SeqImprove: Machine Learning Assisted Curation of Genetic Circuit Sequence Information

Jeanet Mante, Zach Sents, Duncan Britt, William Mo, Chunxiao Liao, Ryan Greer, and Chris J. Myers*

University of Colorado Boulder, Boulder, CO, 80309, United States

E-mail: chris.myers@colorado.edu

Abstract

The progress and utility of synthetic biology is currently hindered by the lengthy process of studying literature and replicating poorly documented work. Reconstruction of crucial design information through post-hoc curation is highly noisy and error-prone. To combat this, author participation during the curation process is crucial. To encourage author participation without overburdening them, an ML-assisted curation tool called SeqImprove has been developed. Using named entity recognition, named entity normalization, and sequence matching, SeqImprove creates machine-accessible sequence data and metadata annotations, which authors can then review and edit before submitting a final sequence file. SeqImprove makes it easier for authors to submit sequence data that is FAIR (findable, accessible, interoperable, and reusable).

Keywords

named entity recognition, named entity normalization, SBOL, ontologies, machine learning

Introduction

Synthetic biology has vast potential applications in numerous fields.[1] However, the progress and utility of synthetic biology are currently hindered by the lengthy process of studying literature and replicating poorly documented work. The reuse of genetic components is currently low.[2,3] More complete data records makes the data more reusable and the database to which they are submitted more valuable.[3,4]

Much of the data that is submitted is not *findable, accessible, interoperable, and reusable* (FAIR).[5] The *Synthetic Biology Knowledge System* (SBKS) attempted to address this problem by creating an integrated knowledge system built using data generated with post-hoc curation.[6] The curation consisted of two parts: (1) text mining to perform automatic annotation of the articles using *natural language processing* (NLP) to identify salient content such as key terms, relationships between terms, and main topics;[7] and (2) a data mining pipeline that performs automatic annotation of the sequences extracted from the supplemental documents with the genetic parts used in them.[8] The curation allows the linkage of knowledge, genetic parts, and the context in which they are used to the papers describing their usage. In order to process vast amounts of data, automated tools are employed to analyze unstructured text and identify relevant keywords, while attempting to derive their intended meaning from the surrounding context. This approach tested the limits of NLP methods, such as *named entity recognition* (NER) and *entity classification*.[9] Furthermore, sequences provided as supplemental information in publications are typically poorly annotated, incomplete, and provided in non-machine accessible formats (e.g. PDFs). The SBKS project demonstrated that reconstruction of important design information through post-hoc curation is extremely noisy and error prone.[6,8]

The idea of author based curation (having the submitters curate their own data) is becoming increasingly popular,[10] and it would help address the issues encountered by the SBKS project. Author curation requires intuitive interfaces to ensure standardization and completeness in their metadata. We developed the SeqImprove curation interface to enable

authors to curate machine generated metadata and annotations and save this in a machine accessible format. This paper presents the capabilities and underlying architecture of SeqImprove.

Results

SeqImprove is designed to aid authors in creating machine accessible sequence data with complete metadata. It consists of a user-interface that was built using modular code. It can be reused by others to work as the front-end for their curation software. Additionally, the back-end consists of a series of tools that automate NER, *named entity normalization* (NEN), sequence annotation, and protein prediction. The functions are accessed by users via the front end. The backend has two main machine aided curation functions:

1. **Annotate Sequence**: This is the method used to suggest sequence annotations. It is based on SYNBICT.[11] It uses the feature libraries found in our Github Repository. These include libraries from parts-rich papers.[12–15]

2. **Annotate Text**: This method is used to suggest keyword annotations. It uses BERN2 for NER and NEN.[16] Additional fuzzy matches are carried out to catch potential misspellings using the fast-fuzzy package.

The first step is sequence data input using an existing sequence file in the *Synthetic Biology Open Language* (SBOL),[17] GenBank, or FASTA file format, or a link to a sequence already stored in SynBioHub.[18] It is also capable of providing an empty template for the user to manually copy-and-paste a DNA sequence of interest. Next, it takes authors through four sections of metadata.

The first section, as shown in Figure 1, provides the description of the part with hyperlinks for recognized terms, allows users to select the role or function of the sequence via a drop down menu of *sequence ontology* (SO) terms,[19] designate any target organisms of sequence

insertion in machine accessible formatting based on the *National Center for Biotechnology Information* (NCBI) Taxonomy,[20] and link relevant papers or pre-prints using a DOI.

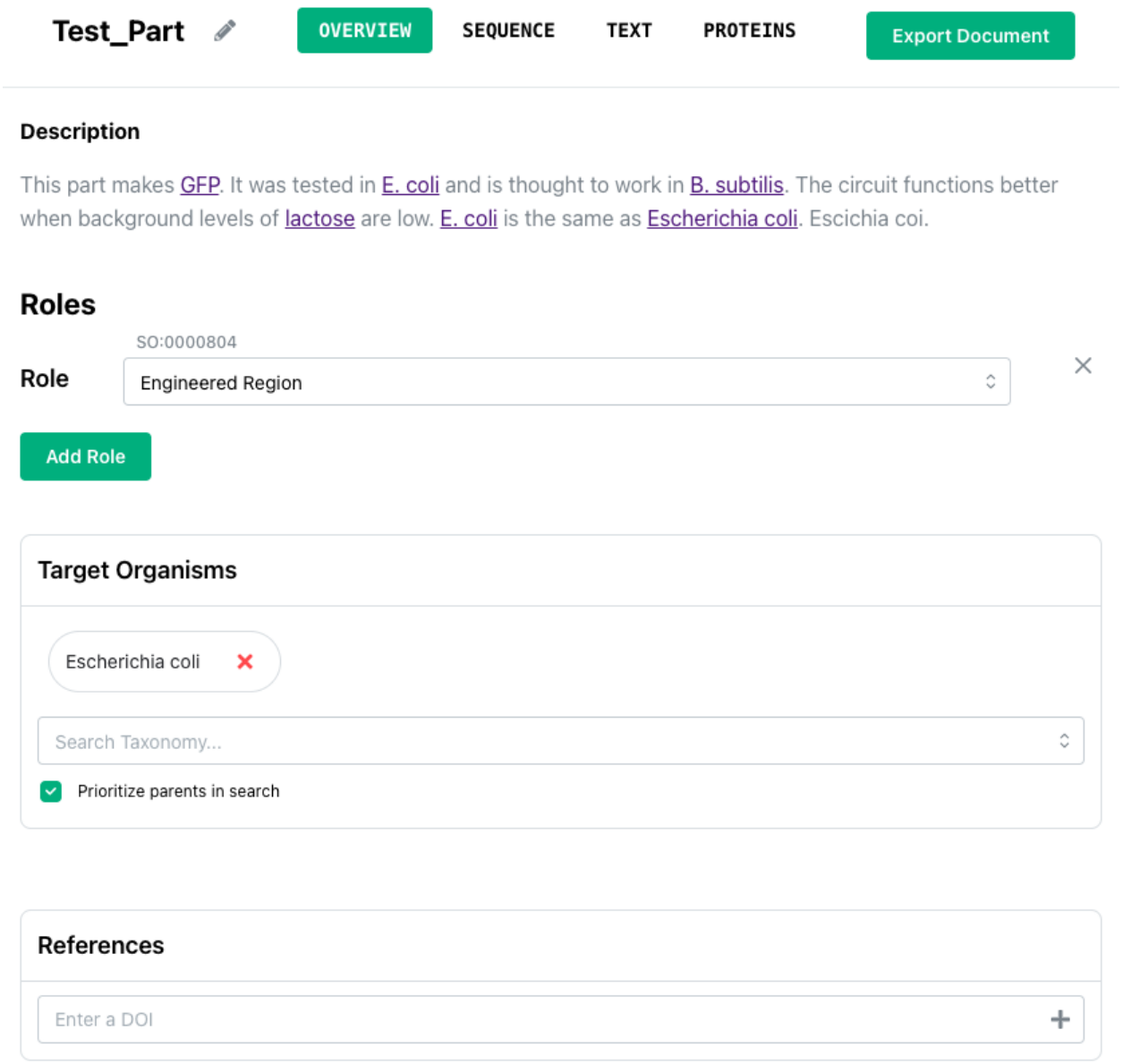


Figure 1: Overview Tab. Shows description with hyperlinks, role selection menu, target organisms with search, and references.

The second section, as shown in Figure 2(a), displays the sequence annotated with sub-components. The "Analyze Sequence" button can be used to generate suggestions of sub-components based on a SeqImprove library of frequently used components. The suggestions

may be accepted by selecting the checkbox next to the sub-component's name. Alternatively, the user can manually add and label their own annotations.

The third section, as shown in Figure 2(b), is where the description can be edited and machine accessible keywords are selected within the description. The "Analyze Text" button uses machine learning to suggest keywords, group similar ones together, and suggest a machine accessible ontology term for the keyword. Like in the previous section, users can approve a keyword with checkboxes, or manually add annotations using the "Create Text Annotation" button.

In the final section, as shown in Figure 2(c), proteins associated with the sequence are added to the metadata. There is a suggestion box where proteins obtained from Uniprot by querying with a text annotation list are provided. For example, *E. coli* in the description field leads to the suggestion of common *E. coli* proteins. The user can also add further proteins by directly searching the UniProt database.[21] These proteins are added as components to the SBOL output, but are not automatically connected to genetic parts that may code for or interact with them.

After completing curation and annotation to their liking, the user can export the final SBOL file. This can be done either as a download to their local directories or as an upload to a SynBioHub collection, if they log in as a registered SynBioHub user.

Discussion

We have presented SeqImprove, a platform for machine-assisted author curation of genetic sequences. SeqImprove helps authors submit sequence data and associated metadata in machine accessible formats. It prompts authors to consider metadata such as role, target organism, reference papers, sub-sequences, protein production, and keywords. It makes the information machine accessible by using existing ontologies to structure the metadata. Authors are helped by suggestions of keywords, proteins, and sequence annotations. They

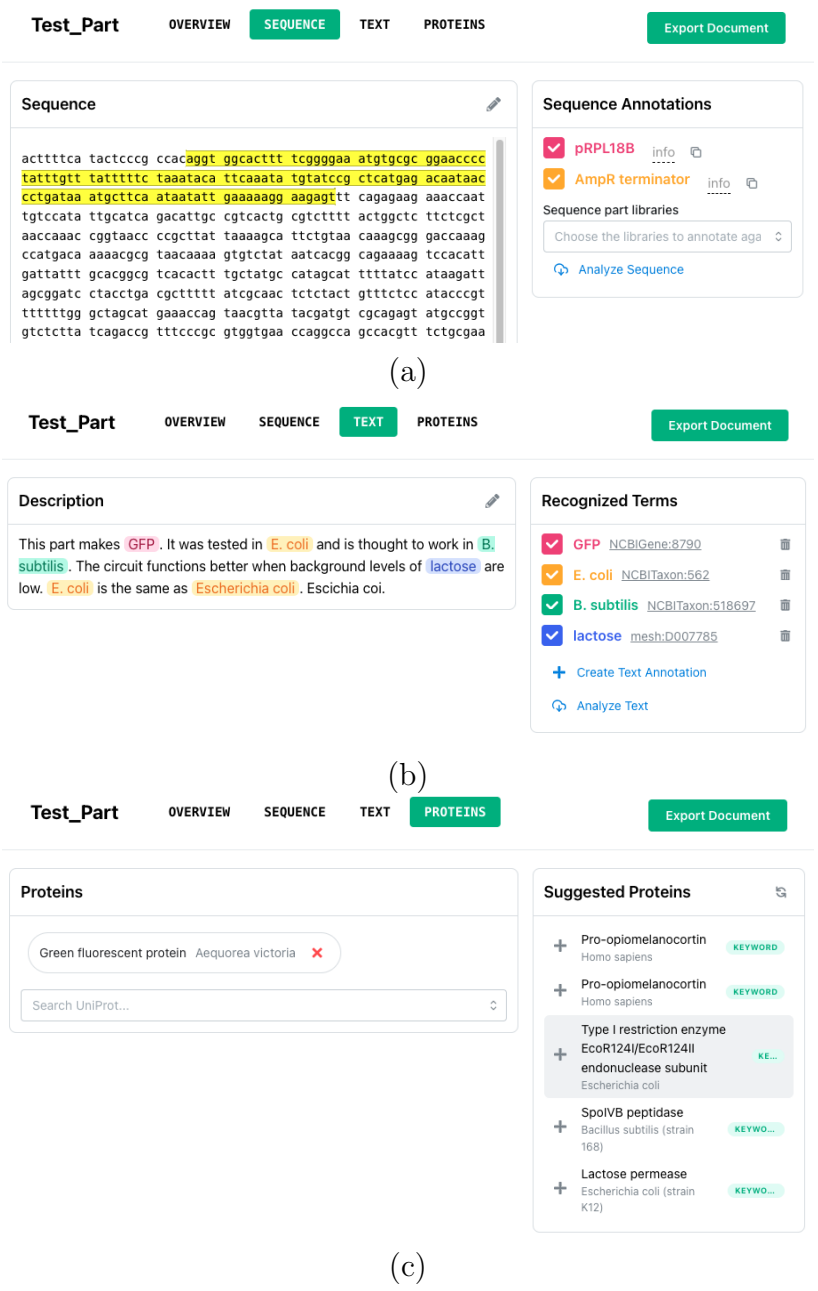


Figure 2: Additional tabs for detailed machine-assisted curation of sequence and metadata. (a) The sequence tab shows the sequence and annotations of parts within a sequence. The dropdown menu below the listed annotations allows the users to control what part libraries are used for the automatic part annotation process. (b) The text tab shows the description and various keyword annotations. Clicking on the gray underlined portion of each annotation will open a new tab to a link explaining the term. Automatic annotations will link to a standard biological database entry. (c) The proteins tab shows associated proteins and suggests additional ones. Users can click on the suggestions to automatically add them as associated proteins, or manually add proteins found via the search function.

can review and edit the suggestions in a user friendly interface. The interface was also designed to be modular so it could be reused for similar curation in other contexts.

ChatGPT is an important alternative machine-learning platform that is intuitive and familiar to many researchers. However, we emphasize that SeqImprove has several advantages over ChatGPT for this specialized task of metadata annotation. As an example, we compared ChatGPT-3.5's annotation attempt against that of SeqImprove on a sample sequence description. ChatGPT was specifically prompted to use NCBI Taxonomy, *Medical Subject Headings* (MeSH) descriptors, and NCBI Gene names to perform NER and NEN on the description field, which is embedded with NCBI or MeSH terms and has some spelling mistakes. ChatGPT was tasked with returning the recognized terms with URLs properly formatted to the ontology terms in the same SBOL recognizable format that SeqImprove generates. Both SeqImprove and ChatGPT are able to ground "GFP" with NCBI Gene, and "E. coli" and "B. subtilis" as NCBI Taxonomy, while failing to recognize the misspelled word "Escichia coi". However, ChatGPT notably fails to detect "lactose" as a MeSH descriptor, whereas SeqImprove can. This discrepancy arises from ChatGPT's training dataset being too general compared to BERN2, and thus viewing certain terms like lactose as too general and not grounding them to proper ontologies. Furthermore, in the realm of sequence annotation, ChatGPT reads all of a DNA or protein sequence as regular text and is entirely unable to annotate because it cannot easily reference part libraries. Even setting aside that SeqImprove pulls ahead in annotation performance, SeqImprove also intrinsically provides a user friendly interface that generates results immediately, whereas ChatGPT must generate the annotation progressively with explicit user guidance. SeqImprove is also capable of directly editing and uploading annotated sequences and metadata to parts repositories, which ChatGPT cannot because it is a separate and more general platform.

While SeqImprove offers many benefits, there are still limitations to the system for future work to address. For example, sequence annotation is currently limited to identifying predefined features and parts in a genetic sequence. Going forward, SeqImprove could be

expanded by using new annotation methods that support other kinds of synthetic biology constructs, and with more expansive and well-characterized part libraries. However, the most important limitation is author participation. SeqImprove only works if authors use it, and it can be difficult to incentivize researchers to participate. Since the top benefit for researchers is faster searching for sequences that others have curated and additional citations for their own, the results of additional effort in curation are indirect. This is particularly the case initially, as there will be little well curated output data to be utilized. This reduces the incentives for researchers to adopt the system, and without adoption, the available data remains scarce. Breaking the initial consensus threshold will be difficult, but would be aided by journal incentives. Fortunately, if a trend towards more author curation is able to take hold, this would result in more reliable information being publicly available and in turn provide better training data for models. This virtuous cycle can lead to SeqImprove itself improving with more reliable prediction and annotation based on existing well-documented work. It will also enable the development of novel LLMs that can invoke SeqImprove and its routines, thus integrating them into an even more user-intuitive interface without the loss of accuracy presently associated with ChatGPT, similar to specialized LLM agents developed in other fields.[22] As more researchers adopt the use of machine learning tools to make their work FAIR, this will both raise the community expectation and lower the technical difficulty to do so.

Methods

SeqImprove is an application for curating genetic designs encoded in SBOL. It can be run standalone or as a SynBioHub curation plugin.[8,23] It is meant to help users easily add metadata to their genetic designs by providing recommendations and a simple interface with which to do so. SeqImprove consists of two applications and a package. The two applications are a React front-end and a Dockerized Flask/Python API that functions as the back-end.

The package is called text-ranger and was developed to make working with text ranges and replacements easier, as this is key functionality for creating and displaying text annotations for genetic designs.

Text-ranger is an internal Node.js package used to aid in the modeling of text annotations. It provides an interface for creating text replacements based on a start and end position. It compiles the replaced text on demand and adjusts replacement positions if the underlying text is edited. This package is used as part of the front-end for text annotation.

Acknowledgement

We thank all members of the Genetic Logic Lab at CU Boulder. This work was supported by the National Science Foundation under Grant No. 1939892 and 2231864, a Draper Doctoral Scholarship, and the CU Summer Program for Undergraduate Research. This document does not contain technology or technical data controlled under either U.S. International Traffic in Arms Regulation or U.S. Export Administration Regulations. 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 agencies.

Author Contributions

JM designed the high-level architecture of SeqImprove. ZS, DB, and RG implemented the design and wrote the bulk of the code. WM and CL provided additional support during the development process. CM provided guidance throughout. All authors contributed to the writing of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Supporting Information Available

- GitHub: `https://github.com/MyersResearchGroup/SeqImprove`

References

(1) Alterovitz, G.; Muso, T.; Ramoni, M. F. The challenges of informatics in synthetic biology: from biomolecular networks to artificial organisms. *Briefings in Bioinformatics* **2010**, *11*, 80–95.

(2) Vilanova, C.; Porcar, M. iGEM 2.0—refoundations for engineering biology. *Nature Biotechnology* **2014**, *32*, 420–424.

(3) Mante, J.; Myers, C. J. Advancing reuse of genetic parts: progress and remaining challenges. *nature communications* **2023**, *14*, 2953.

(4) Howe, D.; Costanzo, M.; Fey, P.; Gojobori, T.; Hannick, L.; Hide, W.; Hill, D. P.; Kania, R.; Schaeffer, M.; St Pierre, S. et al. The future of biocuration. *Nature* **2008**, *455*, 47–50.

(5) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **2016**, *3*, 1–9.

(6) Mante, J.; Hao, Y.; Jett, J.; Joshi, U.; Keating, K.; Lu, X.; Nakum, G.; Rodriguez, N. E.; Tang, J.; Terry, L. et al. Synthetic Biology Knowledge System. *ACS Synthetic Biology* **2021**,

(7) McInnes, B. T.; Downie, J. S.; Hao, Y.; Jett, J.; Keating, K.; Nakum, G.; Ranjan, S.; Rodriguez, N. E.; Tang, J.; Xiang, D. et al. Discovering Content through Text Mining for a Synthetic Biology Knowledge System. *ACS synthetic biology* **2022**, *11*, 2043–2054.

(8) Mante, J. V. Promotion of Data Reuse in Synthetic Biology. Ph.D. thesis, CU Boulder, Boulder Colorado, 2022.

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

(10) Zaveri, A.; Hu, W.; Dumontier, M. MetaCrowd: Crowdsourcing Biomedical Metadata Quality Assessment. *Human Computation* **2019**, *6*, 98–112.

(11) Roehner, N.; Mante, J.; Myers, C. J.; Beal, J. Synthetic Biology Curation Tools (SYNBICT). *ACS Synthetic Biology*

(12) 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.

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

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

(15) 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.

(16) Sung, M.; Jeong, M.; Choi, Y.; Kim, D.; Lee, J.; Kang, J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **2022**, *38*, 4837–4839.

(17) Roehner, N.; Beal, J.; Clancy, K.; Bartley, B.; Misirli, G.; Grünberg, R.; Oberortner, E.; Pocock, M.; Bissell, M.; Madsen, C. et al. Sharing Structure and Function in Biological Design with SBOL 2.0. *ACS Synthetic Biology* **2016**, *5*, 498–506.

(18) 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.

(19) 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.

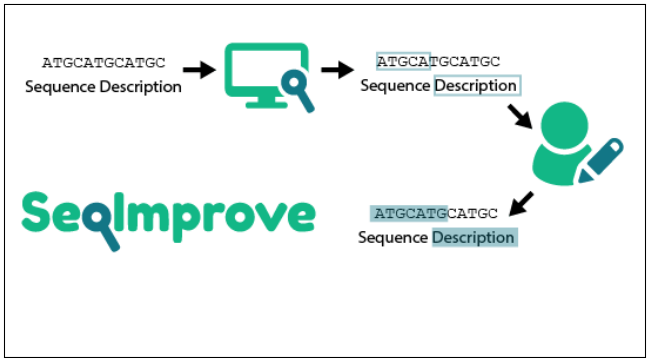
(20) Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation* **2020**, *2020*, baaa062.

(21) Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **2019**, *47*, D506–D515.

(22) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, 1–11.

(23) Mante, J.; Zundel, Z.; Myers, C. Extending SynBioHub's Functionality with Plugins. *ACS Synthetic Biology* **2020**, *9*, 1216–1220.

TOC Graphic


ATGCATGCATGC
Sequence Description
ATGCATGCATGC
Sequence Description
SeqImprove
ATGCATGCATGC
Sequence Description