CHEMISTRY OF MATERIALS

pubs.acs.org/cm

Perspective

Challenges in Information-Mining the Materials Literature: A Case Study and Perspective

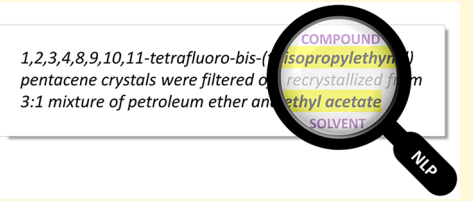
Andrew Smith, Vinayak Bhat, Qianxiang Ai, and Chad Risko*

Cite This: https://doi.org/10.1021/acs.chemmater.2c00445

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The rapid development and application of machine learning (ML) techniques in materials science have led to new tools for machine-enabled and autonomous/high-throughput materials design and discovery. Alongside, efforts to extract data from traditional experiments in the published literature with natural language processing (NLP) algorithms provide opportunities to develop tremendous data troves for these *in silico* design and discovery endeavors. While NLP is used in all aspects of society, its application in materials science is still in the very early stages. This perspective provides a case study on the application of NLP to extract information related to the preparation of organic materials. We present the case study at a basic level with the aim to discuss these technologies and processes with researchers from diverse scientific backgrounds. We also discuss the challenges faced in the case study and provide an assessment to improve the accuracy of NLP techniques for materials science with the aid of community contributions.



INTRODUCTION

Given ever-expanding computational power and data storage resources, we are witnessing a paradigm shift in science due to the abundance of data, the feasibility of data sharing, and the proliferation of data-driven approaches dedicated to providing new insights and even predictive capacities based on "big data".[1,2] The growing trend of transitioning from case-based studies to dataset-based research all but guarantees a tremendous demand for accessible, high-quality scientific data that are machine-digestible, which, at first glance, should not be an issue considering that more than three million scholarly articles are being published electronically every year.[3] However, as the intended audience of research publications is the human-centered scientific community, collecting machine-digestible data from these publications can be rather nontrivial, especially for a highly interdisciplinary field such as materials science.

Big data efforts in materials science, being driven in part by both government agencies and scientific communities, have resulted in initiatives like the Materials Genome Initiative (MGI)[4] in the United States, Materials Genome Engineering[5] in China, Materials Data Platform[6] in Japan, Horizon 2020 in Europe,[7] and NCCR MARVEL[8] in Switzerland, to name but a few. A significant component of these projects is information-mining data from the literature to extract materials properties and synthesis routes with machine learning (ML). Some of the text-mining applications that have resulted from such an undertaking are a database of Curie and Néel temperatures,[9] synthesis protocols for inorganic materials,[10,11] and processing conditions for solid-state batteries,[12] to name a few.

While there is rapid progress in the development of state-of-the-art text-mining algorithms for materials science, there remain several challenges.[13] For instance, the MGI advocates for increased standardization of the reporting of data and metadata—the lack of which currently inhibits many text-mining activities.[14] This perspective illustrates common problems one may encounter in data collection from scientific publications by focusing on a case study of extracting crystallization solvents for organic materials. We aim to present the technical aspects such that those without a computer science background can learn terminologies associated with machine-based data extraction. We demonstrate how a simple rule-based method can extract data from published articles and provide a baseline for more complex methods. We also highlight the challenges associated with text-mining using natural language processing (NLP) algorithms in materials science, no matter how sophisticated the NLP algorithm used. In addition, we also seek to raise awareness on writing and publishing scientific articles in both a human and machine-readable manner, as many of the difficulties in data mining cannot be overcome without collective efforts from research communities.

Received: February 10, 2022
Revised: May 11, 2022

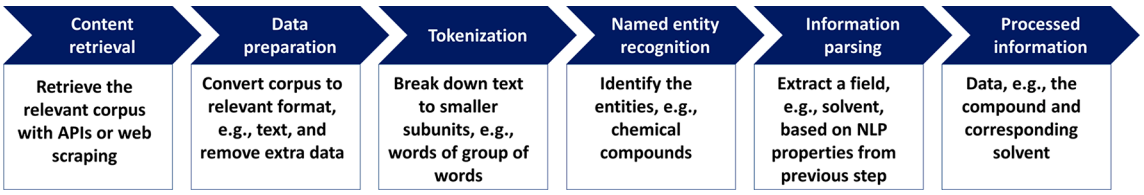


Figure 1. Schema representing the pipeline used to extract data from materials science literature.

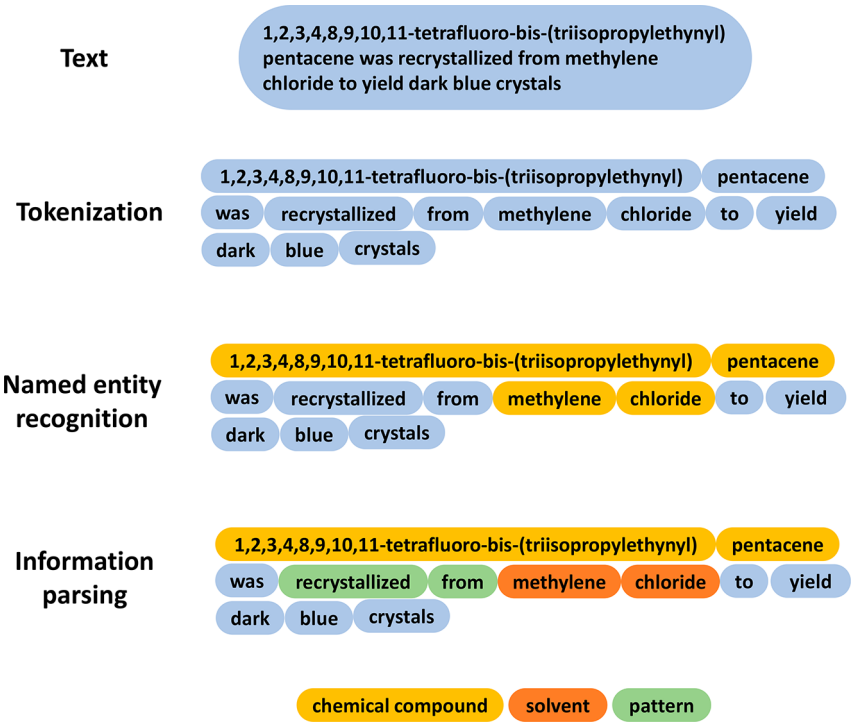


Figure 2. Example of extracting crystallization solvent from text. During information parsing, named entities are classified as solvent or nonsolvent, and then the rule-based method is applied to extract the recrystallization solvent.

CASE STUDY: CRYSTALLIZATION SOLVENTS FOR SMALL ORGANIC MOLECULES

Crystallization is an important step in making and, more commonly, purifying materials, as impurities can significantly affect materials application, such as device performance for organic semiconducting materials.[15] Crystallization is also crucial to polymorphism studies, as variations in crystallization conditions can result in different crystal structures from the same molecule.[16,17] While general rules exist regarding molecular solubility in solvents, the conditions used in recrystallization are not known *a priori* and are often determined via time-consuming trial errors in practice. Thus, converting molecules to materials can be significantly expedited if the recrystallization conditions could be predicted.

While there are established databases for organic molecular crystals,[18,19] unfortunately, crystallization conditions are not required in these databases. For instance, of the 274k single component molecular crystals in the Cambridge Structural Database (version 541), only around 59k structures have the crystallization solvent labeled. Thus, for data-centered projects that aim to derive structure−function relationships that span the chemistries of the molecular building blocks to the properties of crystals, such as those being developed for organic semiconductors,[20−24] it is highly desirable to extract crystallization solvent information from reported articles to train predictive models.[25] The pipeline used in this case study to extract the data is depicted in Figure 1. We note that crystallization can be replaced by other problems regarding material synthesis/properties, and the following discussions could be applied to other fields in materials science.

Content Retrieval. The first step in the data collection process is developing the relevant corpus. At this point, one must consider three main factors impacting corpus gathering: (i) accessibility, (ii) parsability, and (iii) relevance. The data source for this case study was the Supporting/Supplementary Information (SI) in the published literature, as these files are generally freely available, making them more accessible than the main text. SI files are almost exclusively available in PDF, a document format that is difficult for a machine to parse, though this difficulty is made less so by the accessibility and relevance factors. Lastly, in many scientific articles, specific methodological details, such as materials recrystallization, are largely found in the SI.

Once target data has been identified, one must devise a generalized method for retrieving and processing it to a functional form. To retrieve the SI from the article DOI (digital object identifier), we focus primarily on web scraping. This decision is necessary because there currently does not exist a comprehensive application programming interface (API) for retrieving SI. While APIs like CrossRef and Scopus are accurate in retrieving main text, less than 10% of DOIs

https://doi.org/10.1021/acs.chemmater.2c00445
Chem. Mater. XXXX, XXX, XXX−XXX

from our case study had accurate direct SI links when the above-mentioned APIs were used. Hence, we focus on a generalizable web scraper based on a probabilistic expert system, which is largely accurate in retrieving SI.

Content Processing. Converting the data from the publishers' website to a text format can be achieved with computer codes like watr-works,[26] Beautiful soup,[27] and pdfminer,[28] to name a few. In this case study, we used pdfminer through chemdataextractor[29] for PDF processing. The process of converting PDF to text, however, is not perfect, and there can be decoding errors. The majority of problems encountered in this case study were chemical names being inconsistently decoded with different numbers of spaces. For instance, CH_2Cl_2 would be decoded as "*CH2Cl2*", "*C H 2 Cl 2*", "*C H2Cl 2*", or "*CH2Cl 2*". Another common error is the introduction of unnecessary white spaces, such as double or triple spaces, midsentence. Importantly, both types of errors can interfere with NLP. For this reason, it is important to examine the inputs and outputs of any preprocessing steps, such as PDF decoding, for unexpected outputs. Note, however, that some of the errors, such as extraneous white spaces, can be fixed with further processing.

Natural Language Processing (NLP). NLP is a task by which computers attempt a similar level of intrinsic understanding of language to that which humans possess.[30,31] To elaborate, computers have very little understanding of an arbitrary string of characters. Computers can be just as easily told that "The cat sat on the mat" is a threat rather than a statement of fact. It does not know that "cat" is a noun or, more specifically, a furry mammal. With a computer's unfamiliarity with even such a simple sentence, many more complex tasks are presently intractable. Yet, these concepts come almost subconsciously to a human. While computers generally still struggle to grasp the meaning of a text fully, NLP algorithms can endow them with knowledge about the sentence's grammar and sentiment. A more detailed description of NLP for materials science can be found in works by Olivetti and co-workers,[10,32,33] Cole and co-workers,[34] and Hong and co-workers.[13] A core aspect of NLP is tokenization, which consists of breaking a larger text down into smaller subunits (Figure 2). In many cases, these units are words, although they could be groups of words. Furthermore, tokenization removes clutter such as extraneous punctuation and whitespaces. This allows for insight based on the token's position relative to others and its relative abundance.

Extracting the (Re)Crystallization Solvent. To extract the solvent used in the crystallization process from the materials synthesis text, we first need to identify the chemical compound and the solvent from the tokens. This process of identifying entities is called named entity recognition (NER). The named entities can be identified with a simple chemical dictionary like Jochem[35] or more sophisticated ML models like a neural network.[36] In this case study, NER is performed with chemdataextractor, which uses a combination of these methods. After NER, the relationship between the entities can be derived from various methods, e.g., rule-based[37] and supervised learning or semi-supervised learning.[38] In the case study, we use the rule-based method, which is particularly useful to find a pattern in the given text. For instance, if the materials synthesis contains the text "***TIPS-pentacene (12)*** *was recrystallized from* ***dichloromethane***", the pattern to look for is "*<**COMPOUND**> was recrystallized from <**SOLVENT**>*". Hence, the solvent used for recrystallization is the named entity following the text "recrystallized from". However, many articles have differing forms that may contain a mixture of solvent, the crystallization method, temperature, etc. Hence, the rule-based method was modified to use a more diverse sentence structure. The classification of a token as a chemical compound or solvent is derived from the chemical dictionary of chemdataextractor. The F1-score is one of the metrics used to evaluate the performance of NLP. Mathematically, the F1-score is defined by the following equation:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

where,

$$\mathrm{Precision} = \frac{\text{Total true positive}}{\text{Total true positive} + \text{Total false positive}}$$

$$\mathrm{Recall} = \frac{\text{Total true positive}}{\text{Total true positive} + \text{Total false negative}}$$

Results. We evaluated the success of our rule-based model by randomly selecting 100 SI PDFs that contained a report of the crystallization solvent. The validation data used to compute the F1-score was extracted by a human expert. We observed that our simple rule-based model yields an F1-score of 0.76 for the crystallization solvent. The error in automated extraction is due to the diverse styles that authors report the crystallization solvent. For example, "*sample of* **10** *was crystallized two times in a round bottom flask (100 ml) by dissolving in a boiling 1-propanol (20−40 ml) and cautiously diluting...*" and "*A sample of the raw product was recrystallized by slow diffusion of anhydrous ethyl ether into a methanol solution to produce crystals...*". Such errors can be mitigated by using more complex rules or more sophisticated deep learning algorithms like transformer and Bi-LSTM.[39−41]

A major concern, however, is the limited ability to extract the chemical compound and the crystallization solvent. The F1-score for determining both the compound and the solvent is 0.36. The low F1-score results from the complex structure of text describing the process. As most of the text related to crystallization is embedded in the synthesis paragraphs, the chemical compound may be referred to as "product" or "solid" etc. For instance, "*17α-Ethynyl-17β-hydroxyestra-4,9-diene-3-one (***10***): **2** (30.0 g, 0.0882 mol) was dissolved in acetic acid (300 mL) at room temperature. Then perchloric acid (12 mL, 0.1323 mol) was added dropwise while keeping the reaction temperature was under 30 °C. The reaction mixture was stirred for 1 h and then poured into water (600 mL) slowly. The precipitated crystals were filtered off, recrystallized by (petroleum ether/ethyl acetate 3:1) and dried at 50 °C.*" For the above text, our model captures the crystallization solvent as a mixture of petroleum ether and ethyl acetate but fails to identify the chemical compound **10** that is crystallized. This issue accounted for a plurality of 46% of all errors in compound parsing. The second major source of error, 19% of instances where the compound was not retrieved correctly, was that the compound was only represented in the SI as an image. This error cannot readily be resolved as it would require implementing a system to identify the presence of a chemical structure in an image and what chemical that image represents. Though a more complex system could resolve an image to its chemical name, it is notable that this would dramatically increase computational

https://doi.org/10.1021/acs.chemmater.2c00445
Chem. Mater. XXXX, XXX, XXX−XXX

cost. In order to determine if an image is of a structure all images must be checked. This presents a difficulty in the case of SI files that may have many images, particularly of spectral data. In 15% of the failed outcomes, the cause was a coreference resolution problem, which is an inability to relate an abbreviation explicitly stated in the sentence to its compound. Although in this case study we present an easy to implement solution with an F1 score of 0.36, the performance could be improved through the integration of a more complex coreference resolution model such as NeuralCoref.[42] In cases where the solvent was not correctly retrieved, 83% of errors were primarily due to mixtures containing unexpected white space characters, which led to parts of the mixture being left out. Though the text may be easy to comprehend for a domain expert, extracting the relevant information about the chemical compound involved in crystallization with NLP is thus not trivial.

PERSPECTIVE

The case study presented above is a simple illustration of the application of NLP to extract materials science-related data from literature text. There do exist more sophisticated NLP algorithms, like Bi-LSTM[43] and BERT,[44] that can have better performance than the rule-based method used in this case study. Even with these algorithms, however, several challenges remain. On the publishers' side, downloading the articles in machine-friendly formats like XML and HTML is not always allowed. In our case study, we found less than 10% of the records from CrossRef or Elsevier API had a link to the SI. While the publishers do convert the data in the main text into machine-readable formats, the plots and tables in SI are largely left in PDF format. There exist parsers to extract data from tables and optical character recognition algorithms to retrieve the information from figures in PDF,[45,46] but they are still in infancy for materials science. There is no format and checklist enforced for reporting procedures or data and no consensus on the definition of some terms used in an interdisciplinary field like materials science. This, along with a diversity of writing styles, lowers the performance of NLP algorithms to parse data from reported texts.[47] The highly accurate NLP algorithms used for the English language are trained or tested on large, curated, open corpora like SNLI[48] and GLUE.[49] The lack of such an open corpus has hindered the development of materials science-specific NLP algorithms.

To overcome the shortcoming related to increasing the findability of SI, publishers should make the links to SI available through the APIs. For improving the parsability of SI, the figures and tables contained in the SI could be made available as separate files. For instance, high-resolution images or plots may be better presented as a separate JPEG file or even as the raw data used to generate a figure. Furthermore, large tables, especially those spanning many pages, are likely better represented as separate CSV files. These recommendations are consistent with the National Information Standards Organization's (NISO) recommendations for reforming SI.[50]

In fields like biomedical research, there exist checklists like TRIPOD[51] and RECORD[52] to enforce a format for transparently reporting the research. Furthermore, with several prevalent ontologies,[53,54] which are explicit specifications of concepts,[55] there is a rapid development of NLP models in biomedical research.[53] Recently, there are efforts in these directions in materials science—a battery performance checklist is proposed, and a materials ontology is being developed.[56,57] Requirements for reporting SMILES[58] or SELFIES[59] for organic molecules would alleviate the problems with identifying nontrivial chemical names. However, the success of these efforts heavily relies on the materials science community as a whole to accept and adapt these advancements. NLP tools like RepCheck[60] and SynCheck[61] could be leveraged to enforce the reporting standards (Figure 3). While

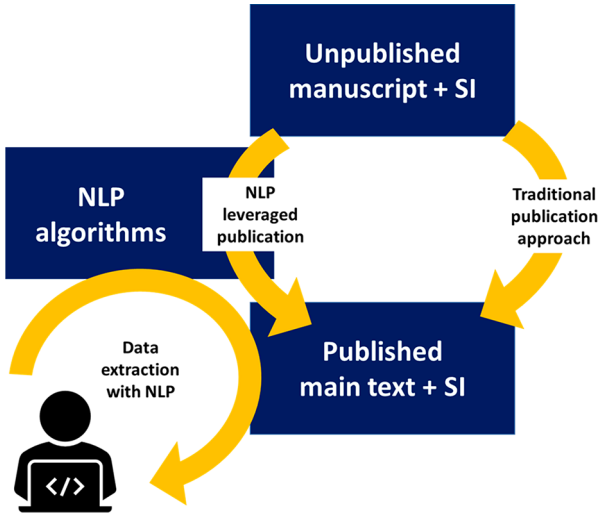


Figure 3. A schematic of leveraging NLP techniques for experiment reproducibility and machine-digestibility before publication.

large databases like Materials Project[62] and AFLOW[63] make data available for training machine learning models to predict property, materials science-specific large, curated, and open corpus should be created to enable training and testing of NLP algorithms. The performance of the NLP models could be improved if they could be coupled with a rich materials ontology.[64]

AUTHOR INFORMATION

Corresponding Author

Chad Risko − *Department of Chemistry & Center for Applied Energy Research, University of Kentucky, Lexington, Kentucky 40506, United States;* orcid.org/0000-0001-9838-5233; Email: chad.risko@uky.edu

Authors

Andrew Smith − *Department of Chemistry & Center for Applied Energy Research, University of Kentucky, Lexington, Kentucky 40506, United States*

Vinayak Bhat − *Department of Chemistry & Center for Applied Energy Research, University of Kentucky, Lexington, Kentucky 40506, United States*

Qianxiang Ai − *Department of Chemistry & Center for Applied Energy Research, University of Kentucky, Lexington, Kentucky 40506, United States;* Present Address: Department of Chemistry, Fordham University, The Bronx, New York, New York 10458, United States

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemmater.2c00445

Author Contributions

A.S., V.B., and Q.A. contributed equally to this project. All authors contributed to the conception of the project and writing of the manuscript.

https://doi.org/10.1021/acs.chemmater.2c00445
Chem. Mater. XXXX, XXX, XXX−XXX

Notes

The authors declare no competing financial interest.

The python code used in the case study is made available at https://github.com/caer200/solvent_nlp. The CSV file with the extracted crystallization data and rules used to extract data can be found in the GitHub repository.

Biographies

Andrew Smith received a B.S. in Chemistry from the University of Kentucky in 2022 and will enter the Ph.D. program in the Department of Chemistry at Duke University. Andrew's research has focused on how data is packaged and used in chemistry, with efforts centering on machine reading, machine learning, and quantum mechanics calculations.

Vinayak Bhat is a Ph.D. candidate in the Department of Chemistry at the University of Kentucky under the supervision of Dr. Chad Risko. Vinayak received a B.S.-M.S. degree in Chemistry from the Indian Institute of Science Education and Research, Thiruvananthapuram, India. His research focuses on creating data architectures and implementing data-driven methods to accelerate the design of organic p-conjugated systems. Vinayak is the current developer of OCELOT (https://oscar.as.uky.edu/).

Qianxiang Ai is a postdoctoral researcher in the Department of Chemistry at Fordham University. Qianxiang received his Ph.D. in 2020 from the University of Kentucky under the direction of Dr. Chad Risko. His Ph.D. studies include computationally aided rational design of organic semiconductors, and he was a lead developer of OCELOT. Qianxiang is currently working on cheminformatics and machine learning technologies for organic−inorganic hybrid materials.

Chad Risko is an Associate Professor of Chemistry at the University of Kentucky. Chad received his B.S. in Chemistry at Baker University and his Ph.D. in Chemistry at the Georgia Institute of Technology under the direction of Dr. Jean-Luc Bredas. Chad's research interests lie at the intersection of computational chemistry and materials chemistry, with the aim of fostering machine-informed materials design and discovery for energy conversion, energy storage, and electronic and optical technologies.

■ ACKNOWLEDGMENTS

This work was sponsored by the National Science Foundation in part through the Designing Materials to Revolutionize and Engineer our Future (NSF DMREF) program under Award Number DMR 1627428 and the Established Program to Stimulate Competitive Research (EPSCoR) Track 2 program under Cooperative Agreement Number 2019574. We acknowledge the University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for their fantastic support and collaboration and use of the Lipscomb Compute Cluster and associated research computing resources.

■ REFERENCES

(1) Hey, A. J.; Tansley, S.; Tolle, K. M. *The fourth paradigm: data-intensive scientific discovery*; Microsoft Research, Redmond, WA, 2009; Vol. *1*.

(2) Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* **2019**, *6* (21), 1900808.

(3) Johnson, R.; Watkinson, A.; Mabe, M. *The STM Report: An overview of scientific and scholarly publishing*; International Association of Scientific, Technical and Medical Publishers: 2018; pp 1−214.

(4) National Science and Technology Council. *Materials Genome Initiative for global competitiveness.*; Office of Science and Technology Policy: Washington, DC, 2011.

(5) O'Meara, S. Materials science is helping to transform China into a high-tech economy. *Nature* **2019**, *567* (567), S1−S5.

(6) Tanifuji, M.; Matsuda, A.; Yoshikawa, H. In Materials Data Platform - a FAIR System for Data-Driven Materials Science; *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*; IEEE: 2019; pp 1021−1022; DOI: 10.1109/IIAI-AAI.2019.00206.

(7) *Horizon*, 2020. https://ec.europa.eu/programmes/horizon2020/ (accessed 2021-11-30).

(8) *NCCR MARVEL*. https://nccr-marvel.ch/ (accessed 2021-08-05).

(9) Court, C. J.; Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data* **2018**, *5* (1), 180111.

(10) 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. *Scientific Data* **2017**, *4* (1), 170127.

(11) Kim, E.; Jensen, Z.; Van Grootel, A.; Huang, K.; Staib, M.; Mysore, S.; Chang, H.-S.; Strubell, E.; Mccallum, A.; Jegelka, S.; Olivetti, E. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (3), 1194−1201.

(12) Mahbub, R.; Huang, K.; Jensen, Z.; Hood, Z. D.; Rupp, J. L. M.; Olivetti, E. A. Text mining for processing conditions of solid-state battery electrolytes. *Electrochem. Commun.* **2020**, *121*, 106860.

(13) Hong, Z.; Ward, L.; Chard, K.; Blaiszik, B.; Foster, I. Challenges and Advances in Information Extraction from Scientific Literature: a Review. *JOM* **2021**, *73* (11), 3383−3400.

(14) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; Analytis, J.; Dabo, I.; DeLongchamp, D. M.; Fiete, G. A.; Grason, G. M.; Hautier, G.; Mo, Y.; Rajan, K.; Reed, E. J.; Rodriguez, E.; Stevanovic, V.; Suntivich, J.; Thornton, K.; Zhao, J.-C. New frontiers for the materials genome initiative. *npj Computational Materials* **2019**, *5* (1), 41.

(15) Diemer, P. J.; Hayes, J.; Welchman, E.; Hallani, R.; Pookpanratana, S. J.; Hacker, C. A.; Richter, C. A.; Anthony, J. E.; Thonhauser, T.; Jurchescu, O. D. The Influence of Isomer Purity on Trap States and Performance of Organic Thin-Film Transistors. *Advanced Electronic Materials* **2017**, *3* (1), 1600294.

(16) Giri, G.; Li, R.; Smilgies, D. M.; Li, E. Q.; Diao, Y.; Lenn, K. M.; Chiu, M.; Lin, D. W.; Allen, R.; Reinspach, J.; Mannsfeld, S. C.; Thoroddsen, S. T.; Clancy, P.; Bao, Z.; Amassian, A. One-dimensional self-confinement promotes polymorph selection in large-area organic semiconductor thin films. *Nat. Commun.* **2014**, *5* (1), 3573.

(17) Sorli, J. C.; Ai, Q.; Granger, D. B.; Gu, K.; Parkin, S.; Jarolimek, K.; Telesz, N.; Anthony, J. E.; Risko, C.; Loo, Y.-L. Impact of Atomistic Substitution on Thin-Film Structure and Charge Transport in a Germanyl-ethynyl Functionalized Pentacene. *Chem. Mater.* **2019**, *31* (17), 6615−6623.

(18) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* **2012**, *40* (D1), D420−D427.

(19) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016**, *72* (2), 171−179.

(20) Ai, Q.; Bhat, V.; Ryno, S. M.; Jarolimek, K.; Sornberger, P.; Smith, A.; Haley, M. M.; Anthony, J. E.; Risko, C. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *J. Chem. Phys.* **2021**, *154* (17), 174705.

(21) Kunkel, C.; Schober, C.; Margraf, J. T.; Reuter, K.; Oberhofer, H. Finding the Right Bricks for Molecular Legos: A Data Mining

https://doi.org/10.1021/acs.chemmater.2c00445
Chem. Mater. XXXX, XXX, XXX−XXX

Approach to Organic Semiconductor Design. *Chem. Mater.* **2019**, *31* (3), 969−978.
(22) Olsthoorn, B.; Geilhufe, R. M.; Borysov, S. S.; Balatsky, A. V. Band Gap Prediction for Large Organic Crystal Structures with Machine Learning. *Advanced Quantum Technologies* **2019**, *2* (7−8), 1900023.
(23) Padula, D.; Omar, Ö. H.; Nematiaram, T.; Troisi, A. Singlet fission molecules among known compounds: finding a few needles in a haystack. *Energy Environ. Sci.* **2019**, *12* (8), 2412−2416.
(24) Yu, M.; Wang, X.; Du, X.-F.; Kunkel, C.; Garcia, T. M.; Monaco, S.; Schatschneider, B.; Oberhofer, H.; Marom, N. Anomalous pressure dependence of the electronic properties of molecular crystals explained by changes in intermolecular electronic coupling. *Synth. Met.* **2019**, *253*, 9−19.
(25) Xin, D.; Gonnella, N. C.; He, X.; Horspool, K. Solvate Prediction for Pharmaceutical Organic Molecules with Machine Learning. *Cryst. Growth Des.* **2019**, *19* (3), 1903−1911.
(26) Saunders, A.; Shastry, A. N. *Watr-Works*. https://github.com/iesl/watr-works (accessed 2022-01-20).
(27) Richardson, L. *Beautiful Soup*. https://www.crummy.com/software/BeautifulSoup/ (accessed 2022-01-20).
(28) Shinyama, Y.; Marsman, P. *pdfminer.six*. https://github.com/pdfminer/pdfminer.six (accessed 2022-01-20).
(29) 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* (10), 1894−1904.
(30) Chowdhury, G. G. Natural language processing. *Annual Review of Information Science and Technology* **2003**, *37* (1), 51−89.
(31) Liddy, E. D. Natural Language Processing. In *Encyclopedia of Library and Information Science*, 2nd ed.; Marcel Decker, Inc.: New York, 2001.
(32) Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and challenges of text mining in materials research. *iScience* **2021**, *24* (3), 102155.
(33) Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews* **2020**, *7* (4), 041317.
(34) Cole, J. M. A Design-to-Device Pipeline for Data-Driven Materials Discovery. *Acc. Chem. Res.* **2020**, *53* (3), 599−610.
(35) Hettne, K. M.; Stierum, R. H.; Schuemie, M. J.; Hendriksen, P. J. M.; Schijvenaars, B. J. A.; Mulligen, E. M. V.; Kleinjans, J.; Kors, J. A. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **2009**, *25* (22), 2983−2991.
(36) Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *J. Chem. Inf. Model.* **2019**, *59* (9), 3692−3702.
(37) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **2020**, *11* (1), 3601.
(38) Huo, H.; Rong, Z.; Kononova, O.; Sun, W.; Botari, T.; He, T.; Tshitoyan, V.; Ceder, G. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Computational Materials* **2019**, *5* (1), 62.
(39) Beltagy, I.; Cohan, A.; Lo, K. SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *arXiv*, 2019-03-26, https://arxiv.org/abs/1903.10676 (accessed 2020-01-20).
(40) Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv*, 2015-08-09. https://arxiv.org/abs/1508.01991 (accessed 2020-01-20).
(41) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018−10−11. https://arxiv.org/abs/1810.04805 (accessed 2022-01-20).
(42) Wolf, T. *NeuralCoref*. https://github.com/huggingface/neuralcoref (accessed 2022-01-20).
(43) He, T.; Sun, W.; Huo, H.; Kononova, O.; Rong, Z.; Tshitoyan, V.; Botari, T.; Ceder, G. Similarity of Precursors in Solid-State Synthesis as Text-Mined from Scientific Literature. *Chem. Mater.* **2020**, *32* (18), 7861−7873.
(44) Trewartha, A.; Walker, N.; Huo, H.; Lee, S.; Cruse, K.; Dagdelen, J.; Dunn, A.; Persson, K. A.; Ceder, G.; Jain, A. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **2022**, *3* (4), 100488.
(45) Mavračić, J. *TableDataExtractor*. https://github.com/CambridgeMolecularEngineering/tabledataextractor (accessed 2022-01-20).
(46) Beard, E. J.; Cole, J. M. ChemSchematicResolver: A Toolkit to Decode 2D Chemical Diagrams with Labels and R-Groups into Annotated Chemical Named Entities. *J. Chem. Inf. Model.* **2020**, *60* (4), 2059−2072.
(47) Kim, E.; Huang, K.; Kononova, O.; Ceder, G.; Olivetti, E. Distilling a Materials Synthesis Ontology. *Matter* **2019**, *1* (1), 8−12.
(48) Bowman, S. R.; Angeli, G.; Potts, C.; Manning, C. D. A large annotated corpus for learning natural language inference. *arXiv*, 2015-08-21. https://arxiv.org/abs/1508.05326 (accessed 2022-01-20).
(49) Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv, 2018-04-20*. https://arxiv.org/abs/1804.07461 (accessed 2022-01-20).
(50) *Recommended Practices for Online Supplemental Journal Article Materials*; National Information Standards Organization: 2013.
(51) Collins, G. S.; Reitsma, J. B.; Altman, D. G.; Moons, K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine* **2015**, *13* (1), 1.
(52) Benchimol, E. I.; Smeeth, L.; Guttmann, A.; Harron, K.; Moher, D.; Petersen, I.; Sørensen, H. T.; Von Elm, E.; Langan, S. M. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine* **2015**, *12* (10), e1001885.
(53) Kersloot, M. G.; van Putten, F. J. P.; Abu-Hanna, A.; Cornet, R.; Arts, D. L. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. *Journal of Biomedical Semantics* **2020**, *11* (1), 14.
(54) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25* (1), 25−29.
(55) Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition* **1993**, *5* (2), 199−220.
(56) Sun, Y.-K. An Experimental Checklist for Reporting Battery Performances. *ACS Energy Letters* **2021**, *6* (6), 2187−2189.
(57) Li, H.; Armiento, R.; Lambrix, P. An Ontology for the Materials Design Domain. In *Lecture Notes in Computer Science*; Springer International Publishing: 2020; pp 212−227.
(58) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(59) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1* (4), 045024.
(60) Bhat, V.; Smith, A.; Ai, Q.; Risko, C. *RepCheck*. https://oscar.as.uky.edu/repcheck (accessed 2022-01-20).
(61) Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. *SynCheck*. https://www.syncheck.org/ (accessed 2022-01-20).
(62) 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 Materials* **2013**, *1*, 011002.

https://doi.org/10.1021/acs.chemmater.2c00445
Chem. Mater. XXXX, XXX, XXX−XXX

(63) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218−226.
(64) Estival, D.; Nowak, C.; Zschorn, A. Towards ontology-based natural language processing. *Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology*; Association for Computational Linguistics: 2004; DOI: 10.3115/1621066.1621075.

https://doi.org/10.1021/acs.chemmater.2c00445
Chem. Mater. XXXX, XXX, XXX−XXX