J_ID: APRPG5 DOI: 10.1063/5.0021106 Date: 4-December-20 Stage: Page: 1 Total Pages: 21

PROOF COPY [APR20-RV-00497]

Applied Physics Reviews REVIEW scitation.org/journal/are

Data-driven materials research enabled by natural language processing and information extraction

Cite as: Appl. Phys. Rev. 7, 000000 (2020); doi: 10.1063/5.0021106
Submitted: 7 July 2020 · Accepted: 19 November 2020 ·
Published Online: 0 Month 0000

AQ1 Elsa A. Olivetti,[1,a)] Jacqueline M. Cole,[2,3,4] Edward Kim,[5] Olga Kononova,[6,7] Gerbrand Ceder,[6,7] Thomas Yong-Jin Han,[8] and Anna M. Hiszpanski[8] Correct

AFFILIATIONS

[1]Department of Materials Science and Engineering, MIT, Cambridge, Massachusetts 02139, USA
[2]Cavendish Laboratory, Department of Physics, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom
[3]ISIS Neutron and Muon Source, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot OX11 0QX, United Kingdom
[4]Department of Chemical Engineering and Biotechnology, University of Cambridge, West Cambridge Site, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom
[5]Science, Evaluation, and Measurement, Xero, Toronto, Ontario M5H 4G1, Canada
[6]Department of Materials Science & Engineering, University of California Berkeley, Berkeley, California 94720, USA
[7]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[8]Materials Science Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

[a)]Author to whom correspondence should be addressed: elsao@mit.edu

ABSTRACT

Given the emergence of data science and machine learning throughout all aspects of society, but particularly in the scientific domain, there is increased importance placed on obtaining data. Data in materials science are particularly heterogeneous, based on the significant range in materials classes that are explored and the variety of materials properties that are of interest. This leads to data that range many orders of magnitude, and these data may manifest as numerical text or image-based information, which requires quantitative interpretation. The ability to automatically consume and codify the scientific literature across domains—enabled by techniques adapted from the field of natural language processing—therefore has immense potential to unlock and generate the rich datasets necessary for data science and machine learning. This review focuses on the progress and practices of natural language processing and text mining of materials science literature and highlights opportunities for extracting additional information beyond text contained in figures and tables in articles. We discuss and provide examples for several reasons for the pursuit of natural language processing for materials, including data compilation, hypothesis development, and understanding the trends within and across fields. Current and emerging natural language processing methods along with their applications to materials science are detailed. We, then, discuss natural language processing and data challenges within the materials science domain where future directions may prove valuable.

TABLE OF CONTENTS

Applied Physics Reviews REVIEW scitation.org/journal/are

I. INTRODUCTION

Data have always been a fundamental ingredient for realizing, accelerating, and optimizing any scientific pursuit. The increasing ubiquity of data science methods, based on improved computing power and algorithm development, has driven significant opportunity and interest in immense, structured datasets. When such data are assembled in a form readily consumed and mined using data science tools, coupled with domain expertise, there is tremendous potential to accelerate discovery,[1] build upon previous findings, rapidly enter a new field, connect individual research efforts, and link across disciplines.

The physics community has long understood the value of curating data in a way that can be comprehended by computer logic. This is particularly true in the domains of high-energy physics, astronomy, and astrophysics, where data emanate from very rare and specialized research machines. For example, the Large Hadron Collider at CERN, in Switzerland, generates a wide range of data from particle collisions, certain types of which can be measured using unique detectors, and enables collaborations among 3000 scientists and engineers for each collaboration. Other examples include the gravitational-wave observatories (LIGO[2] and Virgo[3] collaborations are currently >1000 and >500 members, respectively) and widely shared astronomical mappings from satellites and telescopes. These sources of data tend to be managed by large international research facilities since multinational efforts are needed to fund and build them. Scientists work within large, coordinated, research consortia to produce, process, and analyze the data.[4,5] Raw data are contained within each facility but are accessible, albeit sometimes in normalized form, and their particular data characteristics tend to limit the variety of data types.

However, one aspect of the physical sciences still wanting for more and better organized data to leverage emerging data science tools is in the domain of materials science. Successful examples of application of materials informatics to the discovery of new materials can be found in alloy development,[6] polymer design,[7] organic light emitting diodes,[8] and solar cells.[9,10] However, these cases are still quite limited and suffer most from lack of data. While there are a growing number of open databases that contain materials property information,[11–15] most of these databases are created from computationally calculated properties. As an example, the materials project includes computational information for over 130 000 inorganic compounds, and the analogous experimental databases only contain 9000 materials.[15] Experimentally based, large, and structured materials property databases are still lacking.

Unlike other fields, materials science lacks sufficient incentive to make it practical to centralize its data, not only because the data are so diverse but also because data arise from so many independent scientists and laboratory sources.[16] Figure 1 illustrates this contrast. Data in

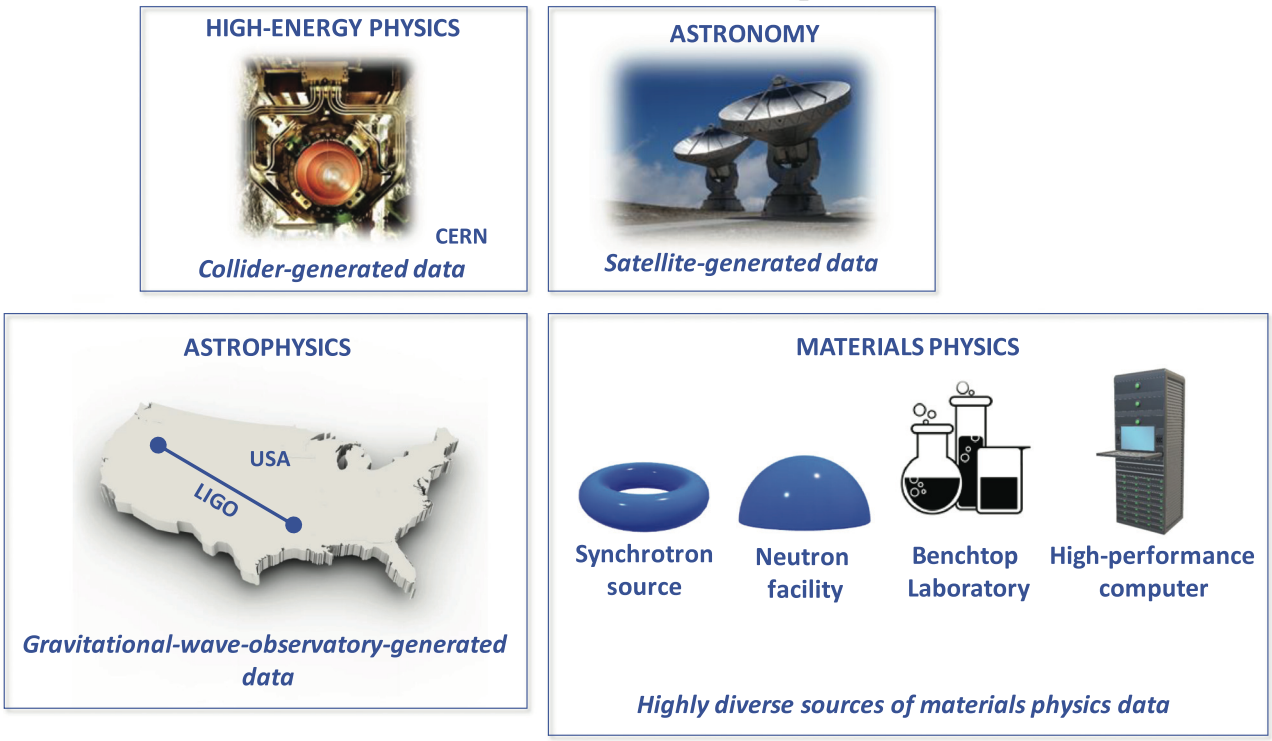

FIG. 1. Comparison of large centralized datasets in high-energy physics, astronomy, and astrophysics compared to heterogeneous, decentralized data in materials physics. Unlike other fields, materials science lacks sufficient incentive to make it practical to centralize the data, not only because the data are so diverse but also because the data arise from a variety of independent scientists and laboratory sources. Data in the field of materials science are particularly heterogeneous due to the wide variety of material classes studied by scientists. The data appear as numerical text or image-based information, which requires quantitative interpretation.

materials science are particularly heterogeneous, based on the significant range in materials classes that are explored and the variety of materials properties that are of interest. This leads to data that range many orders of magnitude, and these data may manifest as numerical text or image-based information, which requires quantitative interpretation. The many length scales of materials science add to this diversity, with data being measured from the atomic structure to massive components that are integrated at a system level, such as airplane wings or turbine blades. In addition, only a small sub-set of specialized materials-physics research needs to be carried out at centralized facilities, such as government-run synchrotrons, neutron and muon sources, nanocenters, high-magnetic field laboratories, laser laboratories, or high-performance supercomputing facilities.[17–21] Some of these facilities archive the raw or normalized ("reduced") data, and some offer their scientific users the option to tag their experimental data with a document object identifier (DOI) to make them traceable.[22] If even once such data become openly available, the metadata generated by the experiment may be missing.[23] Metadata are vital for processing the data to the point where one can interpret their scientific meaning.

Fortunately, there is a prospective approach to address at least some aspects of this data-access quandary in materials science. Scientists will cede control of their processed data if they publish their results, and publications continue to be the primary means of communicating within the materials domain. These data will be spread across various journal articles, patents, or company reports, owing to the variety of ways that scientists can publish their findings. The data will also present in an unstructured form, given the highly diverse way in which scientists write an article and select the most salient results for showcasing their scientific points (i.e., as text or in figures, tables, and schematics). For example, scientists may report the composition of a metal alloy in one table, the processing conditions for that alloy in the body text of the methods, and then the final properties in figures within the results. Despite the distributed nature of these processed data, harvesting them from documents presents a way to retrieve materials-physics data en masse. The manual task of mining information from documents by editors is not practical, given the amount of data that are needed to succeed in the field of materials informatics. A means to automatically extract materials-physics data from scientific documents is, therefore, required. This challenge presents a prime opportunity for information extraction and natural language processing (NLP), whereby "materials-aware" text-mining models can be used to collate processed data that lie within the literature to afford auto-generated materials databases that can be used in materials informatics.

Capturing unstructured information from the vast and ever-growing number of scientific publications has substantial promise to meet this need and enable creation of experimental-based databases currently lacking. This reliance on publications in scientific communication is exemplified by the proliferation of new journals and increased frequency of publication.[24–26] Developing methods to mine the literature for data may also prevent information loss. Without structuring information, scientists cannot make the necessary connections among findings; they may instead be drawn by what the authors of a scientific document have chosen to be highlighted in a journal or individual publicity efforts. Scientific progress relies on hypothesis development, which requires leveraging increased knowledge toward greater understanding, typically based on synthesizing existing information. Scientists are not trained to formalize their findings in a structured way. The rapid growth of scientific knowledge has the potential to provide opportunities to transfer solutions from one domain to address problems in another. However, the underlying relationships largely remain embedded, and groups from disparate domains remain within their own specialties.[24] Limits what one individual can draw relationships between varied concepts, topics, and domains. There is a distinct value to be drawn beyond what is known and what is known as individuals from the collective to broad multidisciplinary knowledge within and across a given domain. This sharing and integration of information across communities is a tall order to accomplish comprehensively, but the ability to automatically extract information from the literature can provide a tool to facilitate this engagement.

A. The scope of this review

In this review, we look at the fully and semi-automated means of assembling and structuring scientific data through NLP and text mining. In the realm of scientific text, methods, tools, and databases of relevance for NLP have been most well developed for the biomedical domain[27,28] where information is sought on genes, proteins, drugs, medical symptoms, and disease. These efforts exist also in the chemistry discipline, which arguably began earlier, but tools for chemistry are less advanced than those in the biomedical domain. Efforts in chemistry have focused on developing comprehensive chemical dictionaries,[29,30] substance and small molecule composition, and structure and property descriptions.[31–34] This review will focus on what has been achieved to date in NLP for the discipline of materials science.

The structure of this article is as follows. We first describe reasons for pursuit of NLP of scientific text given the motivation provided above. Next, we focus on the tasks and methods involved, describing the challenges for the materials science domain including a summary of commonly used tools. Then, we show in detail about particular examples of NLP applications in materials science. Next, we discuss data mining beyond NLP and how this nonetheless tracks back to the cognate need for NLP. Finally, we provide some commentary on the future needs and directions for the use of NLP as a tool for the materials community.

II. THE WAYS THAT NLP CAN BENEFIT DATA-DRIVEN MATERIALS SCIENCE

Leveraging NLP tools in materials science remains in its infancy. The methods used, and the level of accuracy required, vary depending on the inquiring goal. Before diving into the details of how NLP is performed, we briefly mention some of the key benefits that NLP afford for data science. These include generating datasets for mining and visualization across multiple research efforts, as well as contributing to machine learning (ML) predictions and identifying research trends. Examples of the application of NLP in materials science will be provided in Sec. IV.

The use of NLP on scientific text can generate libraries of information to explore, which enables data visualization, mining, and analytics. The primary goals of text extraction can be used to populate databases with quantitative information or make text information summative and interactive in a way that can reveal patterns, gaps, or trends. Advances in data analytics and visualization tools, described in greater detail in Sec. V, have also accelerated the process of information consumption to decision-making. A well-structured database with an interactive and intuitive graphical user interface allows

researchers to perform significant background research, test hypotheses, survey the field, and form a sound basis for designing and performing experimental work, saving hours if not months of labor-intensive literature surveying and wasted experiments. Text extraction can provide data that drive search-engine development in the scientific domain and a beginning of active learning systems tied to automated materials discovery and synthesis platforms.[35–37]

Beyond data extraction and visualization, researchers may also leverage NLP to derive fundamental insight across these data; for example, NLP may be used to find relationships between compounds by mapping materials mentioned in the text to corresponding chemical structures. This identification of relationships and trends is frequently done by using various ML techniques on the extracted data. Scientists can search for similar chemical structures or substructures, meaning that text information can be combined with knowledge from established computational-property databases. For example, this combination of extracted and existing data might allow for exploring and screening the relevance of compounds to a new application as a function of published properties.[38] The ML models used vary in complexity, but the key opportunities for the scientific-language assembly include literature-based knowledge discovery, suggesting novel scientific hypotheses, or predicting the outcomes of reactions.

NLP activities across scientific text can also identify future research trends by predicting emerging associations (co-occurrences) between selected keywords found in the scientific literature. This type of analysis has been done previously for biochemistry,[39,40] neuroscience,[41,42] and human innovations.[43,44] Significant work in this area can also be found in the domain of "the Science of Science."[24] The NLP community presents a nuanced differentiation between "information extraction" and "knowledge-based creation" (traditional and emergent approaches, respectively). Information extraction structures extracted text according to entity recognition and entity relationships, which, then, feed into downstream search and query-based activities. Knowledge-based creation can provide an end unto itself in the form of ontology development where facts and relationships with a discipline are extracted in a form that could be used to annotate area-specific databases or to transfer knowledge between fields. Early efforts in materials science have focused primarily on information extraction. Given the need for expanded datasets in materials science (beyond what is currently available), this is a logical emphasis. As the community refines key tools toward NLP for materials, a broader set of pursuits can be realized.

III. PERFORMING NATURAL LANGUAGE PROCESSING

Before delving into the specific details of methodology, we provide a few key themes related to NLP, which convey the perspective taken in the materials science community. First, there are manual and semi-automated methods of literature-data extraction, which yield insights into "small" datasets (i.e., tens to hundreds of relevant articles), but the focus moving forward (and within this review) must be on the ability to apply methods to create large datasets (i.e., tens of thousands of relevant articles). Generic NLP tools (such as CoreNLP) exist that do not perform well in the materials science domain without modification, as the vernacular, sentence construction, terminology, and chemical semantics are specialized. Therefore, we need to develop and apply materials-specific text mining tools to meet the needs of this community. To reach any sort of economy of scale across such an interdisciplinary field, approaches that transfer effectively within the materials science domain are needed, which requires a balance between accuracy and generalizability. Each application space will have local norms from which rules can be crafted for highly accurate information retrieval in that one domain; however, these rules often breakdown when applied to a different area of inquiry. Challenges with developing generalizable tools are also influenced by the type of document and section within the document. Finally, there is a tension in balancing model development toward the semantic or linguistic structure of the document, while still incorporating critical domain knowledge in how individuals within the field communicate. Natural language carries a high degree of ambiguity, and implicit knowledge plays a significant role in how a field communicates. However, if too much of this implicit knowledge is integrated within models, leveraging the linguistic structure of the text becomes more difficult.

Most natural language extraction pipelines follow a similar overall approach, shown in Fig. 2, which consists of (1) acquiring a relevant corpus of text, (2) processing that text into individual terms, which is called tokenization, (3) segmenting documents and classifying

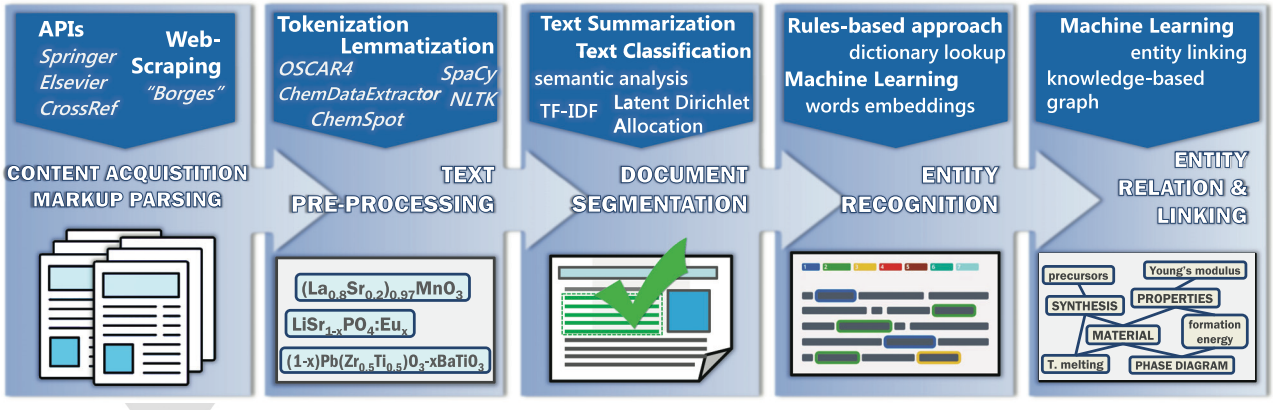

FIG. 2. Schematic of NLP including examples of tools and models at each step. It is visible that most natural language extraction follows similar approaches: (1) acquiring relevant text resources, (2) processing the text into individual terms (also known as tokenization), (3) document segmentation and paragraph classification, (4) recognizing tokens as classes of information, (5) entity relation extraction, and (6) named entity linking.

paragraphs, (4) recognizing tokens as specific classes of information, generally referred to as named entity recognition (NER), (5) entity relation extraction, and (6) named entity linking. Depending on the question being pursued by researchers, pipelines may vary in their methods and approach, including the order in which the above-described steps are performed, the types of information models are provided, and deviations in the models themselves. Using broad strokes, we can describe the continuum of approaches as direct word mapping, defining heuristics, and then to ML-based methods. ML approaches, in and of themselves, can vary within the continuum of unsupervised to supervised, the latter requiring labeled data often in significant volumes. This section describes details on each of these steps with materials-relevant method development presented with each step.

A. Content acquisition

The first step is to develop and acquire a relevant corpus of subject articles of interest from which information will be retrieved. The content varies by the degree of accessibility, the corpus of subject articles of interest, and the kinds of documents (patents vs journal articles, for example). This content can only be digested within the subsequent models if rendered in plain text-accessible format, although there is variety in that format.[45] The older digitized content is available primarily in portable document format (PDF) (introduced in 1993); however, even the older content may be preserved as images, presenting an insurmountable challenge in extracting information at scale. Converting PDF to plain text relies on spatial identification of blocks of text in a layout-aware manner, which is still an area of research.[46] Errors may arise in terms of misplaced blocks of text and font-conversion challenges. Most journals and publishers after the mid-1990s also provide content as hypertext markup language (HTML) or extensible markup language (XML). HTML or XML often has more consistency in their conversion to plain text format, but this format is not ubiquitous across publishers. Given the challenges associated with PDF conversion, nearly all reports of text mining of materials science texts have been on articles available in markup language.[47,48]

Acquiring information from patents provides another way to obtain content, given patent accessibility and centralized hosting by country-specific patent offices. However, patent authors often seek to protect their knowledge from being fully disclosed, and so, these texts may have even more implicit information than journal articles. Patents relevant to materials science have a closely defined structure and style of presentation;[49] in particular, the example section mirrors the synthesis section, so they can be interpreted with a high degree of accuracy. Patents will not be a focus of the methodology discussion going forward in this text, but they have been used in biology and chemistry applications with some frequency.[50]

The downloaded content consists of article text and metadata (journal name, title, abstract, and author names). The metadata provide value in databasing the content, as well as being high-level information that can inform entity recognition as described below; it is typically more structured than the document content.

AQ3 ~~The necessary number of articles gathered within a corpus is also often thought about and whether the database.~~

B. Text preprocessing and tokenization

Once the content has been obtained, three main activities are used to manipulate the information contained within the text: entity extraction, entity relation, and entity linking. This overall flow begins with a series of steps that preprocess the text of the article to enable identification of the desired information. Preprocessing will vary according to the order of events and the tools used for each stage. A low-level, but critical, step is character encoding, establishing the way that the characters are represented. *Tokenization* (a form of preprocessing) segments text into the relevant sentences, phrases, words, or word pieces, to be processed individually or as a sequence. Punctuation marks are the obvious approach to identify sentences, but the language of the scientific domain is often complicated by terms that are composed of multiple words, symbols, and other types of structural entities, which, therefore, requires specialized tokenization pipelines. Some examples of this challenge with chemical and material notation include the uses of commas: $(Y,In)BaCo_3ZnO_7$; periods: $(La_{0.8}Sr_{0.2})_{0.97}MnO_3$ or $CuSO_4{\cdot}5H_2O$; hyphens: $(1-x)Pb(Zr_{0.52}Ti_{0.48})O_{3-x}BaTiO_3$ or Ti-64 (common term for $Ti_{90}Al_6V_4$ alloy); and colons: $LiSr_{1-x}PO_4{:}Eu_x$. Using materials domain-specific tokenization has been shown to be important for successful NLP of materials texts as it can have a significant impact on downstream activities.[47,48,51] Common tokenizers for the scientific literature include those available within the software: *OSCAR4*,[34] *ChemDataExtractor*,[33] *ChemSpot*,[32] and *BANNER's simple tokenizer*.[27] More general tokenizers that may also be used or adapted for the scientific literature include those by SpaCy and the Penn Treebank tokenizer.

Dependency-based parsing of sentences and part-of-speech (POS) tagging identify the syntactic structure of a sentence. Current state-of-the-art approaches use neural algorithms, including sequential and bidirectional modeling; however, these algorithms rely on larger volumes of training data and corpora than is typical for specific cases within materials science.[52] Nonetheless, some models such as bidirectional encoder representations from transformers (BERT)[53] have shown the ability to adapt readily to certain tasks using datasets on the order of thousands of documents, simply by "swapping out" the final layers of the model to a task-specific architecture (e.g., part-of-speech tagging during parsing). Further-distilled models, such as DistilBERT,[54] may improve this ability to adapt to thousands of document-sized datasets, as there are fewer parameters to fine tune during domain adaptation. Note that we will discuss the role of BERT and other word embedding models below.

When compared to general-purpose text, a scientific dependency-parse should learn specific sentence structures and patterns, such as an extensive use of passive and past tense, limited use of pronouns, and depersonalization of a sentence.[55] The accurate construction of dependency-based parse trees is highly sensitive to the punctuation and correct usage of the word forms, especially verb tenses. These aspects of the grammar are often neglected in scientific publications, making it difficult to use standard well-developed algorithms and tools for text mining. To date, there have not been developments to address these caveats for scientific text.

C. Document segmentation and paragraph classification

NLP can afford better accuracy when one operates only within specific parts of the article, such as the abstract, main text body, tables,

J_ID: APRPG5 DOI: 10.1063/5.0021106 Date: 4-December-20 Stage: Page: 6 Total Pages: 21

or figures, depending on the area of inquiry. This approach not only helps computationally but can also increase the uniformity of the desired extracted text. Matching regular expressions to identify section headers provides an easy guide, and this can be done simply using string matching within a set of text or regular expression (regex) coding, although the variation in the application of headers by publishers can present a challenge even in this straightforward activity. Huo and colleagues have recently applied probabilistic methods, such as latent Dirichlet allocation (LDA), across several million articles to use unsupervised approaches to identify experimental steps implied in sentences.[56] LDA provided a probabilistic topic distribution for each sentence. These authors, then, applied random decision forests, using the topic n-gram as the feature, to classify different types of synthesis procedures; this required annotation of only a few hundred paragraphs. Another feature of this work is that the authors were able to construct a Markov chain representation of the material synthesis flow chart.

As an alternative to the unsupervised approaches discussed, Hiszpanski *et al.* used a supervised ML approach to evaluate every sentence within an article and extract solution-based synthesis protocols.[57] Specifically, by iterative rounds of training with human-annotated sentences, they trained a logistic regression classifier that yields the likelihood of a given sentence describing solution-based synthesis protocols based on the words present within the sentence. As may be expected from scientific writing conventions, past tense verbs, unit terms (e.g., ml and min), and chemicals are weighed heavily as being indicative of a synthesis description. Surprisingly, function words such as "the," "of," "then," and "and," which are normally filtered out from text as "stop words" in nonscientific applications, occur more commonly in synthesis protocols and are important in distinguishing sentences that concern synthesis or otherwise. This observation points out again how traditional NLP approaches may need to be modified when these tools are translated in their application from general texts to the scientific literature.

D. Named entity recognition (NER)

Each of the preprocessing steps described above enable the heart of the text-extraction activity, NER, which identifies the objects of semantic value by recognizing and classifying concepts mentioned in the text. Entities are useful in and of themselves for researchers to map to properties, to find similar compounds, or to incorporate in annotation labeling. Historically, immense effort has gone into NER for the medical domain, extracting symptoms, diagnoses, and medications from text.[27] The chemistry domain has expended significant effort in NER, but even state-of-the-art NER systems do not typically perform well when applied to different domains, and effort is required to create quality data for trainable statistical NER systems.[58]

NER is an area where the materials community is clearly in its infancy. There is a need for training data to develop entity-recognition models. Where knowledge bases exist already for a field, training may be done using distant supervision models that map known entities and relations onto unstructured text. In the computer-science community, this activity is supported by "all community" developed learning tasks that are orchestrated through conferences in the field; these tend to tackle significant challenges along a roadmap, thereby making concerted progress as a domain.[59] There is no equivalent yet in the materials space.

The general methods for NER range from dictionary look-ups, rule-based, and machine-learned approaches. Typical pipelines used in the materials science domain include hybrids of all three of these approaches. Hybrid systems provide a balance of precision with computational efficiency, where only those cases that cannot be handled by dictionaries or rules pass to ML approaches to make efficient use of annotated data. Dictionary look-ups include material composition, chemical element names, properties, as well as processes and experimental parameters.

Hand-crafted rule/knowledge-based methods are a collection of rules or specifications defining how to handle relative ordering and matching among those rules. Rules may be developed through corpus-based systems that require examining several cases to obtain the patterns or via domain knowledge understanding of nomenclature convention. To overcome the time intensive nature of rule development, strategies have been developed to learn rules through small collections of seed examples that begin from very high precision rules and learn to generalize or *vice versa*. Examples of these approaches include LeadMine,[60] which uses naming convention rules, ChemicalTagger,[61] which parses experimental synthesis sections of chemistry texts, and portions of ChemDataExtractor, which uses nested rules. For example, when researchers extended ChemDataExtractor for use in magnetic materials, additions were made for domain-specific parsing rules including off-stoichiometry and relevant terms associated with the domain of interest (in this case magnetic materials such as ferroelectrics and ferrites).[51]

Finally, at the other end of the continuum of NER, approaches are ML-based statistical models, which use a feature-based representation of observed data to recognize specific entity names. These models typically depend on sets of annotated documents, which rely on annotated corpora and the development of metrics for inter-annotator agreement where multiple annotators are involved. Given that a sentence is represented as a sequence of words, it is insufficient to consider only the current word class; therefore, sequential (and typically bidirectional) models are necessary to consider the proceeding, current, and following word. While rule-based approaches are tedious to develop and not easily generalized, supervised ML models, in contrast, require substantial expert-annotated data for training along with detailed annotation guidelines. ML models invite careful consideration of the types of classes that are identified and the order in which labels are classified. Initial NER work specific to the materials domain was performed by Kim *et al.*[47] Kononova *et al.* further built upon this work through a two-step materials entity recognition[48] using the bidirectional long short-term memory network with the conditional random field neural network.

As alluded to above, the degree of supervision within NLP is often modulated by word vector representations that capture the syntactic and semantic word relationships, the so-called "word embeddings." Word embeddings are a learned continuous vector representation, which encode the local word context; these can, then, be analyzed to capture distributional similarities of words. These models may be intrinsic, wherein they identify semantic relations, or extrinsic. Character-based word representation models help with "what does the word look like"; these use the individual character of a token to generate the token vector representation and include morphemes (suffixes and prefixes) and morphological inflections (number and tense). The effectiveness of word vectors depends not only on the training

algorithm and hyperparameters but obviously also on the source data. Recent work explored the impact of similarity between pre-training data and target task, particularly in the area of word embeddings.[62] This work proposed to select pre-trained data using the target vocabulary covered rate (percentage of the target vocabulary that is also present in the source data) and language-model perplexity (if the model finds a sentence very unlikely, in other words dissimilar from the data where this language model is trained on, it will assign a low probability and therefore high perplexity). The authors found that the effectiveness of pre-trained word vectors depends on whether the source data have a high vocabulary intersection with target data, while pre-trained language models can gain more benefits from a similar source. We note, therefore, that the choice of corpus used for the training process is critical, pointing to the quality of text and domain-specificity requirements.[38] A range of word-embedding models have been used in the materials community to date and vary with aspects of the corpus that they are trained on; for example, Word2Vec[63] is trained just on the solid-state synthesis paragraphs vs the contextual model, which is trained on full text.[64] Other word embedding models that have been used in the materials science domain include FastText,[65] Embeddings from Language Models (ELMo),[66] and BERT.[53,67]

Materials-specific challenges to NER will vary by the subdomain. These include subtleties associated with the property, context, and reporting of the underlying measurement. For example, within the work from Audus *et al.* on the NLP of polymers,[68,69] the authors have undertaken specific NER efforts related to that domain, termed polyNER. A synthetic polymer is rarely a single entity and is described instead by distributions of molecular weight, often in conjunction with nonstandard naming conventions or trade names.[68] Thus, polyNER focuses on a necessary pretreatment for polymer entity recognition, highlighting the challenges of generalizing NER tools across disparate domains within materials science. As another example, in work pursued by Kononova *et al.* on solid-state synthesis of inorganic materials, material entries were processed with a material parser that converted strings for a material into a chemical formula, which in turn was split into elements and stoichiometric balances. Then, the authors obtained balanced reactions from precursors and target materials by solving a system of linear equations; this included a set of open compounds that can be released or absorbed, which were inferred based on the composition of precursor and target materials.[48]

Often, these approaches require hybrid system development, where the computer automates one aspect of the activity and human intervention enables precise execution. For the polymer extraction work, the NLP-based extraction process identified candidates within the article and subsequent automated and crowd-sourcing curation steps processed these candidates. There are several ways to formalize the role that a human might play in these activities.[70,71] Approaches can leverage word-embedding models to establish entity-rich corpora, the so-called candidate generation, for expert labeling, which feeds into a context-based word-vector classifier.[69] Researchers have also pursued active learning with maximum-entropy uncertainty sampling to achieve valuable annotations from experts to improve performance, but this proved time intensive to pursue.[72] Roles for hybrid systems also include establishing dictionaries for stop words and rules to detect systematic names.

An additional challenge in the materials community is multi-word tokens. Huang and Ling recently proposed multi-word identifying and representing methods. This involves recognizing the multi-word phrases in the chemical literature through unsupervised methods and then representing the phrases in the vocabulary.[73] Typically, word embedding is performed after tokenization with phrase representation obtained based on a post-vector addition. In this method, a new step is incorporated to identify multi-word phrases and add the detected terms to the vocabulary. In this case, word embedding is performed afterwards at the phrase level. Huang and Ling's computationally intense approach starts from tokenized and trimmed single words and sentence context. Then, they use scoring functions to identify bigrams, repeating this process up to n-grams, and then move to phrase-level word embedding.[74,75]

E. Entity relation extraction and linking

Entity relation extraction is the activity that identifies relations between entities mentioned in a given document. It is primarily done in post-processing steps after NER. Entities extracted can also be linked to their properties or co-occurrence with other entities, which allows new knowledge between them to be identified. Efforts have primarily focused on the co-occurrence of entities within a few sentences of each other, although there is a need to extend this to a full document.

Within materials science, most entity-relation extraction occurs through dependency parsing. More direct supervised ML-based approaches would require the development of larger annotated corpora and quantifying similarity by computing representation similarity. One approached used in materials examples are Snowball methods, which include seed examples of known positive relationships. Based on locating sentences with these seed examples, typical patterns are learned using clustering of textual similarity.[61] By comparing unseen sentences to learned patterns, new relationships can be identified based on a threshold minimum level of similarity. These methods have been extended recently within ChemDataExtractor tools using a modified Snowball algorithm.[51] The original Snowball algorithm uses several thousand seed examples.[76] For the modified Snowball algorithm, the quaternary relationships included the property specifier, chemical entity mention, property value, and then unit.[51] Named entity linking, then, connects information extracted from text with data stored in curated databases where the challenges are to delineate entities that are different from those that are synonyms and should be linked to one unique identifier.

There are several issues to consider after initially applying NLP techniques to scientific text. First whether or not the data are extracted accurately. Second are the data reported correctly. Third are data being reported with sufficient details to warrant these efforts. As the process of text mining proceeds down the pipeline shown in Fig. 2, the accuracy of the extracted data decays rapidly, and noise accumulates. Hence, the choice between having higher precision within a set of extracted data vs having a larger dataset size becomes pivotal because this choice will significantly affect the results of the data mining. Kim *et al.* showed that even when using millions of raw papers as a starting position, numbers may drop to just hundreds of thousands of papers depending on the specific topic.[47] Data loss arises not only due to imperfections of the extraction methods but also, oftentimes, due to the misrepresentation of the original information. A prominent example is referencing a previously published procedure or data analysis instead of providing its description in the current paper. The use of

nonstandard abbreviations, acronyms, and terms also significantly affects the amount of false negative outcomes, as these abbreviations complicate the linking of the information from different parts of the text.

All these places for data loss point to the significant role of outliers and how frequently a data point is occurring, as well as how they are treated afterwards. The pursuit of ground truth for supervised ML within NLP is costly and time-consuming since it is based on the limited annotated documents thus far as discussed above. Whether or not accuracy that is sufficient for the spread of goals of NLP and text mining in materials science can be achieved is still an open question.

F. Conceptual network

Separate from the NLP pipeline described above, the use of text-based approaches to generally learn a field has been an area of interest linked to the concept of ontologies, as described above. A high-level workflow for ontology generation is as follows: first, generate concept lists through expert input and comparisons between a curated reference list and a random set of scientific documents. Then, use methods, such as bag-of-words, to populate the ontology. Recent work used a hierarchical LDA, which learned an overall structure from the data and generated a tree of classes that could be used for searching terms, annotation, and standardization of metadata.[77] There are a few interesting ways to generate these concept lists. The work by Krenn and Zelinger analyzed trends in quantum physics by generating a concept list through human-expert input that is expanded by Rapid Automatic Keyword Extraction to a term list; this is, then, fed into a comprehensive corpus to establish links between each of the terms. To project future directions of research, they performed a link-prediction task to ask which new link will be formed between unconnected vertices given the current network. This was done using an artificial neural network with four fully connected layers, which ranked unconnected pairs of concepts and further extended this approach to identify pairs with exceptional network properties.[78]

IV. RESOURCES AND TOOLS FOR NLP

Given the methods described above, a section is provided here, which summarizes some helpful resources and tools, including a coverage of the tools most commonly used in NLP for materials. Table I lists the most common NER toolkits publicly and freely available and the information that they are capable of extracting. Most have been focused on capabilities to extract entities from body text, but many have expanded efforts to extract tables as well. Several also have a focus on extracting biology-relevant information, which stems from the earlier leading NLP efforts in life sciences. The groups that developed these tools have taken varied approaches, tailored to their specific sub-field of literature. The tools typically vary with the tokenizers and techniques that they use to identify chemicals, which often involve a combination of dictionaries, hand-crafted rules/patterns, and POS-tagging methods, as previously discussed.

Researchers are likely to be interested in extracting categories of information, which are specific to their research topic and beyond, for which there are readily available tools shown in Table I. If the category of information that is desired has a formulaic representation, or it has a limited number of possible ways of being expressed, then rather simple pattern- or dictionary-based approaches can be created to extract this new category of information. When these more straightforward methods fail, then ML-based models can be developed, such as the CRF models for chemical-entity recognition, as previously discussed. Common packages for developing such NLP models include Natural Language Toolkit (NLTK),[83] SpaCy,[84] Stanford CoreNLP,[85] AllenNLP,[86] and openNLP.[87]

In addition to the plethora of software packages for NLP, recent developments in word representation research have led to generalized models that may be rapidly fine-tuned to domains of interest. A notable example is BERT,[53] which has been fine-tuned to scientific text to produce SciBERT;[67] such models may ultimately advance the accuracy of entity recognition for chemicals and materials.

Moreover, other advances in NLP research beyond word representation and subsequent supervised tasks (i.e., classification) may

TABLE I. Tools available for natural language processing in the materials discipline.

Entity recognition toolkits	Information capable of extracting	Approach for named entity recognition (chemistry focused)
ChemDataExtractor[33]	Chemicals Tables	CRF (hand-crafted features + unsupervised features) + filtered Jochem dictionary
ChemicalTagger[61]	Chemicals Quantities Synthesis actions and conditions	OSCAR (see below) + pattern-based rules + dictionaries
Chem Spot 2.0[14,79]	Chemicals	CRF (hand-crafted features+ unsupervised features) + ChemIDPlus dictionary
BANNER-CHEMDNER[27]	Chemicals Bio-relevant entities	CRF (hand-crafted features + unsupervised features)
ChemXSeer[80] and TableSeer[81]	Chemicals Tables	CRF (hand-crafted features + unsupervised features) + Jochem and custom dictionaries
OSCAR4	Chemicals Reaction names Bio-relevant entities	Maximum entropy Markov model + ChEBI and custom dictionaries
LeadMine[82]	Chemicals Named reactions Bio-relevant entities	Dictionaries + pattern-based rules
tmChem[31]	Chemicals	CRF (hand-crafted features + unsupervised features)

J_ID: APRPG5 DOI: 10.1063/5.0021106 Date: 4-December-20 Stage: Page: 9 Total Pages: 21

have the potential for rapid domain adaptation to materials science and chemical science. For example, deep-learning approaches to entity resolution[88,89] are largely driven by unsupervised methods and may serve to resolve mentions of materials into canonical, physically meaningful entities.

However, developing ML models requires many examples of human-annotated text for training and testing a model, which can dictate a heavy investment of time. For those embarking on this route, easy-to-use tools for text annotation are needed. Many free and commercial tools exist and continue to be developed for text annotation; as such, a comprehensive review unrealistic but some commonly used tools include brat,[90] Prodigy,[91] WebAnno,[92] and Callisto.[93] A common question that often arises is how many annotated data are enough data to train a good model? The unsatisfying answer is that one cannot concretely say until one tries. ML-model development is often an iterative process involving model training and testing. If the performance of a model does not meet expectations, then a common means of trying to improve the model is to retrain it with additional data, i.e., more annotated text.

The lack of publicly available materials-relevant corpora with human annotations is hindering progress in NLP research within materials science. Having such publicly available datasets would reduce the need for newcomers in the field to engage in the costly annotation exercise previously described. Additionally, such datasets are essential for enabling comparisons of the performance of new entity-recognition models. This comparison is necessary to help the entire field of NLP for materials science better understand our progress and shortcomings. The largest and most materials-relevant publicly available corpus of annotations is the BioCreative IV CHEMDNER corpus, which was created from a community-wide effort in the 2000s to make a "gold standard" for training and testing NLP tools for the life-science literature.[58] The corpus consists of 10 000 abstracts, taken from PubMed in 2013 with 84 355 human-annotated chemical entity mentions, corresponding to 19 806 unique chemical names.

Currently, no large-scale equivalent corpus derived from the materials science literature exists, but smaller and more materials-focused annotated corpora are beginning to be reported, which have annotations beyond only chemicals, as well. For example, Mysore *et al.* released 230 materials-synthesis procedures with annotations of materials, operations, conditions, apparatuses, and units, amongst others.[94] Likewise, Hiszpanski *et al.* recently released "gold standard" annotations of chemicals and wet-synthesis protocols from 99 articles pertaining to materials synthesis that they then used to compare the performance of various chemical entity recognition tools that are identified in Table I.[57] Other recent examples include data related to solid-state electrolytes and fuel cells.[95,96] Though somewhat further afield from materials, Kulkarni *et al.* created an annotated corpus of 622 wet-lab protocols from experimental biology that has labeled actions, conditions, reagents, amounts, and concentrations, amongst others.[97] There have also been attempts to make the annotation process more efficient through improved interfaces that could potentially enable crowd-sourced annotations,[98] although domain expertise has proven critical. There is a paucity of relevant annotated datasets for the field of materials science. Each of these examples required significant domain expertise and time to craft. Continued efforts by the materials community to share annotated corpora will only help further accelerate progress in this field.

To add details around datasets/corpora size, within NLP research, the number of documents is oftentimes provided as an implied proxy for data size, as we have done throughout. The number of documents provides a relevant metric for tasks associated with word embedding models, for example (where the corpora associated with materials science is small relative to the large number of texts in the scientific domain more broadly). However, of relevance beyond the number of documents is the number of tokens of a particular class present in those documents. Ideally, for machine learning, training data are independent and identically distributed, but we know that this is not the case when dealing with tokens within documents for NLP. Rare is it to find a training corpus that has specific entities in nearly equal amounts across the documents. Some documents are of greater relevance to a topic and are more likely to have more tokens, and within the scientific literature, it is expected that published works will influence others' works. Thus, training data for NLP applications are far from being independent and identically distributed. While providing a precise number will vary by tasks, one can surmise an approximation of what a "large enough" dataset constitutes by surveying the material NLP literature. In these works, after filtering documents for relevancy, most have document corpuses on the order of tens-of-thousands where each document has dozens to low hundreds of entities and entity relations for a specific token class.

Finally, a critical but often overlooked category of tools necessary for reaping the full benefits of NLP efforts is data visualization tools. The NLP of the materials literature creates structured datasets from unstructured text, but databases by themselves are of little use if one cannot see and explore the data interactively. While hard-coded plots and graphs can be presented, such fixed visualizations do not allow further exploration of the dataset beyond the presented perspective. The interactive aspect of data visualization is critical to broaden the utility of such databases and enable users to form hypotheses and test them, thereby building their own understanding of trends. Interactive visualization dashboards, which typically have multiple frames of different data representations, are effective tools for this purpose. Custom interactive dashboards can be created using freely available open-source software packages such as Candela,[99] Bokeh,[100] and D3.[101] The increased ubiquity and interest in data science have also spurred many commercial software packages for creating custom interactive visualization of data, which are commonly marketed as business intelligence and analytics tools.

V. EXAMPLES OF NLP BEING USED IN MATERIALS SCIENCE

Based on the motivation for pursuit of NLP within materials, and the detailed methodology provided, we now describe a series of examples of automated text extraction, which are specific to materials science. The reasons for this pursuit include generating data for mining, visualization, contributing to ML predictions, and the identification of research trends. The ultimate goal of NLP in materials science would be to evolve toward a new way of thinking about materials discovery, but this will only become possible as databases that suit a given application are developed.[16] The examples that this section will cover are captured in Fig. 3.

Examples of datasets gathered and curated by NLP-based methods can be found across materials science, although progress is still early in the physical domain. NLP-based curation efforts with

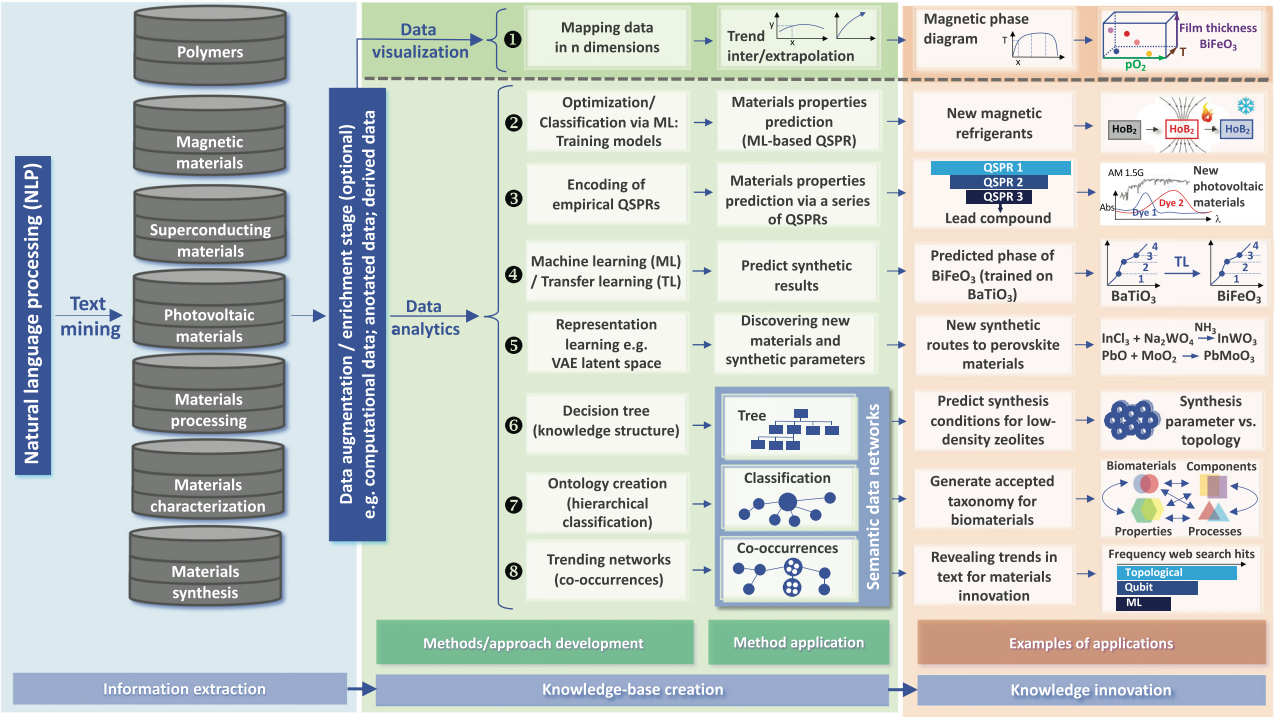

FIG. 3. Overview of the ways that NLP has leveraged data-driven materials science, from information extraction to knowledge base creation and knowledge innovation. The ultimate goal of NLP in materials science would be to evolve toward a new way of thinking about materials discovery, but this will only become possible as databases that suit a given application are developed. Examples that are listed on the right hand side are described within the text.

more of a physical focus include polymers,[68,69,102] Curie and Néel magnetic phase-transition temperatures,[51] and pulsed-laser deposition processing conditions of complex oxides.[103] Efforts that can be linked to physical properties, but are currently focused on materials chemistry, include solid-state reactions for all inorganic materials, synthesis of inorganic oxides,[47,48,104] zeolites,[105] and nanomaterials.[57] Repositories of materials metrology data are also being curated using NLP tools. For example, a database of UV/vis absorption spectral characteristics was auto-generated by mining the experimental values of the wavelength of maximum absorption, λ_{max}, and molar extinction coefficients, ε, of chemicals from the literature.[106] Metrology data offer a more general data platform to serve an entire physics community; the example given will aid a wide range of optical and optoelectronic applications. We offer some specificity around each of these examples.

Within the domain of polymers, leading text extraction efforts are driven by the Polymer Properties Predictor and Database[107] and the NIST Synthetic Polymer MALDI Recipes Database.[108] The former includes semi-automated literature extracted data on Flory-Huggins interaction parameters and glass transition temperatures, T_g, for close to 300 systems. The latter comprises data records for 1250 polymer/matrix combinations. While these datasets are small, they rival those available in relevant, analogous polymer handbooks. Court and Cole have assembled close to 40 000 chemical compounds and associated Curie and Néel magnetic phase-transition temperatures (approximately one-fourth of the data points are Néel temperature records) across almost 70 000 chemistry and physics articles[51] using ChemDataExtractor.[33] These data describe the temperatures for ferromagnetic and antiferromagnetic phase transitions. The work was motivated by the use of ML techniques in magnetism and superconductivity, which has the potential to lead to innovations in data storage devices, quantum information processing, and medicine. Previously, only manually curated databases existed for magnetic materials, designed for single entry lookup. Data have been extracted for pulsed-laser deposition processing conditions of complex oxides[21] (substrate, thickness, growth temperature, repetition rate, and partial pressure of oxygen) and their physical characteristics (critical temperatures, T_c) and functional properties (fluence and remnant polarization); this work leveraged crowd sourcing for error checking.

For the case of solid-state synthesis, just under 20 000 recipes were extracted from over 50 000 paragraphs, and these data include information on the material made, starting compounds, operations, and their conditions.[48] The distinguishing feature about these data, in addition to their breadth (13 000 unique targets and 16 000 unique reactions), was that the authors provide balanced chemical reactions that enable significant informatics work, at a scale not previously obtainable. Earlier work extracted synthesis parameters from the body text of 640 000 journal articles across 30 different oxide systems.[47] For zeolites, an industrially relevant catalysis material, 70 000 relevant articles were fed through an automated pipeline to extract gel-synthesis conditions. This work also included a highly curated set of 1200 synthesis routes that are specific to germanium-based zeolites.[105] These data were used to support comprehensive literature curation in

order to describe inter-zeolite transitions, affording an important opportunity for accessing new zeolitic structures.[109] The work by Hiszpanski and authors extracted synthesis and morphology information from 35 000 articles related to metallic nanomaterials, which enabled them to easily identify the types of nanomaterials that are of higher interest in the field. Furthermore, NLP-based extraction of this information from the broader literature enabled them to identify what specific chemical additions during synthesis result in the morphological differentiation of nanomaterials (i.e., resulting in nanosphere vs nanowire)—information that is otherwise typically gleaned through targeted, time-consuming, and iterative synthesis efforts by individual researchers.[25]

The materials-metrology database of optical absorption spectral characteristics consists of 18 309 records of chemical names, their experimentally determined λ_{max} values, and molar extinction coefficients, ε, where present. These were sourced from just over 400 000k academic papers using ChemDataExtractor.[33] The information density of data extraction (number of data records obtained: number of papers sampled) is quite low in this case, relative to the above examples of text extraction from documents. This is because the data sought on UV/vis absorption spectra nearly always take the form of core materials-characterization data, which support rather than leading to a paper. Accordingly, the information is semi-hidden in a paper or is entirely latent, often being relegated to the supplementary material of a paper. The data that do appear in the main article are highly fragmented and are somewhat elusive to keyword search terms. Moreover, materials-metrology data are reported over a particularly wide range of journals, compared with synthesis or materials-centered data. For example, there are journals that are dedicated to chemical synthesis, materials chemistry, or materials physics, such that it is facile to choose the journals to mine, which are rich in the content required to populate a database that suits a given application. In contrast, UV/vis absorption spectral characteristics will be noted in a paper of any journal that reports a new chemical product, which is optically absorbing, as well as being present in papers that focus on optical properties. The information density of data extraction is thus low, such that NLP tools must track many more papers for the desired outcome. This issue tracks a general trend that despite the highly pervasive nature of core materials-characterization data, such as UV/vis absorption spectra, they can be quite inaccessible to NLP tools.

Beyond, the datasets themselves are the capabilities to visualize them and comment on trends within them. For example, Hiszpanski *et al.* packaged the data that they extracted from 35 000 metallic nanomaterial synthesis articles into a distributable visualization tool that allows users to explore how the chemicals used in protocols vary depending on the targeted nanomaterial morphology and composition. For the case of the pulsed-laser deposition data, the extraction enabled visualization of growth windows, trends, and outliers (Fig. 3, ❶); these serve as an initial pathway for analyzing the distribution of growth conditions to act as feedback for first-principles calculations to link with thermodynamic stability windows. The authors extended their analysis to determine the likelihood of achieving a low, medium, or high T_c through a decision-tree classifier (a predictive modeling approach used in statistics).[21] Kim *et al.* observed that high calcination temperatures are found more frequently in the synthesis of bulk materials with greater elemental complexity.[27] Kononova *et al.* leveraged the reaction dataset for insights related to the nature of solid-state synthesis. For example, alkali and transition-metal cations are typically used in a reaction based on several types of precursors, including binary oxides, nitrides, sulfides, or simple salts such as carbonates, phosphates, and nitrates. They also observed that the counterion in solid-state synthesis controls the temperature of precursor melting or decomposition. This could indicate when the precursor becomes active during synthesis or direct the synthesis method.

The next level of depth within the materials examples that have leveraged NLP are those that perform some degree of ML on the data toward the pursuit of fundamental insights. Within the work by Court and Cole, case studies of perovskite-type oxides and pnictide superconductors demonstrated that magnetic and superconducting phase diagrams could be reconstructed with good accuracy (Fig. 3, ❶), and associated phase-transition temperature predictions could be made, which were relatable to the underlying physical theory of magnetism and superconductivity. Specifically, the authors were able to predict Néel temperatures in rare-earth manganites and orthochromites and document the unconventional superconductivity of ferropnictide superconductors, as well as predict T_c across the lanthanides. The models used elemental and structural features as a basis. While this contribution was for known compounds, the overall approach points to the ability to extend this capability to discovery.[29] Indeed, others have already used this NLP-generated magnetic-materials database, in concert with ML methods, to realize data-driven materials discovery.[110] Thereby, a new magnetic refrigerant, HoB_2, was successfully predicted (Fig. 3, ❷). This is an important discovery since there is currently a world-wide search for a material that exhibits an MCE around the hydrogen liquefaction temperature ($T = 20.3$ K), given the need for hydrogen storage to serve an energy-sustainable fuel industry.[111] AQ4

Methods based on quantitative structure-property relationships (QSPRs) are also being adapted. Such approaches are long-standing on the small scale, but multiple structure-property relationships are now being drawn together to analyze volumes of data. For example, a hierarchical sequence of questions with the generic form "Which data obey this QSPR?" can be set within an inverse pyramid construct of decision making to successively whittle down a large dataset to a few lead candidates that hold all of the requested QSPR requirements that suit a given material application (Fig. 3, ❸). The lead candidates that result from this materials screening process are, then, experimentally validated. For example, the database containing UV/vis absorption spectral characteristics was subjected to this hierarchical QSPR-based decision-making process, to successfully discover five light-harvesting materials for photovoltaic applications.[5] This work also illustrates how the NLP-based provision of materials databases can be embedded within a "design-to-device" pipeline for data-driven materials discovery.[112]

Owing to the nature of extracted data, the structuring of knowledge from an NLP-generated database offers interpretable ways of developing materials insight. For example, decision trees leveraging only extracted data can point to experimental handles that drive particular synthesis outcomes (Fig. 3, ❻). Decision trees have been used to examine the critical parameters that are needed to synthesize titania nanotubes via hydrothermal methods and verify the driving conditions of NaOH and temperature against known mechanisms. For the case of zeolites, data were used to generate a decision tree to predict zeolite synthesis conditions with low framework densities.[105] In addition, this work has demonstrated the capacity for learning across

materials classes using NLP-extracted information. This was done via, so-called, transfer-learning ML approaches, to predict synthesis outcomes on materials systems that were not included in the training set (Fig. 3, ❹); the results outperformed heuristic strategies. For example, in predicting the phase for $BiFeO_3$ (trained on $BaTiO_3$), an SVM, with the synthesis vector for this material as input, performed over 40% better than a heuristic logistic regression whose input was annealing temperature.[47]

More complex ML methods may also be applied to these data. For example, a subset of the authors have generated synthesis parameters based on observations from the literature, conditioned on specific synthesis-relevant parameters using generative ML models.[113] One class of generative models uses an autoencoder, which is a class of neural-network algorithms that learn to reproduce the identity function, while compressing data through a lower-dimensional layer. What makes this particular form of model generative, and therefore useful in making new material predictions, is an additional constraint (variational autoencoder) where the compressed space must also approximate a previous distribution. This model architecture enables a literature-based synthesis-screening technique to generate, for example, suggested synthesis parameters, accelerate positing of driving factors in forming rare phases, and identify correlations among intercalated ions and resulting synthesized polymorphs. These approaches have been applied to $SrTiO_3$, TiO_2, and MnO_2, due to their technological relevance in applications, ranging from energy storage to catalysis.[113] Most recently, the work has been extended to generate syntheses for perovskite materials (Fig. 3, ❺). Using only training data published over a decade prior to their first reported syntheses, the model generated precursors for $InWO_3$ and $PbMoO_3$, which were published in the literature a few years ago (2016 and 2017, respectively).[64] This work demonstrated that the NLP-based model learns representations of materials that correspond to synthesis-related properties, such as aqueous solubility, and that the behavior of the model complements existing thermodynamic knowledge. Data-augmentation strategies using the literature were also applied in this case, demonstrating the value of automated, comprehensive text extraction. Structured data from the literature may also initialize where experimental inquiry should start or seed the design of predictive tools for optimizing reaction procedures. Data that have been assembled in a structured way may lend themselves more effectively to develop reporting standards to inform reproducibility, or they may be made interoperable with other data within materials science or from broader disciplines.[55]

Finally, one might pursue NLP toward knowledge innovation. Linking knowledge discovery and NLP is a relatively new pursuit for the materials community. A recent example was to uncover semantic relations between concepts in a network for quantum physics.[78] This work used the content of 750 000 publications to generate a network of physical concepts where the links between two nodes were drawn when concurrently studied in research articles (Fig. 3, ❽). The authors examined the evolution of the network to identify emerging trends and the rate of those trends. The fastest growing concept found was the qubit, emerging first in 1995, which is the basic unit of quantum information. Another growing topic was found to be research in topological materials and, more recently, the application of machine learning. As far as suggestions of future topics, strong potential links were identified between orbital angular momentum and magnetic skyrmions and spin-orbital coupling. Another example is found in the materials-discovery domain. Taking a largely unsupervised approach, Tshitoyan and coauthors were able to extract implicit knowledge, held within the materials science community around the periodic table, and structure property relationships in materials, perhaps pointing to a way to examine new discoveries. This is a finding that is echoed in the original embedding work that was undertaken on general (nonscientific) text.[114] They have leveraged this capability to point to promising thermoelectric materials.[38]

This use of NLP to develop knowledge bases, from which to derive insight, is not too dissimilar to ontology creation (Fig. 3, ❼); whereby, there has been limited pursuit in the materials community. Ontologies are a formal presentation of a domain, and they provide an account of term meaning and insight into the hierarchical structure of the terms. Ontologies provide and formalize semantics of each entity and their specific domain. Ontologies are organized in formal machine-readable formats. This enables their integration in relation extraction models, and they may provide opportunities to learn ontologies for how materials information should be presented and what needs to be included. A recent effort in biomaterials generated an ontology to attempt to develop an accepted taxonomy for manufactured biomaterials; this captured the complexity of how scaffolds and devices are described and named. Examples of some of the superclasses generated were manufactured objects, biomaterials, material processing, effects on the biological system, and medical applications. The goals of this work were to provide an annotation resource to facilitate "term" (or "entity" in the NLP domain) recognition, outline "accepted" or used language in the field, and offer a common basis for understanding the range of distinct scaffolds with their associated features, beyond just the materials and document discovery.[77]

Table II summarizes some of the open data resources referenced in this section and highlights potential research directions enabled by these data. Despite the early nature of the application of NLP to materials science, these examples illustrate the breadth of what has been accomplished to date and the potential for knowledge creation and innovation as tools and methods mature.

VI. BEYOND BODY TEXT

In addition to extracting information from the main text of documents, valuable data that are embedded in figures and tables should also captured.[116,117] In a given manuscript, figures can include complex images, graphs, and schematics. While these figures, tables, and graphs provide a succinct representation of useful data that are relatively easy for humans to understand, the identification and collection of information from figures and tables to convert them into a structured format are significant challenges.[118,119] Similar to the way that NLP processes identify sections and relevant paragraphs, as mentioned above, the locations of figures and tables have also to be identified and extracted. Once the figures and tables have been extracted, segmentation, classification, and image analysis must be performed to extract relevant information that may need to be reconstructed. Successful extraction of data from the figures can reinforce and validate the information extracted from the main texts, provide additional data points, and aid in building relationships between multiple entities and numerical values. The information from figures and tables will allow the researchers to re-plot, compile, and quickly compare data across multiple sources and add newly obtained data, which can be visualized in

TABLE II. Examples of open data resources for NLP in materials science.

Data resource(s)	Data summary	Example usage
Materials word embeddings[38,64,115]	Word2Vec,[63] FastText,[65] and ELMo[66] word embeddings trained on materials text	Input features for entity recognition models
Annotated materials text[48,94]	(Human and machine) annotated plain-text synthesis paragraphs for materials	Training data for entity relation models or data mining for materials science insights
Text-mined Curie and Néel temperatures[51]	Text-mined database of magnetic compounds and their phase transition temperatures.	Training data for entity linking models that map material mentions and properties

a bigger context. One particularly challenging area of information extraction is from image-based data.

Microscopy images, which characterize the microscopic- to atomic-scale structure of materials, contain a wealth of information that would be useful in the design and understanding of functional materials. Figures in the scientific literature, which arise from image-based metrology, are predominantly sourced from scanning and transmission electron microscopy (SEM or TEM, respectively), as well as atomic force microscopy (AFM). The majority of such images are only discussed qualitatively in their surrounding text, despite the fact that the images contain a wide range of quantitative data on the structure of materials, such as particle size and shape, grain boundaries, crystal habits and crystal facets, material heterogeneity, and morphological diversity. These data could shed light on particularly important research problems that rely on nanotechnology or crystallography. Figure 4 shows the path for extraction of this information from text.

Image-recognition methods based on ML, Bayesian inference, and computer vision have been employed to analyze small datasets that address a bespoke problem in materials science. Efforts in the field of metallurgy are especially noteworthy in this regard. For example, convolutional neural networks (CNNs) have been applied to SEM images of ultrahigh carbon-based steel to analyze grain boundaries therein.[120] Microstructural features of steel, as displayed in SEM and optical microscopy images, have also been classified using CNNs[121] and SVMs.[122] More sophisticated data analytical tools have been applied to individual datasets of STEM and STM images, as befits their

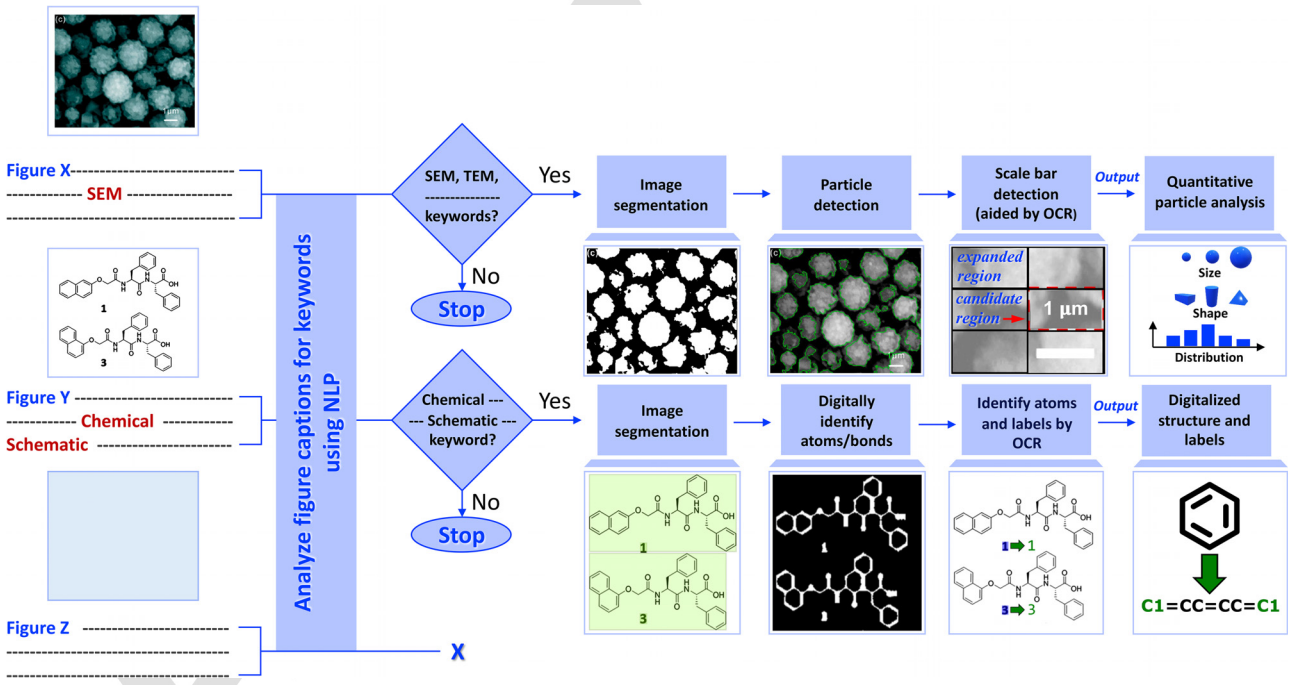

FIG. 4. Image extraction schematic including examples derived from microscopy images or molecular structures. Figures in the scientific literature, which arise from image-based metrology, are predominantly sourced from (scanning) transmission electron microscopy and atomic force microscopy. Most of these images are discussed qualitatively in their surrounding text despite the fact that the images contain a wide range of quantitative data on the structure of materials. These data could shed light on particularly important research problems that rely on, e.g., nanotechnology or crystallography. This figure suggests a path for extraction of this information from text.

greater value in terms of the much greater effort that is expended to produce these types of more specialized data. For example, STEM images that display defects in steels[123] or defects that cause structural transformations in tungsten sulfide[124] have been analyzed quantitatively using deep-learning methods. Interatomic interaction potentials have also been extracted from STM images using Bayesian inference.[125] However, none of these efforts are generalizable or scalable to the high-throughput data extraction and quantitative analysis of microscopy images, which is needed for data-driven approaches to materials physics.

The software tool, ImageDataExtractor,[126] begins to address this issue, shifting from assisting manual analysis of images to a generic tool that auto-extracts and quantifies microscopy images from documents. This tool executes an autonomous pipeline of image-recognition methods to detect particles in a series of microscopy images and quantify them in terms of shape, size, and radial distribution. Particles are detected by a sequential process of image binarization and thresholding, followed by a series of contour-detection algorithms. These algorithms use edge detection to identify all closed contours (particles), excluding any that are occluded by image annotation (e.g., particles that lie beneath the scale bar) or are truncated because they lie at the edge of an image, split apart particles that lie particularly close to each other, and refine contour detection using ellipse fitting where required. Particle sizes are determined via optical character recognition (OCR), which helps to detect and read the text in the scale bar of each image; this scale bar information is normalized with respect to the number of pixels in each image, in order to calculate the particle size. Super-resolution convolutional neural networks (SR-CNNs) are employed to assist the OCR of text in images where the image resolution is too low to identify text solely using the OCR engine, Tesseract 3.0.[127] The standard SR-CNN architecture[128,129] was modified specifically to suit ImageDataExtractor.[126] A radial distribution function that describes the particle-size variation is calculated, pending a sufficient number of particles that are detected on a given image. The shape of each particle is determined by comparing its aspect ratio and contour profile to that of reference data that depict common geometric shapes, using a similarity index.[130]

ImageDataExtractor can function in one of the two operational modes: it can either receive a series of images directly for immediate processing or work in concert with a specially integrated form of ChemDataExtractor[33] that uses its native "chemistry-aware" NLP capabilities to read figure captions of documents to identify microscopy images and then use ImageDataExtractor[126] to process them. If this second operational mode is used, ImageDataExtractor[126] employs a bespoke algorithm that splits apart figures within documents where they constitute panels of multiple images, such that individual microscopy images can be processed in the fashion described above.

More recently, Kim *et al.* have reported an image-recognition tool that identifies the size of nanomaterials and classifies the morphology of each nanomaterial into one of the four categories: nanocubes, nanoparticles, core-shell nanoparticles, and nanorods.[57,131] The particles are located by applying a distance transform-based segmentation process on a binarized form of the image, while their size estimation tracks a similar process to that of ImageDataExtractor.[126] Kim *et al.* identifies and extracted SEM and TEM images from the document via a different route to ImageDataExtractor,[126] employing a convolutional neural network (CNN) with transfer learning. Thereby, a small sample (<100) of SEM and TEM images was fed into the Inception-V3 CNN,[132] which has been pre-trained on pictures from several sources, including ImageNet.[133,134] The image features for SEM and TEM were extracted from the penultimate layer of the CNN, yielding a transfer-learning process with an 89% accuracy in SEM and TEM image classification.

Tatum *et al.* have also recently reported an image-recognition method that provides quantitative analysis of particles appearing specifically in images created by scanning probe microscopy (SPM) techniques, such as STM and AFM.[135] Particles are first detected using feature selection that is enabled by principal-component analysis (PCA); this clusters all data channels into the key representative structure of the image-based information. These clustered data are, then, classified using a Gaussian mixture model (GMM), which segments each pixel into distinct material phases; in the case study, the phases are structural domains of a polymer blend. This semantic segmentation method is, then, complemented by instance segmentation. This involves pixel-by-pixel clustering to characterize the size and distribution of each morphological domain in an image. Tatum *et al.* provided two possible image segmentation options to perform this task: connected component labeling or persistence watershed segmentation (PWS).[135] The former method assigns a domain label to each set of connected pixels, establishes the number of distinct domains that are present, and then places the domains in order of size. The latter method identifies the morphology of each domain using the height channels of the image to help distinguish the particle signal from that of the background. The PWS option tends to better identify isotropic domains, while the connected component method performs best in the characterization of highly anisotropic structural domains.

Another type of material information that is trapped inside figures of documents concerns chemical schematic diagrams (shown in the lower path of Fig. 4). This form of image is often the only means by which one or more organic chemical that is described in a document can be identified. A range of optical chemical structure recognition (OCSR) methods have been developed to interpret such images and convert them into computer-readable output, such as text. Kekulé,[136] CliDE (and its more recent version, CliDE Pro[137]), ChemReader,[138] OSRA,[139] and ChemSchematicResolver[140] all perform such a task. All use a common generic operational pipeline whereby an image figure is segmented into its structures and any surrounding text (e.g., chemical labels). The structure of each chemical schematic is, then, broken down into its bonds and atoms. There are various ways of achieving this goal, the most popular being thinning down the lines of the schematic to one-pixel in width and converting the result into a connected graph of nodes (atoms) and vertices (bonds). Optical character recognition (OCR) is used to interpret any atom names and chemical labels of a given structure. An algorithm may, then, be employed to match up any chemical labels to their associated structures. The resulting digitalized form of the chemical schematic is often converted into a simplified molecular input line entry system (SMILES)[141] text-string to provide the output. Such text output is readily interpretable using NLP tools.

The Kekulé software[136] is quite old, while the newer products, CliDE Pro[137] and ChemReader,[138] are not open-source tools. OSRA[139] is an open-source, but it is not suited to high-throughput data-mining, nor can it resolve generic substituents or atom labels (e.g., R-groups) in a chemical diagram or match chemical labels to the

diagrams. The ChemSchematicResolver[140] was built to incorporate OSRA while overcoming these limitations, as well as provide a framework that intrinsically links up to the NLP-capabilities (ChemDataExtractor).[33] This NLP link-up is important because it enables the ChemSchematicResolver to identify chemical schematic diagrams in the figure captions of documents in an autonomous manner so that they can be processed in a high-throughput fashion.

While the ability to automate domain-aware semantic linkages between figures and text remains an ongoing challenge, attention-based models[142,143] have shown promise for analyzing nonscientific images and describing their contents via captions. Applied to a materials context, such methods may be adapted to identify and annotate phases or locate defects within a micrograph.

VII. CHALLENGES AND OPPORTUNITIES

Many challenges still exist for information extraction and NLP in the materials domain, which stem largely from the complexity and heterogeneity of the text. For NLP specific to materials, there are challenges with transferability across materials domains given the high level of heterogeneity in the discipline, ranging across materials classes, application space, and even fundamental links between chemistry and physics. Since the volume of data within each of these individual domains may be relatively small, the accuracy of the models becomes critical so that data points are not lost as an extraction pipeline progresses. Of course, text extracted from the materials science literature is not and cannot be the only source of data leveraged by the informatics community. High-throughput experimental and computational data ported directly into informatics models still provide the most significant, high quality source of inputs to ML models. Text extracted information provides a supplement to these sources. In general, the challenges in use of NLP to "generate" and compile data are the age variety of quality of texts and the bias within the published literature based on the absence of negative examples.

Despite these challenges, there is potential (and need) to leverage the vast archive of information in published scientific text, toward the generation of new knowledge. For this to be successful, we must continue to push the boundary of what information can be extracted accurately and at scale, but we must also ensure that the extraction is done toward increased synergy with downstream ML algorithm development. One example of this synergy would be improvements in extraction, which are focused on transfer learning, whereby the language representations are pre-trained, in an unsupervised manner, on corpora and fine-tuned on a variety of specific materials questions for which there are fewer data. This will allow each specific area of research within materials science communities to work toward improving accuracy, while sharing the collected data for others to build-off of and to continue to grow the database and the collective information. Advances in entity linking, where entities within a text are automatically linked to databases of information, would also provide distinct synergistic opportunities to leverage fundamental physical knowledge to downstream ML activities.

One critical challenge in NLP is to draw linked information across a document, or the so-called non-local dependencies. To date, information extraction has focused on the use of sequential models that rely primarily on local dependencies. However, as experiments are described throughout a document, this is a significant limitation to reaching at scale accurate, automated extraction from the scientific text, particularly since we aim to extract information across body text, figures, images, and even the supplementary material. To date, this has mostly been done through post-processing activities by constraining the output space during inference, but automatically learning interactions between local and non-local dependencies would provide a significant opportunity to improve learning. One recent effort used a graph-based framework to represent a broad, cross-document set of word or sentence-level dependencies and define a data structure without access to any major processing or external resources.[144] This becomes NER at the discourse level (DiscNER), in contrast to sentence-level NER, where sentences are processed independently. This means that long-range dependencies have a crucial role in the tagging process and that they can be added as a soft constraint to improve information extraction. Given the challenge of labeling long-distance linkages within documents, unsupervised learning may prove useful toward advancing this branch of research. In language translation[145] and entity resolution,[88] the approach of aligning embeddings has proved effective in rapidly computing many unsupervised alignments (e.g., translations between English and Spanish) using a small amount—or sometimes zero—of labeled data.

As the scope and complexity of NLP models used in materials science increase, so too must the evaluation methods adapt. Recent results[146] in invariance testing for commercial NLP models have shown that invariances to typos, names, gender, and so on are not respected by many widely used NLP models. For example, changing the name of the employee in a customer review may affect a model's predicted sentiment, even though the true sentiment should be invariant to this. Such methods could be adapted to materials science: an NER model that correctly labels TiO_2 and $SrCO_3$ as precursors for $SrTiO_3$ should perform equally as well when the metals are exchanged (e.g., Ti with Fe).

Databases that unfold from NLP tasks may also be complemented by high-throughput calculations about materials; these predominantly take the form of electronic-structure calculations. At present, the computationally generated datasets that are afforded by these efforts are separated from experimental data, save for a few exceptions.[11,106,147] One of these exceptions[106] involved concerting NLP-based database auto-generation with high-throughput electronic-structure calculations on the materials that populated this database. This produced pairwise experimental and computational data on chemicals in the database. This synergy stands to be very powerful for a number of reasons. First, the comparison between pairwise experimental and computational data of a given material provides implicit quality control of a database; achieving the quality control of NLP-based auto-generated databases is a matter of concern that has been raised by various agencies.[148,149] Second, a good match between experimental and computational values will assure that wave functions of the electronic-structure calculations are correct; pending that to be the case, computation can, then, be used to calculate many additional properties about a given compound, with an assured reliability, to augment the contents of the materials database. In this sense, computational data have a distinct advantage over experimental data since the latter are naturally limited to the contents of the documents from which they were extracted by NLP. Third, such pairwise data can mitigate the common problem that important experimental data are often not available to suit a particular need in material physics in which case, computation provides a means to combat issues of missing data,

as long as the materials computed are similar to those of calculations that have been benchmarked against their pairwise experimental data. The collation of synergic experimental and computational data and their cohesive deposition into a data repository are nonetheless contingent on the availability of a suitably designed operational pipeline.

In broader terms, data management systems in materials science are starting to be developed for the automated processing and storing of data.[150,151] Some of these efforts involve robotics to aid the automation of materials characterization.[36,152,153] It is also being advocated and regulated that the data management of materials databases needs to attest to findable, accessible, interoperable, and reusable (FAIR) principles.[154] The increasing government regulations toward open-access data will also help journals capture data. These sorts of initiatives will help make data sources themselves more easily processed and analyzed, perhaps in raw data form. This aim is all but a pipe dream on a wide scale, at present, and even if such automation in data processing becomes normal in materials physics, NLP will still be in business for the long term. This is not only because of the huge amount of legacy data that already exist worldwide but also because it will likely never be practical to process raw data from highly specialist experiments automatically since the data analysis will be similarly specialized. NLP, therefore, has a bright future to continue to support automatic extraction from the literature.

VIII. CONCLUDING REMARKS

NLP and information extraction are early in their application to materials science. It will continue to require sustained effort to build domain-relevant extraction algorithms, scientific dependency parsers, annotation sets, and structures for disseminating extracted information. There are domain-specific needs regarding accuracy and ambiguity and tradeoffs to be weighed between the accuracy and degree of generalizability. However, we have shown that there is tremendous potential if we can unlock the troves of information within the primary way that we chose to communicate in the scientific community, through published, unstructured documents.

Throughout discussions of the rise of data in materials science, there is a dialog regarding encouraging researchers to deposit their own data. We must make sure that data continue to be disseminated in a way that provides direct compute operability;[155] infrastructure development within materials science needs to be in lockstep to allow that to happen. Given the potential for data science tools in accelerating the materials development process, data in general, and particularly freely available open data, need to undergo an inversion of priorities. Thus far, materials scientists have only considered humans familiar with their subject material as the audience for their published works. However, with application of NLP to materials science increasing, an entirely new audience should also be considered by authors: software tools. Unfortunately, the writing styles and data presentation formats that are often most interesting to the former can prove quite challenging to the latter. If we shift the pendulum toward data structures that enable compute capabilities, we will not only be able to better leverage the data revolution as materials scientists, and we will increase the reproducibility and comprehension of our output.

ACKNOWLEDGMENTS

E.A.O., O.K., and G.C. would like to acknowledge funding from the National Science Foundation under Award Nos. 1922311, 1922372, and 1922090 and DMREF and support from the Office of Naval Research (ONR) under Contract Nos. N00014–20-1–2280 and N00014–19-1–2114. E.A.O. also acknowledges support from the MIT Energy Initiative. Early work was collaborative under the Department of Energy's Basic Energy Science Program through the Materials Project under Grant No. EDCBEE. J.M.C. is grateful for the BASF/Royal Academy of Engineering Research Chair in Data-Driven Molecular Engineering of Functional Materials, which is partly supported by the STFC via the ISIS Neutron and Muon Source. O.K. and G.C. thank Energy and Biosciences Institute through the EBI-Shell Program and Assistant Secretary of Energy Efficiency and Renewable Energy, Vehicle Technologies Office, U.S. Department of Energy under Contract No. DE-AC02–05CH11231. T.Y.-J.H. and A.M.H. acknowledge the support of Lawrence Livermore National Laboratory, which is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract No. DE-AC52–07NA27344 and acknowledge the support of the LLNL-LDRD Program under Project No. 19-SI-001.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1]National Science and Technology Council, *Materials Genome Initiative for Global Competitiveness* (I[illegible]20[illegible]1).
[2]B. P. Abbott *et al.*, "[illegible] The laser interferometer gravitational-wave observatory," Rep. Prog. Phys. **72**, 76901 (2009).
[3]T. Accadia *et al.*, "Virgo: A laser interferometer to detect gravitational waves," J. Instrum. **7**, P03012 (2012).
[4]G. Longo, E. Merényi, and P. Tiňo, "Foreword to the focus issue on machine intelligence in astronomy and astrophysics," Publ. Astron. Soc. Pac. **131**, 100101 (2019).
[5]K. Albertsson *et al.*, "Machine learning in high energy physics community white paper," J. Phys. Conf. Ser. **1085**, 022008 (2018).
[6]A. O. Oliynyk *et al.*, "High-throughput machine-learning-driven synthesis of full-Heusler compounds," Chem. Mater. **28**, 7324–7331 (2016).
[7]A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for accelerated design of polymer dielectrics," Sci. Rep. **6**, 20952 (2016).
[8]R. Gómez-Bombarelli *et al.*, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," Nat. Mater. **15**, 1120–1127 (2016).
[9]C. B. Cooper *et al.*, "Design-to-device approach affords panchromatic co-sensitized solar cells," Adv. Energy Mater. **9**, 1802820 (2019).
[10]J. M. Cole *et al.*, "Data mining with molecular design rules identifies new class of dyes for dye-sensitised solar cells," Phys. Chem. Chem. Phys. **16**, 26684–26690 (2014).
[11]B. Blaiszik *et al.*, "The materials data facility: Data services to advance materials science research," J. Miner., Met. Mater. Soc. **68**, 2045–2052 (2016).
[12]S. Curtarolo *et al.*, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations," Comput. Mater. Sci. **58**, 227–235 (2012).
[13]A. Dima *et al.*, "Informatics infrastructure for the materials genome initiative," J. Miner., Met. Mater. Soc. **68**, 2053–2064 (2016).
[14]J. O'Mara, B. Meredig, and K. Michel, "Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access," J. Miner., Met. Mater. Soc. **68**, 2031–2034 (2016).
[15]A. Jain *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 11002 (2013).

AQ5

[16]S. Tinkle, D. L. McDowell, A. Barnard, F. Gygi, and P. B. Littlewood, "Sharing data in materials science," Nature **503**, 463 (2013).
[17]National Research Council, *High Magnetic Field Science and Its Applications in the United States: Current Status and Future Direction* (National Academies Press, 2013).
[18]National Science and Technology Council Committee on Technology, *National Nanotechnology Initiative Strategic Plan* (■, 2016).
[19]Basic Energy Sciences Advisory Committee, *Report of the BESAC Subcommittee on Future X-Ray Light Sources* (U.S. Department of Energy, 2013).
[20]■, *Next-Generation Photon Sources for Grand Challenges in Science and Energy: Report of the Workshop on Solving Science and Energy Grand Challenges with Next-Generation Photon Sources* (U.S. Department of Energy, 2009).
[21]National Academies of Sciences, Engineering and Medicine, *Frontiers of Materials Research: A Decadal Survey* (■, 2019).
[22]See https://search.datacite.org/ for DataCite: Find, access, and reuse data; accessed 7 June 2020.
[23]X. Jia *et al.*, "Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis," Nature **573**, 251–255 (2019).
[24]S. Fortunato *et al.*, "Science of science," Science **359**, eaao0185 (2018).
[25]L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," J. Am. Soc. Inf. Sci. Technol. **66**, 2215–2222 (2015).
[26]A. Zeng *et al.*, "The science of science: From the perspective of complex systems," Phys. Rep. **714–715**, 1–73 (2017).
[27]R. Leaman and G. Gonzalez, "BANNER: An executable survey of advances in biomedical named entity recognition," Pacific Symposium on Biocomputing 2008, PSB 2008 (2008), pp. 652–663.
[28]A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," Briefings Bioinf. **6**, 57–71 (2005).
[29]See https://pubmed.ncbi.nlm.nih.gov/ for PubMed.
[30]See https://www.elsevier.com/solutions/reaxys for Reaxys.
[31]R. Leaman, C. H. Wei, and Z. Lu, "TmChem: A high performance approach for chemical named entity recognition and normalization," J. Cheminf. **7**, 1–10 (2015).
[32]T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: A hybrid system for chemical named entity recognition," Bioinformatics **28**, 1633–1640 (2012).
[33]M. C. Swain and J. M. Cole, "ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature," J. Chem. Inf. Model. **56**, 1894–1904 (2016).
[34]D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust, "OSCAR4: A flexible architecture for chemical textmining," J. Cheminf. **3**, 41 (2011).
[35]R. W. Epps *et al.*, "Artificial chemist: An autonomous quantum dot synthesis bot," Adv. Mater. **32**, 2001626 (2020).
[36]B. P. MacLeod *et al.*, "Self-driving laboratory for accelerated discovery of thin-film materials," Sci. Adv. **6**, eaaz8867 (2020).
[37]L. Weston *et al.*, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," J. Chem. Inf. Model. **59**, 3692 (2019).
[38]V. Tshitoyan *et al.*, "Unsupervised word embeddings capture latent knowledge from materials science literature," Nature **571**, 95–98 (2019).
[39]J. G. Foster, A. Rzhetsky, and J. A. Evans, "Tradition and innovation in scientists' research strategies," Am. Sociol. Rev. **80**, 875–908 (2015).
[40]A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans, "Choosing experiments to accelerate collective discovery," Proc. Natl. Acad. Sci. U. S. A. **112**, 14569–14574 (2015).
[41]J. D. Dworkin, R. T. Shinohara, and D. S. Bassett, "The landscape of neuroimage-ing research," NeuroImage **183**, 872–883 (2018).
[42]E. Beam, L. G. Appelbaum, J. Jack, J. Moody, and S. A. Huettel, "Mapping the semantic structure of cognitive neuroscience," J. Cognit. Neurosci. **26**, 1949–1965 (2014).
[43]S. Milojević, "Quantifying the cognitive extent of science," J. Informetrics **9**, 962–973 (2015).
[44]I. Iacopini, S. Milojević, and V. Latora, "Network dynamics of innovation processes," Phys. Rev. Lett. **120**, 48301 (2018).
[45]P. Murray-Rust, J. A. Townsend, S. E. Adams, W. Phadungsukanan, and J. Thomas, "The semantics of chemical markup language (CML): Dictionaries and conventions," J. Cheminf. **3**, 43 (2011).
[46]C. Ramakrishnan, A. Patnia, E. Hovy, and G. A. P. C. Burns, "Layout-aware text extraction from full-text PDF of scientific articles," Source Code Biol. Med. **7**, 7 (2012).
[47]E. Kim *et al.*, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chem. Mater. **29**, 9436–9444 (2017).
[48]O. Kononova *et al.*, "Text-mined dataset of inorganic materials synthesis recipes," Sci. Data **6**, 203 (2019).
[49]D. M. Jessop, S. E. Adams, and P. Murray-Rust, "Mining chemical information from open patents," J. Cheminf. **3**, 41 (2011).
[50]S. A. Akhondi *et al.*, "Automatic identification of relevant chemical compounds from patents," Database **2019**, baz001.
[51]C. J. Court and J. M. Cole, "Auto-generated materials database of Curie and Neél temperatures via semisupervised relationship extraction," Sci. Data **5**, 180111 (2018).
[52]D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Pearson Prentice Hall, 2009).
[53]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805 (2018).
[54]V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv:1910.01108 (2019).
[55]E. Kim, K. Huang, O. Kononova, G. Ceder, and E. Olivetti, "Distilling a materials synthesis ontology," Matter **1**, 8–12 (2019).
[56]H. Huo *et al.*, "Semi-supervised machine-learning classification of materials synthesis procedures," NPJ Comput. Mater. **5**, 1–7 (2019).
[57]A. M. Hiszpanski *et al.*, "Nanomaterials synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge," J. Chem. Inf. Model. **60**, 2876 (2020).
[58]M. Krallinger *et al.*, "CHEMDNER: The drugs and chemical names extraction challenge," J. Cheminf. **7**, 1–11 (2015).
[59]E. F. T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," arXiv:cs/0306050 (2003).
[60]D. M. Lowe and R. A. Sayle, "LeadMine: A grammar and dictionary driven approach to entity recognition," J. Cheminf. **7**, 1–9 (2015).
[61]L. Hawizy, D. M. Jessop, N. Adams, and P. Murray-Rust, "ChemicalTagger: A tool for semantic text-mining in chemistry," J. Cheminf. **3**, 17 (2011).
[62]X. Dai, S. Karimi, B. Hachey, and C. Paris, "Using similarity measures to select pretraining data for NER," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019), pp. 1460–1470.
[63]T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advance Neural Information Processing Systems* (■, 2013), pp. 3111–3119.
[64]E. Kim *et al.*, "Inorganic materials synthesis planning with literature-trained neural networks," J. Chem. Inf. Model. **60**, 1194 (2020).
[65]P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017).
[66]M. Peters *et al.*, "Deep contextualized word representations," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (2018), pp. 2227–2237.
[67]I. Beltagy, A. Cohan, and K. Lo, "SciBERT: Pretrained contextualized embeddings for scientific text," arXiv:1903.10676 (2019).
[68]D. J. Audus and J. J. De Pablo, "Polymer informatics: Opportunities and challenges," ACS Macro Lett. **6**, 1078–1082 (2017).
[69]R. B. Tchoua *et al.*, "Creating training data for scientific named entity recognition with minimal human effort," Lect. Notes Comput. Sci. **11536**, 398–411 (2019).

[70]C. Seifert *et al.*, "Crowdsourcing fact extraction from scientific literature," in *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (Springer, 2013), pp. 160–172.

[71]J. Takis, A. Q. M. S. Islam, C. Lange, and S. Auer, "Crowdsourced semantic annotation of scientific publications and tabular data in PDF," in Proceedings of the 11th International Conference on Semantic Systems (2015), pp. 1–8.

[72]R. Tchoua *et al.*, "Active learning yields better training data for scientific named entity recognition," in Proceedings of the IEEE 15th International Conference on eScience, eScience 2019 (2019), pp. 126–135.

[73]L. Huang and C. Ling, "Representing multiword chemical terms through phrase-level preprocessing and word embedding," ACS Omega **4**, 18510–18519 (2019).

[74]X. Gao, R. Tan, and G. Li, "Research on text mining of material science based on natural language processing," IOP Conf. Ser. Mater. Sci. Eng. **768**, 72094 (2020).

[75]D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (2014), pp. 2335–2344.

[76]E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in Proceedings of the Fifth ACM Conference on Digital Libraries (2000), pp. 85–94.

[77]O. Hakimi *et al.*, "The devices, experimental scaffolds, and biomaterials ontology (DEB): A tool for mapping, annotation, and analysis of biomaterials data," Adv. Funct. Mater. **30**, 1909910–1909913 (2020).

[78]M. Krenn and A. Zeilinger, "Predicting research trends with semantic and neural networks with an application in quantum physics," Proc. Natl. Acad. Sci. U. S. A. **117**, 1910 (2019).

[79]M. Khabsa and C. L. Giles, "Chemical entity extraction using CRF and an ensemble of extractors," J. Cheminf. **7**, S12 (2015).

[80]P. Mitra, C. L. Giles, B. Sun, and Y. Liu, "Chemxseer: A digital library and data repository for chemical kinetics," in Proceedings of the ACM First Workshop on CyberInfrastructure: information Management in eScience (2007), pp. 7–10.

[81]Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "Tableseer: Automatic table metadata extraction and searching in digital libraries," in Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (2007), pp. 91–100.

[82]D. M. Lowe, N. M. O'Boyle, and R. A. Sayle, "Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall," Database **2016**, baw039.

[83]S. Bird, E. Loper, and E. Klein, see http://www.nltk.org for Natural language toolkit, 2009.

[84]See https://spacy.io/ for SpaCy.

[85]See https://stanfordnlp.github.io/CoreNLP/ for CoreNLP.

[86]See https://allennlp.org/ for AllenNLP.

[87]See https://opennlp.apache.org/ for OpenNLP.

[88]M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and N. Tang, "DeepER—Deep entity resolution," arXiv:1710.00597 (2017).

[89]S. Mudgal *et al.*, "Deep learning for entity matching: A design space exploration," in Proceedings of the 2018 International Conference on Management of Data (2018), pp. 19–34.

[90]See https://brat.nlplab.org/ for BRAT.

[91]See https://prodi.gy/ for Prodigy.

[92]See https://webanno.github.io/webanno/ for Webanno.

[93]See http://mitre.github.io/callisto/ for Callisto.

[94]S. Mysore *et al.*, "The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures," in Proceedings of the 13th Linguistic Annotation Workshop (2019).

[95]F. Kuniyoshi, K. Makino, J. Ozawa, and M. Miwa, "Annotating and extracting synthesis process of all-solid-state batteries from scientific literature," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) (2020).

[96]A. Friedrich *et al.*, "The SOFC-Exp corpus and neural approaches to information extraction in the materials science domain," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020), pp. 1255–1268.

[97]C. Kulkarni, W. Xu, A. Ritter, and R. Machiraju, "An annotated corpus for machine reading of instructions in wet lab protocols," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (2018), pp. 97–106.

[98]C. A. Aguirre, S. Coen, M. F. De La Torre, W. H. Hsu, and M. Rys, "Towards faster annotation interfaces for learning to filter in information extraction and search," in CEUR Workshop Proceedings (2018), Vol. 2068.

[99]See https://docs.bokeh.org/en/latest/index.html for Candela.

[100]See https://docs.bokeh.org/en/latest/index.html for Bokeh.

[101]See https://c3js.org/examples.html for D3.

[102]C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, "Polymer genome: A data-powered polymer informatics platform for property predictions," J. Phys. Chem. C **122**, 17575–17585 (2018).

[103]S. R. Young *et al.*, "Data mining for better material synthesis: The case of pulsed laser deposition of complex oxides," J. Appl. Phys. **123**, 1–11 (2018).

[104]E. Kim *et al.*, "Machine-learned and codified synthesis parameters of o materials," Sci. Data **4**, 170127 (2017).

[105]Z. Jensen *et al.*, "A machine learning approach to zeolite synthesis enabled by automatic literature data extraction," ACS Cent. Sci. **5**, 892 (2019).

[106]E. J. Beard, G. Sivaraman, Á. Vázquez-Mayagoitia, V. Vishwanath, and J. M. Cole, "Comparative dataset of experimental and computational attributes of UV/vis absorption spectra," Sci. Data **6**, 1–11 (2019).

[107]R. B. Tchoua *et al.*, "Towards a hybrid human-compu entific information extraction pipeline," in Proceedings of the IEEE 13th International Conference on eScience, eScience 2017 (2017), pp. 109–118.

[108]See https://maldi.nist.gov/ for MALDI.

[109]D. Schwalbe-Koda, Z. Jensen, E. Olivetti, and R. Gómez-Bombarelli, "Graph similarity drives zeolite diffusionless transformations and intergrowth," Nat. Mater. **18**, 1177–1181 (2019).

[110]P. B. de Castro *et al.*, "Machine-learning-guided discovery of the gigantic magnetocaloric effect in HoB_2 near the hydrogen liquefaction temperature," NPG Asia Mater. **12**, 1–7 (2020).

[111]L. W. Jones, "Liquid hydrogen as a fuel for the future," Science **174**, 367–370 (1971).

[112]J. M. Cole, "A design-to-device pipeline for data-driven materials discovery," Acc. Chem. Res. **53**, 599–610 (2020).

[113]E. Kim, K. Huang, S. Jegelka, and E. Olivetti, "Virtual screening of inorganic materials synthesis parameters with deep learning," NPJ Comput. Mater. **3**, 53 (2017).

[114]J. B. Voytek and B. Voytek, "Automated cognome construction and semi-automated hypothesis generation," J. Neurosci. Methods **208**, 92–100 (2012).

[115]E. Kim *et al.*, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chem. Mater. **29**, 9436–9444 (2017).

[116]D. Jung *et al.*, "ChartSense: Interactive data extraction from chart images," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (2017), pp. 6706–6717.

[117]X. Liu, D. Klabjan, and P. NBless, "Data extraction from charts via single deep neural network," arXiv:1906.11906 (2019).

[118]L. Gao, X. Yi, Z. Jiang, L. Hao, and Z. Tang, "ICDAR2017 competition on page object detection," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (IEEE, 2017), Vol. 1, pp. 1417–1422.

[119]L. Gao *et al.*, "ICDAR 2019 competition on table detection and recognition (CTDAR)," in *2019 International Conference on Document Analysis and Recognition (ICDAR)* (IEEE, 2019), pp. 1510–1515.

[120]B. L. DeCost, B. Lei, T. Francis, and E. A. Holm, "High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel," arXiv:1805.08693 (2018).

[121]S. M. Azimi, D. Britz, M. Engstler, M. Fritz, and F. Mücklich, "Advanced steel microstructural classification by deep learning methods," Sci. Rep. **8**, 1–14 (2018).

[122]J. Gola *et al.*, "Objective microstructure classification by support vector machine (SVM) using a combination of morphological parameters and textural features for low carbon steels," Comput. Mater. Sci. **160**, 186–196 (2019).

[123]G. Roberts *et al.*, "Deep learning for semantic segmentation of defects in advanced stem images of steels," Sci. Rep. **9**, 12744 (2019).

[124]A. Maksov *et al.*, "Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS 2," NPJ Comput. Mater. **5**, 12 (2019).
[125]L. Vlcek, A. Maksov, M. Pan, R. K. Vasudevan, and S. V. Kalinin, "Knowledge extraction from atomically resolved images," ACS Nano **11**, 10313–10320 (2017).
[126]K. T. Mukaddem, E. J. Beard, B. Yildirim, and J. M. Cole, "ImageDataExtractor: A tool to extract and quantify data from microscopy images," J. Chem. Inf. Model. **60**, 2492 (2020).
[127]R. Smith, "An overview of the Tesseract OCR engine,"in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (IEEE, 2007), Vol. 2, pp. 629–633.
[128]C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Trans. Pattern Anal. Mach. Intell. **38**, 295–307 (2016).
[129]C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision* (Springer, 2014), pp. 184–199.
[130]M.-K. Hu, "Visual pattern recognition by moment invariants," IRE Trans. Inf. Theory **8**, 179–187 (1962).
[131]H. Kim, J. Han, and T. Y.-J. Han, "Machine vision-driven automatic recognition of particle size and morphology in SEM images," Nanoscale **12**, 19461–19469 (2020).
[132]C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 2818–2826.
[133]X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)* (IEEE, 2017), pp. 783–787.
[134]E. Tran, M. B. Mayhew, H. Kim, P. Karande, and A. D. Kaplan, "Facial expression recognition using a large out-of-context dataset," in *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (IEEE, 2018), pp. 52–59.
[135]W. Tatum *et al.*, "A generalizable framework for algorithmic interpretation of thin film morphologies in scanning probe images," J. Chem. Inf. Model. **60**, 3387 (2020).
[136]J. R. McDaniel and J. R. Balmuth, "Kekule: OCR-optical chemical (structure) recognition," J. Chem. Inf. Comput. Sci. **32**, 373–378 (1992).
[137]A. T. Valko and A. P. Johnson, "CLiDE Pro: The latest generation of CLiDE, a tool for optical chemical structure recognition," J. Chem. Inf. Model. **49**, 780–787 (2009).
[138]J. Park *et al.*, "Automated extraction of chemical structure information from digital raster images," Chem. Cent. J. **3**, 4 (2009).
[139]I. V. Filippov and M. C. Nicklaus, "Optical structure recognition software to recover chemical information: OSRA, an open source solution," J. Chem. Inf. Model. **49**, 740–743 (2009).
[140]E. J. Beard and J. M. Cole, "ChemSchematicResolver: A toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities," J. Chem. Inf. Model. **60**, 2059 (2020).
[141]D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[142]P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 6077–6086.
[143]K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in International Conference on Machine Learning (2015), pp. 2048–2057.
[144]Y. Qian, E. Santus, Z. Jin, J. Guo, and R. Barzilay, "GraphIE: A graph-based framework for information extraction." arXiv:1810.13083 (2018).
[145]A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," arXiv:1710.04087 (2017).
[146]M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with checklist," arXiv:2005.04118 (2020).
[147]See mits.nims.go.jp for NIMS Materials Data Base (MatNavi).
[148]A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," IEEE Intell. Syst. **24**, 8–12 (2009).
[149]J. S. Grosman *et al.*, "Eras: Improving the quality control in the annotation process for natural language processing tasks," Inf. Syst. **93**, 101553 (2020).
[150]A. Zakutayev *et al.*, "An open experimental database for exploring inorganic materials," Sci. Data **5**, 180053 (2018).
[151]P. Nikolaev, D. Hooper, N. Perea-Lopez, M. Terrones, and B. Maruyama, "Discovery of wall-selective carbon nanotube growth conditions via automated experimentation," ACS Nano **8**, 10214–10222 (2014).
[152]Z. Li *et al.*, "Robot-accelerated perovskite investigation and discovery (RAPID): 1. Inverse temperature crystallization," chemRxiv.
[153]R. J. Kearsey, B. M. Alston, M. E. Briggs, R. L. Greenaway, and A. I. Cooper, "Accelerated robotic discovery of type II porous liquids," Chem. Sci. **10**, 9454–9465 (2019).
[154]M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," Sci. Data **3**, 160018 (2016).
[155]A. M. Clark, A. J. Williams, and S. Ekins, "Machines first, humans second: On the importance of algorithmic interpretation of open chemistry data," J. Cheminf. **7**, 9 (2015).