The Synthetic Biology Knowledge System

Jeanet Mante[1], Yikai Hao[3], Jacob Jett[6], Udayan Joshi[3], Kevin Keating[4], Xiang Lu[3], Gaurav Nakum[3], Nicholas E. Rodriguez[5], Jiawei Tang[3], Logan Terry[2], Xuanyu Wu[3], Eric Yu[2], J. Stephen Downie[6], Bridget T. McInnes[5], Mai H. Nguyen[3], Brandon Sepulvado[7], Eric M. Young[4], Chris J. Myers[1*]

[1] *University of Colorado Boulder, Boulder CO 80309, USA,* [2] *University of Utah, Salt Lake City UT 84112, USA,* [3] *University of California San Diego, La Jolla CA 92093, USA,* [4] *Worcester Polytechnic Institute, Worcester MA 01609, USA,* [5] *Virginia Commonwealth University, Richmond VA 23284, USA,* [6] *University of Illinois at Urbana-Champaign, Urbana IL 61801, USA,* [7] *NORC at the University of Chicago Bethesda, Chicago IL 60637, USA*

E-mail: chris.myers@colorado.edu

Abstract

The Synthetic Biology Knowledge System (SBKS) is an instance of the SynBioHub repository that includes text and data information that has been mined from papers published in ACS Synthetic Biology. This paper describes the SBKS curation framework that is being developed to construct the knowledge stored in this repository. The text mining pipeline performs automatic annotation of the articles using natural language processing techniques to identify salient content such as key terms, relationships between terms, and main topics. The data mining pipeline performs automatic annotation of the sequences extracted from the supplemental documents with the genetic parts used in them. Together these two pipelines link genetic parts to papers describing

the context in which they are used. Ultimately, SBKS will reduce the time necessary for synthetic biologists to find the information necessary to complete their designs.

Keywords

text mining, topic modeling, data mining, sequence annotation, SynBioHub, SBOL

Introduction

Synthetic biology has transformative potential for industries such as agriculture, energy, materials, and health.[1–5] Founded on the idea of composable DNA parts, synthetic biology seeks to enable construction of complex genetic devices with predictable functions. This has largely proved possible.[6–8] Yet, even though parts-based design is the enabling approach, there remain few effective means to communicate part information amongst designers. Early in the development of synthetic biology, parts databases were recognized as key for information exchange. Yet, many of these have been ineffective in communicating data and metadata like function, intended host, and assembly method. Thus, databases are rarely used in genetic design, especially for organisms that are not *E. coli*. This leaves the field in a state where relevant part performance data is distributed among results and methods sections, paper supplemental files, and tables of sequences — shifting the work of a genetic designer from design to searching through disparate sources for part information. Many designers simply use a custom set of parts curated from past experience or screen new parts rather than use a database or mine the literature.

The *Synthetic Biology Knowledge System* (SBKS) seeks to combine the richness of information in literature with the searchability and standardization of databases. The end goal is to save the genetic designer search time and reduce the risk of missing valuable information about parts. To do this, the SBKS project combines an instance of the SynBioHub parts database with semantically annotated literature from the *ACS Synthetic Biology* corpus.

The resulting system creates an open and integrated resource that harnesses disparate, heterogeneous data sources to accelerate genetic design. This will ultimately make it easier to construct genetic devices, transfer parts to new hosts, and aid scientific discovery.

SBKS uses an implementation of SynBioHub[9] populated with annotated genetic components that are cross-referenced with the ACS Synthetic Biology corpus that has been semantically annotated. SynBioHub is an existing parts repository which, like other parts repositories, has fallen short of being able to communicate basic part information in a searchable manner.[10] Whilst there have been previous attempts at creating more comprehensive genetic part databases which integrate disparate information sources,[11] the SBKS project attempts this via the integration of semantically annotated literature. This strategy has previously been seen in other disciplines such as bio-geochemistry[12] and bio-medicine.[13] To this end, ACS Synthetic Biology supplemental files are scraped to provide sequences. These sequences are then auto-annotated using curated libraries of genetic parts. The annotated sequences are uploaded to the SBKS SynBioHub instance (`https://synbioks.org`), which also contains the metadata of the papers from the ACS Synthetic Biology corpus. This metadata is enhanced by semantic annotation of the papers which are curated by biological experts. This forms the basis of a semantic search engine for synthetic biology literature and genetic parts.

Results

The core of SBKS is a curation framework (see Figure 1) that integrates knowledge found in both text and data sources. The text mining pipeline applies various *natural language processing* (NLP) techniques to extract salient content from published articles. Each article is provided in the *Journal Article Tag Suite* (JATS) XML format, and it is parsed to extract text and relevant metadata. Each supplemental file is also parsed to extract genetic sequences of parts. The article text is then processed using *named entity recognition* (NER) to identify

terms,[14–16] *relation extraction* (RE) to identify relationships between terms,[17,18] and *topic modeling* (TM) to identify topics.[19,20] These discovered terms and concepts are then used to tag the article, providing a way to link the context of each article to data elements in SBKS.

The data mining pipeline begins with sequences for genetic designs harvested from ACS Synthetic Biology paper supplemental documents and sequences for toolkit part libraries hand curated into spreadsheets. Both sets of sequences, once extracted, are translated into the *Synthetic Biology Open Language* (SBOL),[21] a *resource description framework* (RDF) data standard for genetic design. The genetic design sequences are then automatically annotated using the part libraries and the SYNBICT software tool (`https://github.com/SD2E/SYNBICT`). The result is again encoded into an SBOL Collection that is linked to the ACS Synthetic Biology paper. This SBOL representation is then uploaded to the SBKS instance of the SynBioHub[22] data repository (`https://synbioks.org`). Once deposited, it can be searched and accessed using either a graphical user interface (GUI) or programmatically by its *application programmers interface* (API) (`https://wiki.synbiohub.org/api-docs`).

Text Mining Pipeline: The initial text data set consists of all the articles that have been published in ACS Synthetic Biology as of November 2019. These articles are provided in richly annotated JATS XML format, which includes a rich set of metadata, the full article text, and structured references to cited papers. The project is currently in the process of developing a similar workflow for ingesting articles from Oxford Academic's open access Synthetic Biology journal. The metadata and citation elements of the structured article file are harvested and converted into SBOL-compliant RDF/XML with Dublin Core annotations suitable for ingestion into SynBioHub. Among the steps taken during this process is the employment of python scripts to match article DOIs to corresponding PubMed IDs.

Each XML file is parsed to extract the full article text that is used by all subsequent components in the SBKS Text Mining Pipeline, as described below. XML parsing includes extracting and splitting the article text into sentences grouped by paragraphs. Each paragraph is annotated with its section header and span information. XML parsing also removes

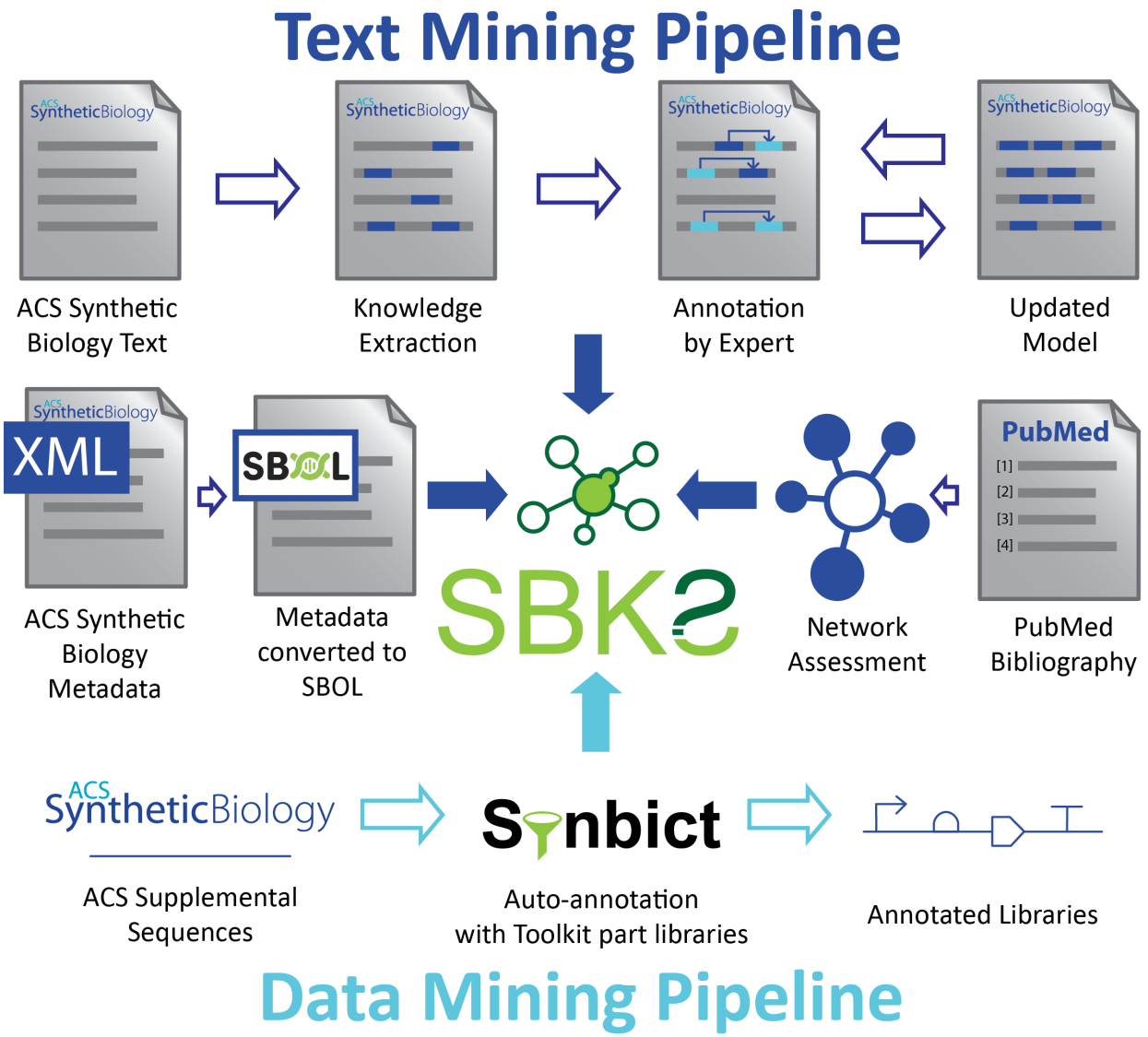

Figure 1: SBKS curation framework. This provides an overview of the information flow from the ACS Synthetic Biology corpus into the SBKS database. **Text Mining Pipeline:** documents are ingested into the pipeline, processed by the NLP components and integrated into the SBKS interface. New expert knowledge is used to refine the NLP models. Alongside NLP analyses, different types of networks (e.g., co-authorship and co-citation) are extracted from PubMed texts and explored to connect synthetic biology topics with relevant ethical issues. In addition to these processes, metadata, topical terms, and references are harvested from the ingested articles, enriched with semantic links, and converted into SBOL-compliant metadata. **Data Mining Pipeline:** Sequences are extracted from ACS Synthetic Biology supplemental files and then annotated using expert curated part libraries from 'toolkit papers'. Annotation is done via the SYNBICT tool. The annotated sequences are then integrated into the SBKS interface where they are matched with the ACS Synthetic Biology metadata of the appropriate paper, which is converted to SBOL as part of the text mining pipeline.

extraneous text such as references and formatting data. The parsed text is then stored in a JSON file for easier downstream processing. We ran the XML parsing on all 1597 ACS Synthetic Biology articles and categorized them into research and non-research articles. There are 907 research articles and 690 of other types such as editorials, letters, reviews, viewpoints, technical notes, etc. Since research articles generally contain more technical details, and therefore would likely have a higher density of entities and relations, the text mining pipeline processes research articles only.

Additionally, supplemental files are parsed to identify and extract genetic sequences of parts. For all 1597 articles, there is at least one supplemental file available for 82 percent of these articles, and there are 10,070 supplemental files in total. These supplemental files are provided in various formats ranging from structured data, such as GenBank files, to general PDF documents. The top 10 most frequent file types are shown in Table 1. The sequence extraction process found at least one sequence from 8 percent of the supplemental files, and 89,620 genetics sequences were found in total. Of these 89,620 entries, 1,732 sequences are skipped due to blank entries or a lack of publications to match them to. As a result, only 87,888 sequences are processed by the SBKS Data Mining Pipeline to link these parts found in supplemental files to SBKS's libraries of parts.

Table 1: Top 10 most common supplemental file types

Rank	File Type	Count	Rank	File Type	Count
1	PDF	1434	6	SBML	475
2	TSD	1091	7	PNG	403
3	XML	1079	8	JPG	337
4	GenBank	707	9	TXT	330
5	HTML	499	10	JS	325

The article text is processed using techniques for NER, which is a sub-task of text mining.[14–16] The goal of NER is to locate and classify named entities present in text into pre-defined categories. For example, *Acinetobacter baylyi* is a *Species*. This is done using deep neural network models to perform NER on these articles, specifically *BioBERT* (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining).[23]

Transformers are designed to learn sequence data. Instead of relying on recurrent connections, however, they use an attention mechanism to weigh the relevance of each input in producing the output.[24] BioBERT is pre-trained on top of BERT,[25] a general-purpose language representation model. This pre-training was conducted over PubMed abstracts and PubMed Central full-text articles to adapt to biomedical text mining tasks. In this work, the BioBERT output is fed into a simple feed-forward neural network for the final NER prediction.

For the initial round of NER, standard biological entity categories (e.g., genes and chemicals) are used, since there is no labeled dataset for synthetic biology entities. Results from this initial round are reviewed and corrected by domain experts to create a more refined dataset with entities more specific to synthetic biology that can be used to fine tune the NER models. Named entities expected to be detected within synthetic biology articles are also added to the articles as suggested annotations, to be confirmed by expert annotators in order to facilitate the creation of gold-standard synthetic biology-specific training data. Table 2 shows the entity types, total number of annotations and average annotation per document for each entity type over the ACS Synthetic Biology dataset. The table also shows the total number of entities and the unique number of mentions annotated by the NER component. Table 3 shows the top ten terms identified by the NER component from the ACS Synthetic Biology dataset for each of the entity types. Figure 2 shows an overall view of the different entities identified by the NER component. These results show that the while NER is not perfect, it is able to identify many of the entity types correctly. The next step in the process is to refine the model based on human corrected annotations. For example, the NER system identified *LB* and *M9* as *Cell Line* whereas they are formulations of media, so therefore should probably be annotated as *Chemical*.

The text mining pipeline also applies RE to the article text.[17,18] RE is the task of classifying the relation between entity mention pairs in a sentence. For RE, the process also uses BioBERT as the base biomedical language representation model and fine tunes it for the RE

Table 2: Entity statistics from named entity recognition component. Entity Type is the entity category. Total Annotations are the number of entities extracted from the dataset. Average per Document is the average number of entities per document over all the documents in the ACS Synthetic Biology dataset. Total Number of Mentions is the total number of entities, and Total Number of Unique Mentions are the number of unique mentions identified in the ACS Synthetic Biology dataset.

Entity Type	Total Annotations	Average per Document
Cell Line	12,084	13.47
Chemical	86,180	96.07
Gene	169,061	188.47
Species	31,785	35.43
Total Number of Mentions	299,110	
Number of Unique Mentions	43,439	

Table 3: Top ten terms extracted from the ACS Synthetic Biology dataset by the NER component for each entity type (Cell Line, Species, Gene and Chemical) and across all of the entity types (All).

All		Cell Line		Species		Gene		Chemical	
#Doc.	Term	#Doc.	Term	#Doc.	Term	#Doc.	Term	#Doc.	Term
7682	E. coli	305	LB	7682	E. coli	4549	GFP	1725	glucose
4549	GFP	292	cells	2965	yeast	1970	Cas9	1391	peptide
2965	yeast	185	HEK293T cells	1202	S. cerevisiae	1298	mCherry	1206	amino acid
1970	Cas9	181	BL21	822	human	1200	dCas9	1180	amino acids
1725	glucose	180	MG1655	564	B. subtilis	1013	CRISPR	835	glycerol
1391	peptide	172	M9	430	P. putida	938	LacI	829	kanamycin
1298	mCherry	125	cultures	367	Escherichia coli	839	sfGFP	789	NaCl
1206	amino acid	118	KT2440	332	C. glutamicum	786	YFP	781	peptides
1202	S. cerevisiae	100	HeLa cells	321	S. oneidensis	748	TetR	737	tetracycline
1200	dCas9	96	strains	299	Synechocystis	673	RFP	657	ampicillin

task. The ChemProt corpus[26] containing chemical-protein interactions, processed to be in the format for input to BioBERT,[27] was used for RE fine tuning. The model was fine tuned on a subset of the chemical-protein relations available in ChemProt which are important for synthetic biology, namely up-regulator (CPR3), down-regulator (CPR4), and substrate (CPR9). The trained RE model is then used for inference on the ACS Synthetic Biology data. For relation extraction on the ACS Synthetic Biology articles, the results of NER are used by the RE model to infer associations between named entities within a sentence as follows: the article text is split into sentences; entity mentions from NER within each sentence are identified; the RE model then processes each sentence with identified entity mentions and classifies each sentence as containing an up-regulator relation, down-regulator

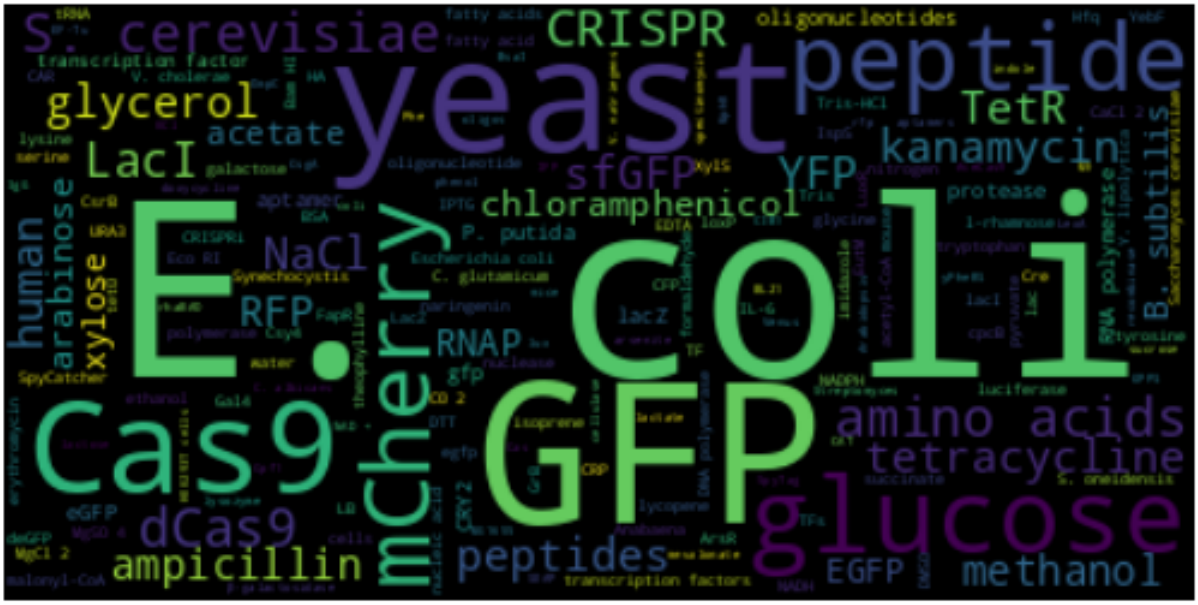

Figure 2: Word cloud of entities. This figure contains a word cloud based on the occurrence of entity mentions within the ACS Synthetic Biology dataset.

relation, substrate relation, or no relation. Table 4 shows the relation types, total number of relations, the average number of relations per document, and the average number of relations of those documents containing at least one relation extracted from the ACS Synthetic Biology dataset.

Table 4: Relation statistics from RE component. Relation Type is the relationship extracted between the Genes and Chemicals. Total Annotations are the number of relations extracted from the dataset. Average per Document is the average number of relations per document over all the documents in the ACS Synthetic Biology dataset. Average per Annotated Document is the average number of relations per document of those documents that contains a Gene-Chemical relationship.

Relation Type	Total Annotations	Average per Doc.	Average per Annotated Doc.
Down-regulator	537	0.59	0.95
Substrate	6,904	7.69	12.22
Up-regulator	840	0.94	1.49

The article text is additionally processed using TM.[19,20] TM uses a probabilistic approach to uncover topics in a collection of documents, where topics are represented by clusters of similar words; we have separate TM pipelines for synthetic biology and ethics-related texts. For the synthetic biology texts, we use Latent Dirichlet Allocation (LDA)[28] as the TM

technique. LDA uses a generative probabilistic approach to model each topic as a mixture of a set of words and each document as a mixture of a set of topics to determine the different topics that a corpus represents and how much of each topic is present in each document. For TM, the article text is pre-processed to convert text to lower case, remove accents, and lemmatize each token. To remove general terms such as 'cell' that do not help to separate topics or documents, we remove the top 0.5 percent of terms found in the ACS Synthetic Biology article dataset. To tailor LDA for synthetic biology text, our process treats named entities found by our NER models as special terms by keeping them intact instead of pre-processing them as other words. The process sets the number of topics to five, which the process found experimentally to be stable and easy to interpret. Since LDA is an unsupervised technique with no available ground truth, a domain expert on the team provided topic interpretations for the abstract topics uncovered by LDA.

For topic modeling, topic interpretation for the abstract topics uncovered by our LDA model for topic modeling was provided by a domain expert on our team, as described in the Methods section. Three of these abstract topics (Genetic Circuit Modeling, Genetic Circuit Design and Metabolic Engineering) are readily recognized as research areas related to synthetic biology. The topics interpreted as "Strain Construction" and "Biosensor / Protein-related Terms" contain terms mostly related to experimental methods. Though authors may not be likely to use these topics as keywords to describe their work, it may be useful to identify papers that contain rich discussions of these techniques. The results of TM are the dominant topic found for each article, which are then used for article tagging. Table 5 lists the topics uncovered by topic modeling. Results from TM can also be used to determine the dominant terms for each topic, composition of topics for each document, and related articles for each document for search and exploration of SBKS.

After the NER models, topic models, and RE models have been applied to the article text to produce lists of named entities, topics, and relationships linking entities (respectively), these lists are reintegrated into the RDF/XML metadata pipeline. In the metadata

Table 5: Topics from topic modeling. These are topics discovered in the ACS Synthetic Biology articles from topic modeling. Each topic is modeled as a mixture of a set of words, and each document is modeled as a mixture of a set of topics. Topic names are provided by a domain expert. The number and percentage of documents shown for each topic are the count and frequency of articles with that topic as the dominant topic.

Topic	Document Count	Document Percentage
Genetic Circuit Modeling	201	22.16
Genetic Circuit Design	168	18.52
Strain Construction	185	20.40
Biosensor / Protein-Related Terms	170	18.74
Metabolic Engineering	183	20.18

pipeline, various quality control measures are applied to de-duplicate, normalize, and disambiguate the various named entities, tokens, and relationships. Once this data cleaning step is completed, the named entities and topics are then inserted into topical metadata associated with the article from which they were identified, while the relationships that have been extracted are used to mint triples that link entities and concepts to one another. In the near future, this process will also integrate a linked data step by grounding the named entities and discovered relationships to concepts in top level ontologies (e.g., NCBI's species taxonomy [https://www.ncbi.nlm.nih.gov/taxonomy], the CHEBI Ontology [https://www.ebi.ac.uk/chebi/], etc.) by adding canonical identifiers for those concepts sourced from the ontologies to the topical metadata. Additionally, the ethics-related texts will be fully integrated into the pipeline for synthetic biology texts.

Alongside the text mining pipeline for synthetic biology generally, SBKS includes analyses of texts related to synthetic biology ethics (synbio-ethics for short) to encourage SBKS users to be informed about ethical debates and associated literature relevant to their queries. The general pipeline for synbio-ethics texts is similar to that for *ACS Synthetic Biology* texts. Articles are ingested in the form of XML files and parsed into JSON files that annotate article paragraphs, sections, and references. Since *ACS Synthetic Biology* is not primarily an ethics publication venue, the process draws from broader literature databases. The synbio-ethics text mining pipeline ingests data from PubMed Central via its API, but the process

has also used the Web-of-Science for exploratory analyses. Before establishing the synbio-ethics text mining pipeline, initial exploratory analyses were conducted to understand the relevant ethical issues being discussed and the structure of the communities engaging in these discussions. The analysis located 15,152 publications in the Web-of-Science pertaining to synthetic biology and then derived from this set of publications a smaller corpus of 562 ethical texts. Although synthetic biology literature began to increase exponentially around 2000, not much attention was devoted to ethics until roughly 2010.

The topic models to identify relevant ethical issues follow a similar pipeline to the synthetic biology texts, though they rely upon a *correlated topic model* (CTM).[29] The CTM identified nine topics within the synbio-ethics corpus. Certain topics tend to be very concrete (e.g., RNA), while others are abstract (e.g., philosophy). Still other texts discuss the social aspects of synthetic biology and governing research practices. Table 6 provides the names of each topic and the proportion of the corpus that engages each topic.

Table 6: Topics from synbio-ethics topic modeling. These are topics discovered in the Web-of-Science synbio-ethics articles. As with the LDA models, each topic is modeled as a mixture of a set of words, and each document is modeled as a mixture of a set of topics. Topic names have been created in consultation with domain experts. Topic proportion refers to the proportion of the corpus that engages each topic.

Topic	Topic Proportion (%)
Philosophy	15.2
Treatment	14.9
RNA	12.7
Science Governance	11.3
Yeast	10.7
Risk / Safety	10.7
Biofuels	10.0
Vaccines / Antibiotics	8.60
Food / Animals	5.93

Another component of the synbio-ethics text mining pipeline employs network analysis to further investigate the conceptual structure of synbio-ethics and to map its social structure. Analyses of co-authorship networks reveal that many disconnected groups of researchers collaborate on synbio-ethics, but there is a large disparity in the extent of certain institutions

in the synbio-ethics literature. Knowing the social structure underlying synbio-ethics can help inform literature recommendations SBKS provides, such as viewpoints associated with certain social groups are not unintentionally omitted. Further, we are currently investigating co-citation networks to understand which synthetic biology texts are frequently cited together. Not only does this approach help us understand the structure of synbio-ethics literature,[30] but we plan to exploit these connections to help link technical synthetic biology papers with those discussing ethical concerns, thus improving our ability to make pertinent synbio-ethics recommendations based upon SBKS user queries.

Data Mining Pipeline: The data mining pipeline extracts sequences from ACS Synthetic Biology supplemental files, annotates them using expert curated part libraries, and links the sequences to the ACS Synthetic Biology metadata of the appropriate paper. This workflow is shown in Figure 3. The initial data set for the data mining pipeline is the synthetic biology parts and designs found in the ACS Synthetic Biology supplemental documents. The sequences used in the pipeline were extracted as part of the text mining pipeline as described above.

The primary source for part collections were parts-rich papers, including "toolkit" papers which have become especially popular for non-conventional chassis. Resources mined for yeast parts included the Yeast Toolkit,[31] Pichia Toolkit,[32] and a combinatorial design paper.[33] Parts for Gram-negative bacteria were drawn from the CIDAR MoClo kit,[34] the CIDAR Extension Kit Volume I,[35] and the Voigt Lab terminator collection.[36] Additionally, the preexisting Cello Parts Library[7] and *Bacillus subtilis* Library[37,38] were also used. These toolkits frequently provide part sequences, characterization data, and information about the progeny of the parts. A description of the part, the chassis it was used in, the role of the part and the sequence were extracted from each paper and entered into a spreadsheet, which was then converted to SBOL and uploaded to the SBKS SynBioHub instance. Sourcing part sequence information from primary papers can also facilitate the accessing of physical parts if they are submitted to a physical repository or otherwise available from the authors. This

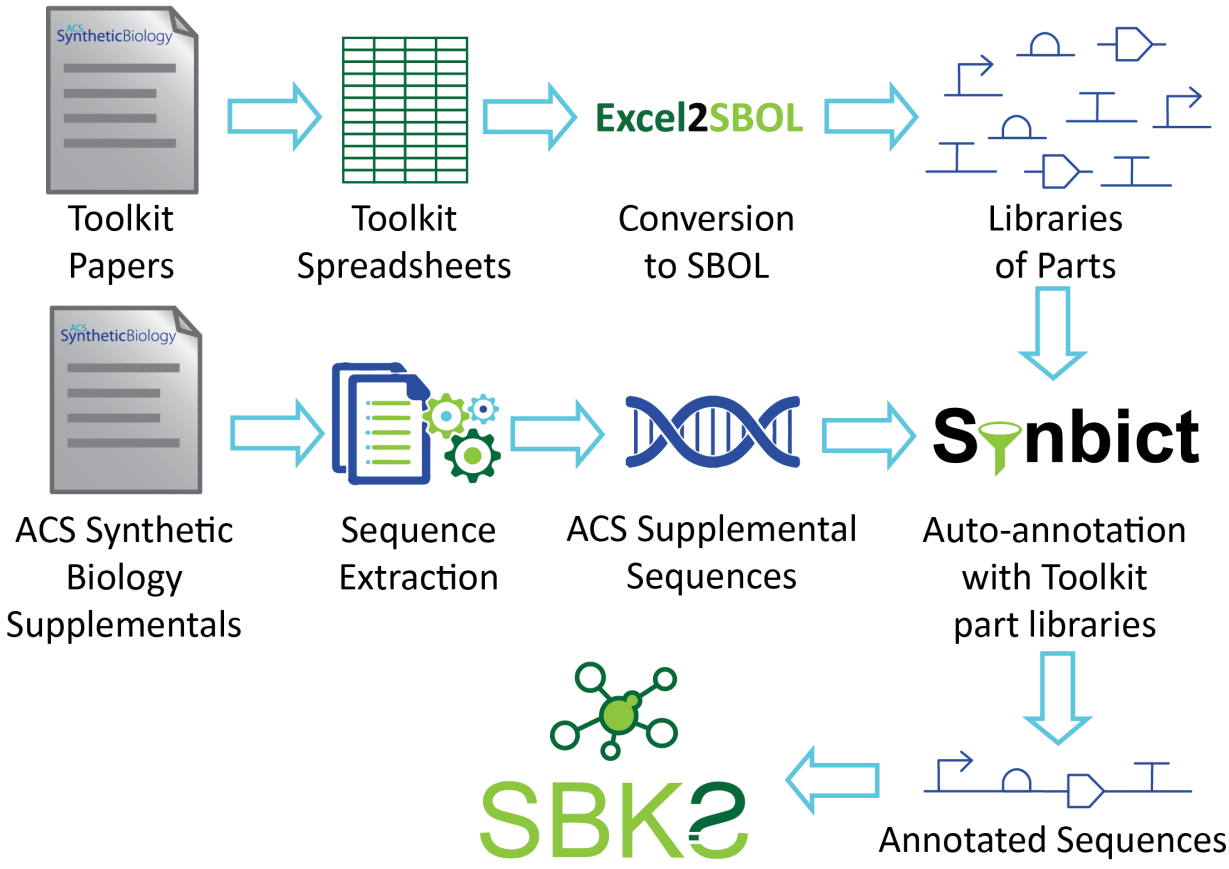

Figure 3: The data mining pipeline annotates sequences extracted from ACS Synthetic Biology supplemental files. Annotation happens using libraries curated by experts using spreadsheets. The spreadsheets are then converted to SBOL by Excel2SBOL. The expert curated libraries are then used by SYNBICT to annotate the sequences. Annotated sequences are uploaded to the SBKS instance of SynBioHub and linked to the ACS Synthetic Biology metadata of the appropriate paper.

provides a rich, well-curated collection of parts in a format which is easily accessible by web users and directly usable in downstream design automation pipelines.

The ACS Synthetic Biology corpus extracted sequences are converted to SBOL so they can be annotated using SYNBICT (`https://github.com/SD2E/SYNBICT`). SYNBICT is a SBOL sequence annotation tool that annotates sub-sequences found in a longer sequence using a library of parts. The libraries used by SYNBICT were those generated from "toolkit papers" as described above.

The annotation of genetic parts was hindered by the accuracy of sequence scraping. Table 7 indicates that some file types do better for scraping than others. Notably PDFs are a very popular supplemental type (93 percent) for presenting sequences, but they fair poorly when automating sequence extraction for annotation. This is partially due to the fragmentation of sequences (any spaces or numbers between rows of a sequence will lead to a sequence being split into several sequences by the scraping method as indicated by the shorter average sequence length). GenBank files put into PDFs were found particularly difficult to scrape. However, when GenBank files are provided in their native file format, they are more effectively annotated. The use of SBOL and FASTA standard formats also significantly increases the machine readability of sequences. SBOL shows annotation within shorter sequences, which may be indicative of the better coverage of the part libraries in these formats or the shorter sequences due to sequences being of parts not whole constructs. The top annotations for the sequences are shown in Table 8.

User Interface: One final aspect of the SBKS project is the development of a user interface that can access the information stored in SBKS to assist a designer of a genetic circuit. One of these tools is SBOLCanvas,[39] a web application for creation and editing of genetic constructs using the SBOL data and visual standard. SBOLCanvas allows a user to create a genetic design from start to finish, with the option to incorporate existing SBOL data from a SynBioHub repository, such as SBKS. While SBOLCanvas is currently able to efficiently create genetic designs for parts selected via searches on SynBioHub, the end goal

Table 7: Annotation of sequences from ACS Synthetic Biology Supplemental Files. There are different supplemental file types used with the majority (93 percent) being PDFs. Unlike FASTA, GenBank, and SBOL file types, PDFs need to be scraped to extract sequences. The lack of standard page breaks and formatting of the sequence leads to errors in sequence extraction meaning a single sequence is often read in as several shorter fragments. This is noticeable as the average sequence length for PDFs is significantly lower than the other file types. The other feature of note is that the percentage annotation is much higher for FASTA, GenBank, and SBOL, which are standard sequence formats. Of these, SBOL shows annotation within shorter sequences which may be indicative of the better coverage of the part libraries in these formats or the shorter sequences due to sequences being of parts not whole constructs.

File Type	No. of Sequences	% of Sequences Annotated	Avg. Sequence Length	Avg. Annotated Sequence Length
FASTA/TXT	679	0.88	397	5014
GenBank	686	79.88	4419	4286
XML/SBOL	5098	26.19	399	988
PDF	81426	0.38	56	1829
Total	87888	2.5	113	1825

Table 8: Top annotations of ACS Synthetic Biology Supplemental sequences. These are the top annotations of different SO:Role types (e.g. Promoter) that were used to annotate the ACS Synthetic Biology supplemental sequences.

All	Promoters	CDS	Terminators	Other
RiboJ (397)	AmpR_promoter (302)	AmtR (173)	L3S2P21 (386)	RiboJ (397)
L3S2P21 (386)	pCONST (288)	PhlF (172)	AmpR_terminator (226)	ColE1 (153)
AmpR_promoter (302)	pTet (157)	SrpR (146)	BBa_B0015 (214)	RiboJ53 (86)
pCONST (288)	pTac (123)	HlyIIR (146)	CamR_Terminator (135)	BydvJ (85)
AmpR_terminator (226)	pBAD (122)	BetI (117)	ECK120033737 (92)	RiboJ51 (73)
BBa_B0015 (214)	CamR_Promoter (116)	CamR (111)	L3S2P55 (86)	RiboJ10 (73)
AmtR (173)	pPhlF (92)	AraC (107)	ECK120033736 (81)	RiboJ57 (57)
PhlF (172)	pSrpR (73)	TetR (98)	ECK120019600 (79)	RiboJ54 (19)

will be to have a design tool that provides a synthetic biology designer a seamless connection to knowledge about the parts that are or could be used in their designs.

Another aspect of the SBKS project is the development of improved search tools. For example, the ability to perform a sequence-based search has been added to SynBioHub via SBOLExplorer and VSEARCH, a distributed search tool and open-source sequence search tool, respectively. Users can upload a compatible file type or a plain text sequence and search for similar sequences within SynBioHub. An API is also provided for users or tools to be able to use this feature programatically.

Discussion

The goal of SBKS is to create a tool that interfaces with the designer to quickly and seamlessly provide parts and part information. This will reduce time spent combing the literature and supplemental information for vital part performance and sequence data. Some example questions that a system such as SBKS could answer include:

- What are some papers about metabolic engineering in *E. coli*? What are some papers about metabolic engineering in *Pseudomonas putida*?
- What parts were used to construct the strains in this paper about a fluoride biosensor?
- Which papers use the pLac promoter sequence?
- What are some inducible promoters for *E. coli*? What are some strong constitutive promoters for *S. cerevisiae*?

Some of these questions are already answerable using the current work and SPARQL queries, examples of which can be found at:
`https://github.com/synbioks/SPARQL-queries-ACS-Synbio`. To achieve the goal of being able to answer all the questions above and more complex questions, there is still much work that needs to be done. In the future, we plan to integrate additional text and data sources into the curation framework. The performance of each of the text mining components is also being improved. The grounding of string-based terms to ontologies is underway. Ultimately, the text mining components will be integrated into an automated turn-key system. The next step for data mining is the development of further libraries for the sequence annotation, and further refinement of the sequence extraction methods, particularly for PDF files. Additionally, a more user friendly interface for searches will be developed.

We would also like to create component families which are functionally the same in different organisms: with a different codon usage, cut out of plasmids differently, or are for use with different assembly methods. We plan to refine and expand the list of columns used

in the Excel2SBOL spreadsheet templates so that parts uploaded via that workflow can be used with many different synthetic biology workflows. Finally, to further test our annotation workflow, we are also running the iGEM registry data set through it.

In the long-term, we need to develop user interfaces to make use of the searching, visualization, and curation of the results of the text and data mining pipelines more accessible to the general synthetic biology community. In particular, we plan to develop knowledge-enabled search for genetic circuit design. This would include structure-based queries, natural language based search, and new search visualizations for linked knowledge. For curation, we plan to partner with journals and funding agencies to develop workflows to support publication of methods, data, keywords, organisms, and provenance information in formats suitable for machine analysis. Till this goal has been achieved, we have submitted our supplemental sequence information as a text file with a comma separated list of URIs pointing to the collections containing our sequences, in an attempt to make further curation easier. Finally, we would like to develop machine learning enabled tools for knowledge enrichment to enable literature-based discovery, knowledge-graph completion, determination of confidence and bias, construction of synthetic biology networks and communities, and a tight integration of ethics, safety, and security.

Conclusion

SBKS is an instance of the SynBioHub repository that includes text and data information that has been mined from papers published in ACS Synthetic Biology. The text mining pipeline performs automatic annotation of the articles using natural language processing techniques to identify salient content such as key terms, relationships between terms, and main topics. The data mining pipeline performs automatic annotation of the sequences extracted from the supplemental documents with the genetic parts used in them. Together these two pipelines link genetic parts to the context of their usage as described in research

papers. Alongside the SBKS pipelines, an analysis of the analyzed ACS Synthetic Biology corpus is presented. The analysis includes: 1) the number of articles with sequence information, 2) the most mentioned species, cell lines, genes, and chemicals, 3) the types of relationships (up/down/substrate) mentioned, 4) the topics of papers, 5) the numbers of sequences found in supplemental files that could be successfully annotated, and 6) the most common sequence annotations. Ultimately, SBKS will reduce the time necessary for synthetic biologists to find the information necessary to complete their designs.

The SBKS project began less than two years ago, and it is being executed by a team that met only a few months before that. While the scope is ambitious, the progress so far is very promising. We look forward to feedback from the community about the needs and potential applications for SBKS.

Methods

Text Mining Pipeline: Each ACS Synthetic Biology article in XML format is parsed to extract the article text. The parsing is performed by a Python script that uses the lxml package.[40] The parsed text is then split into sentences using spacy[41] and the model en_core_sci_sm from scispacy.[42] The code for XML parsing can be found at `https://github.com/synbioks/ACS-XML-to-text/blob/jiawei-dev/acs-xml-parser.ipynb`.

For NER, we use the implementation of BERT from Huggingface.[43] We use case-sensitive BioBERT-Base v1.0 (+ PubMed 200K + PMC 270K)[23,44] as the pre-trained model and fine tune it using the HUNER dataset.[45] Our NER code can be found at `http://web.synbioks.org/author/synbiobert/`.

For RE, we also use the implementation of BERT from Huggingface.[43] We use case-sensitive BioBERT-Base v1.0 (+ PubMed 200K + PMC 270K)[23,44] as the pre-trained model and fine tune it using the ChemProt dataset.[27] Our RE code can be found at `https://github.com/synbioks/Text-Mining-NLP/tree/master/relation-extraction`.

For TM, the BERT BasicTokenizer[46] from Huggingface is used to tokenize the article text by splitting text on white spaces, punctuations, and control characters. We also added preprocessing steps to convert the text to lower case, remove accents, and lemmatize each token using modules from the Natural Language Toolkit (NLTK).[47] The specific implementation of LDA that we use is the LDA Mallet model[48] available in the Gensim package.[49] Our topic modeling code can be found at `https://github.com/synbioks/Text-Mining-NLP/tree/master/topic-modeling`.

For TM with the synthetic biology ethics corpus, our corpus was collected by querying the Web-of-Science for articles related to synthetic biology and then subsetting that corpous to include only those articles pertaining to the social and ethical implications of synthetic biology research. To obtain the synthetic biology texts, we used the query established by Shapira et al.,[50] and then obtained the ethics-related texts by searching article titles, abstracts, and keywords for forms of safety, ethics, security, and dilemma. Preprocessing included conversion to lower case, stemming terms, and removing terms that occurred in fewer than 10 ethics-related documents. We employed the correlated topic model (CTM)[29] rather than LDA for this corpus. CTMs are similar to LDA except CTMs explicitly model correlation between topics. Doing so aligns with the intuition that documents that are more likely to discuss one topic might also be more likely to discuss another topic. Extensive review found nine topics to be the best number of topics, and review by subject matter experts verified that the content of the topics made sense. The CTM code can be found at `https://github.com/synbioks/synbioethics_ctm/tree/main`.

Data Mining Pipeline: ACS Synthetic Biology supplemental files are parsed to identify and extract genetic sequences of parts. The parsing is carried out by a python script which initially attempts to validate the file as SBOL or convert it from GenBank/FASTA to SBOL using the SBOL-Validator.[51] The code for this is found here `https://github.com/synbioks/ACS-XML-to-text/blob/jiawei-dev/acs-xml-parser.ipynb`. Genetic sequences from any file that is successfully converted/validated by the SBOL-Validator are

added to a FASTA formatted text file. One text file is created per supplemental file. For PDF files that are incompatible with the SBOL-Validator, a tool, pdftotext,[52] is used to convert them into plain text, and genetic sequences are scraped from the plain text using the regular expression package that comes with Python 3. The regular expression identifies slices of the plain text that only contain the letters A, T, C, G, U as candidates and extracts the ones that are longer than 6 characters to reduce false positives. Sequences extracted from PDF files are appended into the aforementioned FASTA formatted text file.

Libraries of curated parts were created from Toolkit papers using the Excel2SBOL python library (`https://github.com/SynBioDex/Excel-to-SBOL`). This requires a row to be created for each component and a standard set of columns with information about the component to be completed. Additional columns may be added to capture further information (this information is maintained in the conversion to SBOL). The standard columns used were decided based on an analysis of the iGEM data set and common information encoded as free form text in their registry-part descriptions. The columns used were: Part Name, Role (Sequence ontology term),[53] Design Notes, Altered Sequence (was the sequence pulled directly from the source or was it altered, e.g. codon optimised), Part Description, Data Source (e.g. PubMed: 28252957), Source Organism, Target Organism, Circular, and Sequence. Note that not all of these columns are mandatory. An expansion of columns to be more universal to different synthetic biology workflows would help make a components uploaded reusable in different workflows and thus contribute to a best practices standard that advances the aims of synthetic biology to create libraries of modular and reusable components.

The toolkit papers used to construct the SYNBICT library were chosen because they report widely-used synthetic biology parts, and were not sourced exclusively from the ACS Synthetic Biology corpus. These sequences for these parts were annotated in SBOL and uploaded to the SBKS instance of SynBioHub (`https://synbioks.org/public/a932bf113a/a932bf113a_collection/1`). They were then linked to collections based off of the text mining pipeline metadata SBOL conversion. These collections serve as manually-curated

reference libraries for SYNBICT, and are publicly available for new downstream applications. This manual library of mined parts is distinct from the automated parts mining pipeline described in this paper.

SYNBICT was used to annotate the ACS Synthetic Biology supplemental sequences. The FASTA files constructed from the supplemental files were read in and annotated using the Toolkit Papers. The minimum annotation length was set to 40bp. The python code used to carry out the annotation can by found at `https://github.com/synbioks/sequence_supplementals`.

Acknowledgement

This work was funded by the National Science Foundation under Grants No. 1939892, 1939929, 1939885, 1939887, 1939951, and 1939860. The SynBioHub instance hosting this data (`https://synbioks.org`) is hosted on an Azure server provided by Microsoft Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency.

Author Contributions

YH, UJ, XL, GN, NR, JT, XW, KK, BS, EMY, MN, BM, worked on the text mining pipeline. JJ and JD worked on the ACS Synthetic Biology metadata conversion. BS, JJ and JD worked on the network assessment. JM, JT, and CM worked on the data mining pipeline. EMY and KK provided the annotation libraries. EY, LT, and CM worked on the SBKS user interface. All authors contributed to the writing of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Supporting Information Available

The Synthetic Biology Knowledge System can be accessed at `https://synbioks.org`.

URLs to access sequence libraries found in the supplemental file sequence_library_urls.csv

The annotated part libraries can be accessed at:

`https://synbioks.org/public/Pichia_MoClo_Toolkit_Lu_Lab/Pichia_MoClo_Toolkit_Lu_Lab_collection/1`

`https://synbioks.org/public/CIDAR_MoClo_Extension_Kit_Volume_I_Murray_Lab/CIDAR_MoClo_Extension_Kit_Volume_I_Murray_Lab_collection/1`

`https://synbioks.org/public/EcoFlex_MoClo_Toolkit_Freemont_Lab/EcoFlex_MoClo_Toolkit_Freemont_Lab_collection/1`

`https://synbioks.org/public/MoClo_Yeast_Toolkit_Dueber_Lab/MoClo_Yeast_Toolkit_Dueber_Lab_collection/1`

`https://synbioks.org/public/Anderson_Promoters_Anderson_Lab/Anderson_Promoters_Anderson_Lab_collection/1`

`https://synbioks.org/public/CIDAR_MoClo_Toolkit_Densmore_Lab/CIDAR_MoClo_Toolkit_Densmore_Lab_collection/1`

`https://synbioks.org/public/Itaconic_Acid_Pathway_Voigt_Lab/Itaconic_Acid_Pathway_Voigt_Lab_collection/1`

`https://synbioks.org/public/Natural_and_Synthetic_Terminators_Voigt_Lab/Natural_and_Synthetic_Terminators_Voigt_Lab_collection/1`

References

(1) Roell, M.-S.; Zurbriggen, M. D. The impact of synthetic biology for future agriculture and nutrition. *Current Opinion in Biotechnology* **2020**, *61*, 102–109.

(2) Kelwick, R. J.; Webb, A. J.; Freemont, P. S. Biological materials: the next frontier for cell-free synthetic biology. *Frontiers in Bioengineering and Biotechnology* **2020**, *8*.

(3) Tang, T.-C.; An, B.; Huang, Y.; Vasikaran, S.; Wang, Y.; Jiang, X.; Lu, T. K.; Zhong, C. Materials design by synthetic biology. *6*, 332–350.

(4) Liu, Z.; Wang, K.; Chen, Y.; Tan, T.; Nielsen, J. Third-generation biorefineries as the means to produce fuels and chemicals from CO2. *3*, 274–288.

(5) Kightlinger, W.; Warfel, K. F.; DeLisa, M. P.; Jewett, M. C. Synthetic glycobiology: parts, systems, and applications. *ACS synthetic biology* **2020**, *9*, 1534–1562.

(6) Khalil, A. S.; Collins, J. J. Synthetic biology: applications come of age. *11*, 367–379.

(7) Nielsen, A. A. K.; Der, B. S.; Shin, J.; Vaidyanathan, P.; Paralanov, V.; Strychalski, E. A.; Ross, D.; Densmore, D.; Voigt, C. A. Genetic circuit design automation. *352*, Publisher: American Association for the Advancement of Science _eprint: https://science.sciencemag.org/content/352/6281/aac7341.full.pdf.

(8) Chen, Y.; Zhang, S.; Young, E. M.; Jones, T. S.; Densmore, D.; Voigt, C. A. Genetic circuit design automation for yeast. *5*, 1349–1360.

(9) McLaughlin, J. A.; Myers, C. J.; Zundel, Z.; Mısırlı, G.; Zhang, M.; Ofiteru, I. D.; Goñi-Moreno, A.; Wipat, A. SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synthetic Biology* **2018**, *7*, 682–688.

(10) Urquiza-García, U.; Zieliński, T.; Millar, A. J. Better research by efficient sharing: evaluation of free management platforms for synthetic biology designs. *Synthetic Biology (Oxford, England)* **2019**, *4*, ysz016.

(11) Wang, B.; Yang, H.; Sun, J.; Dou, C.; Huang, J.; Guo, F.-B. BioMaster: An Integrated Database and Analytic Platform to Provide Comprehensive Information About BioBrick Parts. *Frontiers in Microbiology* **2021**, *12*.

(12) Eisenberg, J. D.; Banisakher, D.; Presa, M.; Unthank, K.; Finlayson, M. A.; Price, R.;

Chen, S.-C. Toward Semantic Search for the Biogeochemical Literature. 2017 IEEE International Conference on Information Reuse and Integration (IRI). 2017; p 517–525.

(13) Soto, A. J.; Przybyła, P.; Ananiadou, S. Thalia: semantic search engine for biomedical abstracts. *Bioinformatics* **2019**, *35*, 1799–1801.

(14) Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes* **2007**, *30*, 3–26.

(15) Crichton, G.; Pyysalo, S.; Chiu, B.; Korhonen, A. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics* **2017**, *18*, 368.

(16) Akdemir, A.; Shibuya, T. Analyzing the Effect of Multi-task Learning for Biomedical Named Entity Recognition. *arXiv:2011.00425 [cs]* **2020**, arXiv: 2011.00425.

(17) Weinzierl, M. A.; Maldonado, R.; Harabagiu, S. M. The impact of learning Unified Medical Language System knowledge embeddings in relation extraction from biomedical texts. *Journal of the American Medical Informatics Association* **2020**, *27*, 1556–1567.

(18) Kilicoglu, H.; Rosemblat, G.; Fiszman, M.; Shin, D. Broad-coverage biomedical relation extraction with SemRep. *BMC bioinformatics* **2020**, *21*, 1–28.

(19) van Altena, A. J.; Moerland, P. D.; Zwinderman, A. H.; Olabarriaga, S. D. Understanding big data themes from scientific biomedical literature through topic modeling. *Journal of Big Data* **2016**, *3*, 1–21.

(20) Kavvadias, S.; Drosatos, G.; Kaldoudi, E. Supporting topic modeling and trends analysis in biomedical literature. *Journal of Biomedical Informatics* **2020**, *110*, 103574.

(21) Beal, J.; Nguyen, T.; Gorochowski, T. E.; Goñi-Moreno, A.; Scott-Brown, J.; McLaughlin, J. A.; Madsen, C.; Aleritsch, B.; Bartley, B.; Bhakta, S. et al. Communicating Structure and Function in Synthetic Biology Diagrams. *ACS Synthetic Biology* **2019**, *8*, 1818–1825, PMID: 31348656.

(22) McLaughlin, J. A.; Myers, C. J.; Zundel, Z.; Mısırlı, G.; Zhang, M.; Ofiteru, I. D.; Goñi-Moreno, A.; Wipat, A. SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synthetic Biology* **2018**, *7*, 682–688, PMID: 29316788.

(23) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240.

(24) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* **2017**,

(25) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

(26) Taboureau, O.; Nielsen, S. K.; Audouze, K.; Weinhold, N.; Edsgärd, D.; Roque, F. S.; Kouskoumvekaki, I.; Bora, A.; Curpan, R.; Jensen, T. S. et al. ChemProt: a disease chemical biology database. *Nucleic acids research* **2010**, *39*, D367–D372.

(27) Sun, C.; Yang, Z.; Su, L.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J. Chemical-protein Interaction Extraction via Gaussian Probability Distribution and External Biomedical Knowledge. *Bioinformatics* **2020**, btaa491.

(28) Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research* **2003**, *3*, 993–1022.

(29) Blei, D. M.; Lafferty, J. D. Correlated Topic Models. Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge, MA, USA, 2005; p 147–154.

(30) Raimbault, B.; Cointet, J.-P.; Joly, P.-B. Mapping the Emergence of Synthetic Biology. *PLOS ONE* **2016**, *11*, 1–19.

(31) Lee, M. E.; DeLoache, W. C.; Cervantes, B.; Dueber, J. E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synthetic Biology* **2015**, *4*, 975–986, Publisher: American Chemical Society.

(32) Obst, U.; Lu, T. K.; Sieber, V. A Modular Toolkit for Generating *Pichia pastoris* Secretion Libraries. *ACS Synthetic Biology* **2017**, *6*, 1016–1025.

(33) Young, E. M.; Zhao, Z.; Gielesen, B. E. M.; Wu, L.; Benjamin Gordon, D.; Roubos, J. A.; Voigt, C. A. Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast. *Metabolic Engineering* **2018**, *48*, 33–43.

(34) Iverson, S. V.; Haddock, T. L.; Beal, J.; Densmore, D. M. CIDAR MoClo: Improved MoClo Assembly Standard and New E. coli Part Library Enable Rapid Combinatorial Design for Synthetic and Traditional Biology. *ACS Synthetic Biology* **2016**, *5*, 99–103, Publisher: American Chemical Society.

(35) Addgene: CIDAR MoClo Extension, Volume I. `https://www.addgene.org/kits/murray-cidar-moclo-v1/#protocols-and-resources`.

(36) Chen, Y.-J.; Liu, P.; Nielsen, A. A. K.; Brophy, J. A. N.; Clancy, K.; Peterson, T.; Voigt, C. A. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods* **2013**, *10*, 659–664, Number: 7 Publisher: Nature Publishing Group.

(37) Misirli, G.; Wipat, A.; Mullen, J.; James, K.; Pocock, M.; Smith, W.; Allenby, N.; Hallinan, J. S. BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology. *Journal of Integrative Bioinformatics* **2013**, *10*, 103–116.

(38) Mısırlı, G.; Hallinan, J.; Pocock, M.; Lord, P.; McLaughlin, J. A.; Sauro, H.; Wipat, A. Data Integration and Mining for Synthetic Biology Design. *ACS synthetic biology* **2016**, *5*, 1086–1097.

(39) Terry, L.; Earl, J.; Thayer, S.; Bridge, S.; Myers, C. J. SBOLCanvas: A Visual Editor for Genetic Designs. *ACS Synthetic Biology* **2021**,

(40) lxml - XML and HTML with Python. `https://lxml.de`.

(41) Industrial-Strength Natural Language Processing. `https://spacy.io/`.

(42) allenai, SpaCy Models for Biomedical Text Processing. `https://allenai.github.io/scispacy/`.

(43) Hugging Face: BERT. `https://huggingface.co/transformers/model_doc/bert.html`.

(44) BioBERT Source Code. `https://github.com/dmis-lab/biobert`.

(45) Weber, L.; Münchmeyer, J.; Rocktäschel, T.; Habibi, M.; Leser, U. HUNER: improving biomedical NER with pretraining. *Bioinformatics* **2020**, *36*, 295–302.

(46) Hugging Face: BertTokenizer. `https://huggingface.co/transformers/model_doc/bert.html#berttokenizer`.

(47) NLTK Project: Natural Language Toolkit. `https://www.nltk.org/`.

(48) McCallum, A. K. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`.

(49) Řehůřek, R. Gensim: Topic Modelling for Humans. `https://radimrehurek.com/gensim/index.html`.

(50) Shapira, P.; Kwon, S.; Youtie, J. Tracking the emergence of synthetic biology. *Scientometrics* **2017**, *112*, 1439–1469.

(51) Zundel, Z.; Samineni, M.; Zhang, Z.; Myers, C. J. A Validator and Converter for the Synthetic Biology Open Language. *ACS Synthetic Biology* **2017**, *6*, 1161–1168.

(52) Palmer, J. A. pdftotext: Simple PDF text extraction. `https://github.com/jalan/pdftotext`.

(53) Eilbeck, K.; Lewis, S. E.; Mungall, C. J.; Yandell, M.; Stein, L.; Durbin, R.; Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology* **2005**, *6*, R44.

Graphical TOC Entry