Sidero-Mining: Systematic Extraction of Siderophore Biosynthetic Information Using Large Language Models

Jiqi Shao[1#], Yanzhao Wu[2#], Ningbo Xi[3], Ruolin He[1], Ruichen Xu[4], Peng Guo[5], Shaohua Gu[1,4,6,*], and Zhiyuan Li[1,4,6,*]

[1]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[2]College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Taiyuan, 030600, China

[3]Jiangsu Provincial Key Laboratory for Organic Solid Waste Utilization, Jiangsu Collaborative Innovation Center for Solid Organic Waste Resource Utilization, National Engineering Research Center for Organic-based Fertilizers, Nanjing Agricultural University, Nanjing, China

[4]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[5]School of Life Sciences, Shandong University, Qingdao, 266237, China

[6]Lead contact

#Co-first authors

*Corresponding author: Zhiyuan Li(zhiyuanli@pku.edu.cn); Shaohua Gu (gushaohua@pku.edu.cn)

Abstract

Siderophores are essential secondary metabolites widely distributed across microorganisms, displaying remarkable diversity. Despite extensive research, public databases contain limited information on siderophore biosynthetic gene clusters (BGCs), particularly lacking cross-species distribution and biosynthetic substrate annotations. Systematically collecting and organizing siderophore BGC synthesis data on a large scale would significantly enhance the use of domain knowledge and support data-driven research.

Large language models (LLMs) now offer a practical and scalable approach for mining and curating biological data, especially for converting literature insights into structured datasets In this work, we developed the Sidero-Mining pipeline, using LLMs to efficiently extract siderophore BGC synthesis information. By employing LLMs to screen over 10,000 publications, we identified 1,843 high-quality articles for data mining based on Sidero-Mining framework, manual validation, and data integration.

This effort culminated in the creation of the most comprehensive siderophore BGC dataset to date, containing 728 BGCs and 325 NRPS A domain substrate entries cross various species. Our results highlight LLMs' potential to accelerate secondary metabolite dataset construction, and our methodological framework can be adapted for systematically exploring other secondary metabolites.

Keywords: Large Language Model, Siderophore, Text Mining, Secondary Metabolites, Cross-Species Dataset

Introduction

Natural products serve as the chemical language of microbial-environment interactions, and remain pivotal in drug discovery[1]. Siderophores, secondary metabolites secreted under iron-limited conditions, help bacteria survive and compete by chelating iron from the environement. The ecological roles of siderophores are crucial for deciphering complex microbial interactions, where diverse siderophores mediate both iron-competition and the piracy of siderophores produced by other organisms. In pathogen suppression, siderophores have been harnessed to limit pathogenic microorganisms by depriving them of iron[2,3]. Recent advances in data mining, particularly with the siderophore pyoverdine in *Pseudomonas*, have shifted the focus toward sequence-based approaches for discovering new siderophores and predicting microbial interactions[4,5]. These progresses open up new “sequence-to-ecology” possibilities but calls for detailed information about the BGC of siderophores[6].

However, despite notable advancements in understanding siderophore biosynthesis and their ecological roles, their synthetic diversity and cross-species interactions remain largely unexplored systematically[7]. The most comprehensive siderophore structure databases, SIDERITE, focus primarily on structural information without pinpointing these structures to their biosynthetic gene clusters[7]. Meanwhile, BGC repositories like MIBiG contain some siderophore BGC data, but species-specific annotations, gene details, and chemical structure information remain scattered throughout the scientific literature in various formats[8]. This disconnect between BGC and molecules is problematic, because siderophore structures vary considerably even within the same species. Synthetase gene information is crucial for genome-based data mining and future digital siderophore exploration efforts. However, traditional manual curation approaches struggle to handle this data deluge, with researchers spending disproportionate effort on data cleaning rather than scientific discovery. This has limited comprehensive research on siderophores, including cross-species annotation, evolutionary patterns, and synthetic substrate prediction.

The emergence of large language models (LLMs) like ChatGPT has transformed data mining across disciplines[9-14]. These models excel at parsing unstructured text, demonstrating exceptional semantic comprehension that allows them to identify biological entities such as genes and metabolites, understand their functional relationships, and navigate terminological variations. Unlike conventional text mining approaches, LLMs use contextual reasoning to extract deeper cross-domain knowledge from diverse subfields. Examples include the automated exploration of enzyme-substrate interaction networks[15], identification of key targets and signaling pathways in specific diseases[16], and elucidation of m6A molecular regulatory pathways[17]. In natural product research, such explorations have just started. For example, Kalmer et al. evaluated ChatGPT's ability in extracting natural product bioactivity and biosynthesis data from scientific publications[12]. As these models continue to improve and adapt to specific domains, LLMs are becoming essential tools for integrating biomedical data resources.

This shift addresses the fragmentation of traditional data management while creating scalable knowledge infrastructure in microbial resource exploitation.

In this study, we developed Sidero-Mining workflow, a novel workflow utilizing LLMs to extract siderophore synthesis information from scientific literatures. Using strategically designed prompts, Sidero-Mining efficiently captures synthesis-related details, including siderophore-producing species, synthesis genes, and substrate specificity. We optimized the workflow for substrate extraction within the NRPS pathway through task segmentation and logical integration. Deploying this system at scale, we first used LLMs to screen over 10,000 siderophore-related articles, identifying 1,843 high-quality articles. Through Sidero-Mining's data extraction, manual verification, and data alignment, we assembled a high-quality siderophore BGC dataset, encompassing a knowledge repository of 728 BGCs and 325 NRPS A domain substrate entries cross various species. This work produces not only the most extensive siderophore BGC dataset to date—serving as a gold standard for future siderophore data mining—but also establishes a framework readily adaptable for exploring other secondary metabolites synthetase, facilitating the construction of comprehensive and well-organized secondary metabolite databases.

Result

Utilizing Large Language Models to Extract Information on Siderophore Synthesis

The study of siderophore BGCs is essential for understanding the synthesis mechanisms and ecological roles of siderophores. Nonetheless, a substantial portion of siderophore research remains underutilized and has not been effectively translated into formatted knowledge. A key challenge in siderophore BGC literature mining is identifying which specific information requires extraction. Siderophores are widely distributed across microbial species and show remarkable synthetic diversity[18].

The diversity of siderophore BGCs primarily stems from varied combinations and arrangements of substrate-specific components within their biosynthetic pathways. Around 70% known siderophores are synthesized by the nonribosomal peptide synthetase (NRPS) pathway[7]. NRPS represents a modular assembly line for siderophore synthesis, where each introduces and processes an amino acid substrate. Within each module, the adenylation domains (A domains) exhibits substrate specificity, recognizing distinct amino acid substrates[19-21]. This module configurations fosters product diversity, attracting engineering efforts in reprogramming modules[22,23]. The NRPS-independent siderophore synthetases(NIS) pathway generates distinct classes of siderophores by combining NIS synthetases with differing substrate specificities.

Extracting core synthesis genes and substrate recognition details from siderophore-related literatures constitutes one of the most complex challenges of such mining

endeavors. Within the NRPS pathway, we need to extract not only the core synthesis genes, but also the modules and A domain substrate information from literatures. The extraction of A domains requires attention to the sequential arrangement of the corresponding modules in the final product, a critical aspect for downstream sequence information matching.

Large language models (LLMs) possess powerful natural language comprehension capabilities, making them particularly well-suited for text information extraction. However, they have limitations, including hallucinations and accurately capturing essential information within lengthy contexts[12,24]. To effectively extract siderophore synthesis information, developing appropriate prompt workflows is essential to help LLM s fully understand extraction objects.

In this study, we engineered a series of prompt tasks for mining siderophore-related synthesis information, collectively termed “Sidero-Mining” (Figure 1). This pipeline extracts diverse of information, including siderophore-producing strains with their identifiers, siderophore synthesis types (NRPS, NIS, etc.), chemical formulas, synthesis genes with substrate specificity, as well as accessory biosynthetic genes, transport genes, and other relevant siderophore details. By implementing format output instructions, the extracted content is presented in a markdown format compatible with XMind, allowing for smooth manual verification and corrections, and ultimate exportation in a standardized JSON format. Notably, siderophore synthesis data typically features nested and hierarchical structures, and the integration of XMind and JSON designs not only streamlines the verification process but also allows unified data format management, simplifying the general workflow of data input, organization, and verification, thereby expediting dataset construction (detailed in Methods)

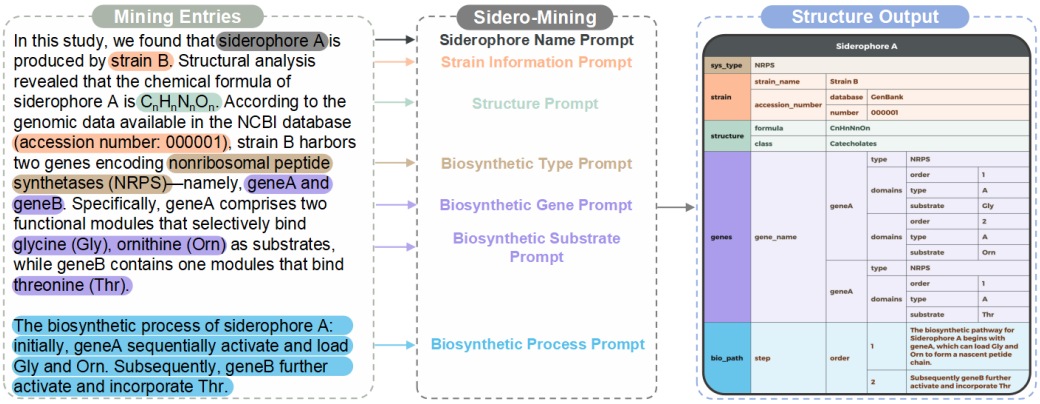

Figure 1. Sidero-Mining Framework for Extraction and Structured Organization of Siderophore Information from Scientific Literature

The figure illustrates the Sidero-Mining workflow for extracting siderophore information from research papers. Left panel ("Mining Entries") shows a sample text entry containing

information about a hypothetical “siderophore A”’s production, chemical formula, and biosynthetic genes. The middle panel ("Sidero-Mining") displays the sequential prompt tasks used for systematic information extraction, including siderophore name, strain information, structure, biosynthetic type, gene identification, and substrate specificity prompts(Figure S2-S10). The right panel ("Structure Output") demonstrates the structured hierarchical output format, organizing the extracted information into a standardized schema that captures relationships between strains, genes, and their functional attributes. This framework enables consistent transformation of unstructured text information into structured, machine-readable data for subsequent analysis.

Improving NRPS Information Extraction Through Task Decomposition in Sidero-Mining

Sidero-Mining has shown remarkable performance in tasks such as siderophore and species identification(Table 1). However, the most critical and challenging aspect of Sidero-Mining is extracting substrate specificity information from NRPS siderophore literature. In our evaluation using a manually curated dataset (detailed in Methods), conventional prompt designs for extracting NRPS siderophore substrate information—including providing background context or examples—did not significantly improve extraction performance. Instead, issues like missing extraction targets, errors, and hallucinations were common. Although the instruction to "extract substrate information of NRPS siderophores" seems straightforward, the task requires understanding complex next information: identifying synthesis genes, determining which genes belong to the NRPS core synthesis pathway, recognizing modules within these core genes, determining the sequence of substrate-specific modules, and finally the substrate specificity itself.

Table1 Final Performance of Sidero-Mining on Downstream Tasks

Mining Elements	Precision	Recall	F1-Score
siderophore-name	0.9833	1.0000	0.9916
species-name	0.9833	1.0000	0.9916
sys-type	0.9259	1.0000	0.9615
gene-info	0.9189	0.8947	0.9067
substrate-info	1.0000	0.7667	0.8679

To address these challenges, we improved the substrate extraction workflow in Sidero-Mining by decomposing complex tasks and logically connecting related steps, significantly improving extraction of synthesis information, including substrate details. We broke down the intricate process into manageable sequential tasks with clear logic (Figure 2A): (1) extracting all genes associated with a specific siderophore, (2) identifying the synthesis type based on gene descriptions, (3) determining the substrate extraction workflow and

core synthesis genes based on the synthesis type, and (4) identifying specific modules and their corresponding substrates within the core synthesis genes.

The power of this approach lies in using information from the previous step to guide subsequent task. For example, extracting genes first and then determining synthesis type based on gene function yields more accurate results than directly querying synthesis type, as genes function logically indicates the siderophore's synthesis pathway. This improved approach dramatically improved the substrate extraction performance, raising precision from approximately 0.8 to 1, recall from 0.63 to 0.77, and F1-score from 0.72 to 0.87.

The logical coupling of extraction tasks also enhanced performance across multiple metrics For synthesis type classification, transitioning from direct queries to gene-based judgment improved precision precision from 0.64 to 0.93, recall from 0.95 to 1, and F1-score by approximately 0.2 to 0.96. Synthesis gene extraction showed more dramatic improvement, with precision increasing from 0.56 to 0.92, recall approximately from 0.3 to 0.89, and F1-score from 0.58 to 0.91(Figure 2 B-D). The improved Sidero-Mining serves as the core component for our large-scale literature mining application.

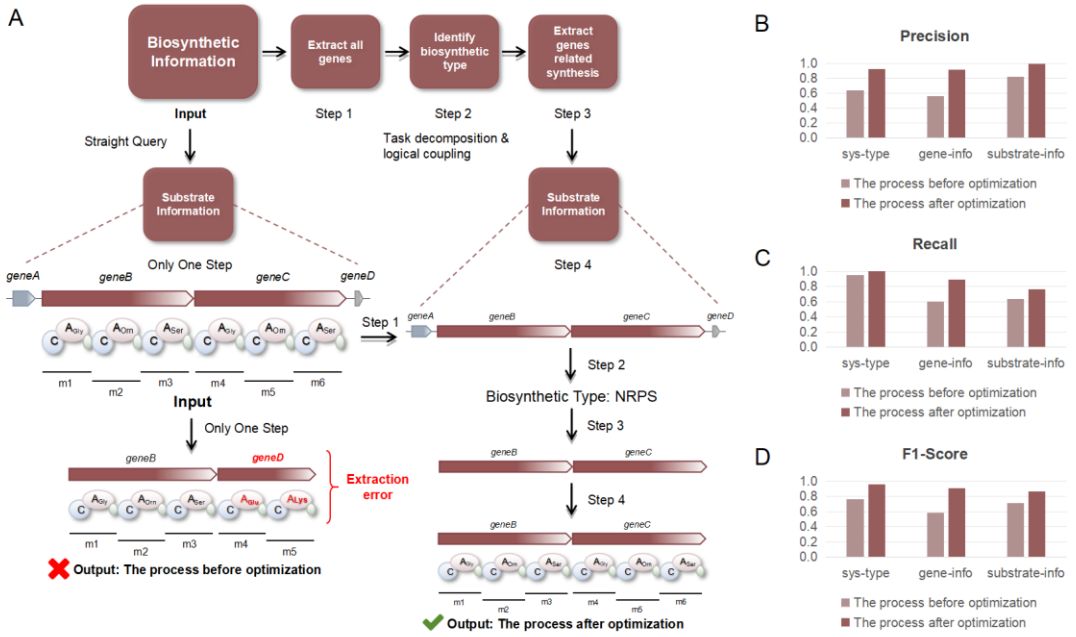

Figure 2. Improving NRPS Substrate Extraction Through Task Decomposition in Sidero-Mining

(A) Schematic comparison of traditional versus optimized substrate extraction workflows. The optimized process breaks down the complex task into four logical steps: extracting all siderophore-related genes (Step 1), determining synthesis type based on gene function (Step 2), identifying genes specifically related to synthesis (Step 3), and extracting substrate specificity information (Step 4). The diagram contrasts this with the traditional single-step approach, which often resulted in incomplete or inaccurate extraction.

(B-D) Performance comparison before and after workflow optimization across three key

metrics: (B) Precision, (C) Recall, and (D) F1-Score. Results are shown for three extraction tasks: synthesis type identification (sys-type), gene information extraction (gene-info), and substrate information extraction (substrate-info). The optimized task decomposition approach shows substantial improvement across all metrics.

Building a Comprehensive Siderophore BGC Dataset through Large-Scale Literature Mining

To construct a high-quality siderophore BGC Dataset, we conducted extensive data mining of siderophore-related literature. Using the Semantic Scholar Dataset as our primary data source, we retrieved all siderophore-related articles published after 1980 (detailed in the Methods section), collecting 12,004 publications. These publications cover diverse aspects of siderophore research, with particular emphasis on synthesis capabilities across various species, necessitating comprehensive classification and filtering.

We first classified articles based on titles and abstracts, focusing on siderophore synthesis, gene information, and chemical structure. This initial filtering identifies 4,174 articles likely to contain synthesis and structural information. We then analyzed the full text of these articles using tailored prompts for different information types (Figure S1), conducting multi-dimensional binary classification synthesis, structure, transport, and species interaction details. This process yielded 1,823 articles containing both structural and synthesis information, which formed our final literature set for data extraction and dataset construction. This approach demonstrates LLMs' effectiveness in condensing datasets to enable more precise and targeted mining of relevant literature.

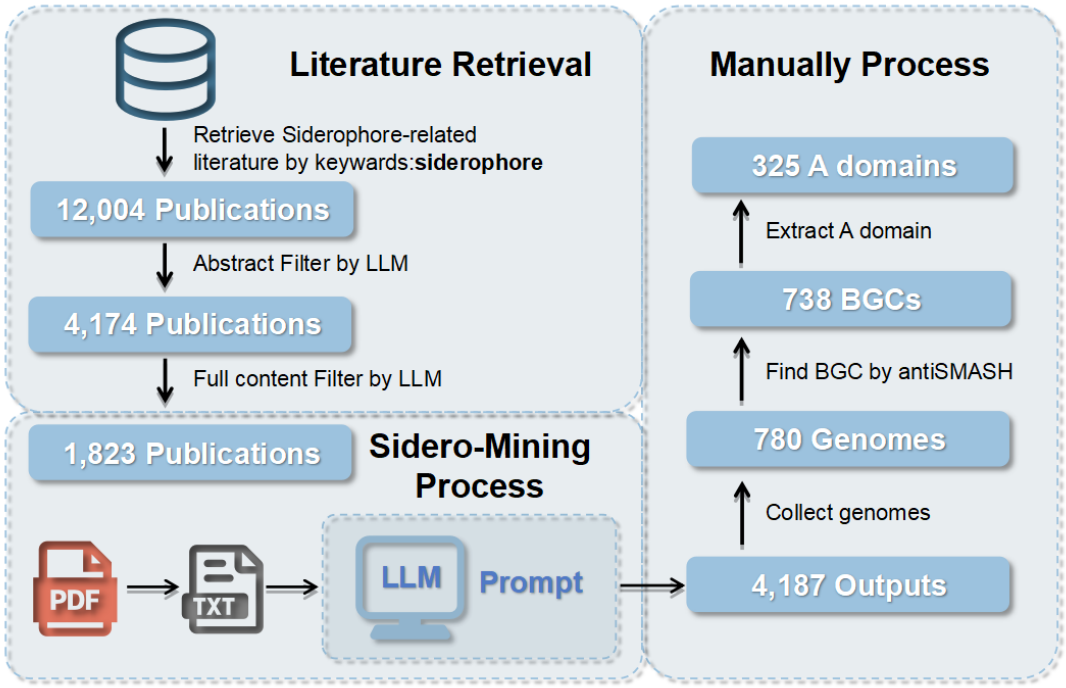

Figure 3. Workflow for Siderophore BGC Dataset through Large-Scale Text Mining and

Manual Curation

The figure illustrates our workflow for building a comprehensive siderophore dataset. Left panel ("Literature Retrieval"): Starting with 12,004 publications retrieved using siderophore keywords, we implemented a two-stage filtering process using LLMs—first on abstracts (yielding 4,174 publications) and then on full text (yielding 1,823 publications). Center panel ("Sidero-Mining Process"): Publications were converted from PDF to text format before processing with our LLM-based extraction system, generating 4,187 structured outputs. Right panel ("Manually Process"): These outputs underwent rigorous validation and enrichment through genome collection (780 genomes), BGC identification using antiSMASH, and A domain extraction, ultimately resulting in 738 curated cross-species siderophore BGCs and 325 validated A domain substrate records. Each BGC or A domain entry corresponds to distinct sequence origins, ensuring that all records represent unique sources. This pipeline demonstrates the effective combination of automated extraction and expert curation to create a high-quality reference dataset.

Our comprehensive siderophore BGC dataset was constructed based on the structured output generated from the aforementioned steps. We processed the 1,823 articles through the Sidero-Mining pipeline powered by ChatGPT-4o (Figure 1), extracting structured data in the format illustrated earlier. This mining yielded 4,187 records, with 1,671 articles contributing 3,296 records specifically related to siderophore synthesis. From these records, we collected 780 genomic datasets using the NCBI API and other sources, then annotated them for BGCs using antiSMASH. Here, we define each BGC derived from different sequence sources as a unique record. Through manual verification and synthesis gene matching, we identified 738 unique siderophore BGCs—a substantial increase compared to the 137 BGC records in the existing MiBiG Dataset (Figure 4A). These BGCs were categorized by synthesis type: 465 NRPS, 187 NIS, 70 NRPS-PKS hybrids, 6 NRPS-NIS hybrids, and several others including PKS, Opine, and CDPS types.

Statistical analysis revealed that 91% of the collected siderophore BGCs were from bacteria, with the remainder primarily from fungi, and one archaeal siderophore, schizokinen (Figure 4 B). Among records with genomic data, enterobactin and pyoverdine had the highest number of unique records (Figure 4 C). Examining cross-species distribution based on both valid records and matched BGC sequence data showed that many identical or related siderophores from the same family were distributed across a wide range of species (Figure 4 D), highlighting the extensive cross-species distribution of siderophore BGCs.

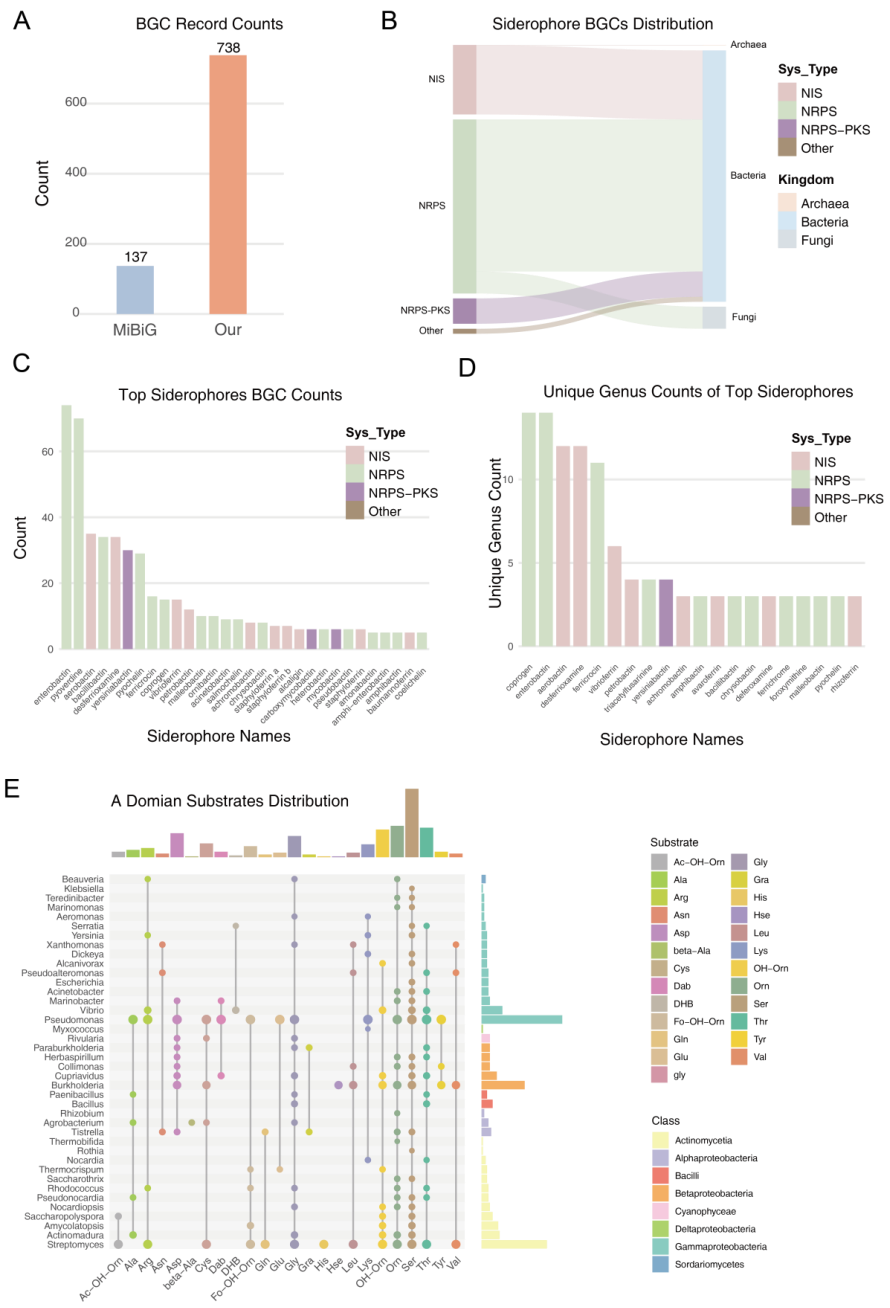

Figure 4. Overview of the Siderophore BGC Dataset

(A) A comparative analysis of the number of siderophore BGCs mined versus those documented in MiBiG.

(B) The taxonomic origins and biosynthetic classification of siderophore BGCs.

(C) The abundance of BGCs associated with distinct siderophores, color-coded by their synthesis types.

(D) The diversity of unique genera harboring different siderophores, color-coded by their synthesis types.

(E) The distribution patterns of A domain substrates and their respective source species. Substrates are differentiated by distinct colors, while species from various genera are color-coded according to their taxonomic class.

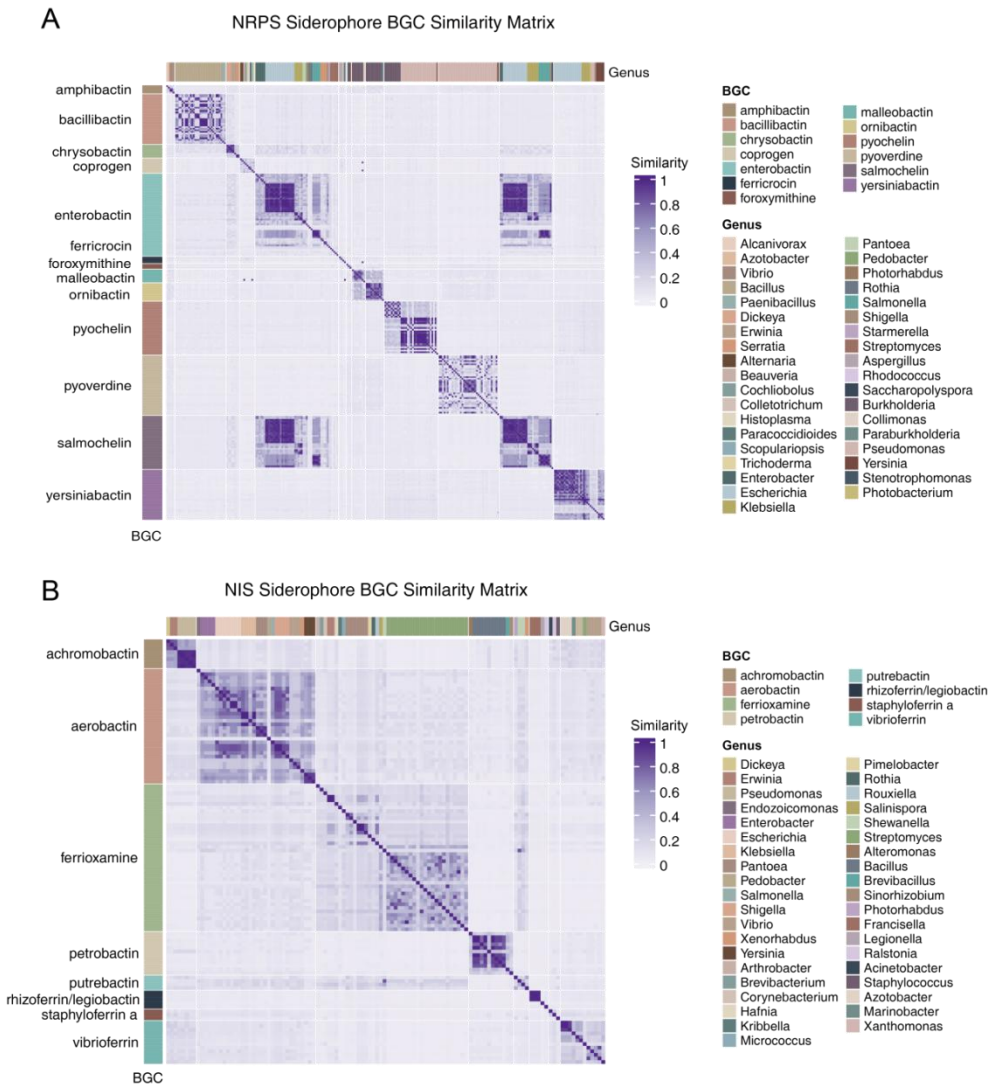

Figure 5. Cross-Genus Similarity Matrix of Siderophore BGCs

(A) Similarity matrix of NRPS siderophore BGCs distributed across genera, computed using BiG-SCAPE.

(B) Similarity matrix of NIS siderophore BGCs distributed across genera, computed using BiG-SCAPE.

The widespread distribution of identical siderophores across diverse species presents significant challenges in their identification and annotation. Homologous siderophores from different species may display considerable sequence divergence due to phylogenetic influences, complicating the identification and clustering of BGCs. In this study, we used BiG-SCAPE to quantify sequence similarities of BGCs encoding the same siderophore across various species[25]. For both NRPS and NIS siderophore BGCs, the sequence similarities of the same siderophore from different genera exhibited substantial variation, making it difficult to cluster them into the same group (Figure 5 A, B). This finding establishes a valuable cross-species benchmark dataset for developing future tools for comparing synthetic gene clusters.

For NRPS siderophore entries with detailed synthesis information, we used the Sidero-Mining pipeline to extract synthesis genes and substrate information. Through manual verification and matching, we compiled 325 unique A domains with their associated substrate specificity from various species (Figure 4 E). These A domain data span 24 distinct substrates distributed across 40 genera, with the most frequently identified substrates including amino acids such as Ser, Orn, and Thr. The genera with the highest abundance of collected A domains were Pseudomonas, Streptomyces, and Burkholderia. This collection serves as a high-quality resource for downstream substrate prediction of siderophore A domains.

Discussion

In the interdisciplinary field of secondary metabolites, where sequence, structure, and function are deeply interconnected, our study demonstrates the potential by developing Sidero-Mining, an advanced LLM-based literature mining pipeline designed to systematically extract siderophore BGC data from extensive literature repositories. Extracting high-quality BGC synthesis information, particularly for NRPS pathways, requires a clear delineation of hierarchical data relationships and the sequential logic of recognition modules. By employing prompt engineering strategies, Sidero-Mining effectively extracts siderophore BGC synthesis data with both precision and efficiency. To validate and align the knowledge extracted by Sidero-Mining pipeline, we utilized LLM-structured output and XMind format conversion, reducing the time required for expert validation and alignment of BGC genes, recognition modules, substrate specificity, and sequence data.

By combining efficient LLM-based data extraction with expert refinement, we have created the most comprehensive cross-species siderophore BGC dataset to date, containing 738 unique siderophore BGCs and 325 NRPS A domain substrate entries with associated sequence data. To our knowledge, this represents the first application of LLMs in constructing a secondary metabolite dataset—an advancement that will significantly enhance both the utilization of siderophore BGC data and the application of LLMs in secondary metabolism research. Analysis of our collected siderophore BGCs reveals several key insights. First, multiple siderophores documented in the literature demonstrate cross-species distribution. Second, homologous BGCs across species often exhibit substantial sequence divergence due to phylogenetic influences, which complicates accurate clustering and annotation with existing tools. Our dataset provides a valuable benchmark for assessing cross-species BGC similarity, facilitating the development of novel methods for BGC comparison[25,26].

Since siderophores influence microbial interactions, our dataset enables more accurate identification of siderophore BGCs and reconstruction of siderophore interaction networks. Additionally, our extensive repository of NRPS-siderophore's A domain data with substrate specificity will support substrate prediction for A domain[27-30], especially for non-

proteinogenic amino acids frequently appears in NRPS-siderophores.

The LLM-based Sidero-Mining pipeline developed in this study can be extended to other categories of secondary metabolites, paving the way for a more comprehensive natural product dataset. This is particularly straightforward for other NRPS-synthesized products, where the workflow logic remains fundamentally similar. Future applications can incorporate diverse types of LLMs. For instance, visual LLMs could extract synthesis information from key figures[10], while reasoning models can augment the understanding of complex relationships. Furthermore, fine-tuning LLMs can improve text extraction and transformation capabilities[31], and specialized AI agents can be designed to invoke various data and tools for processing and generating data[32]. As model capabilities continue to advance, we anticipate significant acceleration in the field of secondary metabolism.

LLMs still face limitations in practical applications. While prompt engineering, retrieval-augmented generation (RAG), and reasoning models can mitigate issues, the problem of hallucination remains unresolved[10]. Also, in processing scientific literature figures, which often contain richer and more intricate information than text, current visual models struggle to understand complex chart data. For instance, our tests showed that visual models could not effectively recognize and extract siderophore chemical structures, lacking the ability to identify complex chemical structures.

The most significant barrier to domain-specific dataset construction is the disconnect between knowledge and data, a challenge shared by other text mining methods and manual data collection and organization. Detailed investigations require sequences, while annotations extracted from scientific texts serve only as indices or pointers to the usable sequences. Significant gaps remain in connecting different datasets and downstream annotations. For genomic data, a large portion is not publicly accessible, and inconsistent naming conventions hinder the extraction of accurate strain data. Across many biological data domains, naming conventions are complex and inconsistent, with some data retrievable only by name in common datasets. The primary bottleneck in sequence dataset construction lies in indexing and matching literature information with actual data. To fully realize the potential of LLMs in scientific data mining, collaborative efforts must establish standardized publishing and data conventions, including uniform methods for data naming and linking. LLMs themselves can serve as tools for data organization and rapid indexing, ensuring seamless connections between different datasets and between literature and data repositories.

Methods

Dataset Construction

By calling the API of Semantic Scholar, using "siderophore" as the keyword and restricting the publication year to be after to post-1980, we retrieved relevant literatures from the

Semantic Scholar Dataset, ultimately obtaining 12004 pieces of literature.

To evaluate the performance of various models in the task of extracting siderophore - related information, we constructed a test dataset composed of 60 articles. This dataset included an equal number of articles with siderophore - related information (positive) and those without such information (negative). To mitigate potential biases of the models and cue words towards specific articles or siderophores, the dataset also included articles published at different periods and a diverse range of siderophore samples. We manually extracted siderophore information from these articles to create the ground - truth labels samples. Based on the model outputs, the responses for each article were classified into true positives, false positives, true negatives, and false negatives, and then the precision, recall, and F1 - score of the entire dataset were calculated.

Core Design of Sidero-Mining Pipeline

In applying Sidero-Mining to extract siderophore-related information, to enhance the model's ability to accurately extract such information from articles, we meticulously designed prompts for each extraction process. The prompts for each task mainly consist of three parts:

1.Requirements for the Extraction Task: This section clearly specifies the siderophore - related information to be extracted from the article. For example, "Extract the name of the siderophore", "Extract the species information capable of producing this siderophore", etc.
2.Output Answer Structure Template: To facilitate subsequent data verification and processing, it is essential that the model to output structured answers. Accordingly, we designed an output structure template in Markdown format to improve the model's ability to deliver structured responses. The Markdown format is not only easily parsed, but also be converted into the XMind format, which is more in line with human visual recognition habits. When there are numerous output levels, converting the Markdown format into the more visual XMind format can significantly increase the speed of manual verification.
3.Reminder of Precautions: To restrict the creativity and prevent hallucination of the model during output, we added warnings to the prompts of each task. For example, "For information that cannot be obtained from the article, please directly output 'null' as the answer. Do not create answers on your own", "When outputting, please strictly follow the output template provided above. Do not output any information other than that provided in the template", etc.

In addition, in the task of extracting siderophore synthesis information, we not only adopted the above - mentioned prompt design method. To further enhance the model's ability to extract siderophore synthesis information, we also carefully designed a new extraction process. By simulating how humans read literature and extract synthesis details, we divided a complete extraction task according to human thinking patterns:

Task 1: Require the model to extract all genes related to the target siderophore from the text and briefly describe the function of each gene in the siderophore synthesis process.

Task 2: Use these genes and their functional descriptions as part of the input and pass them to the model again. Require the model to judge the synthesis type of the target siderophore by combining the literature content and the provided gene information.

Task 3: Pass the synthesis type and the gene information extracted in the first step as input to the model, and prompt the model to screen out the genes directly related to siderophore synthesis from all the genes.

Task 4: Utilize the obtained synthesis gene information and combine it with the article information to guide the model to extract the specific information of the substrate.

This design breaks the synthesis information extraction task into four interconnected subtasks. The output of each subtask serves as the input for the next, enabling the model to focus more on the current task and progressively build a complete reasoning process, thereby improving the model's extraction performance.

Design of Literature Screening for Large - scale Mining

Step 1: The titles and abstracts of articles retrieved in the first step were input into gpt - 4o - mini (gpt - 4o - mini - 2024 - 07 - 18), and six pre-set prompts were used to classify the research fields of the articles. Specifically, we pre-classified the research fields into six categories: siderophore chemical structure, biosynthesis, ecology, pathology, environment and agriculture, and metal ion transport. For the convenience of statistics, during the classification process, by providing the model with an output structure template in the prompt, we required the model to only return "Y" or "N" for easy statistics, so as to judge whether the paper was related to the corresponding field.

Step 2: For the articles screened in the second step, we further filtered the full - text of the articles. The specific process is as follows: First, an automated program was used to download the articles via the download links provided by the Semantic Scholar Dataset, and the remaining articles were downloaded manually. Subsequently, the full - text of these articles was input into gpt - 4o - mini (gpt - 4o - mini - 2024 - 07 - 18), and seven pre - set prompts were used for further filtering. These seven prompts were used to judge whether the article was related to the following contents: single-gene research, genomic research, siderophore transport, iron-affinity determination, siderophore receptors, chemical structure research, and microbial interactions. By requiring the model to return "Y" or "N" responses, we ultimately selected those articles that mainly focused on siderophore synthesis and its synthetic genes as the final dataset for siderophore information mining.

Search for Genomic Data and BGC Annotation

After extracting all the siderophore information mined from the articles by the model, we searched for the corresponding genomes based on the strain information therein. To do this, we designed a genomic search program by leveraging the API provided by NCBI. By inputting the strain names given by the model, this program can automatically download the genomic files of the corresponding strains from the NCBI Dataset. For cases where the strains have been renamed or the strain names in the articles contain redundant characters and thus cannot be matched with the NCBI Dataset, we manually download the corresponding genomic files from the NCBI website.

Once the genomic files are obtained, we annotate the corresponding siderophore biosynthetic gene clusters. In this process, we designed a BGC annotation tool based on antiSMASH 7.0, using the default parameters.

Comparing BGC Similarities

Using BiG-SCAPE as a tool for comparing the similarity of siderophore BGCs, we applied the -mix-type parameter for both NRPS and NIS, while other settings remained default.

Author contributions

Jiqi Shao: Conceptualization; data curation; visualization; writing - original draft. Yanzhao Wu: Data curation; visualization; coding and analysis; writing - original draft. Ningbo Xi: Data curation. Ruolin He: Data curation. Ruichen Xu: Data curation. Guo Peng: Data curation. Shaohua Gu: Conceptualization; funding acquisition. Zhiyuan Li: Conceptualization; funding acquisition; visualization; writing - original draft.

Acknowledgements

This work is supported by the National Key R&D Program of China (2024YFA0919500), National Natural Science Foundation of China (No. T2321001), and in part by the Peking-Tsinghua Center for Life Sciences.

Declaration of Interests

The authors declare no competing interests.

Reference

1. Page, M.G.P. (2019). The Role of Iron and Siderophores in Infection, and the Development of Siderophore Antibiotics. Clinical Infectious Diseases *69*, S529-S537. 10.1093/cid/ciz825.
2. Kramer, J., Özkaya, Ö., and Kümmerli, R. (2020). Bacterial siderophores in community and host interactions. Nature Reviews Microbiology *18*, 152–163. 10.1038/s41579-019-0284-4.
3. Gu, S., Wei, Z., Shao, Z., Friman, V.-P., Cao, K., Yang, T., Kramer, J., Wang, X., Li, M., Mei, X., et al. (2020). Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. Nat. Microbiol. *5*, 1002–1010. 10.1038/s41564-020-0719-8.
4. Gu, S., Shao, Z., Qu, Z., Zhu, S., Shao, Y., Zhang, D., Allen, R., He, R., Shao, J., Xiong, G., et al. (2025). Siderophore synthetase-receptor gene coevolution reveals habitat- and pathogen-specific bacterial iron interaction networks. Science Advances *11*, eadq5038. 10.1126/sciadv.adq5038.
5. Gu, S., Shao, Y., Rehm, K., Bigler, L., Zhang, D., He, R., Xu, R., Shao, J., Jousset, A., Friman, V.-P., et al. (2024). Feature sequence-based genome mining uncovers the hidden diversity of bacterial siderophore pathways. eLife *13*, RP96719. 10.7554/eLife.96719.
6. Gu, S., Shao, J., He, R., Xiong, G., Qu, Z., Shao, Y., Yu, L., Zhang, D., Wang, F., Xu, R., et al. (2025). Forging the iron-net: Towards a quantitative understanding of microbial communities via siderophore-mediated interactions. Quantitative Biology *13*, e84. https://doi.org/10.1002/qub2.84.
7. He, R., Gu, S., Xu, J., Li, X., Chen, H., Shao, Z., Wang, F., Shao, J., Yin, W.B., and Qian, L. (2024). SIDERITE: Unveiling hidden siderophore diversity in the chemical space through digital exploration. iMeta *3*, e192.
8. Zdouc, M.M., Blin, K., Louwen, N.L.L., Navarro, J., Loureiro, C., Bader, C.D., Bailey, C.B., Barra, L., Booth, T.J., Bozhüyük, K.A.J., et al. (2025). MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Res *53*, D678-d690. 10.1093/nar/gkae1115.
9. Chen, Y., Gao, J., Petruc, M., Hammer, R.D., Popescu, M., and Xu, D. (2024). Iterative Prompt Refinement for Mining Gene Relationships from ChatGPT. International Journal of Artificial Intelligence and Robotics Research, 2450005 %U https://www.worldscientific.com/doi/2450010.2451142/S2972335324500054.
10. Zheng, Z., He, Z., Khattab, O., Rampal, N., A. Zaharia, M., Borgs, C., T. Chayes, J., and M. Yaghi, O. (2024). Image and data mining in reticular chemistry powered by GPT-4V. Digital Discovery *3*, 491-501 %U https://pubs.rsc.org/en/content/articlelanding/2024/dd/d2023dd00239j.
11. Zheng, Z., Rong, Z., Rampal, N., Borgs, C., Chayes, J.T., and Yaghi, O.M. (2023). A GPT-4 Reticular Chemist for Guiding MOF Discovery. Angewandte Chemie

International Edition *62*, e202311983 %* © 202312023 Wiley-VCH GmbH %U https://onlinelibrary.wiley.com/doi/abs/202311910.202311002/anie.202311983.

12. Kalmer, T.L., Ancajas, C.M.F., Cheng, Z., Oyedele, A.S., Davis, H.L., and Walker, A.S. (2024). Assessing the ability of ChatGPT to extract natural product bioactivity and biosynthesis data from publications. bioRxiv %* © 2024, Posted by Cold Spring Harbor Laboratory. This pre-print is available under a Creative Commons License (Attribution 4.0 International), CC BY 4.0, as described at http://creativecommons.org/licenses/by/4.0/ %U https://www.biorxiv.org/content/10.1101/2024.08.01.606186v1.

13. Azam, M., Chen, Y., Arowolo, M.O., Liu, H., Popescu, M., and Xu, D. (2024). A Comprehensive Evaluation of Large Language Models in Mining Gene Interactions and Pathway Knowledge. bioRxiv %* © 2024, Posted by Cold Spring Harbor Laboratory. This pre-print is available under a Creative Commons License (Attribution-NonCommercial-NoDerivs 4.0 International), CC BY-NC-ND 4.0, as described at http://creativecommons.org/licenses/by-nc-nd/4.0/ %U https://www.biorxiv.org/content/10.1101/2024.01.21.576542v1.

14. Castro, A., Pinto, J., Reino, L., Pipek, P., and Capinha, C. (2024). Large language models overcome the challenges of unstructured text data in ecology. bioRxiv %* © 2024, Posted by Cold Spring Harbor Laboratory. This pre-print is available under a Creative Commons License (Attribution-NoDerivs 4.0 International), CC BY-ND 4.0, as described at http://creativecommons.org/licenses/by-nd/4.0/ %U https://www.biorxiv.org/content/10.1101/2024.01.23.576654v2.

15. Smith, N., Yuan, X., Melissinos, C., and Moghe, G. (2024). FuncFetch: An LLM-assisted workflow enables mining thousands of enzyme-substrate interactions from published manuscripts. bioRxiv %* © 2024, Posted by Cold Spring Harbor Laboratory. The copyright holder for this pre-print is the author. All rights reserved. The material may not be redistributed, re-used or adapted without the author's permission. %U https://www.biorxiv.org/content/10.1101/2024.07.22.604620v1.

16. Yu, J., Wang, Y., Wang, H., Wei, Z., and Pei, Y. (2024). Decoding Critical Targets and Signaling Pathways in EBV-Mediated Diseases Using Large Language Models. Viruses *16*, 1660 %* http://creativecommons.org/licenses/by/1663.1660/ %U https://www.mdpi.com/1999-4915/1616/1611/1660.

17. Wu, X., Zeng, Y., Das, A., Jo, S., Zhang, T., Patel, P., Zhang, J., Gao, S.-J., Pratt, D., Chiu, Y.-C., and Huang, Y. (2024). reguloGPT: Harnessing GPT for Knowledge Graph Construction of Molecular Regulatory Pathways. bioRxiv %* © 2024, Posted by Cold Spring Harbor Laboratory. This pre-print is available under a Creative Commons License (Attribution-NonCommercial-NoDerivs 4.0 International), CC BY-NC-ND 4.0, as described at http://creativecommons.org/licenses/by-nc-nd/4.0/ %U https://www.biorxiv.org/content/10.1101/2024.01.27.577521v1.

18. Reitz, Z.L., Butler, A., and Medema, M.H. (2022). Automated genome mining predicts combinatorial diversity and taxonomic distribution of peptide metallophore structures. 10.1101/2022.12.14.519525.

19. Baunach, M., Chowdhury, S., Stallforth, P., and Dittmann, E. (2021). The Landscape of Recombination Events That Create Nonribosomal Peptide Diversity. Molecular Biology

and Evolution *38*, 2116-2130. 10.1093/molbev/msab015.

20. Bozhüyük, K.A.J., Präve, L., Kegler, C., Schenk, L., Kaiser, S., Schelhas, C., Shi, Y.-N., Kuttenlochner, W., Schreiber, M., Kandler, J., et al. (2024). Evolution-inspired engineering of nonribosomal peptide synthetases. Science *383*, eadg4320. 10.1126/science.adg4320.

21. Stachelhaus, T., Mootz, H.D., and Marahiel, M.A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chemistry & Biology *6*, 493-505. https://doi.org/10.1016/S1074-5521(99)80082-9.

22. He, R., Zhang, J., Shao, Y., Gu, S., Song, C., Qian, L., Yin, W.-B., and Li, Z. (2023). Knowledge-guided data mining on the standardized architecture of NRPS: Subtypes, novel motifs, and sequence entanglements. PLOS Computational Biology *19*, e1011100. 10.1371/journal.pcbi.1011100.

23. Jenner, M., Hai, Y., Nguyen, H.H., Passmore, M., Skyrud, W., Kim, J., Garg, N.K., Zhang, W., Ogorzalek Loo, R.R., and Tang, Y. (2023). Elucidating the molecular programming of a nonlinear non-ribosomal peptide synthetase responsible for fungal siderophore biosynthesis. Nature Communications *14*, 2832. 10.1038/s41467-023-38484-8.

24. Tai, R.H., Bentley, L.R., Xia, X., Sitt, J.M., Fankhauser, S.C., Chicas-Mosier, A.M., and Monteith, B.G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. bioRxiv %* © 2024, Posted by Cold Spring Harbor Laboratory. The copyright holder for this pre-print is the author. All rights reserved. The material may not be redistributed, re-used or adapted without the author's permission. %U https://www.biorxiv.org/content/10.1101/2023.07.17.549361v2.

25. Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., Los Santos, E.L.C.d., Yeong, M., Cruz-Morales, P., Abubucker, S., et al. (2020). A computational framework to explore large-scale biosynthetic diversity. Nature Chemical Biology *16*, 60–68. 10.1038/s41589-019-0400-9.

26. Kautsar, S.A., van der Hooft, J.J.J., de Ridder, D., and Medema, M.H. (2021). BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience *10*. 10.1093/gigascience/giaa154.

27. Mongia, M., Baral, R., Adduri, A., Yan, D., Liu, Y., Bian, Y., Kim, P., Behsaz, B., and Mohimani, H. (2023). AdenPredictor: accurate prediction of the adenylation domain specificity of nonribosomal peptide biosynthetic gene clusters in microbial genomes. Bioinformatics *39*, i40-i46. 10.1093/bioinformatics/btad235.

28. Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Research *39*, W362-W367. 10.1093/nar/gkr323.

29. Chevrette, M.G., Aicheler, F., Kohlbacher, O., Currie, C.R., and Medema, M.H. (2017). SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics *33*, 3202-3210. 10.1093/bioinformatics/btx400.

30. Stanišić, A., Svensson, C.-M., Ettelt, U., and Kries, H. (2022). Defining a Nonribosomal Specificity Code for Design. 10.1101/2022.08.30.505883.

31. Zhang, W., Wang, Q., Kong, X., Xiong, J., Ni, S., Cao, D., Niu, B., Chen, M., Li, Y.,

Zhang, R., et al. (2024). Fine-tuning large language models for chemical text mining. Chemical Science *15*, 10600-10611 %U https://pubs.rsc.org/en/content/articlelanding/12024/sc/d10604sc00924j.

32. Kang, Y., and Kim, J. (2024). ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. Nature Communications *15*, 4705 %* 2024 The Author(s) %U https://www.nature.com/articles/s41467-41024-48998-41464.