PanKB

An interactive microbial pangenome knowledgebase for research, biotechnological innovation, and knowledge mining

Sun, Binhuan; Pashkova, Liubov; Pieters, Pascal Aldo; Harke, Archana Sanjay; Mohite, Omkar Satyavan; Santos, Alberto; Zielinski, Daniel C.; Palsson, Bernhard O.; Phaneuf, Patrick Victor

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gkae1042

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Nucleic Acids Research, 2025, **53**, D806–D818
https://doi.org/10.1093/nar/gkae1042
Advance access publication date: 22 November 2024
Database issue

PanKB: An interactive microbial pangenome knowledgebase for research, biotechnological innovation, and knowledge mining

Binhuan Sun [1,†], Liubov Pashkova [1,†], Pascal Aldo Pieters [1], Archana Sanjay Harke [1], Omkar Satyavan Mohite [1], Alberto Santos [1], Daniel C. Zielinski [2], Bernhard O. Palsson [1,2,3,4] and Patrick Victor Phaneuf [1,*]

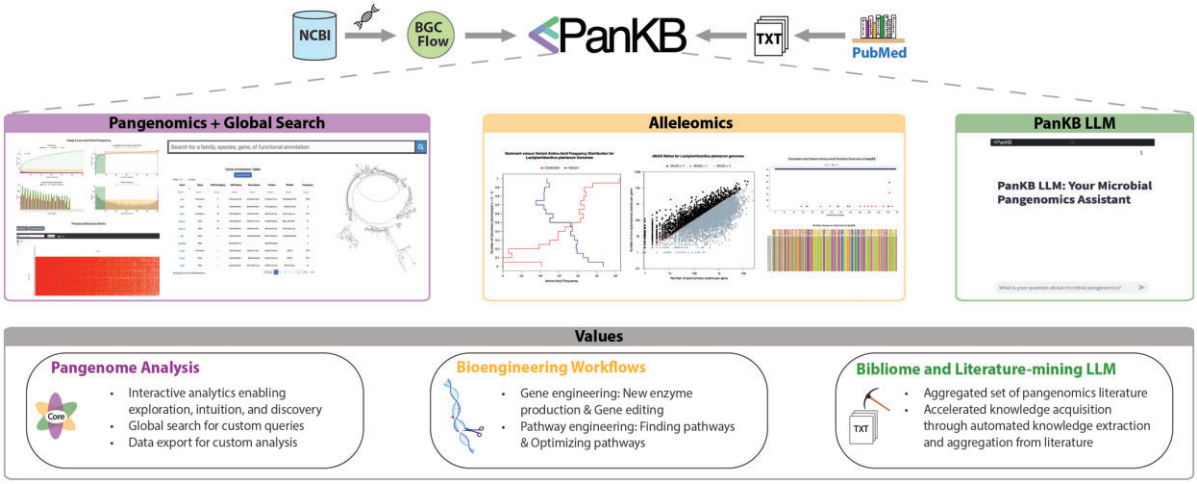
[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Building 220 Søltofts Plads, 2800 Kongens, Lyngby, Denmark
[2]Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, United States
[3]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, California 92093, United States
[4]Department of Pediatrics, University of California, San Diego, La Jolla, California 92093, United States

[*]To whom correspondence should be addressed. Email: phaneuf@biosustain.dtu.dk
[†]These authors contributed equally.

Abstract

The exponential growth of microbial genome data presents unprecedented opportunities for unlocking the potential of microorganisms. The burgeoning field of pangenomics offers a framework for extracting insights from this big biological data. Recent advances in microbial pangenomic research have generated substantial data and literature, yielding valuable knowledge across diverse microbial species. PanKB (pankb.org), a knowledgebase designed for microbial pangenomics research and biotechnological applications, was built to capitalize on this wealth of information. PanKB currently includes 51 pangenomes from 8 industrially relevant microbial families, comprising 8402 genomes, over 500 000 genes and over 7M mutations. To describe this data, PanKB implements four main components: (1) Interactive pangenomic analytics to facilitate exploration, intuition, and potential discoveries; (2) Alleleomic analytics, a pangenomic-scale analysis of variants, providing insights into intra-species sequence variation and potential mutations for applications; (3) A global search function enabling broad and deep investigations across pangenomes to power research and bioengineering workflows; (4) A bibliome of 833 open-access pangenomic papers and an interface with an LLM that can answer in-depth questions using its knowledge. PanKB empowers researchers and bioengineers to harness the potential of microbial pangenomics and serves as a valuable resource bridging the gap between pangenomic data and practical applications.

Graphical abstract



Introduction

Since the assembly of the first complete microbial genome in 1995 (1), the emergence of efficient and inexpensive sequencing technologies has driven the rapid expansion of publicly available microbial genome data, resulting in almost 2 million microbial genome assemblies in public databases (2).

Received: August 15, 2024. **Revised:** October 11, 2024. **Editorial Decision:** October 14, 2024. **Accepted:** October 19, 2024

The rapid accumulation of microbial genome data presents an opportunity to mine the vast potential of these ubiquitous organisms, and the burgeoning field of pangenomics offers a framework for extracting value from this large biological dataset. A pangenome represents the complete collection of genes found across a species' strains, constructed by comparing the strain genomes and resulting in the gene categories of core (genes present in >99% of strains), accessory (genes present in <99% to ≥15% of strains), and rare (genes present in <15% of strains) genes (3–5). By comparing genomes across strains, pangenomics reveals the genomic basis for diverse phenotypes, leading to a deeper understanding of the functional diversity of microorganisms, such as niche adaptation, pathogenicity or antibiotic resistance (6–8). Additionally, pangenome analysis provides extra approaches for microbial taxonomy classification and characterization of evolution (5,9–11).

Historically, microbes have played an important role in industry, with early examples including the use of yeast in brewing and baking, and the production of antibiotics like penicillin from mold (12). The advent of recombinant DNA technology marked a significant milestone, allowing for the creation of insulin-producing bacteria (13), which revolutionized diabetes treatment. Today, microbes are leveraged across a wide variety of applications, including medicine (14), biofuels (15), bioplastics (16), detergents (17), novel materials (18), cosmetics (19) dietary supplements (20), food processing (21), bioremediation (22), biopesticides (23) and biofertilizers (24). While model bacterial like *Escherichia coli* and *Bacillus subtilis* are commonly used in biomanufacturing due to their well-understood physiology and rapid growth, their inefficiency for certain products highlights the potential of non-model microorganisms, which offer unique metabolic traits, diverse genetic backgrounds, and robustness in extreme conditions (25,26). Better visibility into entire species through pangenomes and their derivatives should help bioengineers search for strains with specific capabilities (27), select optimal strains for valuable functions, and better understand the sequence solution space for genes and their feasible variations (28–30). To further harness the potential of microbes in industrial applications, there is a need for a comprehensive resource to investigate microbial functions within the context of pangenomes.

Several microbial pangenome databases have been developed to facilitate knowledge extraction from microbial pangenomes. panX (31) combines an automated pipeline for pangenome analysis with pre-computed pangenomes, which emphasizes in-depth phylogenetic analysis of orthologous gene clusters and allows researchers to investigate the evolutionary history of genes and potential horizontal gene transfer events. MetaRef (32) focuses on storing, visualizing, and exploring pangenome analysis results. ProPan (33) specializes in data mining pangenome dynamics to provide insights into metabolism and antimicrobial resistance across prokaryotic species. These databases are valuable resources for leveraging microbial genome data. However, the lack of an efficient global search function in these databases prevents users from querying information on genes, pathways, functions or other information of interest across species and families, thereby restricting their utility. Additionally, they overlook the importance of allele variants, which provide deeper insights into species evolution and genetic diversity.

Scientific databases typically represent experimental results leading to scientific knowledge, and sometimes link to publications that contain insights on their data. These publications provide critical context for data interpretation, experimental design and hypothesis development. Recent research in microbial pangenomics has generated a substantial amount of literature (34), presenting a unique opportunity for large-scale literature mining. However, extracting knowledge from scientific literature has traditionally required time-consuming manual review. Large language models (LLMs), a branch of natural language processing, have proven valuable for automated information extraction, aggregation and summarization (35–39), offering a potential solution for rapid knowledge extraction. However, LLMs cannot be updated with new content without costly retraining, and their context window length limit prevents the input of an entire collection of literature. Retrieval-augmented generation (RAG) offers a solution to this challenge by automatically consolidating relevant literature content into an LLM query, improving response accuracy, and reducing hallucinations (40–42). Modern databases can benefit from the integration of a RAG-LLM system, enabling users to efficiently query a topic's literature while interacting with the database. At the moment, few public databases are leveraging the newly available potential of LLMs.

To address these needs, we present PanKB, the Pangenome Knowledgebase (pankb.org), a comprehensive web tool with modern interactive pangenomic analytics. PanKB focuses on industrially relevant microbial families and species. It currently encompasses pangenomes of 51 species from 8 selected industrially important families, comprising 8402 genomes and over 500 000 genes, with plans for continued expansion. The platform's interactive analytics facilitate exploration, intuition, and potential discoveries, adding value beyond the static figures in recent pangenomic publications (5,8). PanKB features a global search for rapid navigation of pangenomic entities (genes, pathways, functions, etc) across species and families, and enables dataset export for custom analyses. In addition to pangenomic analytics, PanKB offers alleleomic analytics that describe the genetic variants within a pangenome, encompassing over 7 million mutations. This provides deeper insights into intra-species sequence variations beyond gene presence/absence (28) and demonstrates unique value in narrowing the solution search space for feasible genetic variants (43). Combined, this accessible platform empowers strain engineers to leverage microbial functions beyond model organisms, enabling valuable workflows for enzyme and strain engineering. These include identifying genes for new enzyme production or reintroduction into strains, pinpointing precise gene edits to modify activity, discovering and optimizing valuable pathways and selecting optimal starting strains. Additionally, to facilitate literature mining, PanKB incorporates a curated bibliome of 833 open-access pangenomics articles and integrates a RAG-enhanced LLM interface (AI Assistant). This AI-powered system responds to user queries using the bibliome's pangenomic knowledge, provides response source references, and minimizes response hallucinations. In sum, PanKB provides an integrated platform for comprehensive pangenomic analysis, facilitating microbial research and application through interactive tools, bioengineering workflows and AI-assisted knowledge extraction.

Table 1. Data acquisition and curation of PanKB[1]

Microbial family	Data retrieval date	Number of genomes retrieved from NCBI	Number of genomes in PanKB
Enterobacteriaceae	2023–10-17	3299	3224
Lactobacillaceae	2022–01-15	4783	2447
Bacillaceae	2022–10-19	1681	1353
Mycobacteriaceae	2023–08-08	1986	884
Pseudomonadaceae	2023–08-23	512	288
Streptomycetaceae	2023–06-30	2371	176
Anoxybacillaceae	2023–12-18	37	16
Burkholderiaceae	2023–12-19	207	14

[1]This table presents the genome retrieval date, the total number of retrieved genomes from NCBI and the number of genomes after QC for each microbial family.

Materials and methods

Data collection and quality control

The PanKB pangenome collection comprises a curated subset of industrially important microbial families and species (44–46). This initial collection was manually selected to prioritize organisms of industrial relevance, rather than including all highly sequenced or pathogenic species. All genome data in PanKB was retrieved from the NCBI database (47). The BGCflow pipeline (48) was used for data quality control (QC) and pangenome construction.

The general PanKB data retrieval and QC process consisted of the following steps:

1) Genomic sequences for selected microbial families were retrieved from the NCBI RefSeq across all assembly levels (contig, scaffold, chromosome and complete).
2) All retrieved genomes were re-annotated for taxonomy using the Genome Taxonomy Database Toolkit (GTDB-Tk) (49).
3) Genome QC: Number of contigs <200, Completeness >95%, Contamination <5, N50 (contigs) >50 000.
4) For each microbial family, species with at least 30 genomes were selected for downstream pangenome construction.

The retrieval date and the number of genomes before and after QC for each microbial family are summarized in Table 1. Most microbial families in PanKB followed these data selection and QC steps. However, for specific research needs, customized selections were made for three families:

Enterobacteriaceae: Only *Escherichia coli* was selected. Due to the abundance of publicly available *E. coli* genomes, only high-quality genome assemblies (chromosome and complete assembly levels, which represent nearly complete or fully assembled genomes) were used to construct the *E. coli* pangenome.

Anoxybacillaceae: Only Parageobacillus thermoglucosidasius was selected. Due to its industrial importance, we selected it despite there being fewer than 30 publicly available genomes.

Burkholderiaceae: Only *Cupriavidus necator* was selected. Due to its industrial importance, we selected it despite there being fewer than 30 publicly available genomes.

Pangenome construction and alleleome analysis

We used BGCFlow (48) to annotate genomes and construct pangenomes. The calculation of pangenomes' openness (Heap's Law) is implemented with code from previous research (5,8), and the code is available on the Github repo (https://github.com/biosustain/pankb_data_prep). Alleleome analysis methodology is based on previous studies (28,29) and the code is available on the Github repo (https://github.com/biosustain/Alleleome).

Website and database implementation

The PanKB website is a scalable web project built using the microservices architecture, which means it consists of several independent applications deployed as separate services. It consists of two web applications (the PanKB website and AI Assistant), two databases, and three pipelines (Supplementary Figure S1).

The front-end (user interface) part of the PanKB website is implemented in HTML, CSS, and JavaScript. On the back-end, the PanKB website is written in Python 3.8. Django 3.0.8 (https://www.djangoproject.com) is used as a Python web framework. The website is connected to a NoSQL database that stores information about pangenomes. This data is primarily contained in the PanKB website tables and used to generate dynamic content (e.g. search results). Static data used to generate the PanKB plots and diagrams is stored on Azure Blob Storage (https://azure.microsoft.com/en-us/products/storage/blobs) primarily in JSON, CSV, Newick (for phylogenetic trees) and FASTA (for multiple sequence alignment (MSA) plots) files.

The PanKB AI Assistant is implemented as a separate web application using Streamlit 1.35.0 (https://streamlit.io/) and LangChain 0.2.1 (https://www.langchain.com/). The PanKB AI Assistant web application is connected to a vector database. The data from the vector database are retrieved every time the chatbot is asked a question. The Vector Database creation pipeline populates the vector database.

The PanKB website and vector database are deployed on Azure Cosmos DB for MongoDB vCore (https://learn.microsoft.com/en-us/azure/cosmos-db/mongodb/vcore/) cluster. The M40 tier is chosen due to its support of the NHSW vector index (https://devblogs.microsoft.com/cosmosdb/introducing-vcore-based-azure-cosmos-db-for-mongodb-latest-ai-features/). The PanKB website database is populated using output written by the PanKB pipeline to the Microsoft Azure Blob Storage.

All the web applications and pipelines are executed and deployed on Microsoft Azure Virtual Machines with Ubuntu 20.04 as the operating system (https://azure.microsoft.com/en-us/products/virtual-machines/linux) and docker, docker-compose and git tools installed. Additionally, the inputs for all the data processing pipelines are downloaded and stored on these virtual machines.

Development of PanKB LLM

A RAG-LLM system consists of two elements: an external vector database (pangenomic papers database) and a pre-trained LLM.

Collection and pre-process of microbial pangenomic papers

PanKB features a bibliome of 833 open-access papers on the microbial pangenomic domain, which forms the basis of its

pangenomic papers database (a vector database). The paper list was retrieved using the following boolean query: ((microbial pangenome) OR (bacteria pangenome)) OR (prokaryotic pangenome)) NOT (Eukaryotic pangenome) on PubMed (https://pubmed.ncbi.nlm.nih.gov/). A total of 877 papers were retrieved on April 10, 2024. After excluding 16 non-open-access papers and 28 preprints, plain text files of the rest 833 papers were obtained through publisher-provided application programming interfaces (APIs) or manually scraped with permission. Subsequently, these files were then processed to retain only the main body of each paper, including the 'Abstract', 'Introduction', 'Materials and Methods', 'Results' and 'Discussion/Conclusion' sections.

Pangenomic papers database development

The processed plain text files were divided into chunks using RecursiveCharacterTextSplitter implemented in LangChain 0.2.1. The voyage-large-2-instruct embedding model (https://www.voyageai.com/) was used to convert all chunks into vector representations stored in an Azure Cosmos DB, forming the final pangenomic papers database.

Model selection

Seven pre-trained LLMs were selected as candidates for the base model of PanKB LLM: two OpenAI models (GPT-4 Turbo, and GPT-4o), one Anthropic model (Claude 3 Opus), two Google models (Gemini 1.5 flash and Gemini 1.5 pro) and two open-source models (Llama 3 70B and Mixtral 8 × 22B). At the time of writing this paper, GPT-4 Turbo, GPT-4o, Claude 3 Opus, Gemini 1.5 pro, and Gemini 1.5 Flash are among the top 10 LLMs determined by the LMSYS Chatbot Arena Leaderboard (https://chat.lmsys.org/?leaderboard), an open platform using over 1 000 000 human pairwise comparisons to rank LLMs. Llama 3 70B and Mixtral 8 × 22B are two of the top open-source LLMs determined by the Open LLM Leaderboard (https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard), an open-source LLM evaluation platform developed by huggingface.co, where the LLMs are evaluated by six key benchmarks using the Eleuther AI Language Model Evaluation Harness framework.

The seven selected LLMs were subsequently integrated with the vector database using the LangChain framework, and seven RAG-LLM systems were constructed for evaluation.

Model evaluation

Due to the absence of established benchmarks in this specific domain, a microbial pangenome exam was conducted to evaluate the knowledge extraction capabilities of the seven selected LLMs and their corresponding RAG-LLM systems. The exam used a set of 50 microbial pangenome objective questions generated by the authors. These questions were carefully created based on knowledge extracted from relevant articles in the field (5,9–11,48,50–64). For each question, the original text segment of the paper containing the answer was retained as the standard answer for evaluation.

To minimize the model randomness, during the evaluation, both inference parameters 'temperature' and 'top_p' were set to 0. Additionally, to constrain the LLM to answer questions only based on the provided context (content from the microbial pangenomic papers), the system prompt for the seven RAG-LLM systems was configured as follows:

"""You are PangenomeLLM. You are a cautious assistant proficient in microbial pangenomics. Use the following pieces of context to answer user's questions.
Please check the information of context carefully and do not use information that is not relevant to the question.
If the retrieved context doesn't provide useful information to answer user's question, just say that you don't know.
Please give a clear and concise answer.
Question: {question}
Context: {context}
Answer:"""

The system prompt of the seven selected LLMs (base models, without integrating RAG) was configured as follows:

"""You are PangenomeLLM. You are a cautious assistant proficient in microbial pangenomics.
Please answer user's questions.
If you don't know the answer, just say that you don't know.
Please give a clear and concise answer.
Question: {question}
Answer:"""

Each question was asked three times, and the answers of base models and their corresponding RAG-LLM systems were classified into four categories: (1) Correct. All three answers were correct. (2) Partially Correct. If any of the three answers included incorrect information, i.e. the answers were mixed with correct and incorrect information, then it was marked as 'Partially Correct'. (3) Rejection. The RAG-LLM systems rejected to answer the question, with responses similar to 'Sorry, I don't know'. (4) Incorrect. All three answers are incorrect. The correct rate of 50 questions was used to measure the accuracy of the RAG-LLM systems in the exam.

The complete question set is available in the Supplementary Table S1. The evaluation method above and the question set are inspired by the evaluation method of knowledge recall in another domain-specific LLM (65). While these questions are carefully designed, this examination may not fully cover the scope of knowledge within the microbial pangenomics domain.

Results

The initial release of PanKB incorporates pangenomes of 51 species selected from 8 industrially important microbial families, comprising 8402 genomes, over 500 000 genes, and more than 7M amino acid (AA) mutations (Table 2). Among the eight bacterial families, *Enterobacteriaceae* contains only the *Escherichia coli* species, but has the largest number of genomes of all species within the database (3224). *Lactobacillaceae* exhibits the greatest diversity, encompassing 26 species with approximately 200 000 genes and over 2 million amino acid mutations.

Pangenome analysis dashboard: comprehensive and interactive analytics facilitating pangenome exploration

A key value of PanKB is that it provides modern, comprehensive and interactive analytics for efficient exploration of microbial pangenomes. To explore a species's pangenome analysis, users first select a species from the Organisms page (Figure 1B), which navigates them to a Pangenome Analysis Dashboard (Figure 1B–J). The Pangenome Analysis Dashboard

Table 2. Statistic of the size of the dataset in PanKB[1]

	Species	Genomes	Genes	AA Mutations
Enterobacteriaceae	1	3224	100282	1117235
Lactobacillaceae	26	2447	192998	2698553
Bacillaceae	4	1353	61844	874545
Mycobacteriaceae	10	884	50168	758272
Pseudomonadaceae	5	288	81726	827696
Streptomycetaceae	3	176	51748	952479
Anoxybacillaceae	1	16	7111	75153
Burkholderiaceae	1	14	22582	407964
Total	**51**	**8402**	**568 459**	**7 711 897**

[1]This table presents the number of species, genomes, genes and coding alleleomes for each of the eight bacterial families in PanKB.

presents a complete set of standard pangenomic analytics, distributed across three interconnected pages: Overview (Figure 1B–J), Genes (Figure 1K), and Phylogenetic Tree (Figure 1L). Each page features a left sidebar that includes a navigation panel (Figure 1B) for switching between these pages, and an information panel (Figure 1C) describing basic features of the pangenome (species, number of genomes, etc).

As the first page of the Pangenome Analysis Dashboard, the Overview page summarizes the primary characteristics of a pangenome. Four figures at the top illustrate essential pangenome characteristics: openness (Figure 1B), gene frequency distribution (Figures 1C, D) and COG annotation distribution (Figure 1E), elucidating the gene composition and diversity of a species. Below these, An interactive heatmap (Figure 1H) visualizes the gene presence/absence matrix across MASH-based phylogroups, enabling researchers to identify patterns of gene gain and loss across lineages, thereby contributing to the characterization of species evolution trajectories. Additionally, a histogram (Figure 1I) and a scatter plot (Figure 1J) present the allelome analysis, a pangenomic-scale analysis of gene variants, providing insights into an alleleome's conservation and the evolutionary forces shaping the pangenome (28).

Annotating the genes of a pangenome offers deeper insights into species evolution, niche adaption, and characterization of microbial phenotypes (5,8). The Genes page (Figure 1K) contains an interactive table with comprehensive annotations on genes, including COG and PFAM annotations. Additionally, the table features column-specific search bars and sortable columns, facilitating in-depth exploration of genetic and functional diversity within a species.

The Phylogenetic Tree page (Figure 1L) displays an interactive phylogenetic tree based on MASH distances, illustrating genomic similarity among strains. A toolbar above the tree allows users to toggle between linear and radial layouts, zoom and select specific branches. This interactive design enables detailed exploration of the species' phylogenetic structure, revealing insights into its genetic diversity and evolutionary history.

Gene pages: comprehensive allele analysis with sequence, pathway and variant data

Genes in pangenomes provide broad insights into species-level genetic diversity, revealing the distribution of functional categories within a species. Alleles, representing sequence polymorphisms within individual genes, offer higher-resolution pictures of genetic diversity. Comprehensive characterization of alleles reveals nucleotide and amino acid substitutions, potential functional modifications and strain-specific genomic adaptations, providing insights for microbial research and applications (28,43). To maximize the utilization of pangenomic data, PanKB features dedicated gene pages, presenting detailed information on all gene alleles.

Users can access a Gene page (Figure 2A-C) by selecting a gene from the Genes page (Figure 1K). An interactive table with searching and sorting functions (Figure 2A) presents locus tags, genome IDs, protein annotations, sequences and pathway associations, enabling rapid assessment of alleles and their functional potential. The integrated Pathway Info Page (Figure 2D), accessed via the 'Pathways' column, offers essential metabolic context of a gene by connecting related alleles to the KEGG Pathway Database (66) and displaying related genes and products below the table, a dot plot (Figure 2B) illustrates the frequency of dominant and variant AAs at each position in the gene, providing an overview of variant locations and frequencies. Following the dot plot, an AA MSA plot (Figure 2C) provides a detailed, multi-faceted view of sequence conservation and variation.

Global search and bioengineering workflows: how to advance biotechnology with pangenomics

Another key feature of PanKB is the global search function. Unlike conventional pangenome databases that typically limit searches within individual species or a specific taxonomy group, PanKB global search enables users to search for genes, pathways, products and other information across species or families. This greatly facilitates the navigation and utilization of microbial pangenomics data. The PanKB global search bar (Figure 3A) is available in each page's navigation bar except for the AI Assistant page. To demonstrate how to leverage PanKB data for biotechnology two examples are provided: gene engineering and pathway engineering workflows.

The PanKB gene engineering workflow (Figure 3B) begins with defining the target function or enzyme, followed by searching for it using the global search bar. Lactate dehydrogenase (LDH) is used as an example to demonstrate the workflow. Lactic acid (LA) is a valuable compound with widespread applications in food, pharmaceuticals, and biodegradable plastics (67–73), and LDH is a crucial enzyme in the bacterial LA synthesis pathway (74). The resulting gene table (Figure 3C) displays all genes related to the query 'Lactate dehydrogenase', including gene source (species), gene category (core/accessory/rare), gene annotations and other information. Users can review the retrieved genes and extract whole allele sequences or individual mutations for reintroduction toward gene optimization (Figure 3D).

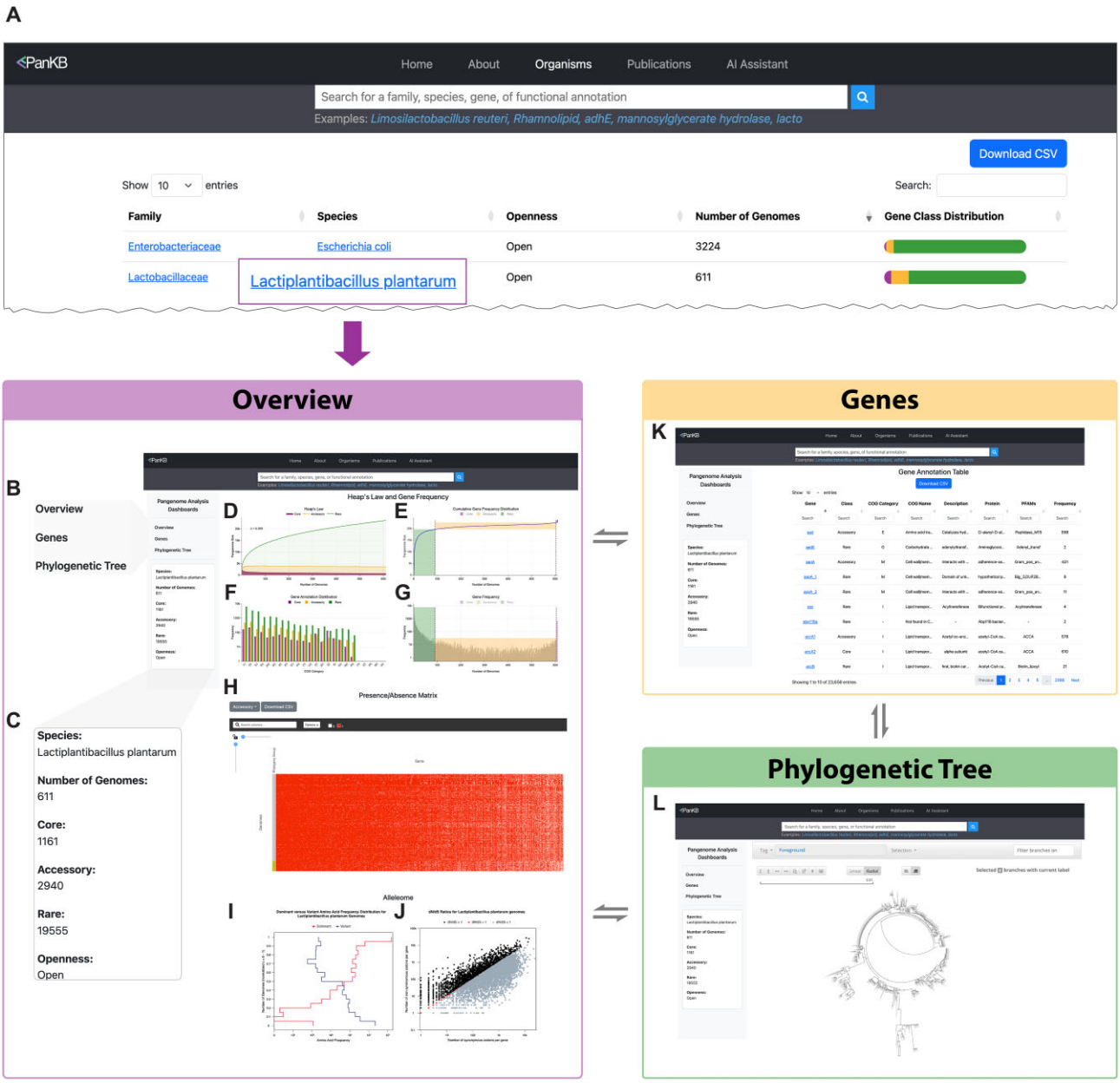


Figure 1. Screenshots of the Organisms page (**A**) and Pangenome Analysis Dashboard of Lactiplantibacillus plantarum (**B–L**). (**A**) Organisms page. This screenshot displays the top 2 pangenomes with the highest number of genomes. The Organisms page can be accessed via 'Select an organism' at the bottom of the home page or 'Organisms' in the navigation bar. (**B**) Links in the left sidebar of the Pangenome Analysis Dashboard, enabling users to switch between the Overview, Genes, and Phylogenetic Tree pages. (**C**) Species information panel displaying basic information about the pangenome, including species name, number of genomes, gene counts in three categories and pangenome openness. (**D**) Heaps plot. Three curves are shown, representing the core, accessory and rare genes. (**E**) Cumulative gene frequency distribution curve. The curve is divided into three parts, representing the core, accessory and rare genes. Dashed lines indicate the cutoff of gene categories. (**F**) Bar chart illustrating the COG annotation distribution across core, accessory and rare genes. (**G**) Histogram depicting the gene frequency distribution. (**H**) Heatmap representing the gene presence/absence matrix (PAM). Rows denote genomes, and columns represent genes. The annotation bar shows the phylogroup of the genomes based on MASH distance. By default, the heatmap displays core genes, and users can explore accessory and rare genes via the 'Select Gene Class' button at the top left of the heatmap. (**I**) A consolidated histogram illustrating frequencies of the dominant and variant AA in all AA positions of *Lactiplantibacillus plantarum* genes comprising the '*L. plantarum* alleleome'. The number of genomes (Y-axis) is normalized from 0 to 1. (**J**) A scatter plot showing the ratio of non-synonymous to synonymous codon substitutions (dN/dS) for each gene within a species. (**K**) A Genes page containing an interactive gene annotation table that includes all genes of Lactiplantibacillus plantarum. (**L**) The Phylogenetic Tree page. This page contains an interactive phylogenetic tree built on MASH distance.

Figure 2. Screenshots of Gene page (**A–C**) and Pathway Info Page (**D**). (**A**) Table of accA2 gene alleles, showing locus tags, genome IDs, protein annotations, sequences, and related pathways. Clicking on the links in the 'Pathways' column navigates users to corresponding Pathway Info Pages. (**B**) Dot plot showing the variant locations and frequencies. (**C**) MSA plot displaying allele alignments. (**D**) Pathway Info Page presents details of a related pathway, including pathway ID, a link to the original entry in the KEGG Pathway Database, species, strain, a list of genes from the same strain that is related to the same pathway and pathway products.

Similarly, the pathway engineering workflow (Figure 3D) begins with defining a target pathway or product. Tyrosine, an aromatic amino acid, is widely used as a dietary supplement (75) and serves as a valuable precursor for various pharmaceutical applications, such as the production of L-DOPA, an important medication for treating Parkinson's disease (75–77). The pathway table (Figure 3E) presents all pathways related to the query 'Tyrosine'. Each pathway entry contains pathway IDs, names, strain and species information, related genes and products. Selecting a pathway of interest (column 'Pathway ID') navigates users to its Pathway Info Pages (Figure 3G), where the related gene set of the pathway (highlighted in green) can be extracted for reintroduction. Alternatively, users can extract gene set variants (Figure 3H) to optimize the pathway.

PanKB LLM: automating knowledge extraction from pangenomic literature

Over the past 20 years, the number of pangenomic publications has been steadily growing (Figure 4A), offering a unique opportunity for large-scale literature mining in the microbial pangenome domain. An LLM-powered AI Assistant (PanKB LLM) was developed to automate knowledge extraction from a collection of 833 open-access microbial pangenomic papers. All papers are accessible through the 'Publications' link in the PanKB navigation bar.

LLMs have demonstrated remarkable capabilities in text mining. However, they face limitations in domain-specific tasks (78) and may generate fabricated information, a phenomenon known as 'hallucination' (79). Although retraining or fine-tuning a model can mitigate these issues, such ap-

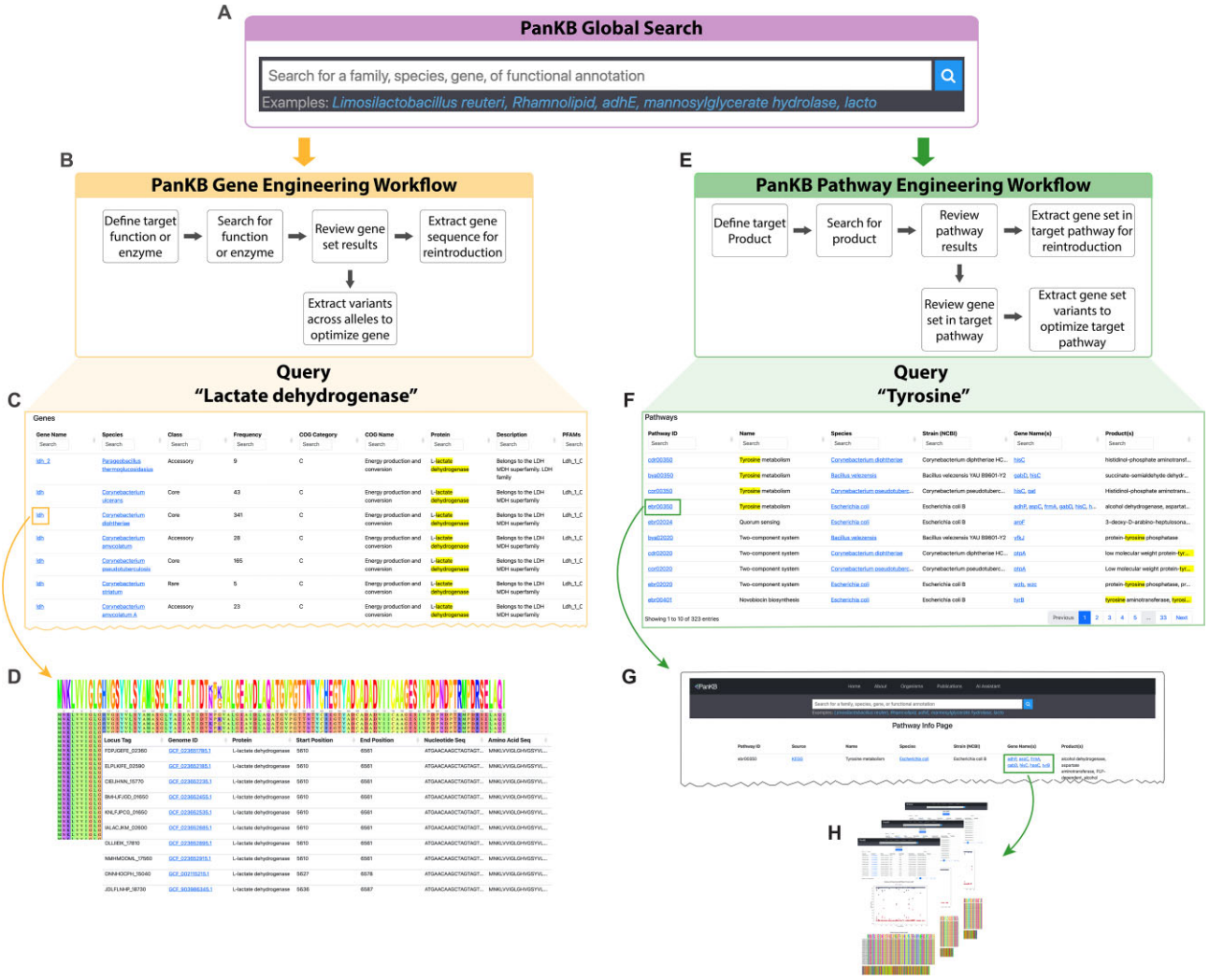


Figure 3. PanKB global search and two bioengineering workflows based on that. (**A**) The PanKB global search bar. (**B**) PanKB gene engineering workflow (**C**) 'Genes' results for querying 'Lactate dehydrogenase' using PanKB global search (**D**) The MSA plot and the allele table for a selected gene from the Genes' results (**E**) PanKB pathway engineering workflow (**F**) The 'Pathways' results of querying 'Tyrosine' using PanKB global search. (**G**) The Pathway Info Page for a selected pathway from the 'Pathways' results. (**H**) Gene pages of genes for a selected pathway.

proaches are costly. RAG is considered an effective and cost-efficient alternative. RAG can reduce hallucinations and enhance LLMs by retrieving query-relevant documents from external databases and incorporating them as contextual input (40). Additionally, RAG can maintain the LLM with up-to-date information through periodically updating the external databases. Importantly, documents retrieved by RAG are traceable; therefore, it can provide references to all documents used to generate responses.

PanKB LLM is implemented as an extension of an available LLM service using a RAG framework. The PanKB LLM question-answer process comprises five components: User Query, Pangenome Papers Vector Database, Enhanced Prompt, Constrained LLM and Response. The workflow proceeds as follows (Figure 4B):

1. User Query Vectorization: Upon receiving a user query, the RAG-LLM system transformed it into a numerical vector representation using a text embedding model.
2. Retrieval of Relevant Documents: This step involved searching for potentially relevant documents for the user query within the pangenome papers vector database. To build this database, we collected 833 pangenome papers in plain text format and segmented them into text chunks of 500 characters each, referred to as documents. Each document was vectorized using the same text embedding model as in Step 1, and its vector representations were stored in the vector database. The vectorized query was matched against this database, and each document was assigned a similarity score between 0 and 1 based on the distance between its vector representation and the query vector—the smaller the distance, the higher the similarity score. Documents close to the query in the vector space indicated potential relevance. The top 20 documents with the highest similarity scores were selected, excluding those with similarity scores below 0.5. These documents were considered potentially relevant to the user's query and were retrieved.
3. Synthesis of the Enhanced Prompt: The system prompt, original user query and the retrieved documents from Step 2 were synthesized to form the enhanced prompt. The system prompt (detailed under the Method—Model

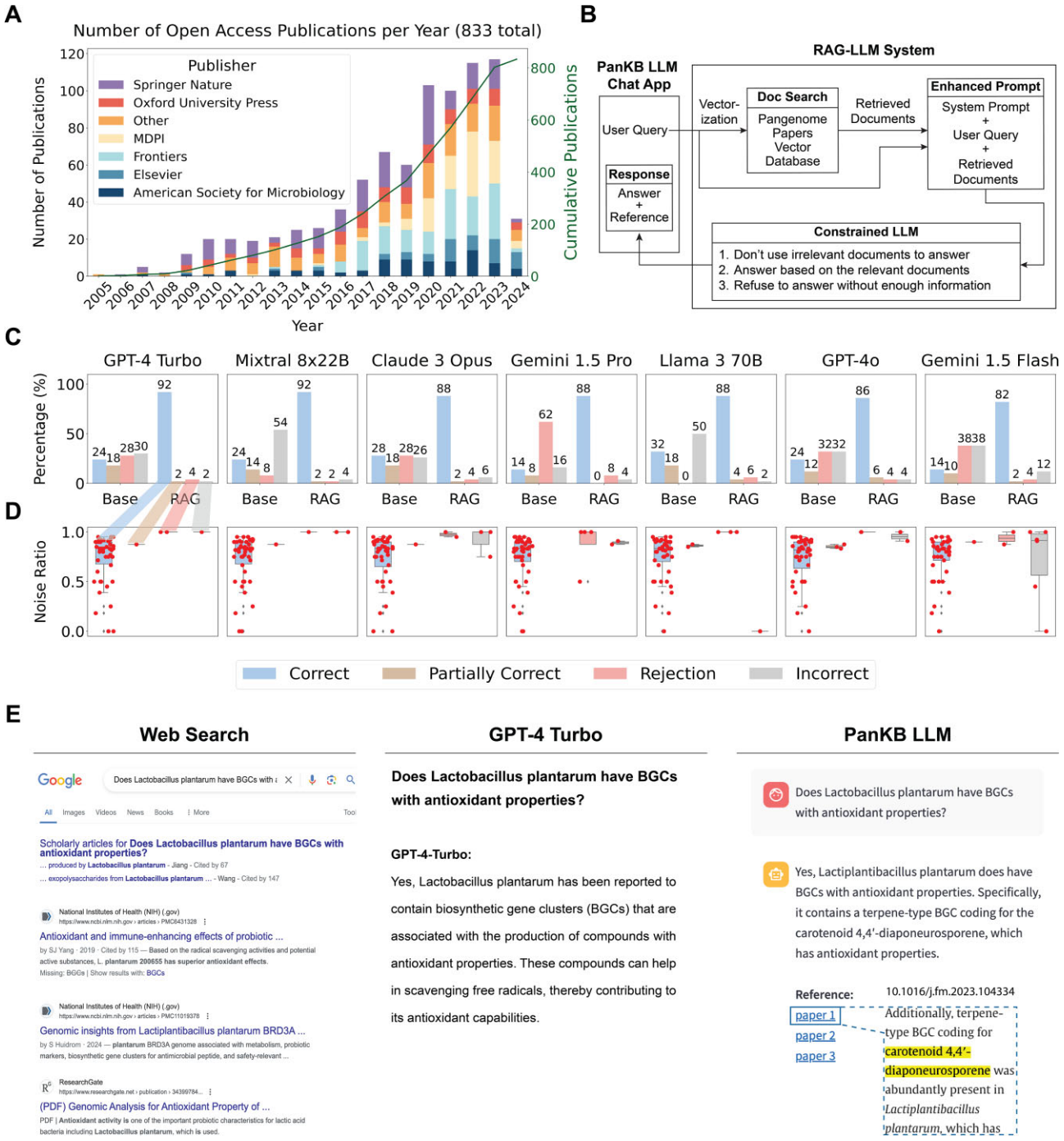


Figure 4. Overview of PanKB LLM and its evaluation. (**A**) Distribution of articles in the vector database by publication year (2005–2024) and publishers. The superimposed line depicts the cumulative number of publications across all represented publishers over time. (**B**) Workflow of PanKB LLM. The PanKB LLM question-answer process comprises five components: User Query, Pangenome Paper Vector Database, Enhanced Prompt, Constrained LLM and Response. (**C**) Performance comparison of seven LLMs (Base) and their corresponding RAG-LLM systems (RAG). Answers are categorized as Correct, Partially Correct, Rejection or Incorrect. (**D**) Noise ratio of all questions, and their distribution across four answer categories in the seven RAG-LLM systems. The order of the boxes, from left to right: Correct, Partially Correct, Rejection or Incorrect. (**E**) Comparison of query responses regarding antioxidant biosynthetic gene clusters (BGCs) in Lactobacillus plantarum using web search, GPT-4 Turbo (base model) and PanKB LLM (GPT-4 Turbo RAG system).

evaluation section) constrained the LLM's behavior by instructing it to:
   a. Determine whether the retrieved documents were relevant to the user's query and avoid using irrelevant documents.
   b. Answer based on the relevant documents.
   c. Refuse to answer if there were no relevant documents/insufficient information.
4. Response Generation by the Constrained LLM: The enhanced prompt was input into the LLM to generate a response.
5. Returning the Response to the User: The generated response was returned to the user and consisted of two parts:
   a. The LLM's answer.
   b. A list of papers from which the retrieved documents originated. Regardless of whether the LLM used the retrieved documents from Step 2 in its answer, their corresponding papers were always provided to the user.

Evaluation of PanKB LLM

The RAG-LLM system comprises two main components: an external database (pangenomic papers database) and a pretrained LLM. The development of the pangenomic papers database and the selection of LLMs are detailed in the 'Materials and methods' section. By integrating the pangenomic papers database with selected LLMs using the LangChain framework, seven RAG-LLM systems were constructed for evaluation.

The performance of the seven RAG-LLM systems was evaluated using a set of 50 paper-specific microbial pangenome questions simulating user queries. The evaluation methodology, inference parameters and system prompt are detailed in the 'Materials and Methods'. The performance comparison between the seven RAG-LLM systems and their corresponding base models is presented in Figure 4C. All base models performed poorly, with an average accuracy of 22.4%. In contrast, their corresponding RAG-LLM systems achieved an average accuracy of 88%. Among all RAG-LLM systems, GPT-4 Turbo and Mixtral 8 × 22B RAG systems outperformed the other RAG-LLM systems, achieving a 92% accuracy on the question set.

Despite sharing a common RAG framework, the seven RAG-LLM systems demonstrated varying performance on the same question set. This variability likely stems from differences in the LLMs' noise robustness. The RAG system is not perfect and can introduce noise. For example, some RAG-retrieved contexts (quantified as documents) are related to the query but lack the necessary information for generating correct answers (irrelevant documents). An effective LLM should be robust to noise, which is the ability to distinguish between relevant and irrelevant information for the given context. The noise ratio is defined as the ratio of irrelevant documents to the total number of retrieved documents for a given query, ranging from 0 to 1. The noise ratio distribution across four response categories (Correct, Partially Correct, Rejection and Incorrect) for seven RAG-LLM systems (Figure 4D) illustrates their respective noise robustness. All seven RAG-LLM systems exhibited good noise robustness, achieving an average accuracy of 88% with mean noise of 75% across 50 questions. Notably, GPT-4 Turbo and Mixtral 8 × 22B demonstrated exceptional noise robustness, achieving the highest accuracy of 92%.

When confronted with completely noisy contexts (noise ratio = 1), an ideal LLM should refuse to answer the question due to insufficient information, a behavior known as negative rejection (80). In this aspect, Llama 3 70B exhibited optimal negative rejection, refusing to answer all three completely noisy-context questions. GPT-4 Turbo also performed well, declining to provide answers for two out of three such questions. Overall, the GPT-4 Turbo demonstrated superior performance and was selected as the final base model for PanKB LLM.

Case study: comparison among web search, GPT-4 Turbo and PanKB LLM

RAG-LLM systems demonstrate high accuracy for paper-specific questions, highlighting their effectiveness in rapidly extracting knowledge from domain-specific articles. A comparison (Figure 4E) was conducted using the Google search engine, GPT-4 Turbo, and PanKB LLM, querying each on a specific microbial pangenomics question:

> "Does Lactobacillus plantarum have BGCs with antioxidant properties?"

The results generated by the traditional web search are mostly links to relevant papers, requiring further reviewing by users. GPT-4 Turbo offers a direct affirmative response, confirming the presence of BGCs with antioxidant properties in L. plantarum, but lacks specificity and citations. In contrast, PanKB LLM provides a concise answer with specific details, identifying a terpene-type BGC in L. plantarum that codes for the antioxidant carotenoid 4,4′-diaponeurosporene (5). Notably, PanKB LLM's response also includes references, enhancing the credibility of its response and allowing for traceability of the information to primary sources.

Other major features: about and data download

An informative 'About' page is available for researchers new to PanKB, presenting an overview of all features, guidance on navigating the PanKB analysis dashboard, and details on the tools used in its creation. Additionally, contact information is provided for users to submit feedback or requests.

PanKB allows users to download results from data tables for custom analysis. Data download can be initiated through the download button located by each data table. Available datasets include the species table from the Organisms page, presence/absence matrices from individual species' Pangenome Analysis Dashboards, gene annotation tables from the Genes page and gene tables from individual gene pages.

Discussion

PanKB is a comprehensive microbial pangenome knowledgebase concentrating on industrially relevant species. It integrates interactive pangenomic analytics, enables bioengineering workflows, and implements AI-assisted knowledge extraction to facilitate microbial research and applications. In addition to providing comprehensive pangenome analytics, PanKB includes alleleome analysis, a pangenome-scale analysis of genetic variants. The alleleome is available for each gene, offering the unique value of narrowing the solution search

space for feasible genetic variants. Unlike existing pangenome databases, PanKB's global search function enables users to query genes, pathways, functions, products and other information of interest across species and families, enhancing the utility of pangenomic data. These features collectively support valuable workflows for enzyme and strain engineering, such as identifying genes for novel enzyme production or strain reintroduction, pinpointing precise gene edits to modify activity, optimizing valuable pathways and selecting optimal starting strains. To further leverage pangenomic knowledge, PanKB also integrates an LLM-powered chatbot for automated knowledge extraction from a large collection of pangenomic papers. Taken together, PanKB's features are expected to be a unique resource that bridges the gap between pangenomic data and practical applications.

Data availability

PanKB is freely available at http://pankb.org and can be accessed with a JavaScript-enabled web browser.

The source code of all the PanKB components is open and available on GitHub with detailed development and deployment instructions included: The PanKB website: https://github.com/biosustain/pankb; The PanKB website database: https://github.com/biosustain/pankb_db; The PanKB AI Assistant application: https://github.com/biosustain/pankb_llm; The pangenomes data generation and processing: https://github.com/NBChub/bgcflow, https://github.com/biosustain/pankb_data_prep; Alleleome analysis: https://github.com/biosustain/Alleleome.

Their corresponding DOIs of Zenodo: The PanKB website: https://doi.org/10.5281/zenodo.13939209; The PanKB website database: https://doi.org/10.5281/zenodo.13939155; The PanKB AI Assistant application: https://doi.org/10.5281/zenodo.13939177; The pangenomes data generation and processing: https://doi.org/10.5281/zenodo.13939352, https://doi.org/10.5281/zenodo.13941149; Alleleome analysis: https://doi.org/10.5281/zenodo.13939283.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors gratefully acknowledge Matin Nuhamunada for their technical support.

Author contributions: Conceptualization: B.O.P., P.V.P., A.S.D. and D.Z.; Implementation: L.P., B.S. and P.A.P.; Data processing and generation: P.A.P., A.S.H., B.S. and O.S.M.; Writing: B.S., L.P. and P.V.P.; Supervision: B.O.P., P.V.P. and A.S.D.

Funding

Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (NNF Grant Number NNF20CC0035580).

Conflict of interest statement

The authors declare no competing financial interest.

References

1. Land,M., Hauser,L., Jun,S.-R., Nookaew,I., Leuze,M.R., Ahn,T.-H., Karpinets,T., Lund,O., Kora,G., Wassenaar,T., *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.
2. Hunt,M., Lima,L., Shen,W., Lees,J. and Iqbal,Z. (2024) AllTheBacteria - all bacterial genomes assembled, available and searchable. bioRxiv doi: https://doi.org/10.1101/2024.03.08.584059, 15 November 2024, preprint: not peer reviewed.
3. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
4. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S., *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome.". *Proc. Natl. Acad. Sci.*, **102**, 13950–13955.
5. Rajput,A., Chauhan,S.M., Mohite,O.S., Hyun,J.C., Ardalani,O., Jahn,L.J., Sommer,M.O.A. and Palsson,B.O. (2023) Pangenome analysis reveals the genetic basis for taxonomic classification of the Lactobacillaceae family. *Food Microbiol.*, **115**, 104334.
6. Wood,S., Zhu,K., Surujon,D., Rosconi,F., Ortiz-Marquez,J.C. and van Opijnen,T. (2020) A pangenomic perspective on the emergence, maintenance, and predictability of antibiotic resistance. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing, Cham, pp. 169–202.
7. Innamorati,K.A., Earl,J.P., Aggarwal,S.D., Ehrlich,G.D. and Hiller,N.L. (2020) The bacterial guide to designing a diversified gene portfolio. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing, Cham, pp. 51–87.
8. Hyun,J.C., Monk,J.M. and Palsson,B.O. (2022) Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, **23**, 7.
9. Zhong,C., Han,M., Yu,S., Yang,P., Li,H. and Ning,K. (2018) Pan-genome analyses of 24 Shewanella strains re-emphasize the diversification of their functions yet evolutionary dynamics of metal-reducing pathway. *Biotechnol. Biofuels*, **11**, 193.
10. Samanta,D., Rauniyar,S., Saxena,P. and Sani,R.K. (2024) From genome to evolution: investigating type II methylotrophs using a pangenomic analysis. *Msystems*, **9**, e00248-24.
11. Liu,Y., Pei,T., Du,J., Yao,Q., Deng,M.-R. and Zhu,H. (2022) Comparative genomics reveals genetic diversity and metabolic potentials of the genus qipengyuania and suggests fifteen novel species. *Microbiol. Spectr*, **10**, e01264-21.
12. Genilloud,O. (2014) The re-emerging role of microbial natural products in antibiotic discovery. *Antonie Van Leeuwenhoek*, **106**, 173–188.
13. Goeddel,D.V., Kleid,D.G., Bolivar,F., Heyneker,H.L., Yansura,D.G., Crea,R., Hirose,T., Kraszewski,A., Itakura,K. and Riggs,A.D. (1979) Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proc. Natl. Acad. Sci*, **76**, 106–110.
14. Suez,J. and Elinav,E. (2017) The path towards microbiome-based metabolite treatment. *Nat. Microbiol.*, **2**, 17075.
15. Paul,D., Arora,A. and Verma,M.L. (2021) Editorial: advances in microbial biofuel production. *Front. Microbiol.*, **12**, 746216.
16. Castilho,L.R., Mitchell,D.A. and Freire,D.M.G. (2009) Production of polyhydroxyalkanoates (PHAs) from waste materials and by-products by submerged and solid-state fermentation. *Bioresour. Technol.*, **100**, 5996–6009.
17. Santos,D.K.F., Rufino,R.D., Luna,J.M., Santos,V.A. and Sarubbo,L.A. (2016) Biosurfactants: multifunctional biomolecules of the 21st century. *Int. J. Mol. Sci.*, **17**, 401.
18. Humenik,M., Smith,A.M. and Scheibel,T. (2011) Recombinant spider silks—biopolymers with potential for future applications. *Polymers*, **3**, 640–661.

19. Kiki,M.J. (2023) Biopigments of microbial origin and their application in the cosmetic industry. *Cosmetics*, **10**, 47.
20. Averianova,L.A., Balabanova,L.A., Son,O.M., Podvolotskaya,A.B. and Tekutyeva,L.A. (2020) Production of vitamin B2 (riboflavin) by microorganisms: an overview. *Front. Bioeng. Biotechnol.*, **8**, 570828.
21. Gholami-Shabani,M., Shams-Ghahfarokhi,M., Razzaghi-Abyaneh,M., Gholami-Shabani,M., Shams-Ghahfarokhi,M. and Razzaghi-Abyaneh,M. (2023) Food microbiology: application of microorganisms in food industry IntechOpen. https://www.intechopen.com/chapters/86078.
22. Ayilara,M.S. and Babalola,O.O. (2023) Bioremediation of environmental wastes: the role of microorganisms. *Front. Agron.*, **5**, 1183691.
23. Vero,S., Garmendia,G., Allori,E., Sanz,J.M., Gonda,M., Alconada,T., Cavello,I., Dib,J.R., Diaz,M.A., Nally,C., *et al.* (2023) Microbial biopesticides: diversity, scope, and mechanisms involved in plant disease control. *Diversity*, **15**, 457.
24. Kumar,S., Diksha,S.,S.S. and Kumar,R. (2022) Biofertilizers: an ecofriendly technology for nutrient recycling and environmental sustainability. *Curr. Res. Microb. Sci.*, **3**, 100094.
25. Lu,L., Shen,X., Sun,X., Yan,Y., Wang,J. and Yuan,Q. (2022) CRISPR-based metabolic engineering in non-model microorganisms. *Curr. Opin. Biotechnol.*, **75**, 102698.
26. Hwang,S., Joung,C., Kim,W., Palsson,B. and Cho,B.-K. (2023) Recent advances in non-model bacterial chassis construction. *Curr. Opin. Syst. Biol.*, **36**, 100471.
27. Ardalani,O., Phaneuf,P., Mohite,O.S., Nielsen,L.K. and Palsson,B.O. (2023) Pangenome reconstruction of Lactobacillaceae metabolism predicts species-specific metabolic traits. bioRxiv doi: https://doi.org/10.1101/2023.09.18.558222, 20 September 2023, preprint: not peer reviewed.
28. Catoiu,E.A., Phaneuf,P., Monk,J. and Palsson,B.O. (2023) Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks. *Proc. Natl. Acad. Sci*, **120**, e2218835120.
29. Harke,A.S., Josephs-Spauling,J., Mohite,O.S., Chauhan,S.M., Ardalani,O., Palsson,B. and Phaneuf,P.V. (2023) Genomic insights into Lactobacillaceae: analyzing the "Alleleome" of core pangenomes for enhanced understanding of strain diversity and revealing Phylogroup-specific unique variants. bioRxiv doi: https://doi.org/10.1101/2023.09.22.558971, 22 September 2023, preprint: not peer reviewed.
30. Palsson,B., Catoiu,E. and Hyun,J. (2023) Alleleomes characterize the survivors of 3.5 billion years of bacterial evolution. ResearchGate doi: https://doi.org/10.21203/rs.3.rs-3168663/v1, July 2023, preprint: not peer reviewed.
31. Ding,W., Baumdicker,F. and Neher,R.A. (2018) panX: pan-genome analysis and exploration. *Nucleic Acids. Res.*, **46**, e5.
32. Huang,K., Brady,A., Mahurkar,A., White,O., Gevers,D., Huttenhower,C. and Segata,N. (2014) MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids. Res.*, **42**, D617–D624.
33. Zhang,Y., Zhang,H., Zhang,Z., Qian,Q., Zhang,Z. and Xiao,J. (2023) ProPan: a comprehensive database for profiling prokaryotic pan-genome dynamics. *Nucleic Acids. Res.*, **51**, D767–D776.
34. Medini,D., Donati,C., Rappuoli,R. and Tettelin,H. (2020) The pangenome: a data-driven discovery in biology. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing, Cham, pp. 3–20.
35. Xiao,Z., Li,W., Moon,H., Roell,G.W., Chen,Y. and Tang,Y.J. (2023) Generative artificial intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology. *ACS Synth. Biol.*, **12**, 2973–2982.
36. Dagdelen,J., Dunn,A., Lee,S., Walker,N., Rosen,A.S., Ceder,G., Persson,K.A. and Jain,A. (2024) Structured information extraction from scientific text with large language models. *Nat. Commun.*, **15**, 1418.
37. Zhao,J., Huang,S. and Cole,J.M. (2023) OpticalBERT and OpticalTable-SQA: text- and table-based language models for the optical-materials domain. *J. Chem. Inf. Model.*, **63**, 1961–1981.
38. Huang,S. and Cole,J.M. (2022) BatteryBERT: a pretrained language model for battery database enhancement. *J. Chem. Inf. Model.*, **62**, 6365–6377.
39. Van Veen,D., Van Uden,C., Blankemeier,L., Delbrouck,J.-B., Aali,A., Bluethgen,C., Pareek,A., Polacin,M., Reis,E.P., Seehofnerová,A., *et al.* (2024) Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.*, **30**, 1134–1142.
40. Gao,Y., Xiong,Y., Gao,X., Jia,K., Pan,J., Bi,Y., Dai,Y., Sun,J., Wang,M. and Wang,H. (2024) Retrieval-augmented generation for large language models: a survey. arXiv doi: https://arxiv.org/abs/2312.10997, 27 March 2024, preprint: not peer reviewed.
41. Zhao,P., Zhang,H., Yu,Q., Wang,Z., Geng,Y., Fu,F., Yang,L., Zhang,W., Jiang,J. and Cui,B. (2024) Retrieval-augmented generation for AI-generated content: a survey. arXiv doi: https://arxiv.org/abs/2402.19473, 21 June 2024 preprint: not peer reviewed.
42. Li,J., Yuan,Y. and Zhang,Z. (2024) Enhancing LLM factual accuracy with RAG to counter hallucinations: a case study on domain-specific queries in private knowledge-Bases. arXiv doi: https://arxiv.org/abs/2403.10446, 15 March 2024, preprint: not peer reviewed.
43. Phaneuf,P.V., Jarczynska,Z.D., Kandasamy,V., Chauhan,S.M., Feist,A.M. and Palsson,B.O. (2023) Using the E. coli alleleome in strain design. bioRxiv doi: https://doi.org/10.1101/2023.09.17.558058, 17 September 2023, preprint: not peer reviewed.
44. Chaudhary,R., Nawaz,A., Fouillaud,M., Dufossé,L., Haq,I.u. and Mukhtar,H. (2024) Microbial cell factories: biodiversity, pathway construction, robustness, and industrial applicability. *Microbiol. Res.*, **15**, 247–272.
45. Steensels,J., Gallone,B., Voordeckers,K. and Verstrepen,K.J. (2019) Domestication of industrial microbes. *Curr. Biol.*, **29**, R381–R393.
46. Di Lorenzo,R.D., Serra,I., Porro,D. and Branduardi,P. (2022) State of the art on the microbial production of industrially relevant organic acids. *Catalysts*, **12**, 234.
47. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S., *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids. Res.*, **50**, D20–D26.
48. Nuhamunada,M., Mohite,O.S., Phaneuf,P.V., Palsson,B.O. and Weber,T. (2024) BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. *Nucleic Acids. Res.*, **52**, 5478–5495.
49. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. and Parks,D.H. (2020) GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, **36**, 1925–1927.
50. Otani,H., Udwary,D.W. and Mouncey,N.J. (2022) Comparative and pangenomic analysis of the genus Streptomyces. *Sci. Rep.*, **12**, 18909.
51. Zhong,C., Qu,B., Hu,G. and Ning,K. (2022) Pan-genome analysis of campylobacter: insights on the genomic diversity and virulence profile. *Microbiol. Spectr.*, **10**, e01029–e22.
52. Gaba,S., Kumari,A., Medema,M. and Kaushik,R. (2020) Pan-genome analysis and ancestral state reconstruction of class halobacteria: probability of a new super-order. *Sci. Rep.*, **10**, 21205.
53. Brito,P.H., Chevreux,B., Serra,C.R., Schyns,G., Henriques,A.O. and Pereira-Leal,J.B. (2018) Genetic competence drives genome diversity in Bacillus subtilis. *Genome Biol. Evol.*, **10**, 108–124.
54. Rahman,M.S., Shimul,M.E.K. and Parvez,M.A.K. (2024) Comprehensive analysis of genomic variation, pan-genome and biosynthetic potential of Corynebacterium glutamicum strains. *PLoS One*, **19**, e0299588.

55. Bosi,E., Monk,J.M., Aziz,R.K., Fondi,M., Nizet,V. and Palsson,B.Ø. (2016) Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci*, **113**, E3801–E3809.
56. Hassan,A., Naz,A., Obaid,A., Paracha,R.Z., Naz,K., Awan,F.M., Muhmmad,S.A., Janjua,H.A., Ahmad,J. and Ali,A. (2016) Pangenome and immuno-proteomics analysis of Acinetobacter baumannii strains revealed the core peptide vaccine targets. *BMC Genomics*, **17**, 732.
57. Norsigian,C.J., Fang,X., Palsson,B.O. and Monk,J.M. (2020) Pangenome flux balance analysis toward panphenomes. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing, Cham, pp. 219–232.
58. Wu,H., Wang,D. and Gao,F. (2021) Toward a high-quality pan-genome landscape of Bacillus subtilis by removal of confounding strains. *Brief. Bioinform.*, **22**, 1951–1971.
59. Vernikos,G.S. (2020) A review of pangenome tools and recent studies. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing, Cham, pp. 89–112.
60. Edwards,S., León-Zayas,R., Ditter,R., Laster,H., Sheehan,G., Anderson,O., Beattie,T. and Mellies,J.L. (2022) Microbial consortia and mixed plastic waste: pangenomic analysis reveals potential for degradation of multiple plastic types via previously identified PET degrading bacteria. *Int. J. Mol. Sci.*, **23**, 5612.
61. Liu,N., Liu,D., Li,K., Hu,S. and He,Z. (2022) Pan-genome analysis of Staphylococcus aureus reveals key factors influencing genomic plasticity. *Microbiol. Spectr*, **10**, e03117-22.
62. Ma,X., Sun,T., Zhou,J., Zhi,M., Shen,S., Wang,Y., Gu,X., Li,Z., Gao,H., Wang,P., *et al.* (2023) Pangenomic study of fusobacterium nucleatum reveals the distribution of pathogenic genes and functional clusters at the subspecies and strain levels. *Microbiol. Spectr.*, **11**, e051842-22.
63. Kim,Y., Koh,I., Young Lim,M., Chung,W.-H. and Rho,M. (2017) Pan-genome analysis of Bacillus for microbiome profiling. *Sci. Rep.*, **7**, 10984.
64. Surachat,K., Deachamag,P., Kantachote,D., Wonglapsuwan,M., Jeenkeawpiam,K. and Chukamnerd,A. (2021) *In silico* comparative genomics analysis of *Lactiplantibacillus plantarum* DW12, a potential gamma-aminobutyric acid (GABA)-producing strain. *Microbiol. Res.*, **251**, 126833.
65. Luu,R.K. and Buehler,M.J. (2024) BioinspiredLLM: conversational large language model for the mechanics of biological and bio-inspired materials. *Adv. Sci.*, **11**, 2306724.
66. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic. Acids. Res.*, **28**, 27–30.
67. Corma,A., Iborra,S. and Velty,A. (2007) Chemical routes for the transformation of biomass into chemicals. *Chem. Rev.*, **107**, 2411–2502.
68. Gao,C., Ma,C. and Xu,P. (2011) Biotechnological routes based on lactic acid production from biomass. *Biotechnol. Adv.*, **29**, 930–939.
69. Alves de Oliveira,R., Komesu,A., Vaz Rossell,C.E. and Maciel Filho,R. (2018) Challenges and opportunities in lactic acid bioprocess design—From economic to production aspects. *Biochem. Eng. J.*, **133**, 219–239.
70. Juturu,V. and Wu,J.C. (2016) Microbial production of lactic acid: the latest development. *Crit. Rev. Biotechnol.*, **36**, 967–977.
71. Abdel-Rahman,M.A., Tashiro,Y. and Sonomoto,K. (2013) Recent advances in lactic acid production by microbial fermentation processes. *Biotechnol. Adv.*, **31**, 877–902.
72. Okano,K., Tanaka,T., Ogino,C., Fukuda,H. and Kondo,A. (2010) Biotechnological production of enantiomeric pure lactic acid from renewable resources: recent achievements, perspectives, and limits. *Appl. Microbiol. Biotechnol.*, **85**, 413–423.
73. Tian,X., Chen,H., Liu,H. and Chen,J. (2021) Recent advances in lactic acid production by lactic acid bacteria. *Appl. Biochem. Biotechnol.*, **193**, 4151–4171.
74. Augustiniene,E., Valanciene,E., Matulis,P., Syrpas,M., Jonuskiene,I. and Malys,N. (2022) Bioproduction of l- and d-lactic acids: advances and trends in microbial strain application and engineering. *Crit. Rev. Biotechnol.*, **42**, 342–360.
75. Lütke-Eversloh,T., Santos,C.N.S. and Stephanopoulos,G. (2007) Perspectives of biotechnological production of l-tyrosine and its applications. *Appl. Microbiol. Biotechnol.*, **77**, 751–762.
76. Min,K., Park,K., Park,D.-H. and Yoo,Y.J. (2015) Overview on the biotechnological production of l-DOPA. *Appl. Microbiol. Biotechnol.*, **99**, 575–584.
77. Surwase,S.N. and Jadhav,J.P. (2011) Bioconversion of l-tyrosine to l-DOPA by a novel bacterium Bacillus sp. JPJ. *Amino Acids*, **41**, 495–506.
78. Kandpal,N., Deng,H., Roberts,A., Wallace,E. and Raffel,C. (2023) Large language models struggle to learn long-tail knowledge. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 15696–15707.
79. Zhang,Y., Li,Y., Cui,L., Cai,D., Liu,L., Fu,T., Huang,X., Zhao,E., Zhang,Y., Chen,Y., *et al.* (2023) Siren's song in the AI Ocean: a survey on hallucination in large language models. arXiv doi: https://arxiv.org/abs/2309.01219, 24 September 2023, preprint: not peer reviewed.
80. Chen,J., Lin,H., Han,X. and Sun,L. (2024) Benchmarking large language models in retrieval-augmented generation. *Proc. AAAI Conf. Artif. Intell*, **38**, 17754–17762.

Received: August 15, 2024. **Revised:** October 11, 2024. **Editorial Decision:** October 14, 2024. **Accepted:** October 19, 2024