Title

Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities

Journal

Scientific Data, 9(1)

ISSN

2052-4463

Authors

Cruse, Kevin
Trewartha, Amalie
Lee, Sanghoon
et al.

Publication Date

2022

DOI

10.1038/s41597-022-01321-6

Peer reviewed

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities

Kevin Cruse[1,2], Amalie Trewartha[2], Sanghoon Lee[1,3], Zheren Wang[1,2], Haoyan Huo[1,2], Tanjin He[1,2], Olga Kononova[1,2], Anubhav Jain[3] & Gerbrand Ceder[1,2] ✉

Gold nanoparticles are highly desired for a range of technological applications due to their tunable properties, which are dictated by the size and shape of the constituent particles. Many heuristic methods for controlling the morphological characteristics of gold nanoparticles are well known. However, the underlying mechanisms controlling their size and shape remain poorly understood, partly due to the immense range of possible combinations of synthesis parameters. Data-driven methods can offer insight to help guide understanding of these underlying mechanisms, so long as sufficient synthesis data are available. To facilitate data mining in this direction, we have constructed and made publicly available a dataset of codified gold nanoparticle synthesis protocols and outcomes extracted directly from the nanoparticle materials science literature using natural language processing and text-mining techniques. This dataset contains 5,154 data records, each representing a single gold nanoparticle synthesis article, filtered from a database of 4,973,165 publications. Each record contains codified synthesis protocols and extracted morphological information from a total of 7,608 experimental and 12,519 characterization paragraphs.

Background & Summary

The synthesis of gold nanoparticles has been practiced for centuries, and their modern applications are widespread, which include *in vitro* diagnostics[1], semiconductor technology[2], and cosmetics[3]. The application of gold nanoparticles often depends on their morphology and size;[4] yet, despite their ubiquity, only relatively recently has the control of these properties been interrogated systematically[5].

While many theories and models exist for the mechanisms that determine nanoparticle morphology[6–8], most of the exploration of this synthesis space is driven by heuristics. For nanorod growth in particular, it appears that the simultaneous presence of many reagents affects the final characteristics of a sample of gold nanorods[9]. While factorial experiments can offer some insights into how varying certain precursor concentrations affects final particle morphology, size, or aspect ratio, it is impractical to perform enough experiments to cover a large enough portion of the synthesis space to produce an effective model, even with state-of-the-art high-throughput synthesis methods.

Beyond empirical modeling and experiment, computational methods exist that either simulate the energetics of the formation of nanoparticles or interrogate the nucleation and growth steps traversed by nanoparticles. However, these approaches come with inherent tradeoffs between the resolution of atomic interaction and computational tractability. For example, calculations from first-principles have been conducted using density functional theory (DFT) that probe the energetic landscape of potential gold nanoparticle shapes[10], including the effects of various surface ligands[11], which are vital for the synthesis of solution-phase noble metal nanoparticles[12]. However, such a technique does not take into account the intricacies of nucleation and growth competition in solution-based nanoparticle synthesis. On the other hand, continuum-level models can represent real-time growth and dispersity dynamics[13], though sacrificing the small-scale energetics highlighted by techniques such as DFT.

In a third paradigm of scientific investigation, the volume of data-driven approaches to understanding chemistry and materials synthesis is accelerating. These approaches represent a resourceful complement to

[1]Department of Materials Science and Engineering, University of California, Berkeley, CA, 94720, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA. [3]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA. ✉e-mail: gceder@berkeley.edu

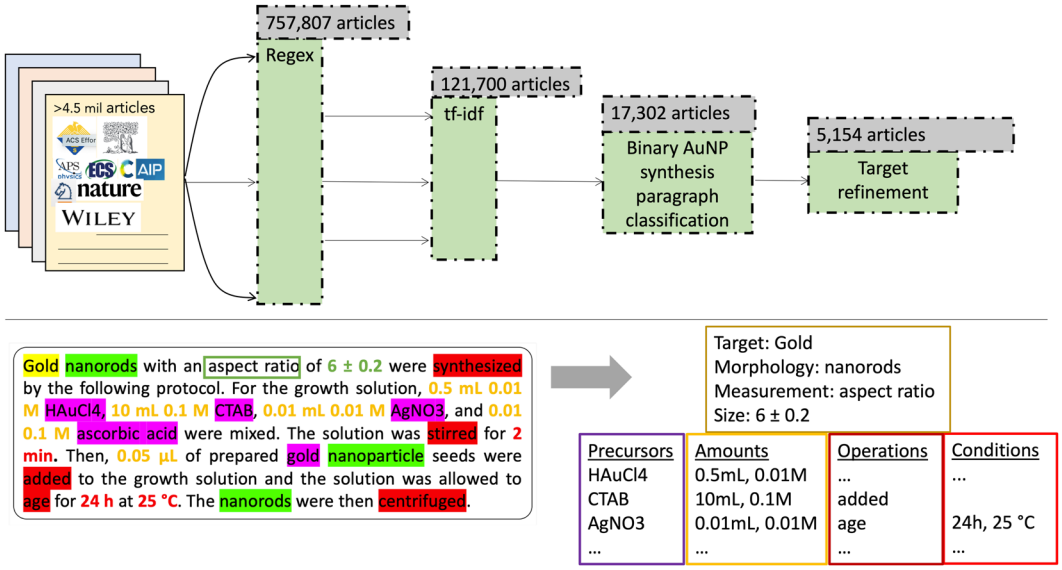

Fig. 1 AuNP synthesis publication extraction pipeline. Starting with the >4.5 million article materials science literature database, parsed paragraphs from the articles are funneled through progressively finer-meshed filters to identify those related to the synthesis of gold nanoparticles. The first two steps include a regex search for "nano" phrases followed by the vectorization of that corpus using TF-IDF, similar to the method used in Hiszpanski *et al.*[20]. Those articles where the TF-IDF scores for gold are higher than other noble metal nanoparticle compositions are accepted. Each paragraph from those articles is then passed through a binary classifier which determines whether or not the paragraph describes the synthesis of gold nanoparticles. Finally, after extracting the synthesis recipes from those relevant paragraphs, 5,154 articles with synthesis paragraphs containing gold- or gold nanoparticle-related targets are collected. An example synthesis paragraph along with a sample of the extracted information is shown in the bottom panel.

established computational methods and raw experimentation, and have been proven successful in applications such as materials discovery[14,15], synthesis protocol querying[16], and the simulation and interpretation of characterization results[17,18]. However, data-driven approaches are limited by the completeness and substance of the data resource(s) used. For the nanotechnology field, Yan *et al.*[19] constructed a table of available nanotechnology databases, their volume, as well as a description of their proposed usages. Each of these, including the dataset created by Yan *et al.*, provide useful and carefully curated information on the characteristic features of various nanostructures; however, to our knowledge, none provide validated protocols for their synthesis.

There is a wealth of gold nanoparticle synthesis and characterization data available for data-driven approaches in unstructured form in the scientific literature[20]. This source remains largely untapped because manually extracting such information is both tedious and unrealistic due to the volume of the literature. Natural language processing (NLP) and text-mining techniques have been successfully employed and established in scientific fields related to materials science[21–23], and the sub-field of text-mining in materials science is budding[24–27]. Here we employ natural language processing (NLP) and text-mining techniques on a collection of nearly 5 million materials science publications[28] to extract gold nanoparticle synthesis recipes and their outcomes. We introduce this dataset as an additional element, along with direct experimentation and computational modeling, in the effort to capture the mechanisms for metal nanoparticle growth.

In this data descriptor, we present an open-source dataset of 5,154 records filtered down from a database of 4,973,165 publications, each representing a single article and containing automatically extracted gold nanoparticle synthesis recipes and morphological information from the contained synthesis paragraphs and characterization paragraphs. Within a synthesis paragraph, the precursors used and their amounts are extracted, and the synthesis actions and conditions are extracted and codified into a procedure graph. Within both synthesis paragraphs and characterization paragraphs, morphological (e.g. "spherical", "nanorod", "dogbone-like") and size (e.g. diameter, aspect ratio) entities were extracted. In total, this dataset encompasses 7,608 synthesis paragraphs and 12,519 characterization paragraphs. A schematic of the pipeline devised for this extraction is shown in Fig. 1.

Methods

Content acquisition. Over 4.5 million publications from material science journals have been scraped from the web and parsed by the process described in Kononova *et al.*[28]. These publications were obtained through agreements with publishers Elsevier, Wiley, the Royal Society of Chemistry, Nature Publishing Group, the American Institute of Physics, Springer, the American Chemical Society, the American Physical Society, and the Electrochemical Society. Journals from each publisher related to materials science were selected manually. Articles were scraped from these journals using a custom web scraping tool. Only articles published after the year 2000 were collected because prior publications are often provided in PDF and not HTML/XML format, the latter

of which is more straightforward to parse[24]. Parsers were built for each publishing group using custom parsing toolkits. From each article, the full text and relevant metadata were parsed and are stored across MongoDB database collections (www.mongodb.com) maintained on an internal server.

AuNP synthesis publication collection. Publications containing gold nanoparticle synthesis procedures were collected using a series of progressively finer-meshed filters, combining unsupervised and supervised text processing methods. This progression is illustrated in Fig. 1, along with the corresponding yield of articles at each step. The stages of the pipeline were built to facilitate adaptation to other nanomaterial compositions in the future. Each of the steps is described in detail below.

Identifying AuNP publications. The early stages of the pipeline cast a wide mesh for Au nanoparticle (AuNP) publications using unsupervised text-mining methods. These first two steps were adapted from the methods used for nanomaterial synthesis publication extraction performed by Hiszpanski *et al.*[20]. The first step consists of a simple regular expression query. To accomplish this, we imported the paragraphs of all papers from our MongoDB database (described in "Content acquisition") into an Apache Solr search engine instance (https://solr.apache.org/), which is powered by Apache Lucene and supports fast full-text search. We used the default English Lucene analyzers and tokenizers in Solr. The full text of every article was queried for words starting with "nano", followed by any number of characters or whitespace. The query was written in Solr-specific query syntax and returned identifiers of the matching paragraphs. We used the query results to lookup and fetch related papers in the MongoDB database. This step yielded 757,807 nanomaterial articles.

Next, all of the documents in this corpus were vectorized in a term frequency-inverse document frequency (TF-IDF) representation, where each element in the vector represents the frequency of a given token normalized by the frequency of that term across documents[29]. We implemented this vectorization using the scikit-learn `TfidfVectorizer` module (https://scikit-learn.org/stable/), with tokenization from ChemDataExtractor's `ChemWordTokenizer`[23], default English stopwords eliminated, and a minimum document frequency of 100 for each token. With the TF-IDF vectorized corpus, articles were collected whose TF-IDF values for the words "gold" or "Au" were larger than any of "silver", "Ag", "copper", "Cu", "palladium", "Pd", "platinum", or "Pt". 121,700 gold nanomaterial articles were collected in this step, each being related to gold nanomaterials but not necessarily containing protocols for their synthesis.

AuNP synthesis paragraph classification. To isolate those AuNP publications that contain synthesis protocols, we trained a transformer-based binary classification model using the Simple Transformers NLP library (https://github.com/ThilinaRajapakse/simpletransformers). We first pre-trained a BERT[30] (Bidirectional Encoder Representations from Transformers) model specializing in materials science text, referred to as "MatBERT"[31]. The pre-training data for MatBERT were 2 million randomly sampled papers from our publications database. Following the original BERT, we trained two WordPiece tokenizers of vocabulary size 30,522 (cased and uncased) from scratch on this full-text to optimize tokenization for materials science terminologies. After all the papers were tokenized, paragraphs with less than 20 or more than 510 tokens were removed. Out of 61 million paragraphs from the 2 million sampled papers, roughly 17% contained less than 20 tokens and about 2% contained more than 510 tokens, for both the cased and uncased set of texts. This yielded around 50 million paragraphs and 8.8 billion tokens. During pre-training, MatBERT was trained for the masked language modeling (MLM) task, which requires MatBERT to predict the original tokens in a paragraph after they are masked. This pre-training step helps MatBERT to develop a general understanding of the language and better learn the classification of synthesis protocols. The training codes and pre-trained MatBERT models can be found at https://github.com/lbnlp/MatBERT.

To gather positive training paragraphs, we first modeled the topics of every paragraph in the aforementioned gold nanoparticle publication collection using latent Dirichlet allocation (LDA)[32]. Then, we collected and manually validated those paragraphs whose dominant topic was related to synthesis (topic words including "synthesized", "solution", "ml", "addition", etc.). A range of negative training paragraphs were collected manually from various parts of a typical publication, including the introduction, results, discussion, and characterization sections. Annotations were accomplished using SpaCy's Prodigy interface (https://prodi.gy). The training data ultimately included 739 training examples, with 242 positive examples and 497 negative examples. Because synthesis paragraphs are far less common in literature than non-synthesis paragraphs, we included more negative than positive training examples to ensure that most kinds of these non-synthesis paragraphs were covered in the training data. Using the `ClassificationModel` module from Simple Transformers, training data was split into 80/10/10 train/validation/test sets and trained over 20 epochs. Articles were then identified that contained at least one paragraph classified as being related to gold nanoparticle synthesis. This step yielded 21,989 AuNP synthesis paragraphs from a total of 17,302 articles.

Synthesis recipe extraction. Synthesis targets, precursors, their amounts, synthesis actions, and action conditions were all extracted using synthesis procedure extraction and codification tools described in[28] and[33]. A sample of an example extraction from a synthesis paragraph is shown in the bottom panel of Fig. 1. Each step is described in detail below.

Materials entity recognition (MER). To identify and classify targets, precursors, and other materials from synthesis paragraphs, we implemented a two-step model. In the first step, each word token was transformed into an embedding vector with the MatBERT model (see "AuNP synthesis paragraph classification"). Then, the embedding vector was passed to a bi-directional long-short-term memory neural network with a conditional random-field top layer (BiLSTM-CRF) to identify whether the corresponding token was a materials entity or

Data description	Data Key Label	Data Type
DOI of the original paper	`doi`	*string*
List of constituent paragraphs and extracted data	`paragraphs`	*list* of Objects (*dict*)[1]
Year of publication	`publication_year`	*int*
Number of citations	`times_referenced`	*int*

Table 1. Format for highest article-level of each data record: description, key label, data type. [1]Contents of paragraphs shown in Table 2.

a regular word. In the second step, each materials entity was replaced with a keyword <*MAT*> and classified as either a *target*, *precursor*, or *other* material using another BERT-based BiLSTM-CRF network with a similar structure. In total 1,281 synthesis paragraphs from 1,155 papers were annotated by labeling each word token as *material*, *target*, *precursor*, or *outside*. The annotated dataset was split into training/validation/test sets with a paper-wise ratio of 700/150/305 to train the aforementioned two neural networks.

Synthesis actions and their attributes. To recognize and classify synthesis actions described in a paragraph, we implemented an algorithm that combines a recurrent neural network (RNN) and rule-based parsing of sentence dependency trees. Sentences were tokenized using `ChemDataExtractor's ChemWordTokenizer`. The RNN performed classification of sentence tokens into 5 categories: *start-synthesis* (general actions that signify that something was synthesized, e.g. "synthesized", "prepared", etc.), *mixing*, *heating*, *drying*, and *cooling*, which are the basic actions in nanoparticle synthesis. The RNN was trained on a set of 3,040 synthesis sentences from 535 synthesis paragraphs (classified according to the paragraph classifier described in Huo *et al.*[34]). Finer details for the development of this model are described in Wang *et al.*[35] In brief, 3,781 sentences were taken from 199 solid-state, 51 sol-gel, 148 hydrothermal, and 137 precipitation synthesis paragraphs. 3,040 sentences in this set were determined to be synthesis sentences (as opposed to characterization or miscellany). The tokens in these 3,040 synthesis sentences were annotated by human experts in NLP and materials synthesis science according to their type of synthesis action, with the actions relevant to nanoparticle synthesis listed above. The tokens' feature vectors were generated using a Word2Vec model[36]. The embeddings were trained on ~400,000 synthesis paragraphs of different synthesis types using the Gensim library[37]. The sentences of paragraphs were lemmatized, all the quantity tokens were replaced with the keyword <*NUM*>, and all the chemical formulas were replaced with the keyword <*CHEM*> using rule-based algorithms. The SpaCy library[38] was used to grammatically parse each sentence and obtain linguistic features of the tokens, such as their part of speech and their dependency on root tokens. For training, validation and testing, the annotated set was split into a 70/10/20 fraction, respectively. Synthesis action attributes, such as temperature, time, and environment were extracted by using dependency tree parsing and a rule-based regular expression approach[29].

Material quantities extraction. To correlate the numerical values of material quantities, such as molarity, concentration, or volume, to materials entities extracted by the MER model (see "Materials entity recognition (MER)"), we applied a rule-based approach. First, we used the NLTK library[39] to build syntax trees[29] for each sentence in a paragraph, where every word is represented as a leaf node. Then, the syntax tree for each sentence was cut into the largest sub-trees for every material, with each sub-tree having only one material entity. To do this, we first identified the materials on leaf nodes. Then, starting from each material, we identified the largest sub-trees (i.e., we traversed the syntax tree upwards until there was more than one material leaf node descending from the same node). Finally, the largest sub-tree for a given material was defined as the sub-tree formed by the node and its descendants identified in the previous step. Next, we searched for the quantities in each sub-tree and assigned the quantities associated with the unique material entity in the sub-tree.

Gold-related target refinement. With all relevant information from synthesis paragraphs extracted and codified, we implemented a final target refinement step to identify all of those papers that contain synthesis procedures explicitly targeting "gold" or gold nanoparticle-related entities, a list for which is provided in the /rsc folder of the GitHub codebase (see "Code availability"). Of the 18,101 articles collected from the binary classification step described above, 5,154 contained a target entity extracted by MER (see "Materials entity recognition (MER)") that was related to gold or a gold nanoparticle-related entity.

Tagging Seed-mediated growths. Each paragraph containing a recipe was tagged as being related to either a seed-mediated or seedless synthesis approach. Seed-mediated approaches are those in which some method is used to create small colloidal seeds that act as nucleation sites for larger growths, often with interesting morphology. These methods are common for rod-based nanoparticle growths and are abundant in the literature[40], so we wanted to make those particular recipes easily queryable. The tags for this field were determined by keyword matching for "seed" and related lemmas for seed-mediated methods as well as "seedless" or the absence of seed-related text for seedless methods. The binary tag for seed-mediated growth in a given recipe paragraph is included in the `seed_mediated` field as a boolean in the provided dataset (see Table 2).

AuNP synthesis outcome extraction. To complement the extracted gold nanoparticle synthesis protocols from the collected 5,154 articles, we also extracted relevant morphological and characterization information. A two-step process was implemented for this extraction.

Data description	Data Key Label	Data Type
Unique paragraph hash	`_id`	*string*
Whether paragraph contains characterization information	`contains_characterization`	*bool*
Whether paragraph contains synthesis recipe	`contains_recipe`	*bool*
Materials and quantities contained in paragraph[1]	`materials_and_quantities`	*list* of Objects (*dict*):
		`-material`: *string*
		`-amount`: *list* of Objects:
		`-value`: *float*
		`-unit`: *string*
Condensed morphological entities in paragraph[2]	`morphological_information`	Object (*dict*):
		`-descriptors`: *list*
		`-measurements`: *list*
		`-morphologies`: *list*
		`-sizes`: *list*
		`-units`: *list*
Morphological entities and token locations in paragraph[2]	`morphology_ner_tokens`	*list* of Objects (*dict*):
		`-annotation`: *string*
		`-start`: *int*
		`-end`: *int*
		`-text`: *string*
Whether or not the AuNP synthesis is seed-mediated[1]	`seed_mediated`	*bool*
List of constituent sentences and extracted data[1]	`sentences`	*list* of Objects (*dict*)[3]
Synthesis actions and conditions[1]	`synth_actions`	*list* of Objects (*dict*):
		`-conditions`: Object (*dict*):
		`-temperature`: Object(*dict*)[4]
		`-time`: Object(*dict*)[4]
		`-string`: *string*
		`-subject`: *string*
		`-type`: *string*
Snippet of paragraph text	`text`	*string*

Table 2. Format for lower paragraph-level of each data record: description, key label, data type. [1]Only if `contains_recipe` is `true`. [2]Only if `contains_characterization` or `contains_recipe` is `true` [3]Contents of paragraphs shown in Table 3. [4]{`value`: *list*, `unit`: *string*, `max_value`: *float*, `min_value`: *float*}.

Characterization paragraph classification. To focus on paragraphs that contain information on the morphology of synthesized gold nanoparticles, we trained a binary transformer-based gold nanoparticle characterization paragraph classifier, using a similar approach to the binary synthesis paragraph classifier described above (see "AuNP synthesis paragraph classification"). Positive training paragraphs were collected by manually selecting characterization-related and morphology-related paragraphs modeled from LDA (with topic words including "morphology", "tem", "size", "diameter", "nm", etc.). The training data for this classifier included 299 training examples, with 69 positive examples and 230 negative examples. Using an 80/10/10 train/validation/test split, a model was trained over the course of 20 epochs using Simple Transformer's `ClassificationModel` Module. This classification yielded 12,519 paragraphs containing morphological characterization information for gold nanoparticles.

Characterization entity recognition (MorphER). To extract relevant gold nanoparticle morphological information, we developed a transformer-based named entity recognition (NER) model specializing in the recognition of entities related to nanoparticle morphology and size ("MorphER"). To train this model, we annotated a set of 119 characterization-classified paragraphs from 91 articles on gold nanoparticle synthesis. The entities labeled for this model include specific morphological information for the synthesized gold nanoparticles, including: **MOR**, noun phrases related to morphology, such as "nanoparticles" or "AuNRs"; **DES**, descriptive terms for morphologies, such as "dumbbell-like" or "spherical"; **MES**, measurements, such as "aspect ratio" or "diameter"; **SIZ**, the value of the measurement; **UNT**, unit, if applicable (i.e. not for aspect ratios). Entities related to nanoparticles (e.g. "NP", "nanoparticle") but not necessarily their shape were labeled as **MOR** entities since the shape of the particle is not always mentioned explicitly, though the size is usually mentioned. This way, one could attribute extracted size information to at least some target entity. We chose to use NER to extract size information as well to deal with cases where we cannot use units as an anchor for rule-based methods, as in aspect ratios for nanorods. The model was fined-tuned over the pretrained MatBERT model described earlier (see "AuNP synthesis paragraph classification") on the paragraph-level with an 80/10/10 train/validation/test split over 20 epochs and deep fine-tuning. This entity recognition model was run on any paragraph that the AuNP synthesis

Data description	Data Key Label	Data Type
All material entities	`all_materials`	*list* of Objects (*dict*)[1]
Non-precursors and non-target material entities	`other_materials`	*list*
Precursor material entities	`precursors`	*list* of Objects (*dict*)[1]
Sequence of synthesis operations and conditions	`procedure_graph`	*list* of Objects (*dict*)
		`-env_toks`: *list*
		`-op_token`: *string*
		`-op_type`: *string*
		`-ref_op`: *bool*
		`-subject`: *string*
		`-temp_values`: *list* of Objects (*dict*)[2]
		`-time_values`: *list* of Objects (*dict*)[2]
Target material entities	`target`	*list* of Objects (*dict*)[1]

Table 3. Format for lowest sentence-level of each data record: description, key label, data type. [1]{`material`: *string*, `amount`: [{`value`: *float*, `unit`: *string*}]}. [2]{`max`: *float*, `min`: *float*, `tok_ids`: *list*, `units`: *string*, `values`: *list*}.

paragraph classifier predicted to be a synthesis paragraph or that the characterization paragraph classifier predicted to be a characterization paragraph.

This is the final extraction step in the pipeline constructed to build this dataset. Thus, the dataset does not contain any entity linking (e.g. particle size to specific nanoparticle morphology, morphological entity to synthesis procedure, etc.). In attempts to address this for the next iteration of the dataset, we have implemented, with moderate success, both rule-based linking through dependency tree parsing as well as simultaneous extraction and linking using more powerful language models such as GPT-3.

Briefly, we address our decisions for and differences in model choice and architecture for the text entity extraction tools described above. Materials Entity Recognition[33] and the synthesis actions extraction model were first trained for the extraction of inorganic solid-state synthesis procedures[28]. The development of our MatBERT model (described in "AuNP synthesis paragraph classification") was more recent and coincided with the development of the MorphER model. Since our development of MatBERT, we have incorporated its embeddings into the Materials Entity Recognition model since this tool is used on paragraphs outside of synthesis paragraphs. Because the extraction task for synthesis actions is linguistically simpler than for materials' names, we continue to use the Word2Vec embeddings trained for synthesis action extraction. Using MatBERT is also significantly more time consuming (as determined by He *et al.*[33]) and Word2Vec embeddings are sufficient for modeling word similarity. Additionally, the RNN model used for synthesis action extraction is capable of capturing contextual differences for certain vocabulary.

Data Records

The dataset, with 7,608 synthesis paragraphs and 12,519 characterization paragraphs from 5,154 articles, is provided as a JSON file, available publicly at https://doi.org/10.6084/m9.figshare.16614262.v3[41]. Each record corresponds to a publication, represented as a JSON object in a top-level list. Within each record is a list of paragraphs, with some containing a codified recipe, extracted morphological information, both, or neither. Metadata contained in the dataset for an article include: article DOI, the year of publication, and the number of times the article has been cited as of August 2021. For each paragraph within an article, metadata include: a unique paragraph hash, a boolean indicating whether or not the paragraph contains synthesis, a boolean indicating whether or not the paragraph contains characterization information, a boolean indicating whether or not the paragraph contains a seed-mediated growth, and a snippet of the paragraph text. Expanded details for the format of the dataset are given in Tables 1–3.

Technical Validation

The quality and content of this dataset is evaluated below through a description of the data extraction model metrics as well as a comparison of the dataset demographics to established heuristics in the field.

Extraction accuracy. We use the 35k article nanomaterial dataset developed by Hiszpanski *et al.*[20] as a benchmark with which to compare our regex/TF-IDF article filtering. We note that this dataset is comprised only of publications from Elsevier, so we only evaluate model performance on this set, which comprises 69% of our total materials science literature collection. The 35k article gold standard nanomaterial dataset contains 10,229 articles predominantly related to gold, of which 2,577 are not contained in our original MongoDB collection and 602 were not captured by our TF-IDF method. Inspection showed that the volume of articles not contained in our database is largely due to our journal selection during the scraping and parsing of articles, which focuses on materials science-specific journals (whereas Hiszpanski *et al.* selected from all publications in Elsevier journals). For the 602 articles not captured by our data filtering, it was found on manual inspection that many only mentioned "nano–" or "gold"-related vocabulary once or twice throughout the article or only in the abstract. Such articles are not considered valuable for this dataset since they likely do not contain recipes for gold nanoparticle synthesis,

Pipeline Component	ML Method	F1: (precision|recall)
Article filtering	regex/TF-IDF	0.96: (1.00|0.92)
Synthesis paragraph classification	BERT classification	0.90: (0.96|0.85)
Characterization paragraph classification	BERT classification	0.90: (0.93|0.87)
Materials Entity Recognition	BiLSTM+CRF (MatBERT embeddings)	
		0.95:(0.95|0.95) - materials[1]
		0.90:(0.89|0.91) - precursors[1]
		0.85:(0.86|0.83) - targets[1]
Morphology Entity Recognition	Fine-tuned MatBERT NER model	0.87:(0.89|0.84) - Micro average
		0.92:(0.90|0.95) - MOR (morphology)
		0.56:(0.70|0.52) - DES (descriptor)
		0.70:(0.83|0.64) - MES (measurement)
		0.69:(0.81|0.62) - SIZ (size value)
		0.91:(0.94|0.91) - UNT (unit)
Synthesis actions[2]	BiLSTM (Word2Vec embeddings)	0.89 (0.90|0.88)
Synthesis conditions[3]	Rule-based	
– Temperature		0.94: (0.97|0.92)
– Time		0.93: (0.98|0.89)
Material quantities[3]	Rule-based	0.87: (0.90|0.85)
Seed-mediated tag	Rule-based	1.00: (1.00|1.00)

Table 4. Text extraction model accuracies. [1]Metrics from He *et al.*[33]. [2]Metrics from associated manuscript on synthesis actions extraction[35]. [3]Metrics from accepted publication on solution synthesis extraction[42].

so their absence is appropriate. No false positives (i.e. articles that our pipeline determined to be related to gold nanomaterials but that Hiszpanski *et al.* determined to be related to another composition) were found from our extraction.

Manual validation was previously performed for 100 solution-based synthesis paragraphs for another recently accepted dataset manuscript[42]. This was done to determine the extraction accuracy of the rule-based methods used in the extraction pipeline, which included synthesis action conditions (time and temperature) as well as the amounts of materials used. These metrics are included for reference in Table 4 as well. We accepted scores with higher precision than recall for these rule-based methods in order to avoid contaminating the dataset with incorrect information, though potentially sacrificing completeness of a given codified recipe.

Manual checks on the validity of the seed-mediated growth tag for 50 paragraphs were performed, including 25 on paragraphs determined to contain a seed-mediated growth method and 25 on paragraphs containing seedless growth. 49 out of the 50 checks were determined to be valid and true. 1 paragraph was labeled as "seedless", though it only contained purchasing information. We still considered this tagging valid since the incorrect classification is due to the synthesis paragraph classifier earlier in the pipeline. The accuracy for this tagging method is shown in 4.

Finally, the F1 score, precision, and recall for each of the paragraph classifiers and the MorphER model (along with the F1 score, precision, and recall for each of the constituent entities) are also shown in Table 4. For the binary classification models, similarly to the rule-based methods discussed above, we accepted scores with higher precision than recall in order to avoid erroneous classifications of paragraphs that should be data-rich, and thus avoiding inflating the breadth of the present dataset.

Dataset mining. The statistical breakdown of the recipes contained in this dataset are visualized and some basic correlations are explored. These are then discussed in the context of current knowledge in the field.

First, we present an overview and statistical breakdown of the precursors used for gold nanoparticle synthesis across the literature in Fig. 2. Some measures were first taken to standardize the information extracted by MER (see "Materials entity recognition (MER)") from each of these paragraphs. This included manually normalizing all synonyms for a given precursor to a single precursor name, as well as investigating and mapping variances of their token representations (e.g. from text-scraping errors, similar unicode characters, typos, etc.) to a single precursor name. The map consists of appropriately curated regex strings capturing these variations for a given precursor. This mapping is provided as a JSON file in the associated GitHub repository for this dataset (see "Code availability"). The presence of each precursor among seed-mediated and seedless growths is also reflected in this breakdown. The overwhelming presence of $HAuCl_4$ is expected since this is the most prevalent gold source for synthesizing nanoparticles, with $AuCl_3$ and $NaAuCl_4$ following. 20 synthesis paragraphs were inspected that did not show any of these gold sources extracted as precursors. From these, it appeared that 15 paragraphs contained incomplete synthesis descriptions in the text. These were most often brief statements regarding the method of synthesis (e.g. "AuNPs were synthesized through the Turkevich method…") followed by a description of their resultant size. Although the synthesis information for these paragraphs was incomplete, they still often included successfully extracted morphological information that we consider valuable for the purposes of this dataset. The remaining 5 paragraphs showed issues with materials entity parsing (see "Materials entity recognition (MER)") due to unusual syntactic structure (here, the gold precursor would be extracted and classified as *other_material*

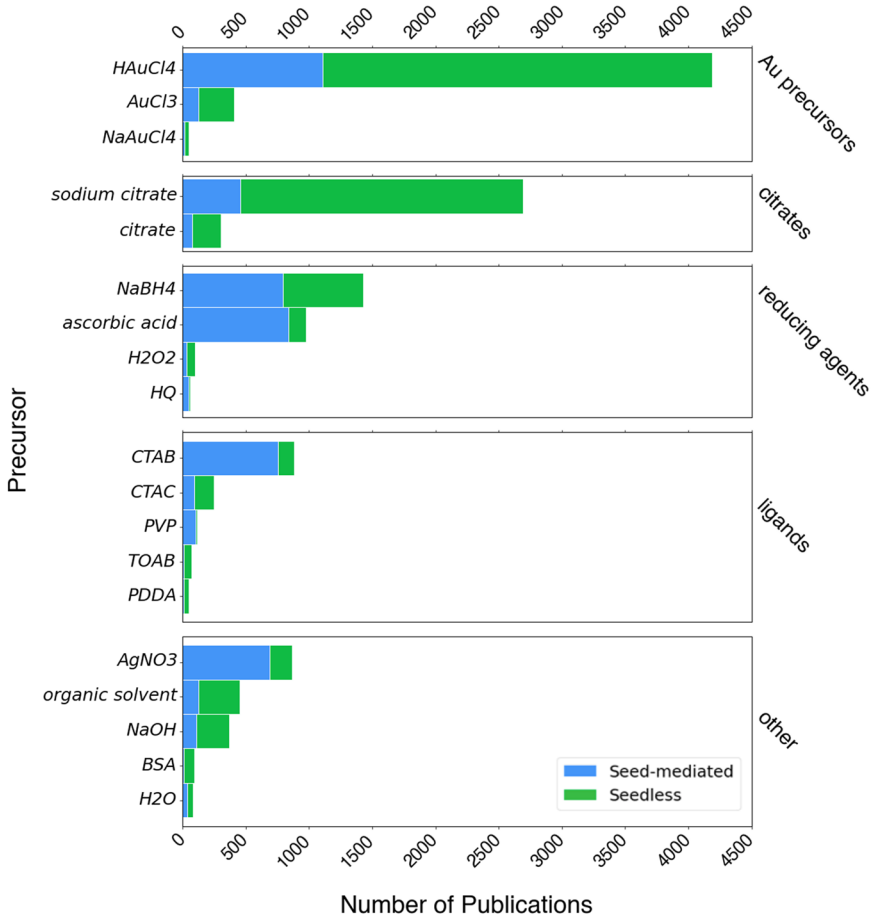

Fig. 2 Frequencies of most common AuNP synthesis precursors. The most frequently extracted precursors using materials entity recognition (MER, "Synthesis recipe extraction") were inspected and compiled into a regular expression-based synonym map, which is housed in the available code repository (see "Code availability"). The precursors are binned by their function in AuNP synthesis and their presence in the number of publications employing seed-mediated growth or seedless methods is distinguished. Citrates are considered in their own category since their function varies depending on the method used, for instance citrates are used as a reducing agent and a ligand in Turkevich or Frens reduction, but only as a ligand in other reduction methods. Only the precursors appearing in more than 50 articles are shown. Precursors were counted once per article within the seed-mediated and seedless growth categories for this analysis to avoid double counting precursors which may be mentioned in a purchasing paragraph and a synthesis paragraph, both of which can sometimes be classified as a synthesis paragraph by our binary gold nanoparticle synthesis paragraph classifier.

as opposed to *precursor*, according to the data structure in Table 2). To better organize the distribution of precursors, we binned each according to their function in a given synthesis. Citrates were given their own bin since they can be used as either a ligand or both a reducing agent and ligand (as in Turkevich[43] or Frens[44] reduction), and because it is currently difficult to extract the specific role of a precursor using our language processing tools. Strong and weak reducing agents were binned together, where $NaBH_4$ is used as a strong reducing agent while the other three are considered weak. This breakdown also indicates which precursors are frequently used for seed-mediated growths, like CTAB and $AgNO_3$ for the growth of nanorods[45]. The common precursors used for seedless growths are often based on attested reduction methods like Turkevich or Frens reduction in this dataset, which both incorporate citrate-based precursors[46]. The lower frequency for several of the precursors is likely due to their relatively recent introduction into the field, such as PVP which was used first in gold nanoparticle synthesis in 2017 to limit growth of nanorods as a capping agent[4]. The low presence of water is likely due to the manner in which precursors are extracted for this dataset using MER, which can extract precursor entities like "water" and "H_2O", but cannot infer water as a precursor from descriptions of solutions like "aqueous".

Moving beyond synthesis details, we also analyze the breakdown of the morphologies discussed in the literature and how those have varied cumulatively across time. Fig. 3 represents the proportion of the most discussed morphologies in the gold nanoparticle literature published between 1998 and 2021. For the purposes of this breakdown, only articles that discuss a single morphology are considered. Through this filtration, the breakdown consists of 1,744 articles out of the 5,154 in the dataset. Morphologies were determined using the `morphologies` and `descriptors` fields within the `morphological_information` field, which combines

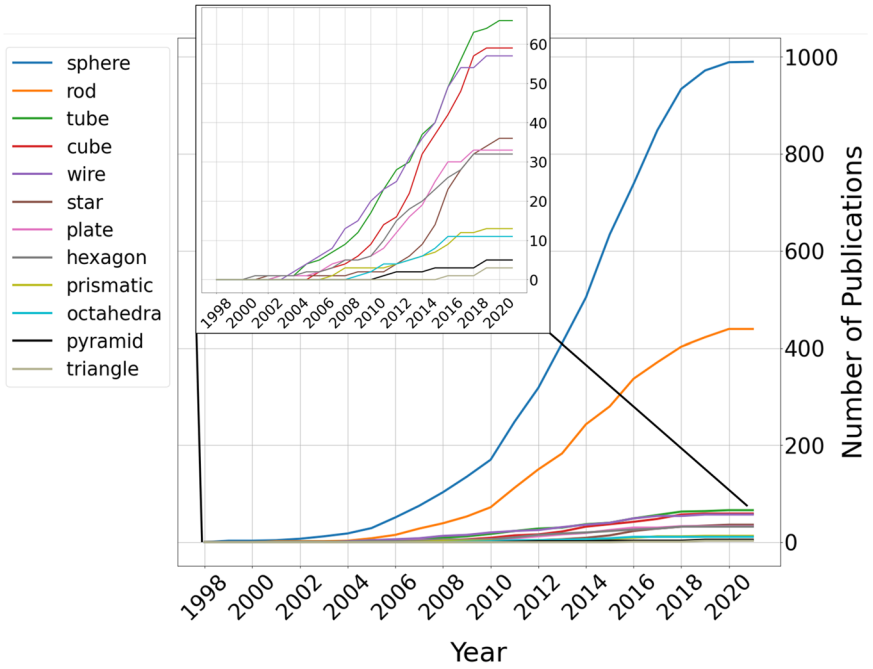

Fig. 3 Breakdown of reported AuNP morphologies discussed by year. Similarly to the frequent precursors analysis, we compiled a regular expression-based synonym map for the most frequently extracted morphological entities and descriptors. The timeline represents the cumulative number of publications discussing each of the specified morphologies from 1998–2021. The entire plot represents 1,744 articles from the dataset, each of which discuss only one of the morphologies in the set depicted.

multiple-entity strings from the MorphER extraction results. The synonym map used to normalize the extracted entities is available in the associated codebase, which was constructed in a similar manner to the precursors synonym map. The strong presence of spherical particles across all years is due to their longevity in the field, being synthesized through a formal procedure by Faraday as early as 1857[47]. Spherical particles are also straightforward to synthesize, with facile methods being pioneered by Turkevich[43] and Frens[44] in the 1950s and 1970s, respectively. Rods are discussed in a quarter of collected publications, more than any of the other anisotropic shapes combined. This reflects the trends in the literature[9], mostly due to their highly tuneable optical properties and more recently developed convenient wet synthesis methods[48].

Finally, we explore correlations between the use of certain precursors and the target morphology of a given synthesis. Using the filtration process described above to consider only single morphology publications related to spheres, rods, tubes, cubes, wires, and stars yielded 1,647 publications. Assuming the one mentioned morphology is indeed the target, we developed a heat map presenting the proportion of select precursors and common precursor ions ($AuCl_4^-$, citrate, CTAB, BH_4^-, ascorbic acid, and Ag^+) mentioned in publications with each target morphology (Fig. 4). In this plot, "Citrate" also contains sodium citrate precursors. The extracted precursors used in a given synthesis were matched against the precursor synonym bank discussed previously. With this additional filtration step, a total of 1,511 publications with only one of the select set of morphologies mentioned and also having at least one of the select precursors are shown in the heat map. A few general trends are reflected in this illustration. First, the frequent mentions of CTAB, Ag^+, ascorbic acid, and BH_4^- are distinct for nanorod synthesis publications. In particular, the use of $AgNO_3$ and CTAB to control the quality and characteristics of gold nanorod growth is well-known to the nanoparticle synthesis field[45]. $AgNO_3$ is used to control the aspect ratios of the rods and there was a recent shift in seed-mediated growth from citrate-capped gold seeds to CTAB-capped gold seeds because the latter showed an improvement on earlier particle formation limitations (e.g. noncylindrical rods, spherical impurities, etc.). Second, citrate is most prominantly used in the synthesis of spherical particles. As was discussed regarding the precursor breakdown earlier (Fig. 2), citrate was used in the seminal experimental works by Turkevich[43] and later by Frens[44] as both a reducing and stabilizing agent. These methods are still among the most prominent for synthesizing spherical gold nanoparticles, as is reflected by Fig. 4.

As was discussed in "Synthesis recipe extraction", the language processing tools and methods used to create this dataset were adapted from tools previously developed for the extraction of solid-state[28] and solution-based[42] recipes. The experimental methods used for nanoparticle synthesis are distinct from other materials synthesis methods. This is particularly so from solid-state methods, but even holds for other solution-based synthesis methods. Because of this, we built additional extraction methods (see "AuNP synthesis paragraph classification", "Characterization paragraph classification", and "Characterization entity recognition (MorphER)"), on top of those that were used for the construction of text-mined solid-state and solution-based recipe datasets, to better handle such synthesis details. However, there are still some pitfalls in this combination of extraction methods

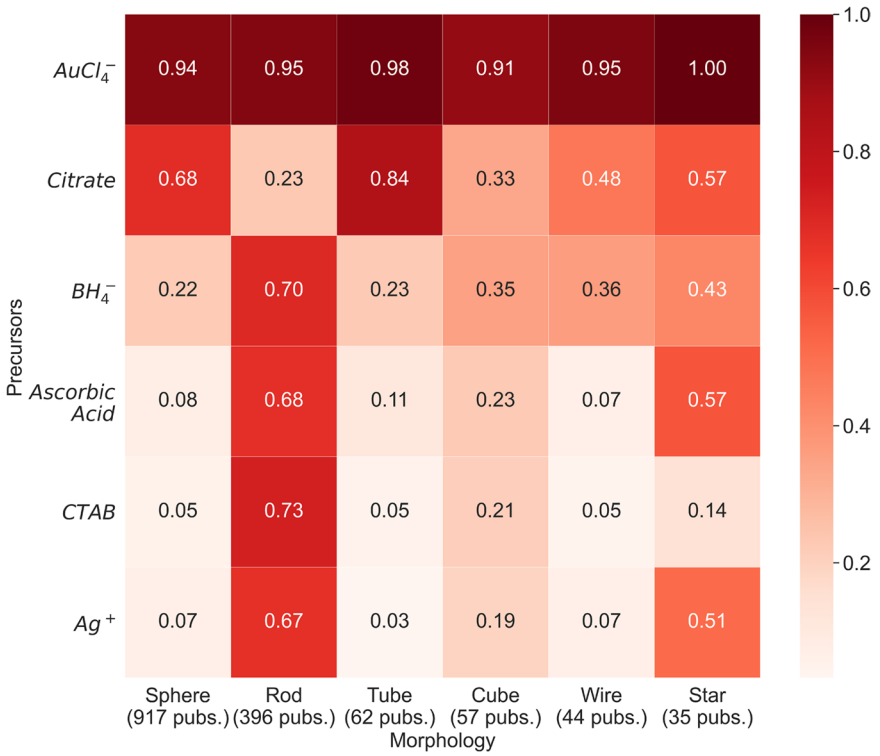

Fig. 4 Heatmap depicting correlation between precursors and resultant AuNP morphologies. The heat illustrated in a given cell represents the fraction of morphologically-targeted articles (say, the fraction of sphere-related articles) which use that particular precursor among one or more precursors it uses in the recipe. For instance, the top left cell shows that more than 90% of purely sphere-related AuNP synthesis papers use $AuCl_4^-$ as a precursor. "Citrate" also includes sodium citrate precursors. The entire heatmap describes 1,511 single morphology-targeted articles with at least one of the precursors or precursor ions shown on y-axis.

that we are addressing for future iterations of this dataset. First, the order of synthesis actions is particularly important in seed-mediated nanoparticle synthesis, which is found to represent a substantial fraction of the major synthesis methods found in the literature (see Fig. 2). This method is comprised of a seed solution preparation step, a growth solution preparation step, and a step that combines these two. Currently, our synthesis action extraction method cannot distinguish these three steps as separate synthesis procedures. Therefore, isolating specific synthesis procedures for the components of a given seed-mediated synthesis is difficult. Because the seed and growth solutions are often described as ingredients in the text, they can be captured in the subject field of the `procedure_graph` (Table 3), which is determined through dependency tree parsing. To address this issue, noun-phrases parsed in the `subject` field can be used to define the relevant synthesis constituent being manipulated or prepared, and thus separate the synthesis procedures into components for seed-mediated growth. Second, the current materials entity recognition model does not detect entities that do not contain specific material formulae or chemical names. Thus, neither "AuNP seed solution" nor "growth solution" would be detected in the sentence "…3 mL of AuNP seed solution was mixed with 5 mL of growth solution to produce the final nanorods." Because of this, the corresponding amounts for each component of the synthesis cannot be extracted. Such information is important for seed-mediated growth, so we plan to address this by using the results of the aforementioned `subject` field and use seed and growth solution-related noun phrases as anchors for an additional material amounts extraction step if the paragraph describes seed-mediated synthesis. Finally, there is currently no way to distinguish extracted morphologies as either the desired target morphology or just a morphology mentioned off-hand by the author. To address this, we plan to develop an additional layer on top of the current morphology entity recognition model that classifies those entities predicted to be **MOR** into either target (**TGT**) or miscellaneous morphologies (**MIS**), similar to the strategy used for materials entity recognition (see "Materials entity recognition(MER)").

Usage Notes

The present dataset is provided as a single JSON file that can be read using all major programming languages (e.g. Python, Matlab, R, etc.). It is publicly available at https://doi.org/10.6084/m9.figshare.16614262.v3[41]. No dependencies are required to access the contents of the dataset.

We invite users to utilize this dataset, among other applications, for the purposes of gold nanoparticle synthesis literature reviews or to query specific recipe protocols that achieve a desired morphology or size.

This data descriptor defines a static version of the gold nanoparticle synthesis and characterization dataset; however, we intend to update the dataset in the repository below on a regular basis here: https://github.com/

CederGroupHub/text-mined-aunp-synthesis_public. This will soon include updates addressing the issues discussed at the end of "Dataset Mining" as well as for morphological entity linking and linking target morphologies to specific recipe protocols.

Code availability

Scripts developed for the generation of this dataset as well as notebooks for example data analysis are available at https://github.com/CederGroupHub/text-mined-aunp-synthesis_public, along with an acknowledgement for this paper. The libraries use for this project are: *ChemDataExtractor*, *SpaCy*, *scikit-learn*, *gensim*, *Tensorflow*, *Keras*, *PyTorch*, and *Simple Transformers*.

Received: 30 September 2021; Accepted: 8 April 2022;
Published online: 26 May 2022

References

1. Liu, X. *et al.* A one-step homogeneous immunoassay for cancer biomarker detection using gold nanoparticle probes coupled with dynamic light scattering. *J. Am. Chem. Soc.* **130**, 2780–2782 (2008).
2. Dawson, A. & Kamat, P. V. Semiconductor−metal nanocomposites. photoinduced fusion and photocatalysis of gold-capped Ti O_2 (Ti O_2/gold) nanoparticles. *J. Phys. Chem. B* **105**, 960–966 (2001).
3. Kaul, S., Gulati, N., Verma, D., Mukherjee, S. & Nagaich, U. Role of nanotechnology in cosmeceuticals: A review of recent advances. *Journal of Pharmaceutics* **2018** (2018).
4. Requejo, K. I., Liopo, A. V., Derry, P. J. & Zubarev, E. R. Accelerating gold nanorod synthesis with nanomolar concentrations of poly(vinylpyrrolidone). *Langmuir* **33**, 12681–12688 (2017).
5. De Souza, C. D., Nogueira, B. R. & Rostelato, M. E. C. Review of the methodologies used in the synthesis gold nanoparticles by chemical reduction. *J. Alloys Compd.* **789**, 714–740 (2019).
6. Grzelczak, M., Pérez-Juste, J., Mulvaney, P. & Liz-Marzán, L. M. Shape control in gold nanoparticle synthesis. *Chem. Soc. Rev.* **37**, 1783–1791 (2008).
7. Personick, M. L. & Mirkin, C. A. Making sense of the mayhem behind shape control in the synthesis of gold nanoparticles. *J. Am. Chem. Soc.* **135**, 18238–18247 (2013).
8. Agunloye, E., Panariello, L., Gavriilidis, A. & Mazzei, L. A model for the formation of gold nanoparticles in the citrate synthesis method. *Chem. Eng. Sci.* **191**, 318–331 (2018).
9. Lohse, S. E. & Murphy, C. J. The quest for shape control: A history of gold nanorod synthesis. *Chem. Mater.* **25**, 1250–1261 (2013).
10. Mukhamedzyanova, D. F., Ratmanova, N. K., Pichugina, D. A. & Kuz'menko, N. E. A structural and stability evaluation of Au_{12}. *J. Phys. Chem. C* **116**, 11507–11518 (2012).
11. Domingo, M., Shahrokhi, M., Remediakis, I. & Lopez, N. Shape control in gold nanoparticles by n-containing ligands: Insights from density functional theory and wulff constructions. *Top. Catal.* **61**, 412–418 (2018).
12. Chakraborty, I. & Pradeep, T. Atomically precise clusters of noble metals: Emerging link between atoms and nanoparticles. *Chem. Rev.* **117**, 8208–8271 (2017).
13. Talapin, D. V., Rogach, A. L., Haase, M. & Weller, H. Evolution of an ensemble of nanoparticles in a colloidal solution: Theoretical study. *J. Phys. Chem. B* **105**, 12278–12285 (2001).
14. Ren, F. *et al.* Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, 4 (2018).
15. Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mat.* **5**, 641–646 (2006).
16. Weston, L. *et al.* Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **59**, 3692–3702 (2019).
17. Wang, X. *et al.* AutoDetect-mNP: An unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles. *JACS Au* **1**, 316–327 (2021).
18. Szymanski, N. J., Bartel, C. J., Zeng, Y., Tu, Q. & Ceder, G. Probabilistic deep learning approach to automate the interpretation of multi-phase diffraction spectra. *Chem. Mat.* **33**, 4204–4215 (2021).
19. Yan, X., Sedykh, A., Wang, W., Yan, B. & Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Comm.* **11** (2020).
20. Hiszpanski, A. M. *et al.* Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *J. Chem. Inf. Model.* **6**, 2876–2887 (2020).
21. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* **117**, 7673–7761 (2017).
22. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminformatics* **3**, 17 (2011).
23. Swain, M. C. & Cole, J. M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
24. Kononova, O. *et al.* Opportunities and challenges of text mining in materials research. *iScience* **24**, 3 (2021).
25. Olivetti, E. *et al.* Data-driven materials research enabled by natural language processing. *Appl. Phys. Rev.* **7**, 041317 (2020).
26. Kim, E. *et al.* Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater* **29**, 9436–9444 (2017).
27. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
28. Kononova, O. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).
29. Jurafsky, D. & Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence (Pearson Prentice Hall, 2009).
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
31. Trewartha, A. *et al.* Quantifying the advantage of domain-specific pre-training on named entityre cognition tasks in materials science. *Patterns* **3**, 4, 100488 (2022).
32. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. of Mach. Learn. Res.* **3**, 993–1022 (2003).
33. He, T. *et al.* Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem. Mat.* **32**, 7861–7873 (2020).
34. Huo, H. *et al.* Semi-supervised machine-learning classification of materials synthesis procedures. *Npj Comput. Mater.* **5**, 62 (2019).

35. Wang, Z. *et al.* ULSA: Unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discovery*, Advance online publication 10.1039/D1DD00034A (2022).
36. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality (2013).
37. Řehůřek, R. & Sojka, P. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50 (ELRA, Valletta, Malta, 2010).
38. Honnibal, M. & Johnson, M. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378 (Association for Computational Linguistics, Lisbon, Portugal, 2015).
39. Bird, E. L., Steven & Klein, E. *Natural Language Processing with Python. O'Reilly Media Inc* (2009).
40. Huang, X., Neretina, S. & El-Sayed, M. A. Gold nanorods: From synthesis and properties to biological and biomedical applications. *Adv. Mat.* **21**, 4880–4910 (2009).
41. Cruse, K. *et al.* Text-mined AuNP Synthesis Recipes Dataset *figshare*, https://doi.org/10.6084/m9.figshare.16614262.v3 (2021).
42. Wang, Z. *et al.* Dataset of solution-based inorganic materials synthesis recipes extracted from the scientific literature. *Accepted to Sci. Data*. Preprint at https://doi.org/10.48550/arXiv.2111.10874 (2022).
43. Turkevich, J., Stevenson, P. C. & Hillier, J. A study of the nucleation and growth processes in the synthesis of colloidal gold. *Discuss. Faraday Soc.* **11**, 55–75 (1951).
44. Frens, G. Controlled nucleation for the regulation of the particle size in monodisperse gold suspensions. *Nat. Phys. Sci.* **241**, 20–22 (1973).
45. Nikoobakht, B. & El-Sayed, M. A. Preparation and growth mechanism of gold nanorods (NRs) using seed-mediated growth method. *Chem. Mater.* **15** (2003).
46. Herizchi, R., Abbasi, E., Milani, M. & Akbarzadeh, A. Current methods for synthesis of gold nanoparticles. *Artificial Cells, Nanomedicine, and Biotechnology* **44**, 596–602 (2016).
47. Faraday, M. X. the bakerian lecture. - experimental relations of gold (and other metals) to light (1857).
48. Scarabelli, L., Sánchez-Iglesias, A., Pérez-Juste, J. & Liz-Marzan, L. M. A "tips and tricks" practical guide to the synthesis of gold nanorods. *J. Phys. Chem. Lett* **6**, 4270–4279 (2015).

Acknowledgements

This work was funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (D2S2 program KCD2S2). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0020531. We thank Anna Sackmann, Rachael Samberg, and Timothy Vollmer (Science Data and Engineering Librarians at UC Berkeley) for assistance in obtaining Text and Data Mining agreements with the relevant publishers. We would also like to thank Sam Gleason, Xingzhi Wang, Jakob Dahl, Caitlin McCandler, John Dagdelen, Nicholas Walker, and Akshay Subramanian for valuable advice and discussion regarding the development of this pipeline and analysis of the data.

Author contributions

K.C. developed the extraction pipeline, annotated paragraphs for and trained the relevant paragraph classifiers and MorphER model, analyzed the data, and wrote the manuscript. A.T. provided guidance for approaches in developing the pipeline and developed framework for MorphER training. S.L. performed manual data inspection, advised on updates to extraction methods, and analyzed the data. Z.W. developed the material quantities extraction method. H.H. developed the Apache Solr search engine and trained MatBERT. T.H. developed materials entity recognition. O.K. developed synthesis action and conditions extraction method. A.J. supervised the project and wrote the manuscript. G.C. developed the approach, supervised the project, and wrote the manuscript. All authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.C.