CH-302 Team #1
Group Presentation-2

Team Members :
Parth Shah
Abhyuday Patil
Princy Asawa
Janhavi Kashyap
Ranuva Aashrith Vathsal Rao

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

Authors:

Thomas Yong-Jin Han − Materials Science Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States;
orcid.org/0000-0002-3000-2782;
Email: han5@llnl.gov

Authors:

Anna M. Hiszpanski − Materials Science Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

Brian Gallagher − Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

Karthik Chellappan − Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

Peggy Li − Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

Shusen Liu − Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

Hyojin Kim − Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

Jinkyu Han − Materials Science Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States; orcid.org/0000-0002-6374-116X

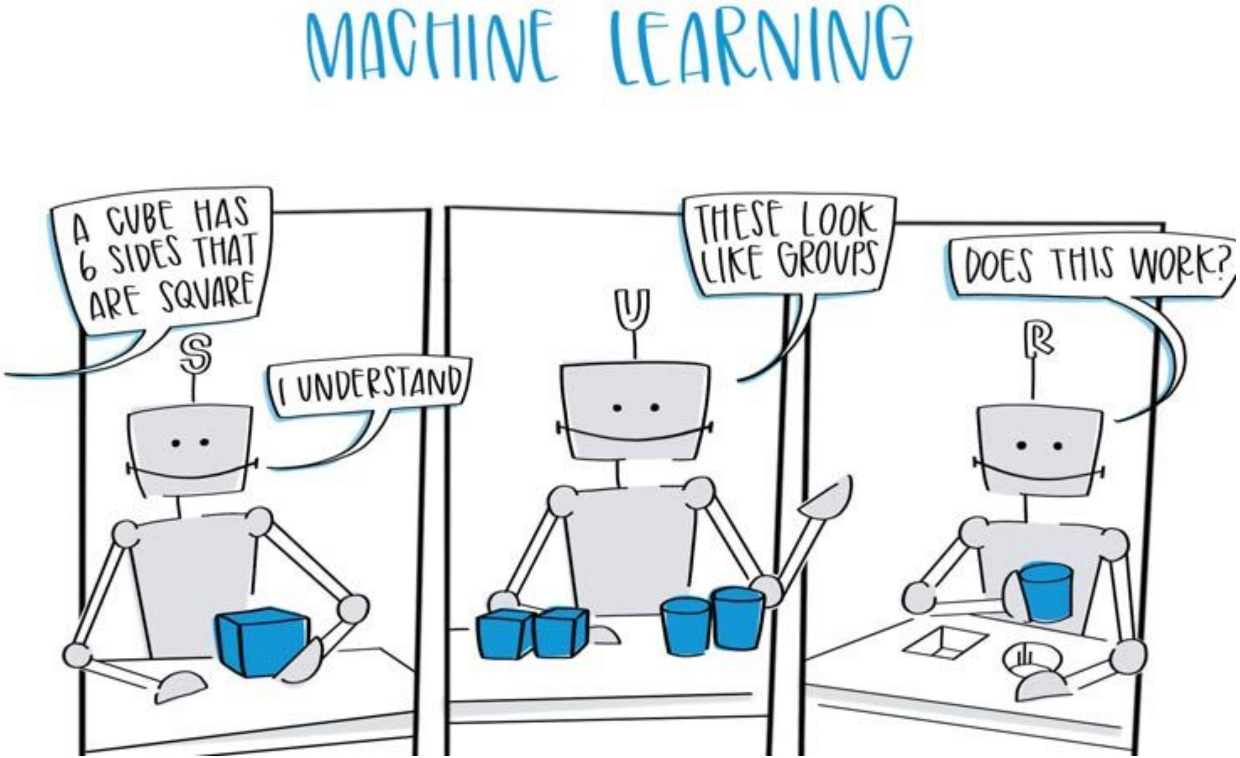
Bhavya Kailkhura − Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

David J. Buttler − Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, United States

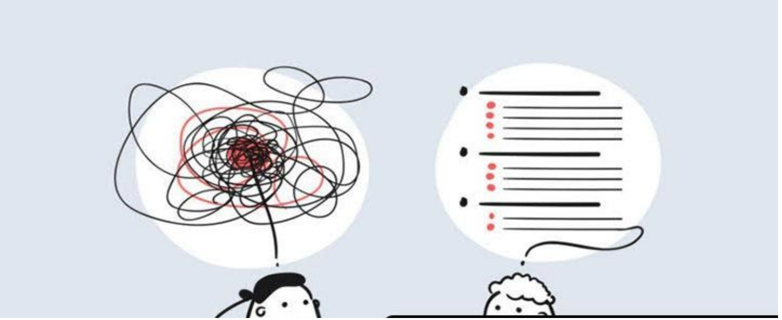
Introduction

Machine Learning:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

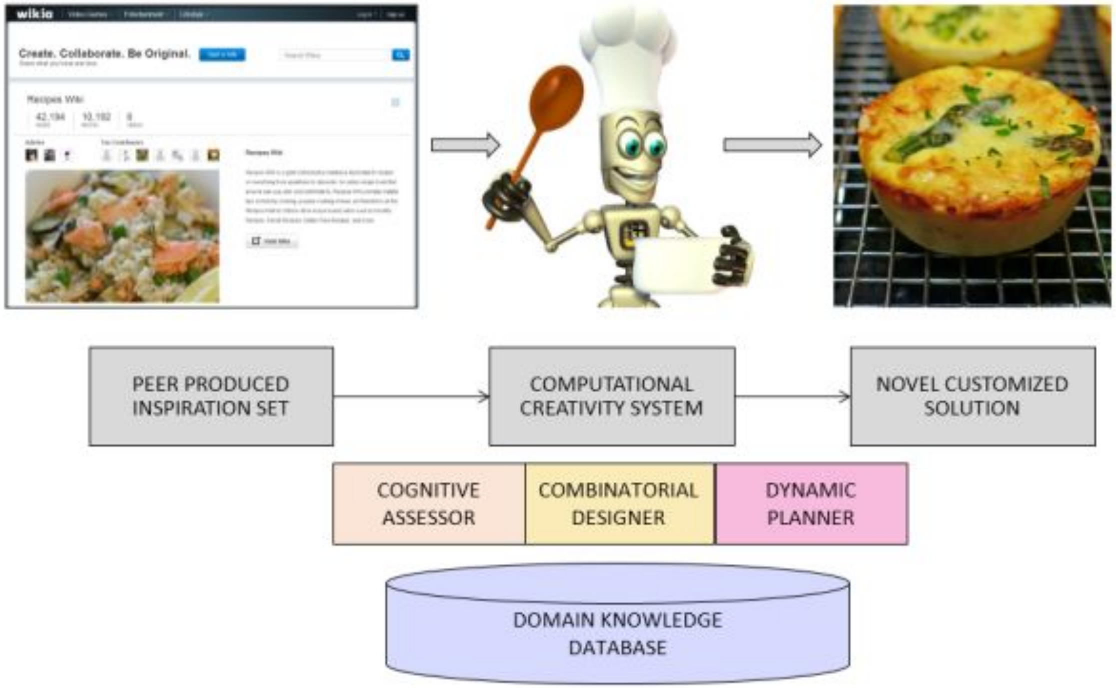


Introduction



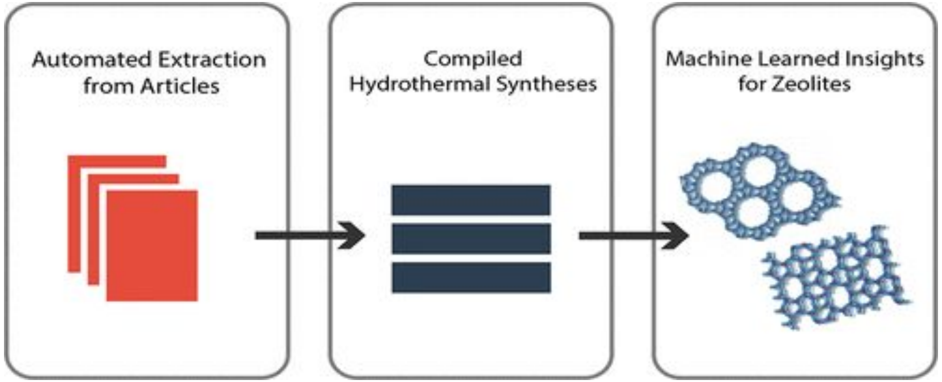
The scientific community has generally agreed on one pseudo standardized form of data across varying subfields: publications. While text is unstructured data, which creates its own complications, scientists have generally agreed on the format (e.g., abstract, experimental details, discussion of
results) and the level of detail to include.

Introduction



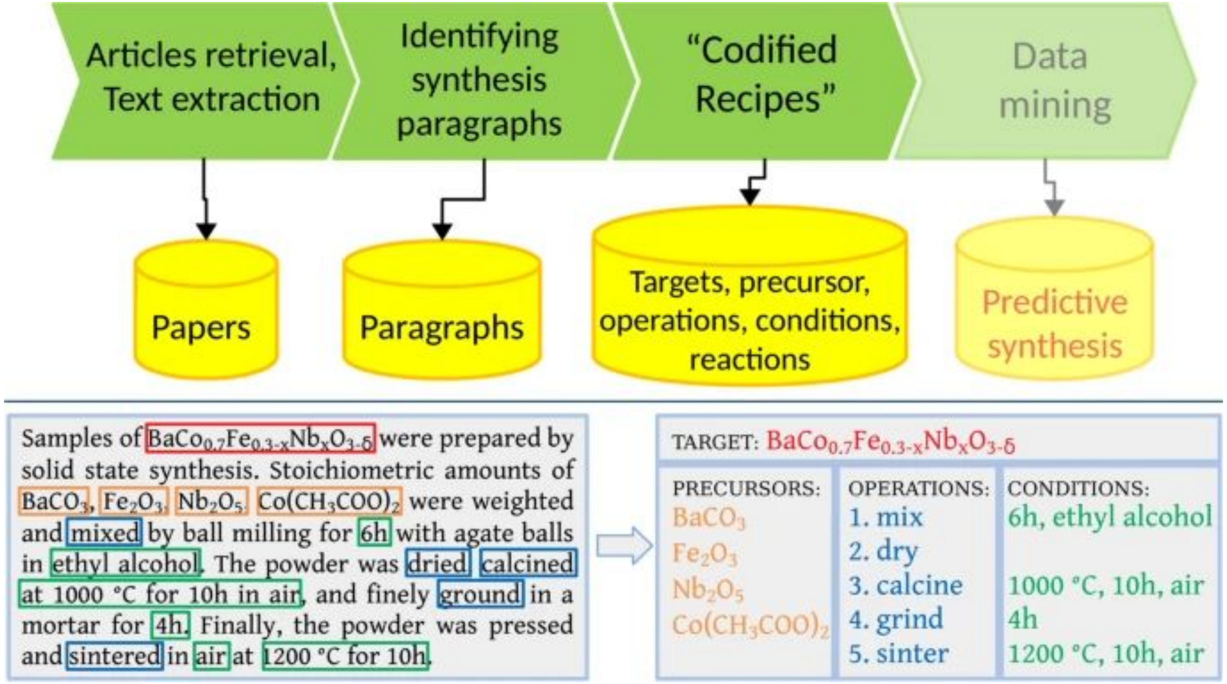
Chef Watson by IBM

Introduction



Zeolite synthesis insights using Machine Learning, Jensen et al

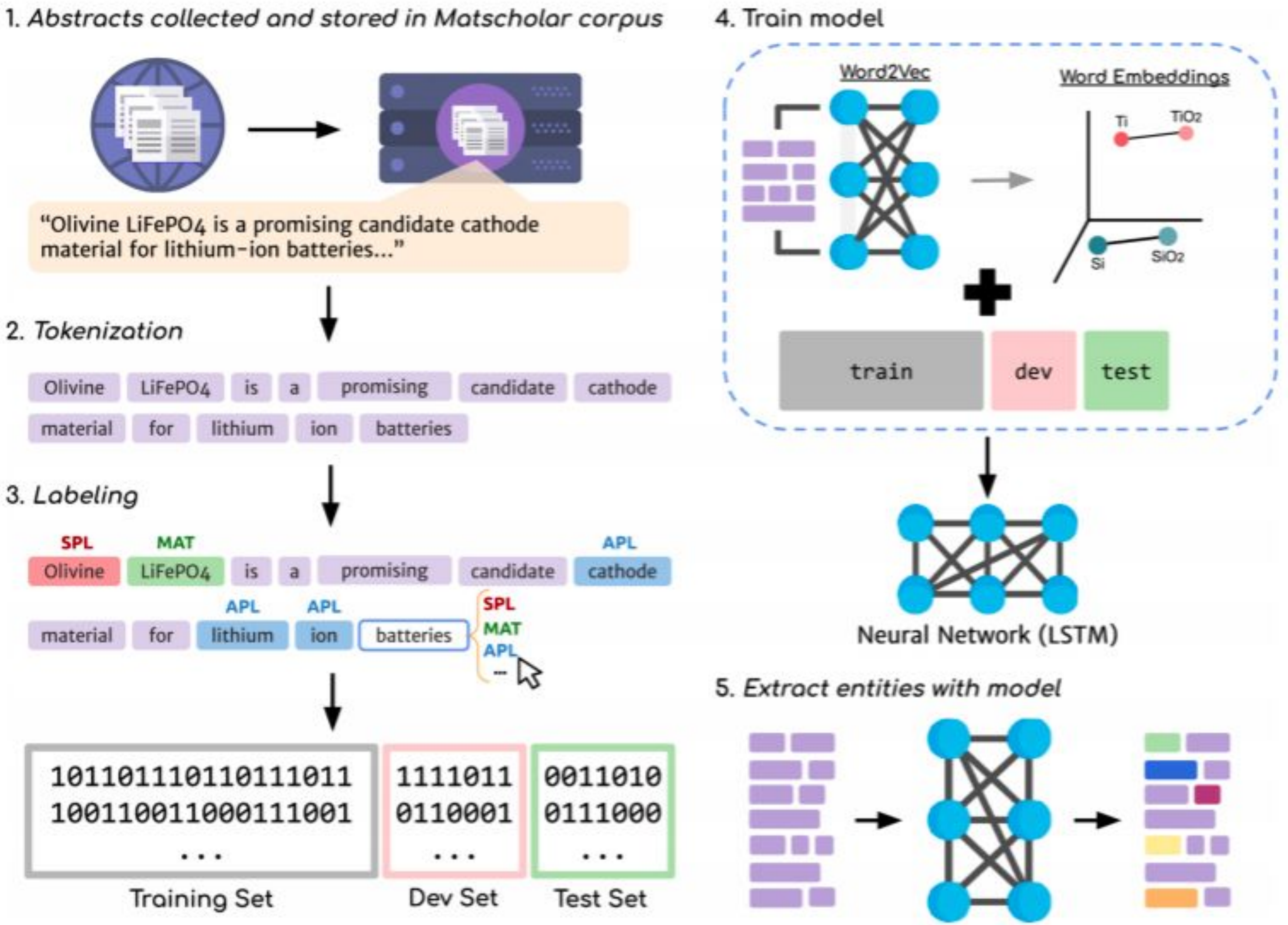
Introduction



Text-mined dataset of inorganic materials synthesis recipes, Konova et al

Introduction

Named Entity Recognition and Normalization Applied to Large Scale Information Extraction from the Materials Science Literature, Wetson et al



Introduction

Nanomaterials are critical for
a number of applications, including

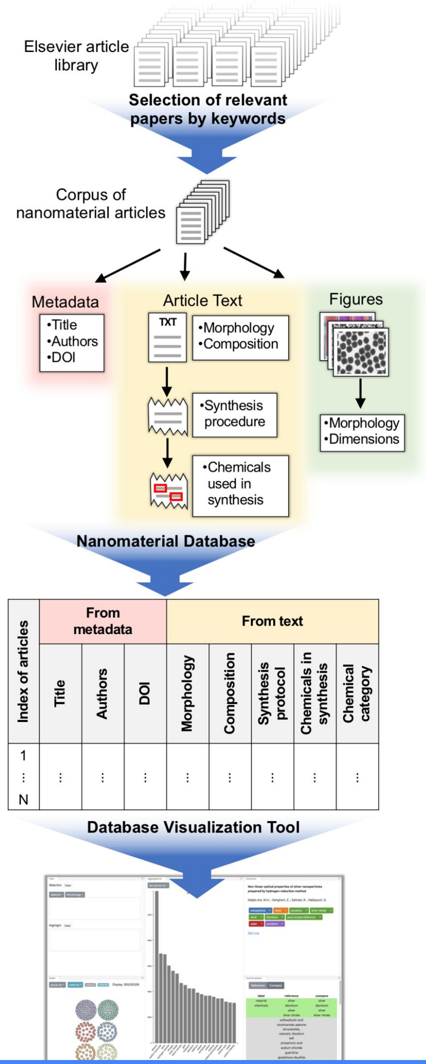
- catalysis,
- Optical components,
- additive manufacturing feedstocks.

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, authors create a suite of tools, for automatically extracting and structuring targeted information for nanomaterials from published scientific articles and demonstrate the kinds of insights such information can provide.

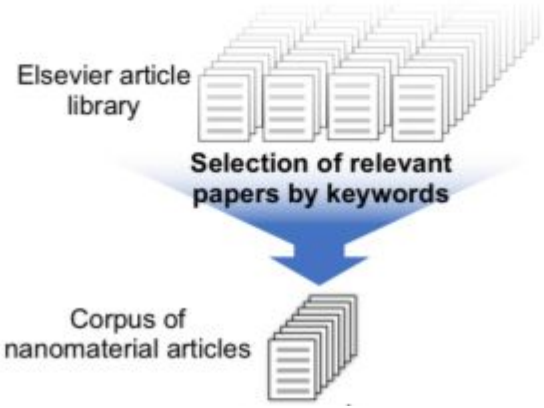


Result & Discussion - A Pipeline

Corpus	Building a Relevant Corpus of nanomaterials articles
Metadata	Extracting metadata on each article (i.e., title, authors, DOI).
Protocols	Processing the text of each article to identify the target Nanomaterials' morphology and composition, synthesis procedure, and chemicals used in its synthesis.
Morphology and Size Distributions	Extracting figures and further processing SEM/TEM figures of nanomaterials to obtain their morphology and size distributions

Building a Relevant Corpus

- Keywords:
  - “X nanoY”, where **X**
    is the nanomaterial compositions of interest, and
    Y indicates the nanomaterial morphology of interest.
  - “Synthesis”
- **35,345** unique papers obtained.



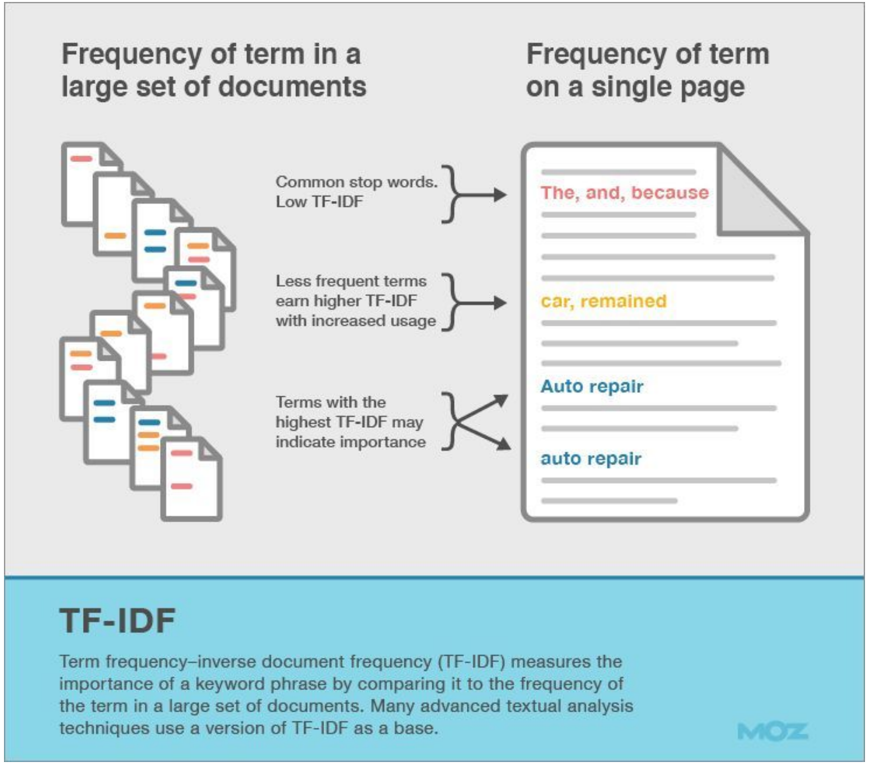
Identifying Composition and Morphology from Text

- TF-IDF (Term Frequency-Inverse Document Frequency) statistic
- high TF-IDF <-> greater relevance
- “gold standard” -> set of 99 hand-labeled papers
    - 100% accuracy on composition prediction
    - 95% accuracy on morphology prediction.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents



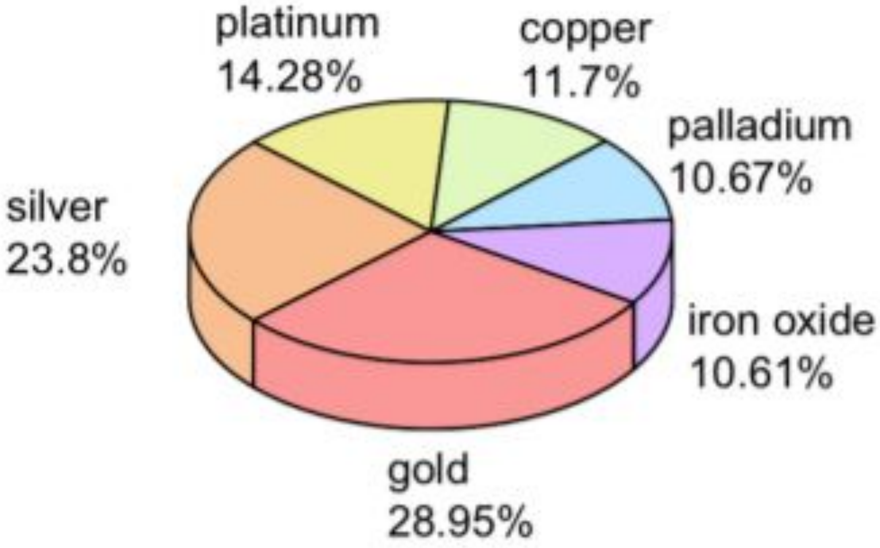
Result & Discussion

- **Overrepresented** combinations -> "hot topics" or commonly synthesized materials.
- **Underrepresented** combinations -> difficult to synthesize material-morphology combinations or areas ripe for exploration

		Nanomaterial composition						
		gold	silver	platinum	iron oxide	copper	palladium	SUM
Nanomaterial morphology	particle	6,748	5,931	2,922	3,082	2,416	2,612	23,711
	tube	669	393	721	170	367	324	2,644
	structure	717	521	390	119	320	207	2,274
	wire	512	498	223	40	290	128	1,691
	rod	610	274	116	83	187	79	1,349
	sheet	265	254	237	55	139	129	1,079
	crystal	284	242	178	103	202	129	1,138
	sphere	190	117	98	50	77	60	592
	cube	48	78	59	21	41	58	305
	flower	50	27	48	5	20	19	169
	octahedral	30	34	30	17	53	18	182
	star	61	10	1	2	5	1	80
	ribbon	29	23	24	2	19	6	103
	triangle	18	9	0	0	0	1	28
	SUM	10,231 28.95%	8,411 23.80%	5,047 14.28%	3,749 10.61%	4,136 11.70%	3,771 10.67%	35,345

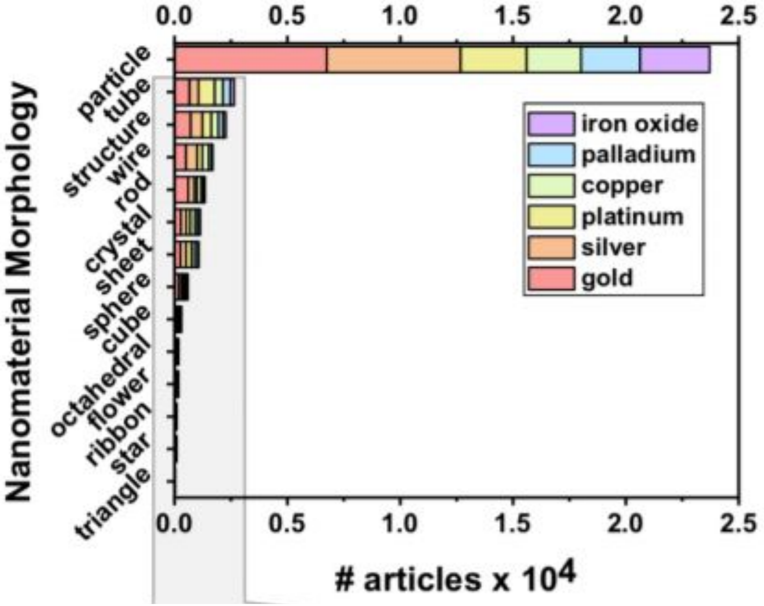
Papers in corpus by nanomaterial morphology and composition.

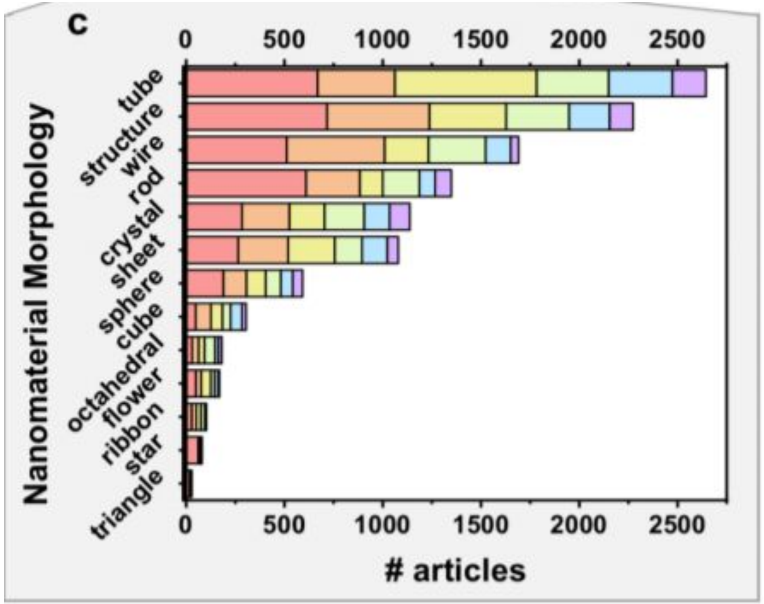
Result & Discussion



Distribution of the Corpus on the basis of the materials composition (independent of the specific nanomorphology)

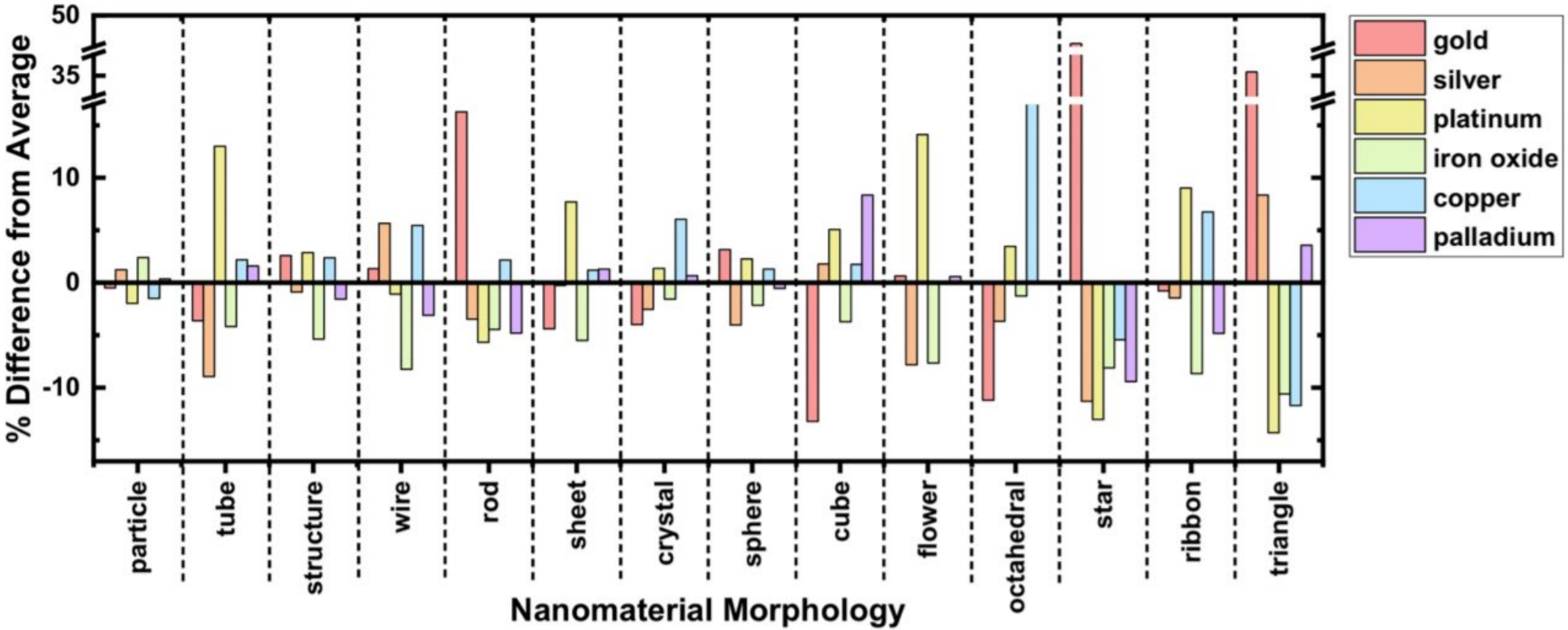
Result & Discussion





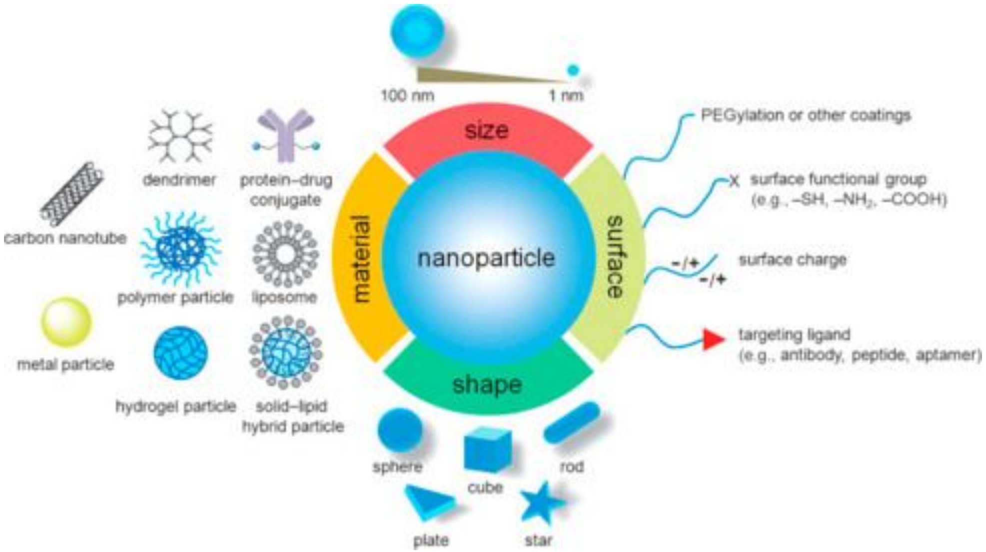
Nanomaterial Morphology vs its occurrence in unique articles and material-wise distribution

Result & Discussion



Identifying Nanomaterials' Synthesis Protocols from Text.

Each sentence in every article was analysed individually to determine whether it contains details relevant to the synthesis procedure.



Logistic Regression Classifier

Phase 1

Web-based Brat annotation tool was used to hand-annotate the synthesis-related sentences in 18 nanomaterials synthesis articles

Phase 2

The model was refined the model iteratively, using an active learning approach

Phase 3

The trained model was applied to automatically identify synthesis sentences in the 99-article gold standard data set

Phase 4

A final model was trained based on narrowed focus on synthesis protocol

Identifying Nanomaterials' Synthesis Protocols from Text.

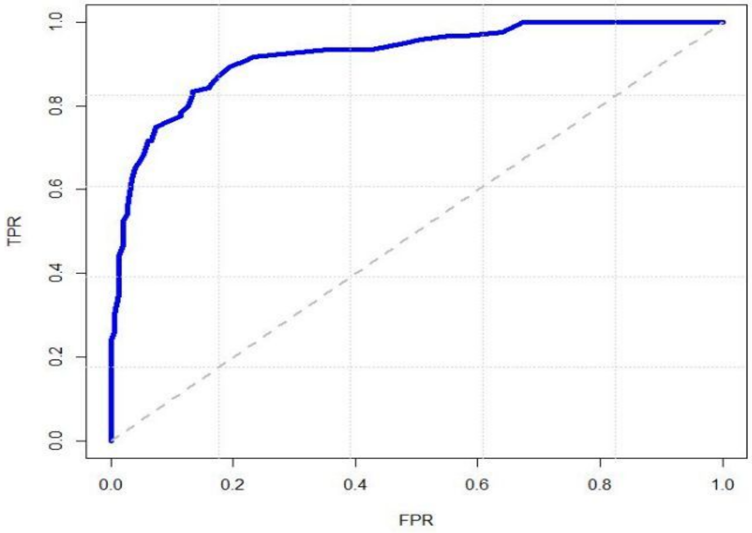
Gold Standard Set	Logistic Regression Classifier	Evaluation
27,125 sentences = 629 positive + 26,496 negative examples	Sentences classified as either relevant or non relevant to nanomaterials' synthesis	Leave-one-article-out cross-validation on the labeled data set

Identifying Nanomaterials' Synthesis Protocols from Text.

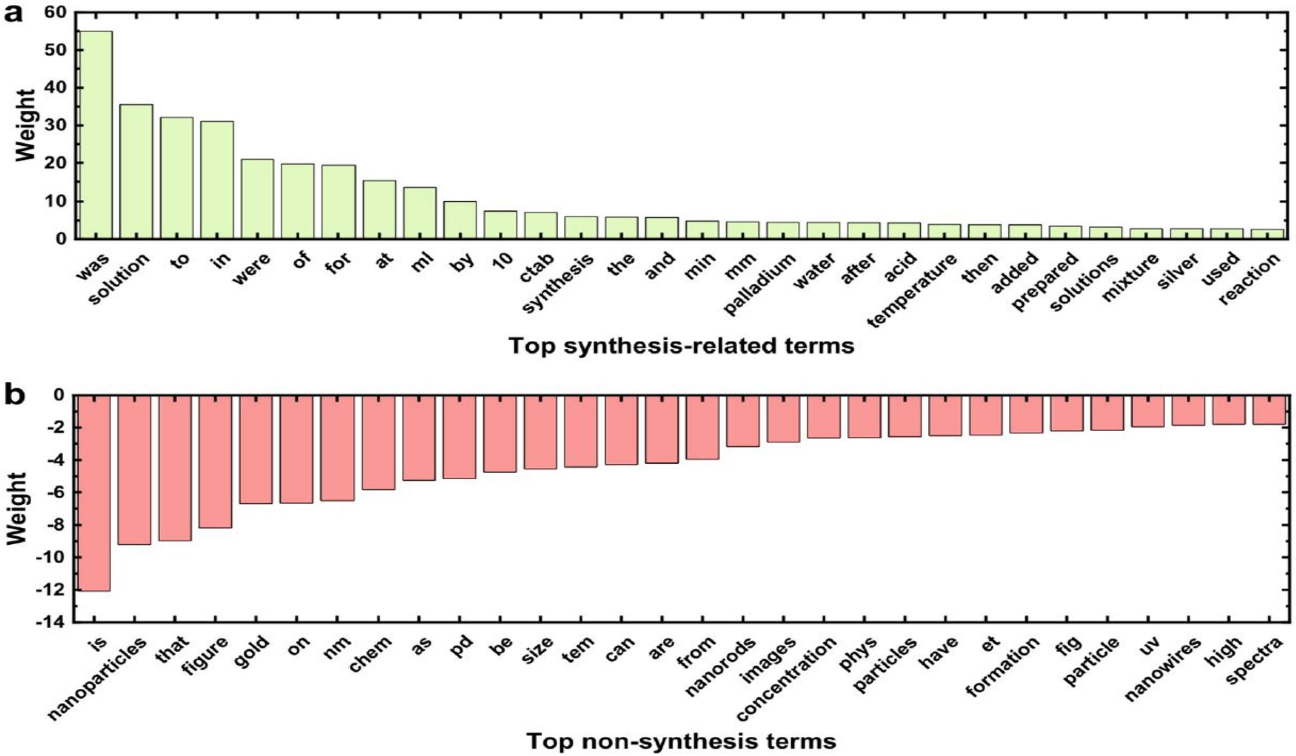
Evaluation Metrics

METRIC	OBSERVATION
AUC	0.99
Precision	52%
Recall	90%

ROC curve



Identifying Nanomaterials' Synthesis Protocols from Text.



Identifying Chemicals Used in Nanomaterials' Synthesis

Other than a library of Nanomaterial synthesis procedures, it will be helpful to know the generalizations and differences between the procedures

Synthesis Procedures

NER

1. Chemicals used
2. Process conditions

Chemical Entity Recognition

Tokenization → Chemical Identification → Testing

Tokenization

Tokenization → Chemical Identification → Testing

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.
- For chemistry related texts, a number of chemical text tokenizers have been developed
- 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.

OSCAR4

ChemSpot

Banner

Chemical Identification

Tokenization → Chemical Identification → Testing

Dictionary → Rules/Patterns → Context and parts of speech → CRF Models

1. Stanford's CoreNLP
2. AllenNLP

Unsupervised Learning from training corpus, to extract features such as n-gram frequencies, word shape features, Word2Vec, character-level encodings, etc.

Open Source Tools:

1. ChemSpot
2. ChemXSeer
3. BANNERCHEMDNER

Tokenization → Chemical Identification → Testing

Testing

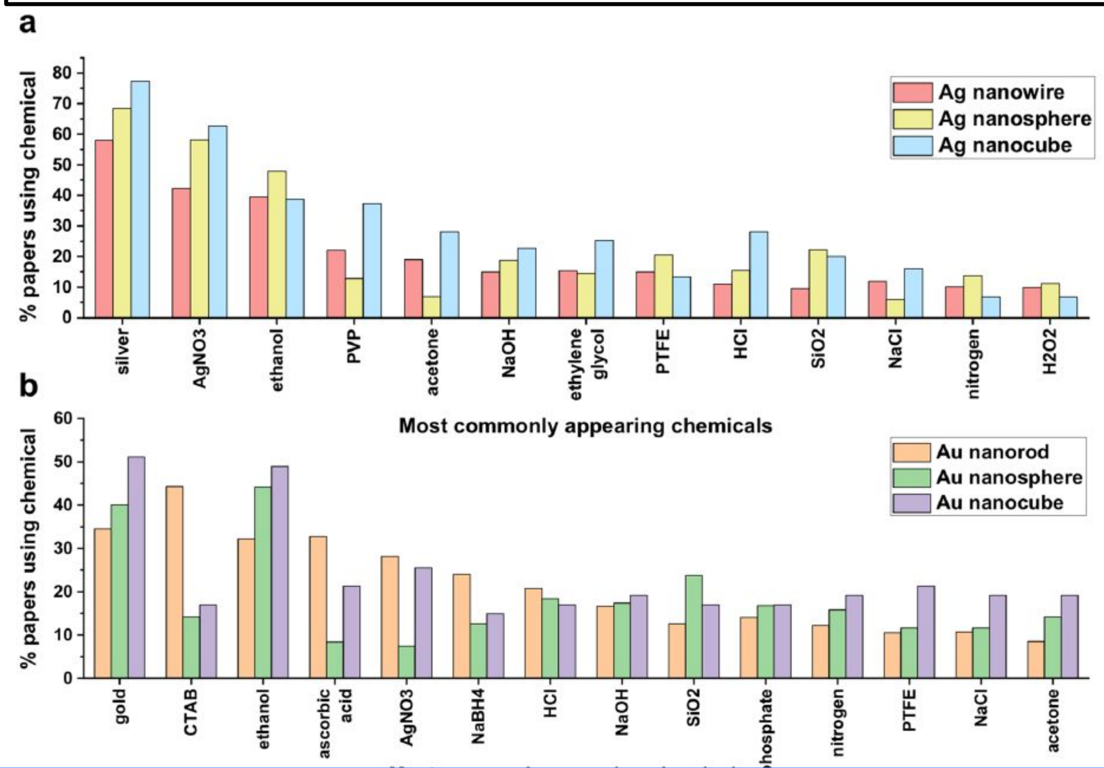
The performance of these various CER tools tested with the 99-article hand-labeled, "gold standard" papers

CER Tool	Precision (%)	Recall (%)	F1 (%)
ChemDataExtractor	97.1	79.1	87.2
StanfordNLP	90.4	80.1	84.9
Chem Spot 2.0	93.2	76.0	83.7
BANNERCHEMDNER	95.0	74.1	83.2
ChemXSeer	96.9	70.0	81.3
OSCAR4	64.6	94.6	76.8
AllenNLP	63.6	70.7	67.0

Selected as the best model

Demonstration

Most commonly occurring chemicals in papers involving the synthesis of Ag nanowires, nanospheres, and nanocubes and Au nanorods, nanospheres, and nanocubes



Key Insights

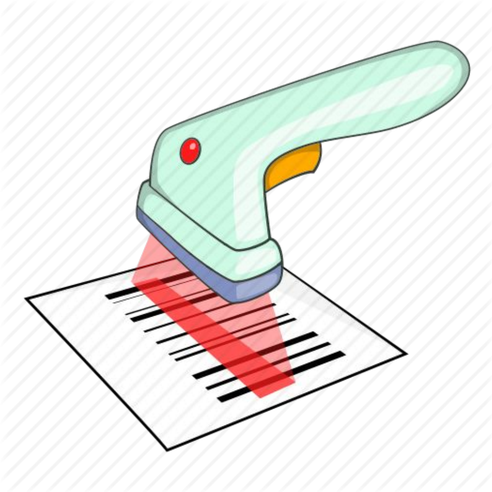
1. Many chemicals appear commonly regardless of nanomaterial morphology or composition, like ethanol, which is often used for washing, and elemental silver and gold.

2. polyvinylpyrrolidone (PVP) and hydrochloric acid (HCl) both appear nearly twice as frequently in the syntheses of Ag nanocubes than Ag nanospheres or nanowires

3. Within the Au nanomaterial articles, hexadecylcetyltrimethylammonium bromide (CTAB) is commonly occurring in Au nanorod articles as compared to nanosphere or nanocube articles.

4. Ascorbic acid and AgNO3 also occur more than twice as often in Au nanorod and nanocube synthesis protocols than in Au nanosphere protocols.

Extracting Information from Figures



- Images reported in nanomaterials synthesis-related articles are valuable.
- Scanning Electron Microscopy (SEM)
- Transmission Electron Microscopy (TEM)
- Tools for capturing and processing image information are developed.

Capturing Images

Recognising SEM and TEM images

+

Processing Image Information

Accomplished using a transfer learning approach with a convolutional neural network model

→

Valuable Immediate Perspective

Nanomaterials geometry, dimensions, and polydispersity.

Extracting Information from Figures

- Recognize and extracting SEM and TEM images from figures.
- These images are then analyzed to:
    - Identify the nanomaterials morphology present.
    - Provide dimensional estimates of all the nanomaterials present in the image.

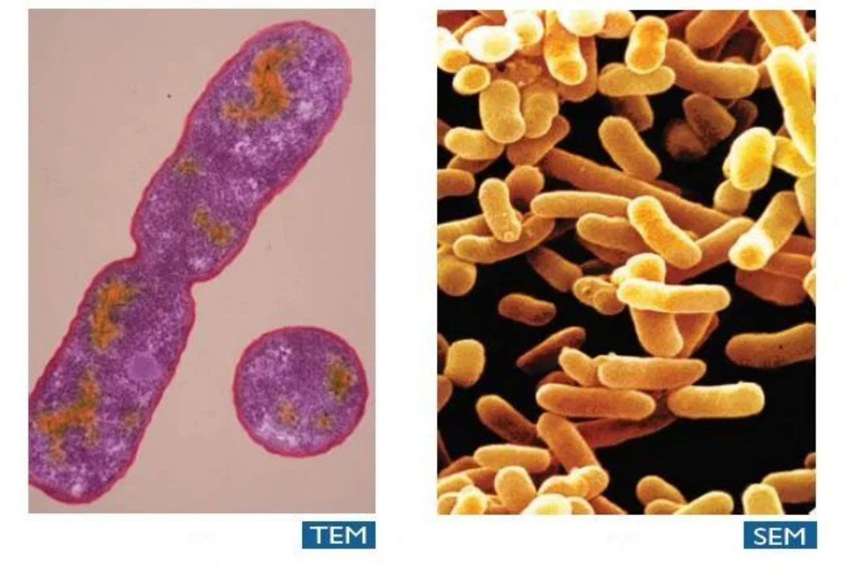


Image Processing Tools

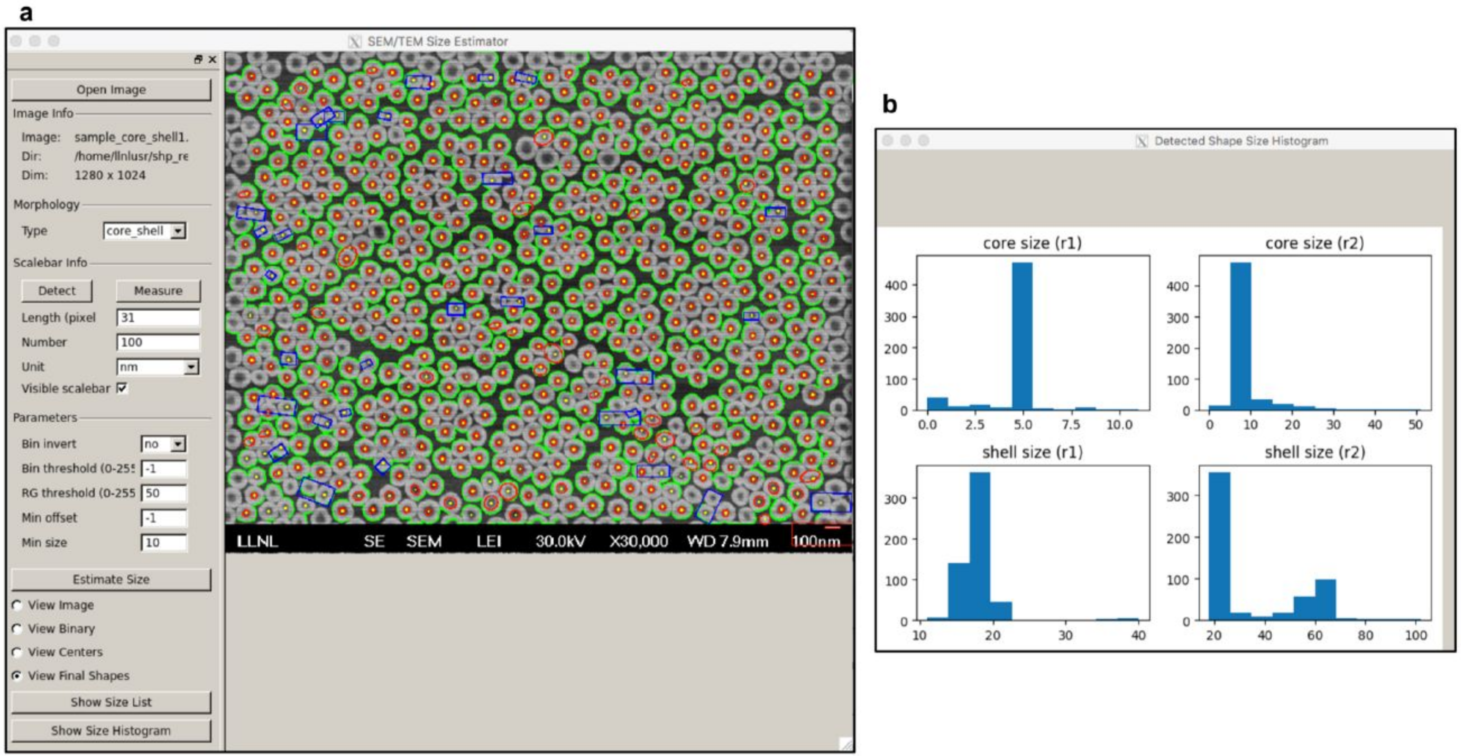
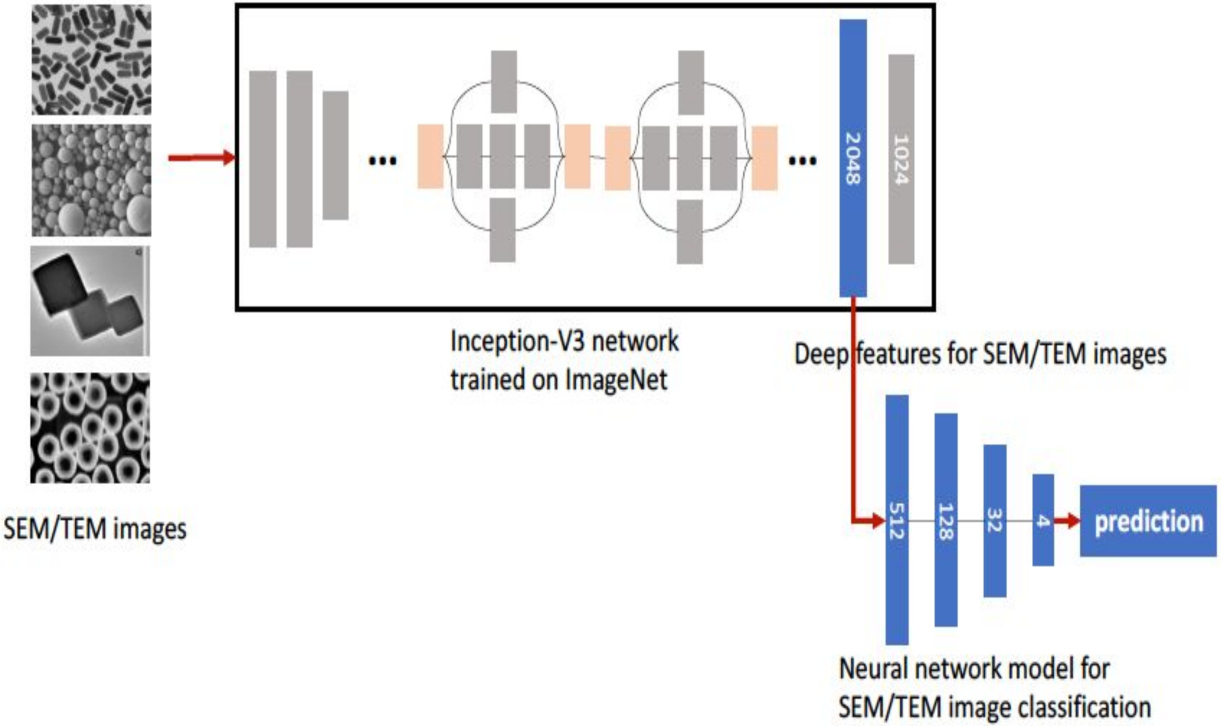


Figure 6. (a) Input and (b) output screenshots of the GUI-form of the pipeline that automatically analyzes nanomaterial SEM and TEM images extracted from articles.

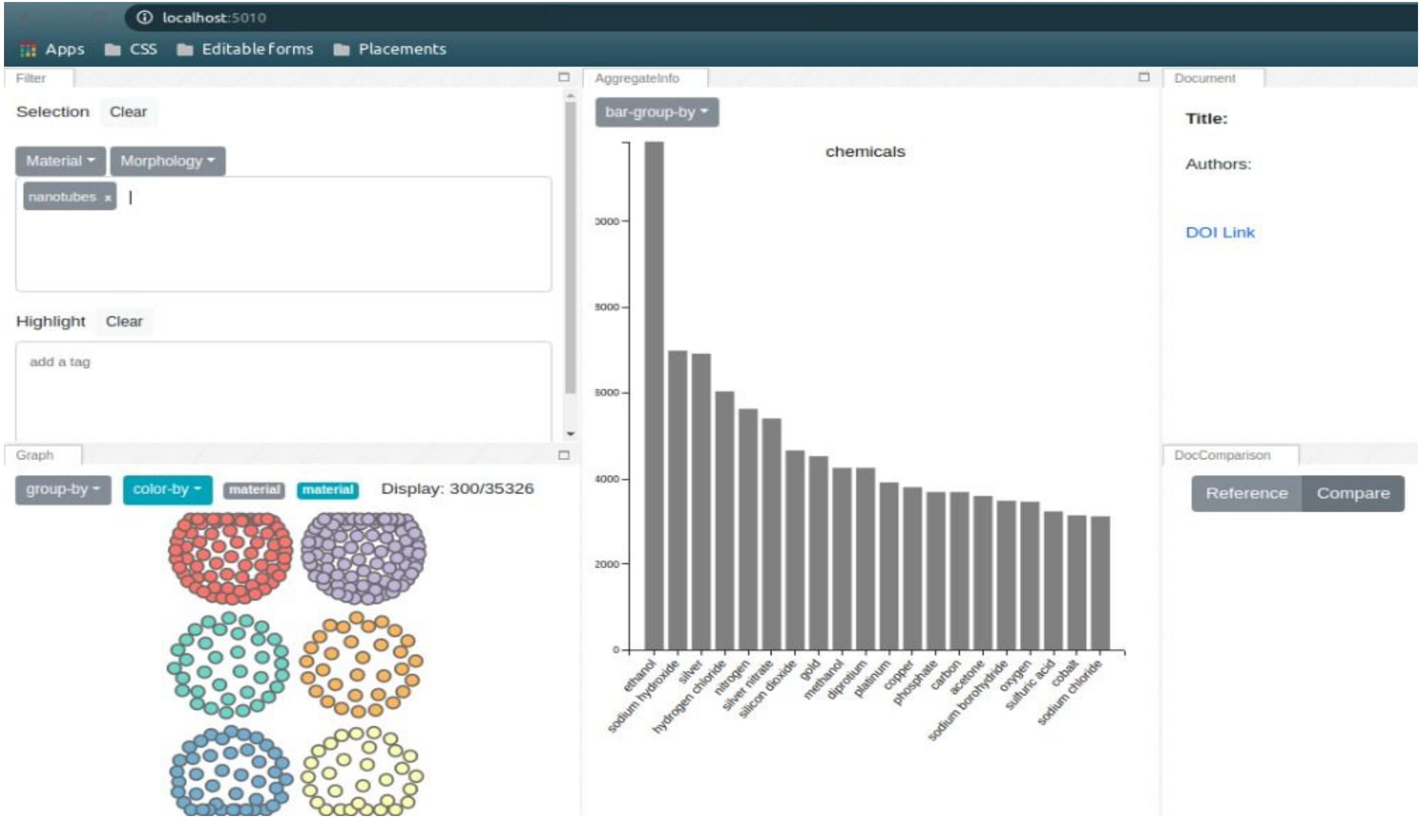
Diagramatic Flow



Flexible Framework for Visualizing Insights

- Interactive, intuitive, and visual content.
- Flexible visual analytics environment.
- Visual analytic approach is exploratory.
- Explore the unknown, uncover the questions/answers.
- Easy experimentation with the data.

Visual Analytics



browser-based visualization tool (available from https://github.com/LLNL/MI-ChemVis/)

First — A suite of tools that extract and structure information from the text and figures of scientific articles.

Second — Classifying the articles according to their target nanomaterial composition and morphology.

Third — Classification of articles can provide insights as to what combination of nanomaterial composition and morphology are over- or under-explored compared to average.

Fourth — **This paper can potentially help inform the development of new synthesis protocols for as-of-yet unrealized nanomaterials**.