

Leveraging LLMs and Web-based Visualizations for Profiling Bacterial Host Organisms and Genetic Toolboxes

Gilchan Park^{1*}, Vivek Mutalik², Christopher Neely²,
Carlos Soto¹, Shinjae Yoo¹, Paramvir Dehal²

¹Brookhaven National Laboratory, Upton, New York, USA

²Lawrence Berkeley National Laboratory, Berkeley, California, USA

*Correspondence: gpark@bnl.gov

Abstract

Building genetic tools to engineer microorganisms is at the core of understanding and re-designing natural biological systems for useful purposes. Every project to build such a genetic toolbox for an organism starts with a survey of available tools. Despite a decade-long investment and advancement in the field, it is still challenging to mine information about a genetic tool published in the literature and connect that information to microbial genomics and other microbial databases. This information gap not only limits our ability to identify and adopt available tools to a new chassis but also conceals available opportunities to engineer a new microbial host. Recent advances in natural language processing (NLP), particularly large language models (LLMs), offer solutions by enabling efficient extraction of genetic terms and biological entities from a vast array of publications. This work presents a method to automate this process, using text-mining to refine models with data from bioRxiv and other databases. We evaluated various LLMs to investigate their ability to recognize bacterial host organisms and genetic toolboxes for engineering. We demonstrate our methodology with a web application that integrates a conversational LLM and visualization tool, connecting user inquiries to genetic resources and literature findings, thereby saving researchers time, money and effort in their laboratory work. The code and data are available at: <https://github.com/boxorange/LLM-GeneticTool-Extraction>

1 Introduction

Our planet currently faces significant challenges concerning biological resources, including limited renewable energy sources, lack of innovative treatments for endemic infectious diseases, water pollution, insufficient arable land resulting in food crises, and the degradation of ecosystems (WEF, 2020;

Arkin et al., 2010), among other urgent issues. We postulate – as others have – that the capacity to domesticate and genetically engineer non-model microorganisms from relevant environments could facilitate the development of potential solutions to many of these urgent global problems (Endy, 2005; Stacey, 2017). Although recent technological advances have been made at a rapid pace to address several of these challenges, the information needed for each potential new model organism is dispersed across the literature and is not readily accessible to many practitioners. This situation complicates every new synthetic biology, bioenergy, and biomanufacturing project involving a non-model organism (Mutalik et al., 2013; Council et al., 2015). The disorganized nature of the information not only impedes machine-readable approaches but also hinders the assessment of the scope of work and identification of knowledge gaps, subsequently offering limited guidance for investment to overcome technological barriers. For instance, despite decades of progress in the field of synthetic biology, it remains challenging to pinpoint suitable microbial targets for specific applications and conditions, as well as the genetic tools required for cultivating and engineering non-model microorganisms (Arkin, 2008; Council et al., 2015; Oberhardt et al., 2015; Price and Arkin, 2017).

A comprehensive literature mining tool that monitors emerging technologies and genetic tools critical for biotechnology professionals would be highly beneficial. This envisioned tool would allow us to identify information gaps and detect opportunities concealed within extensive literature. For instance, the tool should efficiently ascertain whether a chosen organism is suitable for laboratory domestication and which genetic tools are available for that organism, streamlining the search process and conserving time, effort, and funding for numerous lab-oriented projects.

Recent advancements in Natural Language Pro-

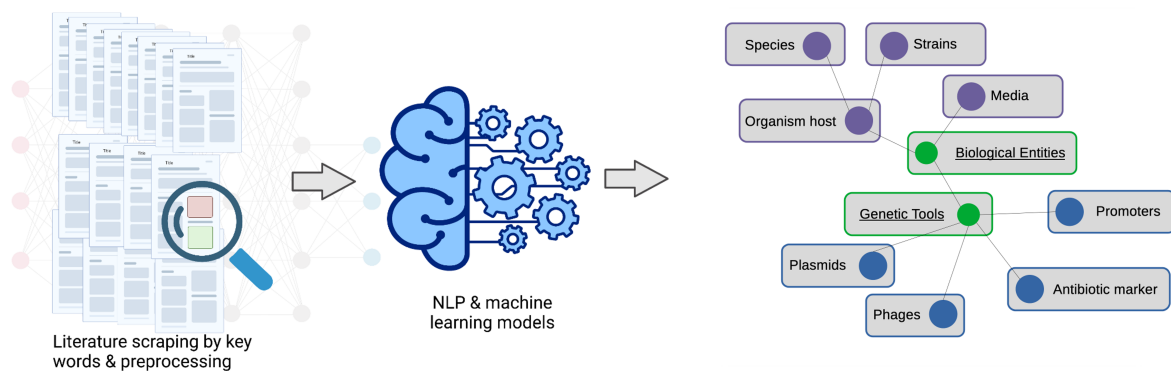


Figure 1: The project objective and workflow: We propose to use NLP and machine learning models to process and learn from literature data about growth characteristics, conditions, traits and available *in silico* models and genetic tools to engineer microorganisms.

cessing (NLP) and Large Language Model (LLM) have enabled the analysis of textual data at an unprecedented scale (e.g., millions of documents), allowing for the extraction of significant contextual information in ways that were previously unachievable. These innovative techniques offer considerable potential to bridge the knowledge gap discussed earlier. In this study, we propose employing NLP on biological literature to identify organism traits (as depicted in Figure 1) and systematically compiling this extracted knowledge. This approach facilitates the development of automated, curated centralized systems crucial for the cultivation and engineering of microorganisms.

This work specifically focuses on the extraction of information about bacterial organism hosts and genetic toolbox for them from literature. As detailed in this paper, our contribution is threefold:

1. We present a curated corpus of bacterial host organisms and genetic toolboxes, classified into 14 distinct labels derived from scientific literature, including plasmids, promoters, reporters, and other entities of interest. The selection of a bacterial host organism is determined by the accessibility and suitability of genetic toolbox for efficient manipulation and engineering. This, in turn, informs the feasibility and ease of engineering the selected host organism for targeted applications within the synthetic biology field.
2. This paper provides an evaluation of various publicly available LLMs for the task of recognizing organism hosts and genetic tools. Our findings demonstrate the efficacy of fine-tuning LLMs on an annotated dataset, which

enhances the models' performance in those entity recognition.

3. This work presents a chatbot interface designed to facilitate discussions between users and a specialized chatbot. The chatbot leverages public biological resource such as NCBI taxonomy, genetic tool databases, and publication information. Users can pose questions regarding the genetic engineerability of biological entities and engage in informative dialogues on the subject.

2 Related Work

LLMs demonstrated significant improvements in addressing a multitude of NLP tasks that are critical to the fields of biology and biomedicine (Chen et al., 2023; Yu et al., 2024). Instructed on a broad spectrum of text corpora, encompassing web crawls, medical records, and rigorously selected datasets, LLMs are equipped with the proficiency to integrate information from diverse sources. These sources range from scientific publications and databases to various other forms of informational repositories. This integrative ability enables LLMs to identify complex interconnections, nuanced contextual aspects, and insights that may remain obscure to traditional methods. The BioMistral model (Labrak et al., 2024), built upon the Mistral foundational model and subsequently pre-trained on PubMed Central, underwent evaluation on a benchmark encompassing 10 medical question-answering (QA) tasks. This assessment revealed superior performance compared to the original model and existing open-source medical counterparts.

They can subsequently be iteratively combined with other Bio-Bricks for the assembly of the desired vectors. BioBrick cloning. Markers from [E. coli](#) plasmids that confer resistance to kanamycin ([Kanr](#)), spectinomycin ([Spcr](#)), and tetra-cycline ([Tetr](#)) were cloned into [pIM1154](#), thereby converting them into BioBricks ([pIM1157](#), [pIM1212](#), and [pIM1265](#), respectively). A series of expression cassette BioBricks, each containing a regulator gene, a promoter, a

•Organism_hos...
•AntibioticMa...
•AntibioticMa...
•AntibioticMa...
•Plasmid
•Plasmid
•Plasmid
•Plasmid

Figure 2: An example of annotation through the Doccano web server.

McInnes et al. (2022) presents the development of a synthetic biology knowledge system, wherein a text processing pipeline utilizes NLP techniques to extract and correlate information from the literature aimed at synthetic biology researchers. The pipeline integrates named entity recognition, relation extraction, concept grounding, and topic modeling methodologies to extract pertinent information from published literature. Subsequently, this extracted information is utilized to establish connections between articles and elements within the knowledge system. The findings demonstrate the effectiveness of each component when applied to synthetic biology literature and propose avenues for further enhancing the pipeline’s capabilities. Gong et al. (2023) explored the potential of various LLMs such as GPT-4, GPT-3.5, PaLM2, Claude2, and SenseNova in addressing conceptual biology questions, including those related to synthetic biology, such as principles of genetic circuit design and CRISPR-based genome editing techniques. While the findings revealed the adeptness of LLMs in logical reasoning and their potential to support biology research by facilitating tasks such as data analysis, hypothesis formulation, and knowledge synthesis, the authors underscored the necessity for further refinement and validation before fully harnessing the potential of LLMs to expedite biological discovery.

In this study, we assessed the effectiveness of LLMs in recognizing and extracting pertinent information regarding host organisms and their associated genetic engineering tools. Of particular emphasis was the evaluation of open-source LLMs, chosen for their heightened adaptability and transparency in contrast to proprietary counterparts, thereby enabling users to exercise greater customization and oversight over model operations. The main objective was to gauge the efficacy of these models in discerning insights from a collec-

tion of biological literature and resources, thereby augmenting our comprehension of LLMs’ applicability in biological inquiry and ability to inform prospective applications within this domain.

3 Host Organisms and Genetic Toolbox Curation

To the best of our knowledge, there are no publicly available datasets specifically tailored for the recognition task involving bacterial host organisms and their associated genetic toolbox by machine learning models. In order to enhance the proficiency of machine learning models in identifying such entities through training on labeled datasets, we undertook an annotation endeavor aimed at labeling both organism hosts and genetic tool types as depicted in biological literature. To facilitate this annotation task, a comprehensive list of terms describing the bacterial genetic toolbox was curated, which is provided in Appendix A. When conducting a search using the genetic toolbox related terms, numerous papers unrelated to our specific focus emerge. These papers span topics ranging from human genetics to plant research. While we may consider including them at a later stage, our current emphasis lies on bacteria. Therefore, we incorporate the compound word “bacteria” along with relevant keywords to refine our search results.

For the annotation process, a corpus comprising 434 PDF papers was assembled, meticulously selected by two domain experts. Additionally, 376 XML articles containing any of the terms from the curated list described in Appendix A were obtained by querying the bioRxiv database within subject areas encompassing *Biochemistry*, *Bioengineering*, *Bioinformatics*, *Microbiology*, *Molecular Biology*, *Synthetic Biology*, and *Systems Biology*. From this corpus, a total of 795 text snippets were extracted from abstracts and main body texts, each compris-

Entity	The number of entities
Plasmid	445
Organism Host	407
Promoter	181
Genome Engineering	169
Cloning Method	158
Reporter	122
Regulator	86
Antibiotic Marker	65
Genetic Screen	40
RBS†	35
Counter Selection	27
Terminator	23
DNA Transfer	14
Operator	5

Table 1: The statistics of 1,777 annotated labels for organism hosts and genetic tools. †RBS stands for Ribosome Binding Site.

ing a target sentence containing one of the bacterial genetic toolbox terms, accompanied by two preceding and two succeeding sentences, all of which are part of the same paragraph. To ensure non-redundant annotation, we eliminated duplicate snippets that contain multiple genetic toolbox terms.

To facilitate the annotation process, a Doccano web server (Nakayama et al., 2018) was employed, thereby streamlining the task of annotating textual data. An annotation sample is depicted in Figure 2. In order to label entities within the text, a framework comprising 14 distinct entity labels was defined. Subsequently, a total of 1,777 annotated entities were obtained across the entire corpus following the completion of the annotation process. Table 1 presents the 14 labels along with the corresponding number of entities.

4 Evaluation of LLMs for Recognizing Host Organisms and Genetic Tool Types

Our study aimed to assess the potential of LLMs for the task of entity type recognition, utilizing annotated datasets. To this end, we employed a selection of LLMs, namely Falcon (Almazrouei et al., 2023), MPT (MosaicML-NLP-Team, 2023), LLaMA 2 (Touvron et al., 2023), SOLAR (Kim et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), and LLaMA 3 (Meta-AI, 2024). Model evaluation was conducted utilizing a question answering formatted prompt paired with text snippets acquired from our annotation task. An

illustrative sample of such a prompt for the entity type recognition task is provided below.

Question: Given the options: “plasmid”, “organism host”, “promoter”, “genome engineering”, “cloning method”, “reporter”, “regulator”, “antibiotic marker”, “genetic screen”, “RBS”, “counter selection”, “terminator”, “DNA transfer”, “operator”, which one is the entity type of J23108 in this text?

Text: Plasmids were cloned using Gibson Assembly or inverse PCR, propagated in *E. coli* TG1 competent cells in LB media, and isolated through miniprep (Qiagen.) Reporter plasmids had a p15A origin of replication, chloramphenicol resistance, and the terminator *trnB* downstream of the sfGFP coding sequence. Plasmids for overexpressing ribosomal proteins in vivo had a ColE1 origin of replication, ampicillin resistance, the synthetic constitutive *E. coli* promoter J23108 from the Registry of Standard Biological Parts, and the terminator *trnB* after the protein expression gene.

Answer: promoter

The 1,777 text snippets underwent partitioning into distinct train, validation, and test sets, maintaining an 8:1:1 ratio. A comparative analysis was then conducted between the original pre-trained models and their fine-tuned counterparts. To facilitate the fine-tuning process, we employed the Low-Rank Adaptation (LoRA) technique (Hu et al., 2021) coupled with quantization (QLoRA) (Dettmers et al., 2024), aiming to enhance memory efficiency and expedite training procedures. The training was performed on all linear layers within the models. The experiments were conducted on 4×NVIDIA A100 80GB GPUs. The configurations for fine-tuning the models were established as follows.

- **Batch size:** 2
- **Training epochs:** 5
- **QLoRA target modules:** all linear layers
- **Quantization technique:** BitsandBytes
- **Quantization:** 4-bit
- **Learning rate:** 1e-4 with AdamW

Table 2 presents the micro and macro F1-scores derived from the evaluation of the original pre-trained LLMs and those fine-tuned with QLoRA adaptation for the entity type recognition task employing zero-shot prompting. The results indicate that the LLaMA 3 (70B) model demonstrated superior prediction capability compared to other original LLMs. Notably, following adaptation to the

Model	Context Length	Original		QLoRA adapted	
		Micro F1	Macro F1	Micro F1	Macro F1
Falcon (7B)	2K	0.4213	0.1795	0.8933	0.7672
Falcon (40B)	2K	0.6966	0.4106	0.8820	0.6857
MPT-Chat (7B)	2K	0.5618	0.3814	0.8708	0.6503
MPT-Chat (30B)	8K	0.7697	0.6152	0.9213	0.8160
LLaMA-2-Chat (7B)	4K	0.5562	0.3703	0.8933	0.6810
LLaMA-2-Chat (70B)	4K	0.7584	0.5701	0.9157	0.8087
SOLAR-10.7B-Instruct (30B)	4K	0.7528	0.5815	0.9045	0.7252
Mistral-7B-Instruct (7B)	8K	0.7022	0.5565	0.9326	0.8351
Mixtral-8x7B-Instruct (46B)	32K	0.7135	0.5499	0.9607	0.7585
LLaMA-3 (8B)	8K	0.6573	0.4969	0.9270	0.6867
LLaMA-3 (70B)	8K	0.8708	0.6194	0.9551	0.8557

Table 2: The original pre-trained and QLoRA fine-tuned LLMs evaluation on the entity type recognition task with zero-shot prompting.

QLoRA framework, the performance of the LLM models exhibited a substantial improvement, with the LLaMA 3 (70B) and Mixtral 8x7B model as the top performer. The macro F1-scores, being lower than the micro F1-scores, suggest challenges encountered by the models in accurately identifying certain classes, such as “RBS” and “genetic screen”. The potential ambiguity or variability in interpretation of these terms may arise particularly in instances where the training data available to the models lacks comprehensive examples of the term’s utilization within the context of genetic techniques. Moreover, the effective application and interpretation of these terms can require specialized knowledge, which may be acquired through access to specific datasets within the respective field. To enhance the models’ capacity to discern and comprehend the intended meaning of such terms, it might be necessary to integrate supplementary sources of information, such as external knowledge repositories, and to provide explicit contextual explanations upon their introduction.

5 Preliminary Evaluation of a LLM for Extracting Host Organisms and Genetic Tools from Raw Articles

Our primary objective is to develop an automated tool capable of extracting host organisms and genetic tools for engineering purposes from unprocessed textual data. Such a tool would offer practitioners and domain experts a more streamlined and convenient means of accessing this information. As a preliminary phase, we initially evaluated the performance of a plain LLM in identifying bio-

logical entities and genetic tools within scientific literature. Employing a question-answering framework, we devised a structured prompt using the full text of each article. The following exemplifies a sample prompt for this specific task, incorporating the entire text of an article.

Question: Which biological entities and genetic tools are mentioned in this paper, and what are their types? Choose from the following options: “plasmid”, “organism host”, “promoter”, “genome engineering”, “cloning method”, “reporter”, “regulator”, “antibiotic marker”, “genetic screen”, “RBS”, “counter selection”, “terminator”, “DNA transfer”, “operator”

Paper: Title: A BioBricks® toolbox for multiplexed metabolic engineering of central carbon metabolism in the tetracenomycin pathway. Full-Text: The tetracenomycins are aromatic anticancer polyketides that inhibit peptide translation via binding to the large ribosomal subunit. Here, we expressed the elloramycin biosynthetic gene cluster ...

Answer: The biological entities and genetic tools mentioned in this paper are: 1. Organism host: "Streptomyces coelicolor M1146, Streptomyces glaucescens GLA.0, Streptomyces olivaceus TTü2353, Amycolatopsis sp. A23, E. coli JM109, E. coli ET12567/pUZ8002 2. Promoter: ermE*_p promoter, sf14_p promoter, LacI promoter 3. Genome engineering: cos16F4iE cluster, actinophage integrases (e.g. fC31, fBT1, SV1, TG1, SAM2, VWB) 4. Cloning method: restriction digestion, ligation, 3A cloning, isocaudomer cloning 5. Reporter: red fluorescent protein (RFP) coding device 6. Regulator: LacI repressor 7. Antibiotic marker: ampicillin, kanamycin, apramycin, viomycin, hygromycin, nalidixic acid 8. Genetic screen: blue-white colony screening 9. RBS: B0034 strong ribosome binding site 10. Counter selection: nalidixic acid 11. Terminator: B0015 transcriptional terminator

As a preliminary evaluation, we employed 233 articles used in our annotation task. Due to the length of a full text, only the Mixtral 8x7B model was adopted for this evaluation, which can have up to 32K input length. The post-processing of the model output generated a JSON formatted data. The following illustrates an example of post-processed model prediction.

```

Title: "A BioBricks toolbox for multiplexed metabolic
       engineering of central carbon metabolism
       in the tetracenomycin pathway"
{
  "antibiotic marker": [
    "ampicillin , kanamycin , apramycin , viomycin ,
    hygromycin , nalidixic acid"
  ],
  "cloning method": [
    "restriction digestion , ligation , 3A cloning ,
    isocaudomer cloning"
  ],
  "counter selection": [
    "nalidixic acid"
  ],
  "dna transfer": [
    "intergeneric conjugation , conjugal matings ,
    electroporation , ..."
  ],
  "genetic screen": [
    "blue-white colony screening"
  ],
  "genome engineering": [
    "cos16F4iE cluster , actinophage integrases
    (e.g. fC31 , fBT1 , SV1 , TG1 , SAM2 , VWB)"
  ],
  "operator": [
    "attB site , attP site , oriT , attP site"
  ],
  "organism host": [
    "Streptomyces coelicolor M1146 ,
    Streptomyces glaucescens GLA.0 , ..."
  ],
  "promoter": [
    "ermEsp promoter , sf14p promoter , LacI promoter"
  ],
  "RBS": [
    "B0034 strong ribosome binding site"
  ],
  "regulator": [
    "LacI repressor"
  ],
  "reporter": [
    "red fluorescent protein (RFP) coding device"
  ],
  "terminator": [
    "B0015 transcriptional terminator"
  ]
}

```

Two domain experts vetted this model prediction for species and tool names/types detection, and the model’s performance is displayed in Table 3. The result shows 0.8816 (micro F1) and 0.8734 (macro F1) for 6,962 entities. The model displays inherent uncertainty when encountering ambiguous terminology. For example, the term “genetic screen” has been utilized across diverse contexts, resulting in confusion within the model. This assertion is supported by the individual accuracy measurements presented in Table 4, where “genetic screen” exhibited the lowest level of precision. A similar observation was made in an earlier experiment, during which the models encountered difficulties in recognizing “genetic screen”.

Model	Micro F1	Macro F1
Mixtral-8x7B-Instruct	0.8816	0.8734

Table 3: The Mixtral 8x7B model’s evaluation on the host organism and genetic toolbox extraction task with zero-shot prompting.

Entity	Count	Accuracy
Plasmid	1485	0.8936
Organism Host	716	0.8282
Promoter	656	0.8872
Genome Engineering	601	0.8602
Antibiotic Marker	525	0.9295
Regulator	501	0.9122
Cloning Method	498	0.8313
Reporter	434	0.8594
Operator	356	0.9719
DNA Transfer	356	0.9129
RBS	233	0.8670
Genetic Screen	221	0.8281
Terminator	197	0.8782
Counter Selection	183	0.8634

Table 4: Individual Entity Accuracy

6 Chatbot for Genetic Tool Engineering

Complementing the development of a LLM to assist research in synthetic biology and biomanufacturing, KBase (Arkin et al., 2018) provides a web application for users to interact with this model through a chatbot interface. Starting from a simple question, users can ask for information about bacteria and genetic tools of interest and receive responses from the trained model. Conversations are logged, allowing users to provide feedback on the efficacy of the chatbot’s responses, which serves as valuable input for refining and enhancing the system in future iterations (see Figure 3).

Using the "outlines" Python package (Willard and Louf, 2023) for structured output generation, we identify any biological entities and their associated tools in the model’s response. The genus of each entity is collected, and a pruned NCBI taxonomy tree (Schoch et al., 2020) is rendered that highlights these organisms in the context of their genus-level neighbors. As a result of the continuous evolution of taxonomic nomenclature, the information output by the model may not reflect the current information in NCBI databases. Therefore, this tool performs additional checks against previous names and synonyms for organisms iden-

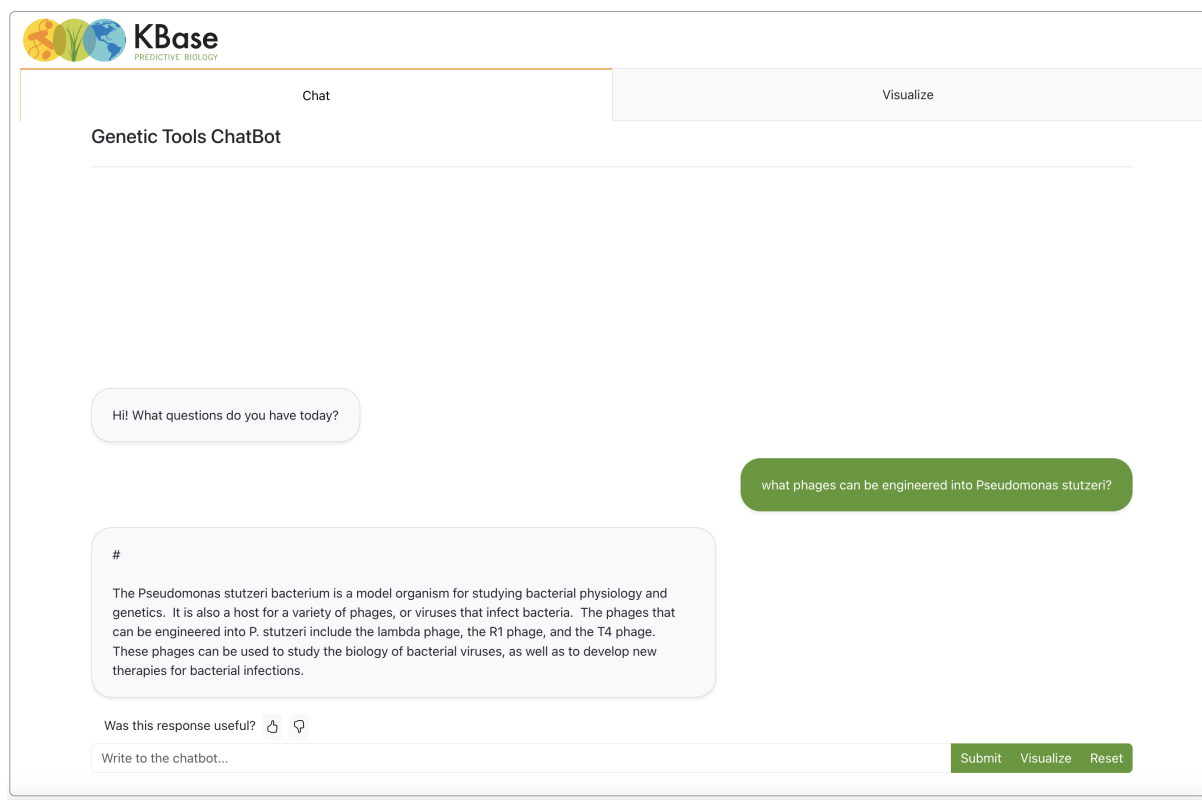


Figure 3: A screen shot of the LLM-powered Chatbot in the web application.

tified in the model output. The web application downloads the most current version of the NCBI taxonomy database at container startup.

Users may select any organisms on the species tree, and the selection will be summarized on the right side of the visualizer. The top right provides summary information from the BacDive database (Reimer et al., 2022) to rapidly identify culture conditions amenable to isolation and growth. The bottom right section summarizes genetic tool database information from the Phage-Host Daily (Albrycht et al., 2022), Virus-Host (Mihara et al., 2016), and Plasmid (Schmartz et al., 2022) databases. Additional information identified by the model is also described here. This combination of prepackaged curated databases in conjunction with information extracted from the model's large text corpora provides a comprehensive summary of the available strains, tools, and publications describing the organism in question (see Figure 4).

To facilitate the identification of organisms that do not have isolation or genetic tool information, users may elect to visualize other relative organisms without entries in the accompanying databases. Similarly, while the tree visualization is species-focused, users may also select to view

all strains for a given species in order to highlight strain-level differences in isolation and genetic tool usage.

The incorporation of this tool into the KBase infrastructure serves as an additional avenue for researchers to access pertinent information, establishing connections not only with biologically relevant organisms for laboratory investigations but also with the broader ecosystem of KBase, facilitating subsequent analyses and dissemination of findings.

Integration with the KBase platform is undertaken in adherence to established standards pertaining to containerization, user-oriented tool development, and deployment protocols. The tool is being developed with the intent of serving as a reusable proof-of-concept that caters to a diverse audience.

7 Conclusion

A significant bottleneck within the domains of synthetic biology and biomanufacturing pertains to the identification of suitable microbial targets tailored to specific applications and environmental conditions, alongside the selection of genetic tools conducive to the cultivation and engineering of non-model microorganisms. This bottleneck poses

- et al. 2010. Grand challenges for biological and environmental research: A long-term vision. Technical report.
- Adam Arkin. 2008. Setting the standard in synthetic biology. *Nature biotechnology*, 26(7):771–774.
- Adam P Arkin, Robert W Cottingham, Christopher S Henry, Nomi L Harris, Rick L Stevens, Sergei Maslov, Paramvir Dehal, Doreen Ware, Fernando Perez, Shane Canon, et al. 2018. Kbase: the united states department of energy systems biology knowledgebase. *Nature biotechnology*, 36(7):566–569.
- Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557.
- National Research Council, Division on Earth, Life Studies, Board on Life Sciences, Board on Chemical Sciences, Committee on Industrialization of Biology, and A Roadmap to Accelerate the Advanced Manufacturing of Chemicals. 2015. Industrialization of biology: A roadmap to accelerate the advanced manufacturing of chemicals.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- Drew Endy. 2005. Foundations for engineering biology. *Nature*, 438(7067):449–453.
- Xinyu Gong, Jason Holmes, Yiwei Li, Zhengliang Liu, Qi Gan, Zihao Wu, Jianli Zhang, Yusong Zou, Yuxi Teng, Tian Jiang, et al. 2023. Evaluating the potential of leading large language models in reasoning biology questions. *arXiv preprint arXiv:2311.07582*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.
- Bridget T McInnes, J Stephen Downie, Yikai Hao, Jacob Jett, Kevin Keating, Gaurav Nakum, Sudhanshu Ranjan, Nicholas E Rodriguez, Jiawei Tang, Du Xi-ang, et al. 2022. Discovering content through text mining for a synthetic biology knowledge system. *ACS synthetic biology*, 11(6):2043–2054.
- Meta-AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/llama3>. Accessed: 2024-04-19.
- Tomoko Mihara, Yosuke Nishimura, Yugo Shimizu, Hiroki Nishiyama, Genki Yoshikawa, Hideya Uehara, Pascal Hingamp, Susumu Goto, and Hiroyuki Ogata. 2016. Linking virus genomes with host taxonomy. *Viruses*, 8(3):66.
- MosaicML-NLP-Team. 2023. [Introducing mpt-30b: Raising the bar for open-source foundation models](#). Accessed: 2023-06-22.
- Vivek K Mutalik, Joao C Guimaraes, Guillaume Cambrey, Colin Lam, Marc Juul Christoffersen, Quynh-Anh Mai, Andrew B Tran, Morgan Paull, Jay D Keasling, Adam P Arkin, et al. 2013. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature methods*, 10(4):354–360.
- Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. [doccano: Text annotation tool for human](#). Software available from <https://github.com/doccano/doccano>.
- Matthew A Oberhardt, Raphy Zarecki, Sabine Gronow, Elke Lang, Hans-Peter Klenk, Uri Gophna, and Eytan Rupp. 2015. Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nature communications*, 6(1):8493.
- Morgan N Price and Adam P Arkin. 2017. Paperblast: text mining papers for information about homologs. *Msystems*, 2(4):10–1128.
- Lorenz Christian Reimer, Joaquim Sardà Carbasse, Julia Koblitz, Christian Ebeling, Adam Podstawka, and Jörg Overmann. 2022. Bac dive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Research*, 50(D1):D741–D746.
- Georges P Schmartz, Anna Hartung, Pascal Hirsch, Fabian Kern, Tobias Fehlmann, Rolf Müller, and Andreas Keller. 2022. Plsdb: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Research*, 50(D1):D273–D278.
- Conrad L Schoch, Stacy Ciuffo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen

O'Neill, Barbara Robbertse, et al. 2020. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062.

counter-selection marker, Cre-lox, genetic toolkit, lambda red system, standard biological parts, synbio reporter system, synbio toolkit

Gary Stacey. 2017. Grand challenges for biological and environmental research: Progress and future vision. a report from the biological and environmental research advisory committee. Technical report, USDOE Office of Science (SC), Washington, DC (United States). Biological and . . .

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

WEF WEF. 2020. The global risks report 2020. In *Davos: World Economic Forum*. Retrieved November, volume 15, page 2020.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Huizi Yu, Lizhou Fan, Lingyao Li, Jiayan Zhou, Zihui Ma, Lu Xian, Wenyue Hua, Sijia He, Mingyu Jin, Yongfeng Zhang, et al. 2024. Large language models in biomedical and health informatics: A bibliometric review. *arXiv preprint arXiv:2403.16303*.

A A List of Genetic Toolbox Keywords

bacteria antibiotic resistance marker, bacteria auxotrophic vector, bacteria bicistronic design, bacteria biosensors, bacteria broad-host range, bacteria chassis, bacteria counter-selection, bacteria CRISPR toolbox, bacteria CRISPR tools, bacteria degradation tags, bacteria fluorescence reporter, bacteria fluorescent marker, bacteria Fosmid system, bacteria genetic elements, bacteria genetic engineering toolbox, bacteria genetic modification, bacteria genetic toolbox, bacteria golden gate parts, bacteria heterologous expression, bacteria inducible promoters, bacteria plasmid replicon, bacteria RBS part, bacteria recombineering tools, bacteria riboswitch, bacteria ribozyme, bacteria selection marker, bacteria shuttle vector, bacteria strain engineering, bacteria suicide vector, bacteria Tn5, bacteria Tn7, bacteria TnSeq, bacteria transformation method, bacteria transposons, bacterial conjugative, bacterial genetic parts, bacterial genetic tools, bacterial genome editing, bacterial inducer, bacterial integrative vector, bacterial molecular toolbox, bacterial plasmid collection, bacterial promoter library, bacterial replicative vector, bacterial reporter, bacterial transcription terminator, bacterial vectors, bglbrick, biobrick,