs.org/journal/acscii

A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction

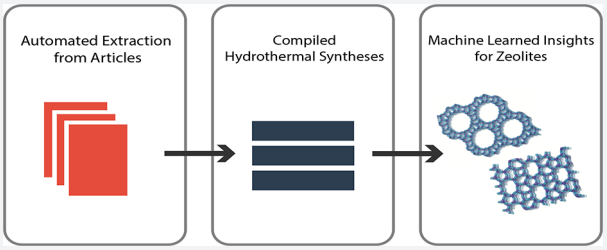
Zach Jensen,[†] Edward Kim,[†] Soonhyoung Kwon,[‡] Terry Z. H. Gani,[‡] Yuriy Román-Leshkov,[‡] Manuel Moliner,[§] Avelino Corma,[§] and Elsa Olivetti*[,†]

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

[§]Instituto de Tecnología Química, Universitat Politècnica de València-Consejo Superior de Investigaciones Científicas, Avenida de los Naranjos s/n, 46022 Valencia, Spain

Supporting Information

ABSTRACT: Zeolites are porous, aluminosilicate materials with many industrial and "green" applications. Despite their industrial relevance, many aspects of zeolite synthesis remain poorly understood requiring costly trial and error synthesis. In this paper, we create natural language processing techniques and text markup parsing tools to automatically extract synthesis information and trends from zeolite journal articles. We further engineer a data set of germanium-containing zeolites to test the accuracy of the extracted data and to discover potential opportunities for zeolites containing germanium. We also create a regression model for a zeolite's framework density from the synthesis conditions. This model has a cross-validated root mean squared error of 0.98 T/1000 Å^3, and many of the model decision boundaries correspond to known synthesis heuristics in germanium-containing zeolites. We propose that this automatic data extraction can be applied to many different problems in zeolite synthesis and enable novel zeolite morphologies.



INTRODUCTION

Zeolites are microporous, crystalline aluminosilicate materials with a wide range of applications including catalysis, adsorption, separation, and ion exchange.[1,2] Beyond their use as Brønsted acid catalysts in the chemical and petroleum industries,[2−4] zeolites have been utilized for several important environmental improvement and renewable energy applications including biomass conversion, CO_2 capture, NO_x abatement, and water purification.[5] Notably, the topochemical features (i.e., pore structure, framework type, and heteroatom composition) often determine the performance of the zeolite.[6,7] As such, recent efforts in the community have focused on developing rational design strategies to engineer zeolites for targeted applications, such as designing a pore geometry that mimics the transition state of the specific reaction or crystallizing a framework with structural chirality.[8,9]

Zeolite crystallization often occurs through a hydrothermal synthesis pathway governed by a large synthesis parameter space and complex crystallization kinetics that yield metastable structures.[10] In a typical zeolite synthesis, sources of SiO_2, Al_2O_3, and a mineralizing agent (e.g., a source of OH^- or F^- anions) are mixed with water to form an aluminosilicate gel. In addition, inorganic cations or organic structure directing agent (OSDA) molecules are added to direct the formation of the zeolite structure. This gel is aged, reacted, and then crystallized under hydrothermal conditions. The composition of the gel, traditionally parametrized using molar ratios (e.g., OSDA/Si or H_2O/Si), and the synthesis conditions determine the outcome of the crystallization process. Because of these complexities, zeolite synthesis−structure relationships are difficult to understand. Several studies have advanced this understanding;[11−14] however, global methodologies for predicting new zeolite structures from synthesis parameters are still limited. As a result, the synthesis of novel zeolite structures requires a semiempirical process governed mostly by domain heuristics acquired through experience.

The lack of predictive ability to design synthesis routes for zeolites is a major bottleneck for discovering new zeolite structures.[15,16] Using first-principles approaches, researchers have estimated that several million unique zeolite structures are energetically favorable.[17−19] However, currently only 245 zeolites have been synthesized,[20] and far fewer are commercially available.[21] This presents a particular opportunity considering that the global market for zeolite-driven commercial processes exceeds 2 million metric tons per year.[22] This massive gap existing between theoretical and synthetically confirmed structures (also important in crystallization more generally) demonstrates the need for new, cutting-edge approaches to zeolite synthesis.[23]

Data-driven synthesis approaches have found success in a number of domains, including organic[24−27] and inorganic[28−30] materials synthesis. These approaches have the potential to accelerate the development of new materials, as experts can

Received: February 26, 2019
Published: April 19, 2019

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

learn new, complex relationships from existing data resources using visualization and automated data mining algorithms, as well as build fast predictive models coupled to experimental validation.[31] Along with the need for significant volumes of data, a critical aspect of accurate, data-driven models is the inclusion of negative examples,[32,33] for example, synthesis routes that did not yield the desired product. Unlike many other materials science domains, the zeolite community often includes failed syntheses (i.e., amorphous and dense phases) in publications and data sets, thus making the data-driven study of zeolites very promising.

Several zeolite studies have found success using data science to predict the zeolite framework type from crystallographic data[34,35] and modeling the mechanical properties of zeolites.[36] However, only a handful of reports exist that successfully model relationships between synthesis parameters and the resulting structure.[37,38] These studies relied on high-throughput synthesis methods to generate data used to model synthesis parameters. Even with synthesis methods designed for rapid sample generation, each generated less than 150 synthesis routes, thereby limiting the analysis to only a subset of zeolite structures.

Global data-driven zeolite synthesis approaches will require large amounts of data. Given that the field of zeolite synthesis has been very active in both the academic and industrial communities for more than 60 years, one rich source of abundant data is directly from scientific journal articles and patents.[39] However, it is impractical to manually extract data from more than a few hundred publications.[32,40,41] Automatic data extraction from materials science and chemistry text using natural language processing (NLP) techniques greatly increases the amount of available data.[42] Several NLP tools and software pipelines have been developed for automatic data extraction from scientific journal articles.[42−44] These pipelines have been used to extract material property and synthesis information from several different material domains including Curie and Néel temperatures for magnetic materials,[45] synthesis conditions of titania,[39,46] and screening of potential novel perovskite materials.[47] Indeed, this automatic extraction can be applied to capture all published, available zeolite synthesis data into a single data set allowing global comparisons between all types of zeolite structures, but the necessary tools to do so have not been developed.

In this paper, we present an automatic data extraction pipeline to study the crystallization of zeolite structures and suggest ways in which machine learning (ML) can be used to predict synthesis pathways for new zeolite structures. We create tools to automatically extract zeolite synthesis and topology data from multiple locations within a journal article, including tables, captions, and footnotes along with body text, thus greatly extending our previous method developed for metal oxides.[42] We demonstrate the accuracy and usefulness of our extracted data by examining trends in both a global zeolite set and a focused subset comprising germanium-containing zeolites. The latter data set is used to elucidate specific synthesis trends, where a random forest regression model allowed prediction of the framework density of synthesized zeolites. This model moves toward predicting new zeolite topologies from synthesis data.

RESULTS AND DISCUSSION

Zeolite Data Extraction. From our database of 2.5 million journal articles, we filtered down to a set of 70 000 papers relevant to zeolite synthesis through text matching specific zeolite keywords. The papers were processed through a pipeline that consists of extracting precursor information from the text of the paper with NLP algorithms (see Kim et al.[42] for additional information), applying HTML and XML parsing on the relevant synthesis tables, and using Regular Expressions (regex) to locate and extract compositional ratios. The extracted data were combined to reveal trends and train ML algorithms that could be used to gain insight into the effect of different synthesis variables. Figure 1 presents a schematic

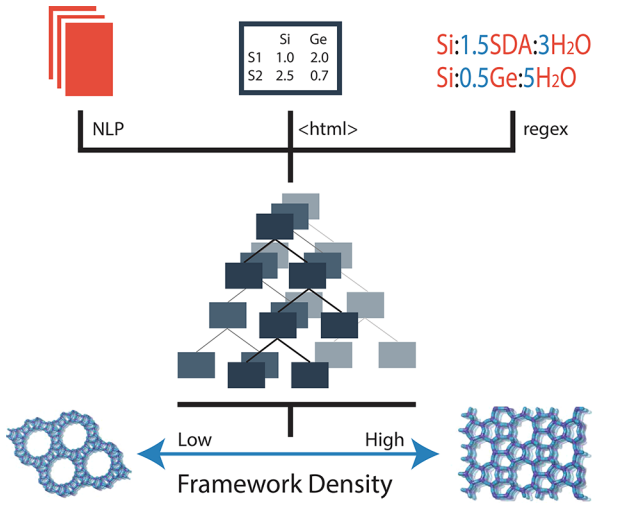


Figure 1. Schematic overview of zeolite data engineering including (1) literature extraction from sources such as NLP from body text, parsing of html tables, and regex matching between text and tables, (2) regression modeling, and (3) zeolite structure prediction.

representation of this data pipeline of extracting and combining zeolite synthesis data. The figure depicts our process to obtain information from multiple aspects of a journal article (including text and table data) and use these data to inform zeolite synthesis through prediction of a structural property such as the framework density, defined as the number of T atoms (Si, Al, Ge, etc.) per 1000 Å^3, which is one of the simplest metrics used to distinguish zeo-type porous materials.

In a typical journal manuscript, zeolite synthesis information in the form of molar ratios and crystallization conditions is often scattered throughout tables, figures, and text within the main, supporting, and methods sections, each requiring a specialized extraction technique. Prior to our work, techniques capable of accurately extracting and correlating data from both tables and text in a journal article had not been developed. For tables, our software extracted information from HTML files, accounting for variation in both difference in HTML implementation and design of the actual table. Tables were converted into data mineable JSON file formats that are both human- and machine-readable. Next, we used our NLP pipeline to locate the target zeolite, type of OSDA, and missing crystallization conditions within the body text of the paper. Finally, the data were featurized into a fixed set of zeolite-relevant synthesis features (such as structural data from the International Zeolite Association (IZA) database) suitable for data analytics and ML (see methods section for extraction and data engineering details).

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

Gel composition is a critical variable in determining the resulting zeolite topology for a synthesis route.[48,49] Using our table extraction software, we extracted gel composition data from the synthesis tables found in our set of 70 000 zeolite papers to identify trends among the synthesis variables and the products of zeolite synthesis. Figure 2 shows these data plotted as pairwise relationships between several of the compositional features.

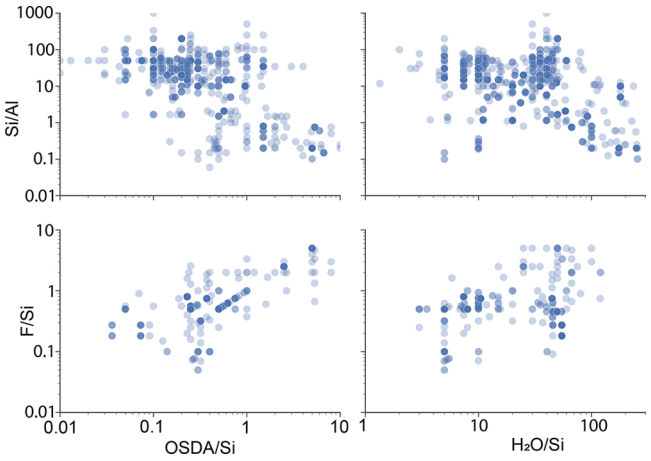


Figure 2. Pairwise plot of gel composition data automatically extracted from zeolite tables.

Zeolites are traditionally synthesized with theoretical molar ratios of Si/Al > 1, OSDA/Si < 1, H_2O/Si < 100, and F/Si < 1. However, the data extracted by our pipeline represented in Figure 2 clearly show that these ranges can be exceeded. This effect is rationalized based on the specific conditions required for the synthesis of related zeotypes, such as silicoaluminophosphates (SAPOs) and Ge-rich silicogermanates.[50] The SAPO framework is formed by alternating tetrahedrally coordinated Al and P atoms, with a few of these heteroatoms isomorphically substituted by Si. Consequently, both SAPOs and Ge-rich molecular sieves have low Si contents, resulting in Si-normalized molar ratios beyond the classical values for typical high-silica zeolites.

Although it is difficult to extract complex synthetic relationships and predictions from the simple compositional features shown in Figure 2, the data obtained from our pipeline can be used to validate general trends in zeolite synthesis. For example, a positive linear trend between the quantity of fluoride ions and OSDA molecules is observed (see bottom-left panel in Figure 2). Fluoride is used as a mineralizer in zeolite synthesis,[51] often resulting in zeolites with a lower concentration of defects by providing more negative species to counterbalance the positive charge of the OSDA cations.[52] These fluoride-based routes are often performed close to neutral pH values, and, consequently, researchers tend to add similar molar amounts of fluoride and OSDA cations to the synthesis,[53−55] which is reflected in the trend we see in Figure 2. Taken together, these data show that the automated extraction algorithms are capable of isolating compositional information from the literature in a reliable fashion, thus allowing us to perform more in-depth analyses of the zeolite synthesis space (vide infra).

Analysis of Germanium-Containing Zeolites. Germanium addition into zeolite framework sites is responsible for the synthesis of many new zeolite structures over the past two decades.[56] Motivated by this success, we constructed a germanium-containing zeolite data set with our automated extraction pipeline. These data enabled us to verify the accuracy of our extracted data against known trends between synthesis variables and structures by providing a concise data set in a zeolite subdomain with a large amount of heuristic synthesis knowledge developed by the community. Besides verification of the data extraction, we also identified potentially interesting areas within the germanium zeolite system that can be explored further with ML and experimental techniques.

Using germanium keyword text matching, we condensed our zeolite data into a set of 238 papers discussing the impact of germanium on zeolite synthesis. Using our automated data extraction pipeline and manually adding data from the supplemental sections of these papers, we created a data set of 1638 unique synthesis routes, an excerpt of which is shown in Table 1. Of these, 1214 synthesis routes successfully result in the creation of a zeolite or germanate, while the remainder result in either a dense crystal or amorphous material. The data contained compositional variables (i.e., Si, Ge, Al, B, alkali cations, H_2O, F^-, and OSDA amounts), conditional variables (e.g., crystallization time and temperature), the type of OSDA used in the synthesis, and the products formed all of which are extracted automatically and manually checked to ensure accuracy. The latter were featurized further with structural information extracted from the IZA Web site (e.g., framework density, secondary building units, and composite building blocks). Note that the OH^-/Si molar ratio could be obtained by a simple postextraction data refining process.

Figure 3a shows the wide range of structural variability in Ge-containing zeolites with medium-, large-, and extra-large pore materials spanning framework densities from 7.5 to 19 T atoms/1000 Å^3. Indeed, the inclusion of Ge, which is an element with a larger nonbonding radius compared to Si and capable of forming smaller OTO angles into the framework of silicates results in the stabilization of small-ring secondary building units (SBUs), including double four-membered rings (D4R), three-membered rings (3MR), and double three-membered rings (D3R).[16,62] The presence of these units gives rise to zeolite topologies with low tetrahedral site densities and large pores. While the use of Ge to stabilize small-ring SBUs is

Table 1. Excerpt of the Data Set of Germanium-Containing Zeolites[a]

Si/Ge	Si/H_2O	Si/F^-	OSDA	product	reference
4	0.08	1.6	1,2-dimethyl-3-(3-methylbenzyl)imidazolium	CIT-13	57
30	0.19	1.9	hexamethonium	ITQ-13	58
2	0.67	2.7	benzyltriethylammonium	ITQ-44	59
1	0.1	1	1-methyl-3-(2′-methylbenzyl)imidazolium	NUD-2	60
7.5	0.13	1.76	pentamethyldiethylenetriamine	amorph	61

[a]The full data set is available online (see Supporting Information).

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

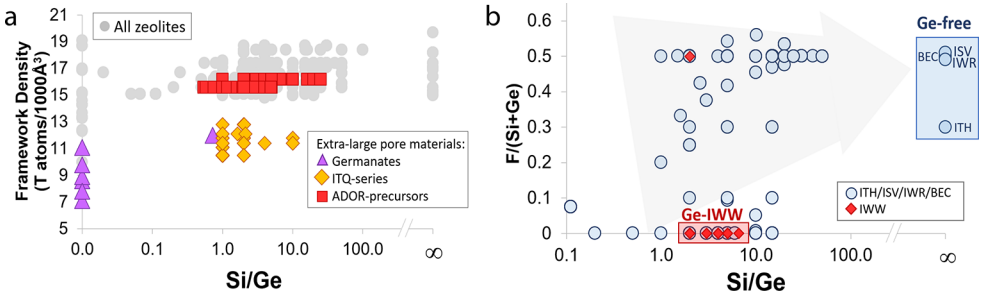


Figure 3. Germanium-containing zeolite data extracted with our pipeline. (a) Framework density clusters corresponding to different classes of germanium-containing zeolites. (b) Trade-off between Ge content and the amount of F^- ions required to stabilize different zeolites. The three letter codes refer to specific zeolite framework structures defined by the IZA. ADOR is an interzeolite transformation synthesis method.[73]

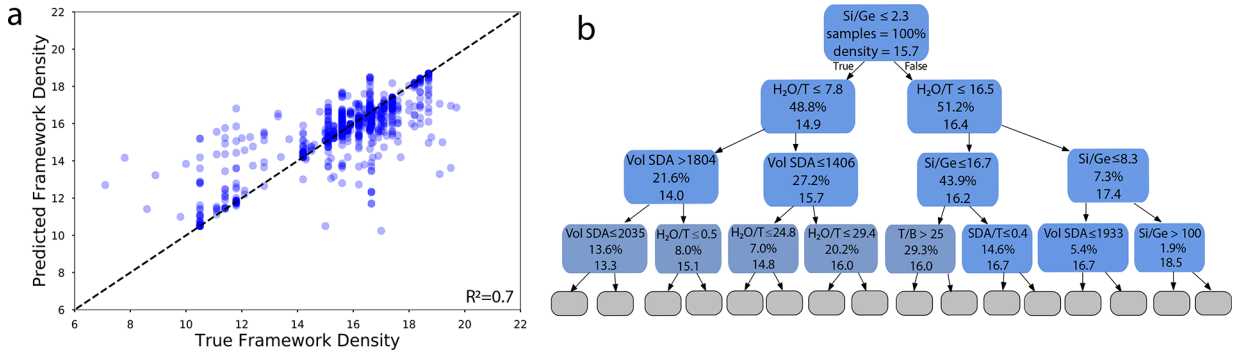


Figure 4. Random forest regression model predicting zeolite framework density from synthesis conditions. (a) Cross-validation results for the random forest model showing the actual experimental vs model predicted values for framework density. (b) A single decision tree regression model trained to predict framework density. Samples values correspond to the percentage of data passing through a node. Density refers to the average framework density value passing through each node. Vol SDA = the volume of the OSDA.

a known effect, the visual representation of all the data extracted with our pipeline gives rise to new insights and trends that were not previously clear. For example, extra-large pore structures are clustered in three areas corresponding to low, intermediate, and high framework densities (see purple triangles, yellow diamonds, and red squares, respectively, in Figure 3a). Further analysis revealed that materials with framework densities less than 10 T atoms/1000 Å^3 correspond to pure nonzeolitic germanates (see germanates in Figure 3a), while materials with densities ranging between 11 and 14 T atoms/1000 Å^3 correspond to topologies with some of the largest pores reported to date including ITQ-33 (18 MR × 12 MR × 12 MR)[16] and ITQ-44 (18 MR × 12 MR × 12 MR)[63] that have only been obtained with Si/Ge less than 4 (see ITQ-series in Figure 3a). Lastly, extra-large pore materials with narrow framework densities ranging from 15.5 to 16.5 T atoms/1000 Å^3 correspond to crystalline structures, including UTL and CTH, where Ge is placed within the D4R units spacing the siliceous layers (see Assembly-Disassembly-Organization-Reassembly (ADOR)-precursors in Figure 3a).[57,64] This feature has been exploited to access new topologies by disassembling the interlayer Ge−O bonds and reorganizing into a new structure (i.e., the ADOR method).[65,66]

Figure 3b depicts the close relationship between Ge and fluoride (F^-) ion contents. The stabilization of small-ring SBUs requires either the presence of Ge as a heteroatom with smaller OTO angles or F^- as a small structure-directing agent that fits within the SBU.[67] Our data clearly reveal that there is a trade-off between Ge content and the amount of F^- ions required to stabilize a particular structure in agreement with well-established synthesis tenets. Thus, zeolites containing large amounts of Ge can be synthesized with simple OSDAs and small amounts, or even in the absence, of F^-, but these structures will not have high hydrothermal stability. For example, polymorph C of Beta (BEC) and IWR zeolites can be synthesized with Si/Ge ratios below 5 using simple OSDA molecules, such as, tetraethylammonium or hexamethonium, under F^- free conditions.[68,69] In contrast, synthesizing more hydrothermally stable zeolites with the same topology that have less Ge content always requires the use of F^- ions (see Figure 3b), in combination with more specific OSDAs, such as large organic molecules synthesized via the Diels−Alder cycloaddition of bulky addends.[70,71] Importantly, visualization of the data obtained with our extraction tool provides new insights by identifying areas of interest for future study. For example, Figure 3b reveals that there exist several cases for Ge-containing zeolites, including ITQ-22 (IWW, see Ge-IWW in Figure 3b), for which an OSDA has not been discovered to crystallize a Ge-free high-silica analogue.[72] We surmise that our data extraction tool combined with ML approaches will be essential to predict the required physicochemical properties to design such OSDAs. This is currently a main research topic in our laboratories.

Germanium Zeolite Framework Density Prediction. Finally, we combined our extracted data with ML algorithms to model the structural properties of a zeolite for a given set of synthesis parameters. While the previous examples verified our extracted data through simple trends, here we aimed to discover less intuitive, more complicated relationships between the synthesis parameters with the ultimate goal of potentially unearthing synthesis routes for new zeolite structures.

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

Specifically, we modeled framework density as a regression problem using a random forest ensemble method (see Methods). In Figure 4a, we evaluated the fivefold cross validation accuracy of the model, where the color hue corresponds to the frequency of data points. The root mean squared error (RMSE) is 0.98 T/1000 Å^3 compared with the standard deviation of framework density in our data, which is 1.76 T/1000 Å^3. The RMSE and the *r*-squared values indicate our model begins to map synthesis conditions to the resulting structure's framework density allowing predictions of synthesis conditions for novel zeolite with both high and low framework densities.

Besides the ability to accurately map synthesis conditions to a zeolite's framework density, an additional benefit of using decision trees to model zeolite synthesis is human interpretability. In Figure 4b, we compared a single decision tree machine learned regression model trained on the data to known synthesis pathways for zeolites with various framework densities. Following the different nodes of this decision tree, it is possible to predict the framework density of the potentially achieved zeolite depending on the synthesis parameters employed (lower framework densities are ordered toward the left side of the tree). The first nodes embrace the more influencing parameters on the target variable (in this case is the framework density of the zeolite). As seen in Figure 4b, the Si/Ge molar ratio, the H_2O/T molar ratio, and the volume of the OSDA, in this particular order, are the more determinant variables to predict the zeolite framework densities of the Ge-containing zeolites. As a simple validation, we note that most of the Ge-containing zeolites featuring a very low framework density reported in the open literature require Si/Ge molar ratios of 1−2, very concentrated gels with H_2O/T less than 5, and bulky OSDA molecules, all parameters that are in good agreement with the variables and their corresponding values presented in our decision tree.[74,75] While some of these heuristics might be evident to an expert in the field of zeolite synthesis, this example represents the first instance of a machine learned decision guideline for zeolites generated from automatically extracted literature synthesis data.

The models in Figure 4 demonstrate the potential of ML for predicting zeolite structural information from synthesis parameters. While not directly related to catalytic performance, predicting framework density represents an important step in tailoring synthesis conditions for zeolites. Combined with models for ring geometry and active-site chemistry, we will continue to progress toward predicting the synthesis conditions required to make new zeolites tailored for specific applications and find the synthesis conditions necessary to yield hypothetical zeolite structures.

CONCLUSION

We have developed an automatic data extraction pipeline that locates, extracts, and formats zeolite synthesis data from tables, ratios, and text. This pipeline is applied to the synthesis of germanium-containing zeolites to study the complex relationships between the synthesis parameters and resulting topology. Beyond looking at existing trends, we have demonstrated a machine learning model that predicts an important structural descriptor of a zeolite's topology from the synthesis conditions. This model represents an important step toward using data to predict synthetic pathways for plausible zeolite structures that have not been crystallized yet.

With relatively small changes in data engineering, this pipeline can be applied to other research questions in zeolite chemistry. The prevalence of unsuccessful synthesis routes provides an opportunity to model the success of potential zeolite synthesis routes. Future directions could also include more complicated models to study OSDA design, more complicated structure representations for new zeolite topology synthesis, or synthesis parameter optimization using active learning.

EXPERIMENTAL SECTION

Data Extraction. *Tables.* Tables from HTML and XML files were converted into hierarchical JSON structures (see Supporting Information for examples). Rule-based approaches based on the placements of number entries in a table determined the correct position of the column and row headers and, by elimination, any header nesting within the table. All words in the row and column headers were classified, and the orientation of the table was determined by the frequency of materials versus properties within the two headers. The extractor then constructed the correct relationship for each cell in the table. We also extracted the table caption and table footers. Any references in the table were linked to the corresponding footer entry as a dictionary key. We extracted full tables from ACS, APS, Elsevier, Wiley, RSC, and Springer. We were only able to extract table captions from Nature and AAAS due to tables being embedded within the paper HTML as external links.

Ratios. We used regular expressions to search the zeolite paper text for compositional ratios. Once the ratio was located in the text, we determined the type of numeric value associated with each compositional element: either a number, range, or variable. If the element was associated with a number, we assumed every data point extracted from the paper had that value. If the element value was a range, we assumed the range described many experiments detailed elsewhere in the paper. If the element value was a variable, we combined all other elements with matching variables to construct algebraic expressions. These expressions were necessary for correctly normalizing compositional information.

Text. Text information filled in gaps in synthesis conditions that existed after table extraction. We searched for crystallization operations by filtering operations by requiring both a time and temperature condition while excluding many incorrect operations such as mix, dry, calcine, and stir. The conditions associated with remaining operations were assumed to be the crystallization time and temperature for all data points associated with the syntheses extracted from the paper. We also searched the text for common OSDA names, again assuming the same OSDA applied to every syntheses.

Data Engineering. *Composition.* For the Ge data, the compositional features are the molar amounts of Si, Ge, Al, B, alkali cations, H_2O, F, and OSDA. Raw extracted values needed to be engineered from their representation in their respective tables, to these standardized features. Other important compositional variables, such as the OH/Si molar ratio, can be achieved by a simple postextraction data refining considering the sources employed in the zeolite syntheses. Ratio values extracted from tables were split into the corresponding features. Next we solve the algebraic expressions extracted from ratios and normalize all species with the condition that Si = 1, unless Si = 0, in which case Ge = 1.

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

OSDA Featurization. All OSDAs were featurized using a multistep procedure starting with the conversion of the text form of each OSDA molecule into its SMILES representation using ChemSpider.[76] OSDA molecules represented by a non-IUPAC name or picture were manually assigned the correct IUPAC name and then converted to SMILES with ChemSpider. The Kier flexibility index and force field-optimized Cartesian coordinates were then obtained from a locally modified version of molSimplify.[77] Finally, ORCA 4.1[78] was used to calculate the volume, surface area, and dipole moment from the molSimplify-generated Cartesian coordinates. More details can be found in the Supporting Information.

Product Featurization. We featurized the products of the synthesis route with structural data from the IZA database. Zeolite materials were matched to the corresponding topology giving access to the framework density, ring configuration, and building units. Several nonzeolite germanate structures were also featurized with framework density and ring configuration provided by ITQ crystallographers.

Manual Data Supplementation and Cleaning. In addition to data extracted automatically, we manually extracted and engineered data from the supplementary sections of the Ge papers. These supplementary sections are highly unstructured PDF files, which prevents us from processing them with our automatic pipeline. After extraction and engineering, all the data were manually checked for inaccuracy, and any incorrect values were fixed.

Random Forest Model Architecture. We trained a random forest regression model using sci-kit learn,[79] a machine learning Python library. The ensemble consisted of 100 decision trees with splits determined by mean squared error. We trained and cross validated the model on syntheses that resulted in a pure phase zeolite or germanate, which includes 898 synthesis routes. We also created support vector regression, simple neural network, and Gaussian process regression models to compare with the random forest model. The random forest model was chosen, as it exhibited the highest accuracy compared to the other models while also having the benefit of human interpretability.

Decision Tree Model Architecture. We trained a single decision tree regression model using sci-kit learn.[79] Decision splits were determined by mean squared error. The model was trained on the 898 pure phase zeolite synthesis routes without cross validation, since we were only concerned with demonstrating machine learned synthesis intuition rather than any predictive ability with this model. The model was able to reproduce the framework density of the training data with an r-squared score of 0.97.

Safety Statement. No unexpected or unusually high safety hazards were encountered.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.9b00193.

Computational details for the OSDA property calculations used in the Machine Learning models. The germanium data set and table extraction code are available at www.github.com/olivettigroup/table_extractor (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: elsao@mit.edu. Phone: +1 617 2530877.

ORCID

Zach Jensen: 0000-0001-7635-5711
Edward Kim: 0000-0002-0781-5531
Terry Z. H. Gani: 0000-0003-0357-6390
Yuriy Román-Leshkov: 0000-0002-0025-4233
Manuel Moliner: 0000-0002-5440-716X
Avelino Corma: 0000-0002-2232-3527
Elsa Olivetti: 0000-0002-8043-2385

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge funding from the National Science Foundation Award No. 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. Early work was collaborative under the Department of Energy Basic Energy Science Program through the Materials Project under Grant No. EDCBEE. This work has also been supported by the Spanish Government through the Severo Ochoa Program SEV-2016-0683 and the Grant No. MAT2015971261-R, and by La Caxia Foundation through the MIT-SPAIN SEED FUND Program (LCF/PR/MIT17/11820002).

■ REFERENCES

(1) Davis, M. E. Ordered porous materials for emerging applications. *Nature* **2002**, *417*, 813.
(2) Martínez, C.; Corma, A. Inorganic molecular sieves: Preparation, modification and industrial application in catalytic processes. *Coord. Chem. Rev.* **2011**, *255*, 1558−1580.
(3) Abdo, S.; Wilson, S. *Zeolites in Catalysis;* The Royal Society of Chemistry, Thomas Graham House Cambridge, 2017; Vol. *28*, pp 310−350.
(4) Degnan, T. F., Jr. Applications of zeolites in petroleum refining. *Top. Catal.* **2000**, *13*, 349−356.
(5) Li, Y.; Li, L.; Yu, J. Applications of zeolites in sustainable chemistry. *Chem.* **2017**, *3*, 928−949.
(6) Csicsery, S. M. Shape-selective catalysis in zeolites. *Zeolites* **1984**, *4*, 202−213.
(7) Weitkamp, J. Zeolites and catalysis. *Solid State Ionics* **2000**, *131*, 175−188.
(8) Gallego, E. M.; Portilla, M. T.; Paris, C.; León-Escamilla, A.; Boronat, M.; Moliner, M.; Corma, A. Ab initio synthesis of zeolites for preestablished catalytic reactions. *Science* **2017**, *355*, 1051−1054.
(9) Brand, S. K.; Schmidt, J. E.; Deem, M. W.; Daeyaert, F.; Ma, Y.; Terasaki, O.; Orazov, M.; Davis, M. E. Enantiomerically enriched, polycrystalline molecular sieves. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 201704638.
(10) Cundy, C. S.; Cox, P. A. The hydrothermal synthesis of zeolites: Precursors, intermediates and reaction mechanism. *Microporous Mesoporous Mater.* **2005**, *82*, 1−78.
(11) Piccione, P. M.; Yang, S.; Navrotsky, A.; Davis, M. E. Thermodynamics of pure-silica molecular sieve synthesis. *J. Phys. Chem. B* **2002**, *106*, 3629−3638.
(12) Corma, A.; Davis, M. E. Issues in the Synthesis of Crystalline Molecular Sieves: Towards the Crystallization of Low Framework-Density Structures. *ChemPhysChem* **2004**, *5*, 304−313.
(13) Serrano, D. P.; van Grieken, R. Heterogenous events in the crystallization of zeolites. *J. Mater. Chem.* **2001**, *11*, 2391−2407.

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

(14) Navrotsky, A.; Trofymluk, O.; Levchenko, A. A. Thermochemistry of microporous and mesoporous materials. *Chem. Rev.* **2009**, *109*, 3885−3902.
(15) Newsam, J. M.; Bein, T.; Klein, J.; Maier, W. F.; Stichert, W. High throughput experimentation for the synthesis of new crystalline microporous solids. *Microporous Mesoporous Mater.* **2001**, *48*, 355−365.
(16) Corma, A.; Díaz-Cabañas, M. J.; Jordá, J. L.; Martinez, C.; Moliner, M. High-throughput synthesis and catalytic properties of a molecular sieve with 18-and 10-member rings. *Nature* **2006**, *443*, 842.
(17) Treacy, M.; Rivin, I.; Balkovsky, E.; Randall, K.; Foster, M. Enumeration of periodic tetrahedral frameworks. II. Polynodal graphs. *Microporous Mesoporous Mater.* **2004**, *74*, 121−132.
(18) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. Computational discovery of new zeolite-like materials. *J. Phys. Chem. C* **2009**, *113*, 21353−21360.
(19) Pophale, R.; Cheeseman, P. A.; Deem, M. W. A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.* **2011**, *13*, 12407−12412.
(20) *IZA Structure Commission*. 2018; http://www.iza-structure.org/.
(21) Zones, S. Translating new materials discoveries in zeolite research to commercial manufacture. *Microporous Mesoporous Mater.* **2011**, *144*, 1−8.
(22) Yilmaz, B.; Müller, U. Catalytic applications of zeolites in chemical industry. *Top. Catal.* **2009**, *52*, 888−895.
(23) Blatov, V. A.; Ilyushin, G. D.; Proserpio, D. M. The zeolite conundrum: why are there so many hypothetical zeolites and so few observed? A possible answer from the zeolite-type frameworks perceived as packings of tiles. *Chem. Mater.* **2013**, *25*, 412−424.
(24) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465−1476.
(25) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **2017**, *3*, 1237−1245.
(26) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.
(27) Grzybowski, B. A.; Bishop, K. J.; Kowalczyk, B.; Wilmer, C. E. The'wired'universe of organic chemistry. *Nat. Chem.* **2009**, *1*, 31.
(28) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J.; 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*, 094104.
(29) Kim, K.; Ward, L.; He, J.; Krishna, A.; Agrawal, A.; Wolverton, C. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. *Physical Review Materials* **2018**, *2*, 123801.
(30) Aykol, M.; Hegde, V. I.; Suram, S.; Hung, L.; Herring, P.; Wolverton, C.; Hummelshøj, J. S. Network analysis of synthesizable materials discovery. *arXiv preprint arXiv:1806.05772* **2018**.
(31) Ward, L.; Aykol, M.; Blaiszik, B.; Foster, I.; Meredig, B.; Saal, J.; Suram, S. Strategies for accelerating the adoption of materials informatics. *MRS Bull.* **2018**, *43*, 683−689.
(32) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73.
(33) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **2019**, *10*, 539.
(34) Carr, D. A.; Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Machine learning approach for structure-based zeolite classification. *Microporous Mesoporous Mater.* **2009**, *117*, 339−349.
(35) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. Identifying zeolite frameworks with a machine learning approach. *J. Phys. Chem. C* **2009**, *113*, 21721−21725.
(36) Evans, J. D.; Coudert, F.-X. Predicting the mechanical properties of zeolite frameworks by machine learning. *Chem. Mater.* **2017**, *29*, 7833−7839.
(37) Corma, A.; Moliner, M.; Serra, J. M.; Serna, P.; Díaz-Cabañas, M. J.; Baumes, L. A. A new mapping/exploration approach for HT synthesis of zeolites. *Chem. Mater.* **2006**, *18*, 3287−3296.
(38) Serra, J. M.; Baumes, L. A.; Moliner, M.; Serna, P.; Corma, A. Zeolite synthesis modelling with support vector machines: a combinatorial approach. *Comb. Chem. High Throughput Screening* **2007**, *10*, 13.
(39) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **2017**, *29*, 9436−9444.
(40) Gaultois, M. W.; Sparks, T. D.; Borg, C. K.; Seshadri, R.; Bonificio, W. D.; Clarke, D. R. Data-driven review of thermoelectric materials: performance and resource considerations. *Chem. Mater.* **2013**, *25*, 2911−2920.
(41) 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.
(42) 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. Data* **2017**, *4*, 170127.
(43) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust. Chemical Tagger: A tool for semantic text-mining in chemistry. *J. Cheminf.* **2011**, *3*, 17.
(44) 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.
(45) Court, C. J.; Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **2018**, *5*, 180111.
(46) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational Materials* **2017**, *3*, 53.
(47) Kim, E.; Jensen, Z.; van Grootel, A.; Huang, K.; Staib, M.; Mysore, S.; Chang, H.-S.; Strubell, E.; McCallum, A.; Jegelka, S. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *arXiv preprint arXiv:1901.00032* **2018**.
(48) Yang, S.; Vlessidis, A. G.; Evmiridis, N. P. Influence of gel composition and crystallization conditions on the conventional synthesis of zeolites. *Ind. Eng. Chem. Res.* **1997**, *36*, 1622−1631.
(49) Uguina, M. A.; de Lucas, A.; Ruiz, F.; Serrano, D. P. Synthesis of ZSM-5 from ethanol-containing systems. Influence of the gel composition. *Ind. Eng. Chem. Res.* **1995**, *34*, 451−456.
(50) Lok, B. M.; Messina, C. A.; Patton, R. L.; Gajek, R. T.; Cannan, T. R.; Flanigen, E. M. Silicoaluminophosphate molecular sieves: another new class of microporous crystalline inorganic solids. *J. Am. Chem. Soc.* **1984**, *106*, 6092−6093.
(51) Kessler, H.; Patarin, J.; Schott-Darie, C. The opportunities of the fluoride route in the synthesis of microporous materials. *Stud. Surf. Sci. Catal.* **1994**, *85*, 75−113.
(52) Koller, H.; Lobo, R. F.; Burkett, S. L.; Davis, M. E. $SiO^{-}\cdots HOSi$ hydrogen bonds in as-synthesized high-silica zeolites. *J. Phys. Chem.* **1995**, *99*, 12588−12596.
(53) Barrett, P.; Camblor, M.; Corma, A.; Jones, R.; Villaescusa, L. Synthesis and structure of as-prepared ITQ-4, a large pore pure silica zeolite: the role and location of fluoride anions and organic cations. *J. Phys. Chem. B* **1998**, *102*, 4147−4155.
(54) Camblor, M. A.; Corma, A.; Valencia, S. Synthesis in fluoride media and characterisation of aluminosilicate zeolite beta. *J. Mater. Chem.* **1998**, *8*, 2137−2145.
(55) Zones, S. I.; Darton, R. J.; Morris, R.; Hwang, S.-J. Studies on the role of fluoride ion vs reaction concentration in zeolite synthesis. *J. Phys. Chem. B* **2005**, *109*, 652−661.
(56) Li, J.; Corma, A.; Yu, J. Synthesis of new zeolite structures. *Chem. Soc. Rev.* **2015**, *44*, 7112−7127.

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899

(57) Kang, J. H.; Xie, D.; Zones, S. I.; Smeets, S.; McCusker, L. B.; Davis, M. E. Synthesis and characterization of CIT-13, a germanosilicate molecular sieve with extra-large pore openings. *Chem. Mater.* **2016**, *28*, 6250−6259.
(58) Li, L.; Chen, Y.; Xu, S.; Li, J.; Dong, M.; Liu, Z.; Jiao, H.; Wang, J.; Fan, W. Oriented control of Al locations in the framework of Al-Ge-ITQ-13 for catalyzing methanol conversion to propene. *J. Catal.* **2016**, *344*, 242−251.
(59) Qian, K.; Wang, Y.; Liang, Z.; Li, J. Germanosilicate zeolite ITQ-44 with extra-large 18-rings synthesized using a commercial quaternary ammonium as a structure-directing agent. *RSC Adv.* **2015**, *5*, 63209−63214.
(60) Gao, Z.-H.; Chen, F.-J.; Xu, L.; Sun, L.; Xu, Y.; Du, H.-B. A Stable Extra-Large-Pore Zeolite with Intersecting 14-and 10-Membered-Ring Channels. *Chem. - Eur. J.* **2016**, *22*, 14367−14372.
(61) Smeets, S.; Koch, L.; Mascello, N.; Sesseg, J.; McCusker, L.; Hernández-Rodríguez, M.; Mitchell, S.; Pérez-Ramírez, J. Structure analysis of a BEC-type germanosilicate zeolite including the location of the flexible organic cations in the channels. *CrystEngComm* **2015**, *17*, 4865−4870.
(62) Corma, A.; Díaz-Cabañas, M.; Jiang, J.; Afeworki, M.; Dorset, D.; Soled, S.; Strohmaier, K. Extra-large pore zeolite (ITQ-40) with the lowest framework density containing double four-and double three-rings. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13997−14002.
(63) Jiang, J.; Jorda, J. L.; Diaz-Cabanas, M. J.; Yu, J.; Corma, A. The Synthesis of an Extra-Large-Pore Zeolite with Double Three-Ring Building Units and a Low Framework Density. *Angew. Chem., Int. Ed.* **2010**, *49*, 4986−4988.
(64) Corma, A.; Díaz-Cabañas, M. J.; Rey, F.; Nicolopoulus, S.; Boulahya, K. ITQ-15: The first ultralarge pore zeolite with a bi-directional pore system formed by intersecting 14-and 12-ring channels, and its catalytic implications. *Chem. Commun.* **2004**, 1356−1357.
(65) Roth, W. J.; Nachtigall, P.; Morris, R. E.; Wheatley, P. S.; Seymour, V. R.; Ashbrook, S. E.; Chlubná, P.; Grajciar, L.; Položij, M.; Zukal, A.; et al. A family of zeolites with controlled pore size prepared using a top-down method. *Nat. Chem.* **2013**, *5*, 628.
(66) Verheyen, E.; Joos, L.; Van Havenbergh, K.; Breynaert, E.; Kasian, N.; Gobechiya, E.; Houthoofd, K.; Martineau, C.; Hinterstein, M.; Taulelle, F.; et al. Design of zeolite by inverse sigma transformation. *Nat. Mater.* **2012**, *11*, 1059.
(67) Kessler, H.; Patarin, J.; Schott-Darie, C. The opportunities of the fluoride route in the synthesis of microporous materials. *ChemInform* **1995**, *26*,
(68) Corma, A.; Navarro, M. T.; Rey, F.; Rius, J.; Valencia, S. Pure polymorph C of zeolite beta synthesized by using framework isomorphous substitution as a structure-directing mechanism. *Angew. Chem.* **2001**, *113*, 2337−2340.
(69) Castañeda, R.; Corma, A.; Fornés, V.; Rey, F.; Rius, J. Synthesis of a new zeolite structure ITQ-24, with intersecting 10-and 12-membered ring pores. *J. Am. Chem. Soc.* **2003**, *125*, 7820−7821.
(70) Moliner, M.; Serna, P.; Cantín, Á.; Sastre, G.; Díaz-Cabañas, M. J.; Corma, A. Synthesis of the Ti- silicate form of BEC polymorph of *β*-zeolite assisted by molecular modeling. *J. Phys. Chem. C* **2008**, *112*, 19547−19554.
(71) Cantín, A.; Corma, A.; Diaz-Cabanas, M. J.; Jordá, J. L.; Moliner, M. Rational design and HT techniques allow the synthesis of new IWR zeolite polymorphs. *J. Am. Chem. Soc.* **2006**, *128*, 4216−4217.
(72) Corma, A.; Rey, F.; Valencia, S.; Jordá, J. L.; Rius, J. A zeolite with interconnected 8-, 10-and 12-ring pores and its unique catalytic selectivity. *Nat. Mater.* **2003**, *2*, 493.
(73) Eliášová, P.; Opanasenko, M.; Wheatley, P. S.; Shamzhy, M.; Mazur, M.; Nachtigall, P.; Roth, W. J.; Morris, R. E.; Čejka, J. The ADOR mechanism for the synthesis of new zeolites. *Chem. Soc. Rev.* **2015**, *44*, 7177−7206.
(74) Sun, J.; Bonneau, C.; Cantín, Á.; Corma, A.; Díaz-Cabañas, M. J.; Moliner, M.; Zhang, D.; Li, M.; Zou, X. The ITQ-37 mesoporous chiral zeolite. *Nature* **2009**, *458*, 1154.
(75) Jiang, J.; Jorda, J. L.; Yu, J.; Baumes, L. A.; Mugnaioli, E.; Diaz-Cabanas, M. J.; Kolb, U.; Corma, A. Synthesis and structure determination of the hierarchical meso-microporous zeolite ITQ-43. *Science* **2011**, *333*, 1131−1134.
(76) Pence, H. E.; Williams, A. *ChemSpider: an online chemical information resource,* 2010.
(77) Ioannidis, E. I.; Gani, T. Z.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106−2117.
(78) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(79) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; Varoquaux, G. In *API design for machine learning software: experiences from the scikit-learn project;* Proceedings of ECML PKDD Workshop Languages for Data Mining and Machine Learning, Prague, Czech Republic, Sept 23−27; Springer, 2013; pp 108−122.

DOI: 10.1021/acscentsci.9b00193
ACS Cent. Sci. 2019, 5, 892−899