Nanomaterials Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge

Anna Hiszpanski, Brian Gallagher, Karthik Chellapan, Peggy Li, Shusen Liu, Hyojin Kim, Jinkyu Han, Bhavya Kailkhura, David Buttler, Yong Han

Submitted date: 09/11/2019 • Posted date: 20/11/2019
Licence: CC BY-NC-ND 4.0
Citation information: Hiszpanski, Anna; Gallagher, Brian; Chellapan, Karthik; Li, Peggy; Liu, Shusen; Kim, Hyojin; et al. (2019): Nanomaterials Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge. ChemRxiv. Preprint.
https://doi.org/10.26434/chemrxiv.10278164.v1

Nanomaterials of varying compositions and morphologies are of interest for many applications from catalysis to optics, but the synthesis of nanomaterials and their scale-up are most often time-consuming and Edisonian processes. Information gleaned from scientific literature can help inform and accelerate nanomaterials development, but again, searching the literature and digesting the information are time-consuming manual processes for researchers. To help address these challenges, we developed scientific article-processing tools that extract and structure information from the text and figures of nanomaterials articles, thereby enabling the creation of a personalized knowledgebase for nanomaterials synthesis that can be mined to help inform further nanomaterials development. Starting with a corpus of ca. 35k nanomaterials-related articles, we developed models to classify articles according to the nanomaterial composition and morphology, extract synthesis protocols from within the articles’ text, and extract, normalize, and categorize chemical terms within synthesis protocols. We demonstrate the efficiency of the proposed pipeline on an expert-labeled set of nanomaterials synthesis articles, achieving 100% accuracy on composition prediction, 95% prediction on morphology prediction, 0.99 AUC on protocol identification, and up to 0.87 F1-score on chemical entity recognition. In addition to processing articles’ text, microscopy images of nanomaterials within articles are also automatically identified and analyzed to determine nanomaterials’ morphologies and size distributions. To enable users to easily explore the database, we developed a complementary browser-based visualization tool that provides flexibility in comparing across subsets of articles of interest. We use these tools and information to identify trends in nanomaterials synthesis, such as the correlation of certain reagents with various nanomaterial morphologies, which is useful in guiding hypotheses and reducing the potential parameter space during experimental design.
Nanomaterials Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge

Anna M. Hiszpanski,1 Brian Gallagher,2 Karthik Chellappan,3† Peggy Li,3 Shusen Liu,2 Hyojin Kim,2 Jinkyu Han,1 Bhavya Kailkhura,2 David J. Buttler,2 T. Yong-Jin Han1*  
1 Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA 94550  
2 1@llnl.gov, Lawrence Livermore National Laboratory, Livermore, CA 94550  
3 Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

ABSTRACT: Nanomaterials of varying compositions and morphologies are of interest for many applications from catalysis to optics, but the synthesis of nanomaterials and their scale-up are most often time-consuming and Edisonian processes. Information gleaned from scientific literature can help inform and accelerate nanomaterials development, but again, searching the literature and digesting the information are time-consuming manual processes for researchers. To help address these challenges, we developed scientific article-processing tools that extract and structure information from the text and figures of nanomaterials articles, thereby enabling the creation of a personalized knowledgebase for nanomaterials synthesis that can be mined to help inform further nanomaterials development. Starting with a corpus of ca. 35k nanomaterials-related articles, we developed models to classify articles according to the nanomaterial composition and morphology, extract synthesis protocols from within the articles’ text, and extract, normalize, and categorize chemical terms within synthesis protocols. We demonstrate the efficiency of the proposed pipeline on an expert-labeled set of nanomaterials synthesis articles, achieving 100% accuracy on composition prediction, 95% prediction on morphology prediction, 0.99 AUC on protocol identification, and up to 0.87 F1-score on chemical entity recognition. In addition to processing articles’ text, microscopy images of nanomaterials within articles are also automatically identified and analyzed to determine nanomaterials’ morphologies and size distributions. To enable users to easily explore the database, we developed a complementary browser-based visualization tool that provides flexibility in comparing across subsets of articles of interest. We use these tools and information to identify trends in nanomaterial synthesis, such as the correlation of certain reagents with various nanomaterial morphologies, which is useful in guiding hypotheses and reducing the potential parameter space during experimental design.

INTRODUCTION

Machine learning promises to accelerate the rate of materials science discovery for a variety of applications – including catalysis, thermoelectrics, photovoltaics, batteries, and pharmaceutics – by helping to reveal the underlying chemistry-processing-structure-function relationships at the heart of materials development.1 However, leveraging these tools requires large and structured materials data sets, which, aside from crystallographic2–5 and computational6 databases, are not broadly available in materials science. Indeed, the “development of shared public datasets and environments for AI training and testing” was identified as one of the seven key components of National Artificial Intelligence Research and Development Strategic Plan7 and was also a key priority of the Materials Genome Initiative. A number of groups and consortia are actively working to address this need with the formation of more general and all-encompassing collaborative materials databases and repositories.8–21 However, while there is strong consensus across the materials science community for the need to aggregate and collect results,18–21 there is less consensus about the precise form that such databases should have, given the diversity of materials data and the diversity of materials’ end-use applications.

Nonetheless, the scientific community has generally agreed on one pseudo-standardized form of data across varying subfields: publications. While text is unstructured data, which creates its own complications, scientists have generally agreed on the format (e.g. abstract, experimental details, discussion of results) and the level of detail to include. If targeted information can be automatically extracted from publications, materials databases can be populated much more easily without requiring manual data entry from individual researchers and general consensus on an all-inclusive and comprehensive data format. To help execute on Department of Energy (DOE) missions, Lawrence Livermore National Laboratory (LLNL) has invested in the development of automated extraction and processing tools for large volumes of unstructured data having various forms,22–23 including news articles,24 images,25–26 and audio.27 Leveraging these tools and expertise, our vision is to create tools that enable scientific users to automatically create their own personal databases from the scientific literature. Data visualization, data mining, and data analytics using machine learning algorithms can then be readily applied to these custom structured databases to meet the needs of the end-user to gain insights and elucidate underlying relationships from complex multimodal data. We envision that these individual and personalized databases can then be shared and form a peer-to-peer network creating large federated databases to accelerate materials science discovery and development.
Indeed, by manually extracting information of interest from the existing literature, several groups have already demonstrated the utility of creating custom databases and gained new insights from them. For example, manually extracting and aggregating information on organic solar cells’ performance from a mere 80 papers, Jackson et al. found that careful engineering of organic semiconductors’ energy levels does not actually correlate with higher power conversion efficiency in devices, opposite to much of the conventional wisdom in the field. By manually extracting the reported properties of various types of thermoelectric materials from 100 papers, Gaultois et al. identified key material attributes that are correlated with high performance, thereby creating helpful guidelines for future materials exploration and discovery. Similarly, Ghadbeigi et al. manually parsed 200 publications to evaluate existing cathode and anode materials in lithium ion batteries and developed guidelines for new materials to improve battery performance and cost.

By automating the processing of data extraction and structuring, orders of magnitude more papers can be searched and larger databases constructed than what is feasible by manual processing, thereby providing an even richer space for data mining and data analytics. IBM provided an illustrative example with Chef Watson, a cognitive computing application that combed 10,000 recipes from Bon Appétit to learn relationships among ingredients and dishes, and that can propose recipes for novel dishes. Though the application spaces are clearly quite different, the challenges of synthesizing new edible dishes and new scientific materials are similar. For example, the Olivetti group demonstrated the extraction of synthesis and processing details of metal oxides from ca. 13,000 papers, which then enabled them to correlate the synthesis of specific metal oxide products with the use of certain solvents and processing temperatures, and extended this analysis to zeolite synthesis as well. Similarly, Young et al. used text data mining from the literature to establish relationships between the synthesis parameters of oxide films formed by pulsed laser deposition and their functional properties. More recently, the Ceder and Jain groups have parsed literature to extract synthesis recipes for inorganics and demonstrated that word embeddings can be used to capture latent knowledge and relationships amongst materials.

While automatic extraction of information from the scientific literature has been demonstrated in the handful of cases discussed above, this approach has not yet been applied to metallic nanomaterials synthesis. Nanomaterials are critical for a number of applications, including catalysis, optical components, and additive manufacturing feedstocks. Beyond affecting the chemical composition of the materials, the details of the synthesis can also affect the nanomaterials’ morphology and size – both of which are often critical for the ultimate function and utility of these materials. In this work, we present a suite of tools, illustrated in Figure 1, for automatically extracting and structuring targeted information for nanomaterials – specifically, synthesis details and nanomaterials characteristics – from published scientific articles and demonstrate the kinds of insights such information can provide.

To gain insights, we find that simply having tools to extract and structuring information into a database is not enough to be of utility for materials scientists; tools that enable easy data visualization, exploration, and hypothesis-testing are essential to realizing the full potential of database creation from the scientific literature. For this reason, we have packaged and are distributing the information extracted from the ca. 35k articles in a database visualization tool that allows a user to look holistically at the state of the nanomaterials field and also dive into the details of specific articles (available from https://github.com/LLNL/MI-ChemVis). We show how extracting from the literature details of what nanomaterials have been synthesized can provide an immediate bird’s eye view of the current state of the field and enable us to quickly ascertain if a target nanomaterial has already been successfully developed or not. By further extracting the details of these nanomaterials’
syntheses, researchers can begin to look for patterns amongst synthesis schemes that may aid in materials optimization. As specific examples, we identify reagents that appear critical to attaining morphological differentiation for Ag and Au nanoparticles (e.g. nanocube vs. nanosphere).

RESULTS AND DISCUSSION

Our process, an overview of which is provided in Figure 1, consists of several steps, including: 1) building a relevant corpus of nanomaterials articles; 2) extracting metadata on each article (i.e., title, authors, doi); 3) processing the text of each article to identify the target nanomaterial’s morphology and composition, synthesis procedure, and chemicals used in its synthesis; and 4) extracting figures and further processing SEM/TEM figures of nanomaterials to obtain their morphology and size distributions. The information extracted from the text of articles is then aggregated to create a nanomaterials database (available from https://github.com/LLNL/MI-ChemVis) that can then be mined to identify patterns and generate hypotheses.

Building a relevant corpus

Elsevier’s catalog of journals served as the basis to building our corpus as Elsevier provides bulk access to full-text articles for text mining purposes. The full text of articles (not limited to keywords or abstract) were searched for those that contain “X nanoY” and “synthesis.” Here X indicates the nanomaterial compositions of interest and specifically included gold, silver, copper, iron oxide, palladium, and platinum; Y indicates the nanomaterial morphology of interest and specifically included wire, cube, triangle, rod, particle, sphere, crystal, sheet, star, structure, ribbon, flower, and tube. These search terms yielded a corpus of 35,345 unique papers. Given that we limited ourselves to Elsevier’s library, we assume that all ca. 35k articles resulting from these search terms are relevant to our interest in nanomaterials synthesis, but should a more catch-all approach be employed, like web scraping, then a classifier may be needed to filter for relevant articles prior to further processing.

Identifying composition and morphology from text

To accurately assign labels for each article’s targeted nanomaterial composition and morphology, we implemented an unsupervised classification algorithm based on the “term frequency-inverse document frequency” or TF-IDF statistic for each word in each article. Detailed in the Supporting Information, this statistic combines a word’s frequency within a single document and its frequency across all documents. A high TF-IDF indicates that a word is frequently and uniquely used in a specific article, suggesting greater relevance. The single composition and morphology terms with the highest TF-IDF weights are then assigned as the topical label for each document.

To evaluate this approach, we used a set of 99 papers that we henceforth refer to as our “gold standard” and that we hand-labeled with the appropriate nanomaterial composition and morphology (see Supplementary Information for list of papers and labels). When applied to the gold standard, the unsupervised TF-IDF approach obtained 100% accuracy on composition prediction and 95% accuracy on morphology prediction.

Using this method, composition and morphology labels were created for all 35k articles in our corpus. We note that 19 papers from our gold standard set overlapped with our 35k corpus, so we removed these 19 papers, resulting in a corpus of 35,326 articles. Interestingly, “nano-octahedral” was not in our original morphology search terms, but its TF-IDF across articles was high enough that we included it as a morphology label. We note that while we are currently using a predetermined list of composition and morphology labels based on our search terms, we have identified automated label generation as an area for further development so that articles with nanomaterials beyond our current scope of composition can also be automatically processed.
are manually selected and read. Ecdotal evidence provided in a narrow selection of papers that level insights that are not susceptible to the perspectives or an-

newcomers to the field quickly gain confidence in community experts or under-

representation of certain combinations of nanomaterial composition and morphology.

Figure 2 depicts the results of our classification on our 35k-article corpus (see Table S1 for tabular form). Based on this classification, we were interested to understand if any combinations of nanomaterial morphology and composition are either overrepresented or underrepresented within our corpus. Overrepresented combinations may be indicative of “hot topics”, while underrepresented combinations may be indicative of difficult to synthesize material-morphology combinations or areas ripe for exploration. For example, approximately 14% of the articles overall pertain to platinum nanomaterials irrespective of their morphology, yet articles pertaining specifically to platinum nanotubes represent 27% of all nanotube articles – a difference of +13%. A manual search of the literature reveals that platinum nanotubes are of particular promise as electrocatalysts for fuel cell applications.

Figure 3 shows this percentage difference for each nanomaterial’s composition and morphology.

Taking another example, silver and copper nanowires are also both overrepresented within the nanowire subset of papers compared to the overall percentage of silver- and copper-related papers by ca. 6% and 5%, respectively. Manually searching the literature, we find that networks of dispersed silver and copper nanowires are of particular interest as transparent electrodes in solar cells and electro-optical devices. Overrepresentation can also potentially signify how easily and reproducibly the materials can be synthesized to be explored by diverse researchers for various applications. Gold nanorods – so-called because their length is much shorter than nanowires’ (i.e., <1 μm) – are also highly overrepresented within the subset of nanorod articles, and this overrepresentation may reflect interest in this material for plasmonic bio-sensing applications or the challenge of synthesizing significantly longer, higher aspect ratio gold nanowires. While these trends may be known to experts working in the field, such analysis helps non-experts or newcomers to the field quickly gain confidence in community-level insights that are not susceptible to the perspectives or anecdotal evidence provided in a narrow selection of papers that are manually selected and read.

Identifying nanomaterials’ synthesis protocols from text

Having sorted the 35k nanomaterial articles by their target nanomaterial, we were next interested to extract the synthesis protocols from these articles, thereby creating a useful reference library for nanomaterial syntheses. Simply searching for and extracting the “experimental section” delineated in many articles is not a robust approach because some articles do not contain a distinct experimental section, and of those that do have such distinct experimental sections, they often contain details beyond the synthesis of the nanomaterials themselves. Additionally, sometimes relevant details to nanomaterials’ syntheses are included outside of the experimental section in other parts of the article, like the results and discussions sections. Thus, our approach is to make a determination about each sentence in an article individually as to whether it contains details relevant to the synthesis procedure.

We implemented this synthesis-related sentence identification using a supervised machine learning approach. We trained a logistic regression classifier using sentences hand-labeled by domain experts as either relevant or irrelevant to nanomaterials’ synthesis. Sentences in which the classifier had high confidence that they were synthesis related (denoted as P(+)) were then presented to domain experts for additional labeling to improve the classifier. This iterative training approach, detailed in the Supporting Information, has several advantages, most notably:

- The original full-document labeling produced relatively few positive examples (i.e., synthesis-relevant sentences) compared to negative examples (i.e., sentences not relevant to synthesis) (approximately 1:20). The subsequent active learning rounds helped to produce more positive examples, since we chose high P(+) sentences for labeling (i.e., sentences where the classifier had high confidence they were synthesis-related).

- Initially the approach suffered from a high false positive rate (i.e., sentences erroneously classified as relevant to the synthesis). The iterative rounds helped to identify and correct the most egregious false positives.
by introducing high P(+) negatives as additional training examples (i.e., sentences that the classifier had high confidence were synthesis-related but were not deemed so by domain experts).

- Although labeling the initial set of 18 articles was labor-intensive and required reading each article in its entirety, in subsequent iterative labeling, a user only has to evaluate a single sentence, which can be done in seconds with a single click using a fast annotation tool we developed for this task. We found that showing a few sentences before and after the sentence in question provides helpful context and aids in evaluation.

- The iterative approach provides a natural way to continue to improve the model as we discover additional errors.

The 99-article gold standard set contains a total of 27,125 labeled sentences (629 positive examples of synthesis-relevant sentences and 26,496 negative examples). We used these examples to train a logistic regression classifier using scikit-learn with default settings (see Supporting Information for details). To evaluate the approach, we ran a leave-one-article-out cross-validation on our labeled data set. Plotting the true positive rate against the false positive rate produces a receiver operating characteristic curve, or ROC, where the area under the curve, or AUC, indicates how well the classifier discriminates between synthesis relevant and irrelevant sentences. The AUC value can vary between 0.5 and 1.0, where 1.0 represents a perfect test and 0.5 indicates random performance. Our test produced an AUC of 0.99, indicating that the ranking of sentences according to their relevancy to synthesis is extremely accurate. In practice, where non-synthesis sentences outnumber synthesis sentences more than 40-to-1, performance is hampered by class skew, but we still achieve 52% precision (true positives divided by total predicted positives) at 90% recall (true positives divided by total actual positives). Although the above-mentioned processes are used for targeted extraction of synthesis sentences, we envision that the application of such a process is agnostic and can be generalized to target information or texts for end-user's needs.

To understand what terms lead the model to classify a sentence as related to the synthesis protocol or not, Figure 4 shows the 30 terms with the highest and lowest weight from our model. The model appears to weigh heavily past-tense verbs (i.e., was, were, added, prepared, used) to classify sentences as synthesis-related and present-tense verbs (i.e., is, are, be, have) for classifying non-synthesis related sentences. These trends agree with science writing convention where synthesis protocols are written in past tense whereas the rest of the article is written in present tense. However, verb tense by itself cannot classify relevant sentences because: 1) often brief descriptions of others’ prior work are included in articles using past tense, and 2) other portions of an experimental section, like material characterization for example, or figure captions are often written in past tense yet not part of the synthesis protocol. Interestingly, the
model appears to have identified key words to aide in discriminating such cases since “et” (as in “Smith et al.”) when describing others’ work) and characterization-related terms, like “TEM,” “images,” “UV,” and “Fig.” are weighted heavily as non-synthesis related terms. Analyzing more closely the terms that are positively correlated with synthesis (Fig. 4a), we note that many are either technical synthesis terms (i.e., solution, mL, synthesis, acid, temperature, mixture) or function words (i.e., to, in, for, of, at, and, after, then). That function words are used as indicators of synthesis-related sentences is particularly interesting because in traditional natural language processing such function words are not informative and filtered from the text. With scientific articles, such words appear to be indicative of procedures and thus helpful in identifying synthesis protocols (e.g., “The solution of PVP then turned yellow, and AgNO₃ was then added and reacted for 12 more minutes, resulting in…..”).

**Identifying chemicals used in nanomaterials’ syntheses**

While having a library of nanomaterials synthesis protocols may be a useful reference in itself, we believe that generalizations and important differences across protocols can be extracted by deconstructing and structuring protocols to allow for comparisons of, for example, chemicals used or processing conditions. Such extraction of specific categories of words or terms from text is generally referred to as named entity recognition (NER). We have initially focused on identifying and extracting chemicals, which is specifically referred to as chemical entity recognition (CER). Following general NER practice (see ref. 29), CER first requires tokenization – a process whereby sentences are divided into their constituent subunits (i.e., words, numbers, and punctuation). General language tokenizers often rely on white spaces and punctuation to identify word tokens. However, because white spaces and various forms of punctuation (i.e., commas, parentheses, hyphens, brackets) are routinely used in chemical nomenclature,29 these general language tokenizers have proven ill-suited for chemistry texts.29 To address these unique challenges in processing chemistry-related texts, a number of chemical text tokenizers have been developed and included in CER tools, like OSCAR4,60 ChemSpot,61 and BANNER.62 For example, OSCAR4 performs a coarse whitespace tokenization, before recursively splitting up the generated tokens using human-defined rules to handle oxidation states, unmatched brackets, trademark symbols, hyphens, etc.60 After sentences are tokenized, CER requires processing tokens to identify those that are chemicals. This chemical identification is generally performed using some combination of the following approaches: 1) use of a dictionary of chemical entities, 2) use of rules/patterns to identify chemical entities (e.g., formula notations), and/or 3) use of context and parts of speech to identify chemical entities using machine learning-based approaches.

For extracting chemicals used in the synthesis of nanomaterials, we evaluated a number of open-source CER tools, including BANNER-CHEMDNER63 ChemXSeer,64 ChemSpot61 OSCAR,60 and ChemDataExtractor10 that use different chemical tokenizers and variations of these three approaches, as summarized in Table 1.

In addition to evaluating the existing set of CER tools, we also developed our own using Stanford’s CoreNLP65 and AllenNLP66 natural language processing toolkits. Similarly to the existing CER tools, we trained conditional random field (CRF)67 models as the basis of our approach where the CRF

<table>
<thead>
<tr>
<th>CER Tools Evaluated</th>
<th>Tokenizer</th>
<th>Technique Used</th>
<th>Dictionary Used</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemDataExtractor</td>
<td>Modified Penn Treebank tokenizer67</td>
<td>CRF (hand-crafted features + unsupervised features)</td>
<td>Jochem (filtered)</td>
<td>Precision: 97.1%</td>
</tr>
<tr>
<td>LLNL-modified StanfordNLP</td>
<td>Modified Penn Treebank tokenizer</td>
<td>CRF (unsupervised features)29</td>
<td>None</td>
<td>Precision: 90.4%</td>
</tr>
<tr>
<td>Chem Spot 2.0</td>
<td>split at every non-letter and non-digit character, as well as all number-letter changes</td>
<td>CRF (hand-crafted features+unsupervised features) + dictionary</td>
<td>ChemIDPlus</td>
<td>Precision: 93.2%</td>
</tr>
<tr>
<td>BANNER-CHEMDNER</td>
<td>BANNER's simple tokenizer</td>
<td>CRF (hand-crafted features + unsupervised features)</td>
<td>None</td>
<td>Precision: 95.0%</td>
</tr>
<tr>
<td>ChemXSeer</td>
<td>OSCAR4 Tokenizer</td>
<td>CRF (hand-crafted features + unsupervised features) + dictionary</td>
<td>Jochem + custom dictionaries</td>
<td>Precision: 96.9%</td>
</tr>
<tr>
<td>OSCAR4</td>
<td>OSCAR4 Tokenizer</td>
<td>Maximum Entropy Markov Model + dictionary</td>
<td>ChEBI + custom dictionaries</td>
<td>Precision: 64.6%</td>
</tr>
<tr>
<td>LLNL-modified AllenNLP</td>
<td>SpaCy Tokenizer61</td>
<td>CRF (unsupervised features: Word2Vec and Character Encoding features integrated using a GRU)</td>
<td>None</td>
<td>Precision: 63.6%</td>
</tr>
</tbody>
</table>

Table 1. Overview of CER tools evaluated and their performance on LLNL’s gold standard.
models make use of the sequence of words, specifically their parts of speech or context, to aid in classification of a given word or token. However, whereas other CER tools also make use of hand-crafted features within their CRF models – essentially customized, expert-defined rules for classifying tokens as chemicals – we omitted doing so with our Stanford NLP- and AllenNLP-based models in an effort to make them more broadly applicable for CER across scientific subfields. Instead, we utilize features that can be learned in an unsupervised manner from a training corpus, such as n-gram frequencies, word shape features, Word2Vec character-level encodings, etc. We trained our chemical identification model using an iterative approach similar to the one described above to identify synthesis-related sentences. To generate our initial training data set of annotated chemical terms, we combined the results of ChemSpot, ChemXSeer and BANNER-CHEMDNER applied to our dataset of 35k nanomaterials articles.

We tested the performance of these various CER tools against our 99-article “gold standard” where experts hand-labelled chemicals within the synthesis protocols (available in Supporting Information). Table 1 reports each model’s precision (true positives divided by total predicted positives), recall (true positives divided by total actual positives), and f1 score (harmonic average of precision and recall) against this gold standard. We find that, despite taking different approaches, ChemXSeer, BANNER-CHEMDNER, ChemSpot 2.0, our LLNL-modified Stanford NLP, and ChemDataExtractor all perform comparably well with precision >90%, recall 70-80%, and f1 scores 81-87%. We choose to implement ChemDataExtractor into our article-processing pipeline as it had the best f1 score. However, separately, we continued to make efforts to improve our LLNL-modified Stanford NLP model by applying different regular expressions to identify chemical formulas, training the model on different datasets, and implementing chemical tokenizers (see Supplementary Information for details). In the best case, we are able to raise the f1 to 87%, on par with ChemDataExtractor.

After extracting all chemicals from the synthesis protocols using ChemDataExtractor, the pipeline normalizes chemical synonyms (e.g., dichloromethane, methylene chloride, DCM, CH₂Cl₂ all refer to the same chemical) using a combination of the JoChem and PubChem dictionaries (see Figure S1 for details). Comparing the frequency of use of chemicals in the syntheses of different nanomaterial morphologies may potentially provide insight as to how important that chemical is to attaining a given morphology. While such trends can be impacted by bias (authors influencing one another's work, for example) and do not by themselves provide any physical insights to the underlying mechanisms, they may help generate hypotheses for targeted studies in a reduced experimental phase space as well as provide the end user a bird’s eye view of the synthesis space for a given targeted materials, especially to those who are new to the field (i.e. graduate students).

To demonstrate the utility of such analysis, Figure 5 shows the most commonly occurring chemicals in papers involving the synthesis of Ag nanowires, nanospheres, and nanocubes (Fig. 5a) and Au nanorods, nanospheres, and nanocubes (Fig. 5b). Many chemicals appear commonly regardless of nanomaterial morphology or composition, like ethanol, which is often used for washing, and elemental silver and gold. However,
polyvinylpyrrolidone (PVP) and hydrochloric acid (HCl) both appear nearly twice as frequently in the syntheses of Ag nanocubes than Ag nanospheres or nanowires (PVP: 37% nanocube, 13% nanosphere, 22% nanowire; HCl: 28% nanocube, 15% nanosphere, 11% nanowire), and indeed, manually searching the literature from all publishers (not only our Elsevier corpus), we found several reports of PVP’s\(^71-73\) and HCl’s\(^74-76\) critical roles in directing Ag nanoparticle to a cube morphology. Within the Au nanomaterial articles, hexadecyltrimethylammonium bromide (CTAB) is commonly occurring, particularly within Au nanorod articles (44%) when compared to nanosphere or nanocube articles (14% and 17%, respectively), and again a targeted manual literature search yields many studies highlighting and surveying the mechanism by which CTAB enables anisotropic crystal growth for nanorod morphologies.\(^77-79\) Ascorbic acid and AgNO\(_3\) also occur more than twice as often in Au nanorod and nanocube synthesis protocols than in Au nanosphere protocols (ascorbic acid: 33% nanorod, 21% nanosphere, 8% nanoscale; AgNO\(_3\): 28% nanorod, 26% nanocube, 7% nanosphere). Sau et al. previously performed detailed experimental studies showing that the addition of high concentrations of ascorbic acid and AgNO\(_3\) tends to yield Au nanocubes, and as the concentration of CTAB is increased, a reaction tends more towards Au nanorods.\(^80\) Without performing any detailed experimental studies but simply surveying the occurrence of chemicals across a broad spectrum of nanomaterial synthesis articles, we infer a similar trend.

Flexible Framework for Visualizing Insights

In the previous sections, we detailed how we extract important nanomaterial synthesis data from scientific articles’ text and images, enabling us to create a nanomaterials knowledge database. However, to be of use to materials scientists, a database by itself is not enough; the database must be easily explored and queried, and information must be presented in an interactive, intuitive, and visual way that helps build understanding. To address these challenges, we also developed a flexible visual analytics environment for our database that enables the user to quickly query the complex information and generate aggregated summaries via intuitive visual encodings.

Extracting Information from Figures

Scanning and transmission electron microscopy images (SEM and TEM, respectively) reported in nanomaterials synthesis-related articles are valuable to analyze as they can often provide an immediate perspective of the nanomaterial’s geometry, dimensions, and polydispersity. Thus, in addition to processing text, we feel it is important to develop tools for capturing and processing image information from articles, as well. The tools we developed recognize and extract SEM and TEM images from figures, which is accomplished using a transfer learning approach with a convolutional neural network model (see Supporting Information for details). SEM and TEM images are then analyzed to identify the nanomaterials morphology present (currently have four categories: nanocube, nanoparticle, core-shell nanoparticle, and nanorod) and provide dimensional estimates of all the nanomaterials present in the image (see Supporting Information for details). Figure 6 shows the input and output of this tool with a graphical user interface, allowing for easier analysis. The detailed methodology for how the tool identifies nanomaterials’ morphologies and size distributions from SEM/TEM images will be described in a future manuscript.\(^81\)

![Figure 6. (a) Input and (b) output screenshots of the GUI-form of the portion of the pipeline that automatically analyzes nanomaterial SEM and TEM images extracted from articles.](image-url)
extracted nanomaterials synthesis information as the exploration-oriented system enables the domain experts, who may not have an extensive background in writing analysis code or database queries, to easily experiment with the data and view the results.

Figure 7 shows a screenshot of our browser-based visualization tool (available from https://github.com/LLNL/MI-ChemVis). The tool consists of five panels – Filter, Graph, Aggregate Info, Document, and Doc Comparison – that provide the user with both a bird’s eye view of the dataset as a whole and the ability to zoom-in to analyze and compare individual articles. Starting with the Filter panel, a user can choose via dropdown menus a material, morphology, and/or chemical of interest to focus exclusively on a subset of the data from the 35k nanomaterial articles. The Graph panel displays up to 300 individual papers from the papers within the selected dataset and has dropdown menus allowing the user to visually group and color the papers by morphology and/or material. In order to maintain reasonable loading times, the Graph panel is limited to display up to 300 individual papers; if more than 300 articles are selected, then 300 are randomly selected for display from the entire dataset. Any data point/article in the graph panel can be selected, which then results in the corresponding article’s information being displayed in the Document panel. Details of two different selected articles can also be compared side-by-side using the Doc Comparison panel. The Aggregated Info panel display via bar graph the number of articles within the selected data set as a function of either material, morphology, or chemicals utilized, as decided by the user from the dropdown menu (can also sub-specify chemicals to analyze solvents, reducing agents, and surfactants). To enable others to easily use and explore the data we have extracted from nanomaterials synthesis articles, we are providing a distributable form of this visualization tool from our GitHub page.

CONCLUSIONS

We have developed a suite of tools that extract and structure information from the text and figures of scientific articles pertaining to nanomaterials synthesis. Utilizing a corpus of 35k articles, we have demonstrated methods of classifying the articles according to their target nanomaterial composition and morphology. While the specific trained models we employ would require re-training to extract different target data, the machine learning methods that we employ are more broadly applicable to other materials science domains. By itself, classification of articles can provide insights as to what combination of nanomaterial composition and morphology are over- or under-explored compared to average, potentially indicating “hot topic” areas. By extracting the synthesis protocols from these classified articles and then extracting the chemicals used in the synthesis, we showed that further correlations can be identified between the use of certain reagents and specific nanomaterial morphologies and compositions. In releasing this dataset to the broader community, we hope it may serve as a useful source of information for generating and testing hypotheses regarding existing nanomaterials syntheses and potentially help inform the development of new synthesis protocols for as-of-yet unrealized nanomaterials. Important for the utility of this data to
others in field, we have packaged it in a visualization tool that provides easy exploration of nanomaterials synthesis trends across thousands of articles while also allowing users to delve into the details of any one paper.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Method details, list of articles included in “gold standard” (PDF)

Morphology and composition labels for nanomaterials article in “gold standard” (XLSX)

AUTHOR INFORMATION

Corresponding Author

* han5@llnl.gov

Present Addresses

† AEye, Inc. 5700 Stoneridge Drive, #102, Pleasanton, CA 94588

ACKNOWLEDGMENT

We thank Alyssa Troksa, Cameron Cornell, and Steve Lan for their help in creating the “gold standard” data set. This work was funded by the Laboratory Directed Research and Development (LDRD) program at Lawrence Livermore National Laboratory (16-ERD-019). Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344. LLNL-JRNL-779959.

REFERENCES


76. Zhang, Q.; Li, W.; Wen, L.-P.; Chen, J.; Xia, Y., Facile Synthesis of Ag Nanocubes of 30 to 70 nm in Edge Length with CF3COOAg as a Precursor. *Chemistry – A European Journal* 2010, 16 (33), 10234-10239.


81. Han, J.; Kim, H.; Han, T. Y.-J., Rapid and automated size and morphology extract of SEM/TEM images. In preparation.