Accelerating Antibiotic Discovery with Large Language Models and Knowledge Graphs

Maxime Delmas[1], Magdalena Wysocka[1,2], Danilo Gusicuma[1], André Freitas[1,2,3]

[1]Idiap Research Institute, Switzerland
[2]National Biomarker Centre (NBC), CRUK Manchester Institute, United Kingdom
[3]Department of Computer Science, University of Manchester, United Kingdom

Abstract

The discovery of novel antibiotics is critical to address the growing antimicrobial resistance (AMR). However, pharmaceutical industries face high costs (over $1 billion), long timelines, and a high failure rate, worsened by the rediscovery of known compounds. We propose an LLM-based pipeline that acts as an alert system, detecting prior evidence of antibiotic activity to prevent costly rediscoveries. The system integrates literature on organisms and chemicals into a Knowledge Graph (KG), ensuring taxonomic resolution, synonym handling, and multi-level evidence classification. We tested the pipeline on a private list of 73 potential antibiotic-producing organisms, disclosing 12 negative hits for evaluation. The results highlight the effectiveness of the pipeline for evidence reviewing, reducing false negatives, and accelerating decision-making. The KG for negative hits as well as the user interface for interactive exploration are available at https://github.com/idiap/abroad-kg-store and https://github.com/idiap/abroad-demo-webapp.

1 Introduction

Antibiotics are naturally occurring chemical compounds produced by organisms, known as natural products, that can inhibit the growth or eliminate bacteria and other microorganisms (Waksman, 1947). However, the introduction, use, and overuse of new antibiotics inevitably lead to the emergence of resistant pathogens (Altarac et al., 2021), and Antimicrobial Resistance (AMR) has been recognized as one of the top ten global public health threats (EClinicalMedicine, 2021). This ongoing cycle drives a continuous race to expand the antibiotic spectrum and treat patients infected with multidrug-resistant pathogens (MRPs) (Ahmed et al., 2024; Iskandar et al., 2022).

The development of new antibiotics is highly challenging (Payne et al., 2007; Altarac et al., 2021). The process has a high failure rate, and the total cost from identifying lead compounds to market approval can exceed $1 billion and take over a decade (Årdal et al., 2020; Wouters et al., 2020). In the initial phase, pharmaceutical companies explore ecosystems (Quinn and Dyson, 2024), searching for exotic organisms that produce novel bioactive compounds (see Figure 1). This phase involves identifying and isolating these compounds and evaluating their activity against MRPs. Identifying promising lead compounds (those with the highest potential for success) can already require over $1 million and years of research (Årdal et al., 2018). A major challenge in this early phase is avoiding rediscovery scenarios, when a potentially active compound has already been reported in scientific literature or patent databases. Such prior knowledge often eliminates the compound's commercial value by removing its novelty. In addition, one can consider that if an active molecule produced by an organism is publicly known but not already commercialized, it is likely that it has already been tested but failed in later clinical stages. Therefore, ensuring comprehensive awareness of existing research is critical to avoid costly investments in non-viable targets. As stated by (Paul et al., 2010), if a candidate has to fail, it is critical to it make fail faster and less expensively.

Preventing rediscoveries requires an extensive review of scientific literature, databases, and patents related to the initial list of target organisms. This task is firstly complicated by the unstable taxonomy and nomenclature of organisms (Beninger and Backeljau, 2019). Many organisms have been repeatedly rediscovered and reclassified under different names. For instance, *Cephalosporium acremonium*, *Hyalopus acremonium*, *Acremonium strictum* and *Sarocladium strictum*, published in 1882, 1941, 1971 and 2011 respectively, all refer to the same organism under the most recent classification. To capture relevant data, literature reviews must

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 693–705
July 28-30, 2025 ©2025 Association for Computational Linguistics

expand the search for such synonyms.

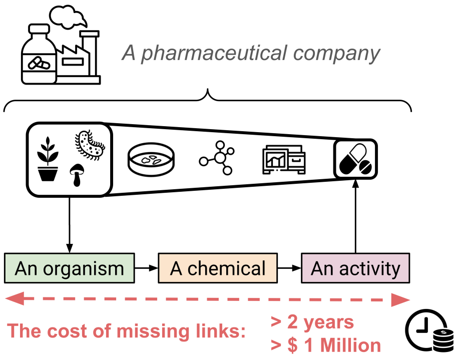


Figure 1: An overview of the early phase of antibiotic development and the cost attached to lead compounds identification.

Evidence of prior activity can appear in diverse forms. Some references from the literature of the organism describe its activity without identifying specific active compounds, e.g., “The culture of A inhibited the growth of *Staphylococcus aureus*." Others may report the isolation of a compound from the organism without detailing its biological activity ("Compound C was isolated from organism A"), requiring a 2-hop search for chemical activity evidence (e.g. Compound C exhibited antibacterial activity against *Staphylococcus aureus*.")

This review process is traditionally manual and extremely time-consuming. Allen and Olkin (1999) previously estimated that over 1,000 hours may be required to review 2,500 citations. There is a need for semi-automation given the expanding scientific literature and the high cost of false negatives. In this context, large language models (LLMs) have emerged as powerful tools for assisting literature reviews, particularly in the biomedical domain (Wysocka et al., 2024; Yun et al., 2023; Liao et al., 2024; Hsu et al., 2024). Beyond review, an effective solution would serve as an alert system, flagging previously reported antibiotic activities associated with target organisms. Compared to novelty detection (Ghosal et al., 2022), we rather seek for non-novelty detection for relations between organisms, chemicals, and activities.

In this work, we propose an LLM-based pipeline to automate the construction of such an alert system. The system is based on a Knowledge Graph (KG), ensuring taxonomic and nomenclature resolution, interoperability between natural product resources, and classification of evidence into three alert levels. We demonstrate the system in a real industrial setting using a private input list of 73 organisms, identifying 12 negative hits that were used to evaluate the system’s performance.

2 Data

Our dataset is composed of an initial private list of 73 organism identifications, from which we disclosed 12 negative hits for evaluation after evidence of already reported activity have been found. This review was conducted by a team of three experts, using public literature (PubMed), databases (eg. LOTUS (Rutz et al., 2022)) and proprietary tools (eg. CAS SciFinder (Gabrielson, 2018)). See details in appendix A. From this analysis, 27 evidence triples *organism-chemical-activity* had been identified for the 12 negative hits by the experts. For the proposed alert system, we excluded proprietary resources and decided to primary focus on two large public resources: PubMed and LOTUS. LOTUS is an open, community-curated database containing over 750,000 structure-organism pairs which is hosted on the Wikidata KG. Taxononomic and nomenclature information of organisms are extracted from the GBIF backbone taxonomy (GBIF Secretariat, 2023), a comprehensive and synthetic classification that integrates taxonomic data from multiple sources.

3 Methodology

This section provides a step-by-step description of the pipeline represented in Figure 2. The input is a list of user-defined organism identifications. Identifications can be specific, at the species level (e.g., *Aspergillus calidoustus*), or unspecific (represented by the abbreviation *sp.*), indicating an undetermined species within a genus[1] (e.g., *Aspergillus sp.*).

In step *(1)*, each identification is aligned with an entity in the GBIF taxonomy. Species-level identifications are expanded to include all known synonyms, while genus-level identifications are expanded to encompass all species within the genus and their respective synonyms. In step *(2)*, abstracts and relevant paragraphs from PubMed full-text articles are retrieved using the NCBI EUtils API[2].

Step *(3)* filters the organism literature to exclude articles irrelevant for antibiotic activity (AA) evidence extraction (e.g., ecology, environmental

[1] A genus is a taxonomic rank grouping species that share common characteristics.

[2] https://dataguide.nlm.nih.gov/eutilities/utilities.html

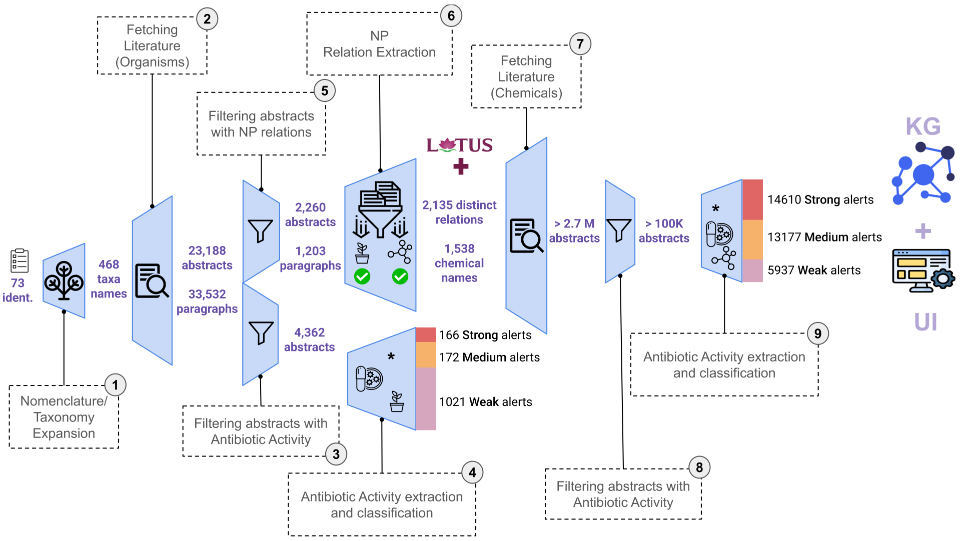


Figure 2: An illustration of the proposed pipeline, step-by-step, from the initial list of organism identifications to the extraction of AA evidence alerts in 3 levels. Intermediary annotations (in **purple**) describe the flow of literature, relations, and evidence that have been processed.

studies, genetics). A lightweight lexical classifier, trained on MeSH[3] annotations, ensures efficient filtering. In step *(4)* we prompt the LLM (Mixtral-8x7b (Jiang et al., 2024)) for Zero-shot extraction of AA evidence from the selected abstracts (Kojima et al., 2022). These evidence, derived solely from the organism's literature, are designated as `OL`-evidence (Organism-Literature). Evidence are then categorized into three alert levels: `Strong` (direct experimental evidence of activity), `Medium` (indirect, imprecise, or minor evidence), and `Weak` (no substantial evidence) using the LLM. More details about the prompting strategy and concrete examples in appendix B.

Steps *(5)* to *(7)* focus on identifying chemicals isolated from the organisms. Similar to *(3)*, step *(5)* filters literature to retain only texts likely to report chemical isolations. Since MeSH annotations are unavailable for this task, we used LLM-generated pseudo-labels to train a second lexical classifier (Wang et al., 2023). Details on the classifiers used for filtering are provided in Appendix C.

In step *(6)*, a natural products Relation Extraction (RE) model (Delmas et al., 2024) (fine-tuned from BioMistral-7B (Labrak et al., 2024)) processes selected passages to extract natural products relations (`NPR`). These relations are sourced from abstracts (`TiabNPR`) or paragraphs (`ChunkNPR`), then augmented with relations from the LOTUS database (`LotusNPR`).

Steps *(7)* to *(9)* mirror steps *(2)* to *(4)*, but use the extracted chemical names as input. This produces a prioritized list of chemical literature evidence (`CL`-evidence), categorized into the same three alert levels.

All processed data, including nomenclature, relations, literature, and alerts, are integrated into a Knowledge Graph (KG) using a dedicated ontology (see appendix E). Figure 3 provides a snapshot centred on the example of *Sarocladium strictum* and its active metabolite *Cephalosporin C*. The KG supports transparent resolution of taxonomic and synonym relations (e.g. *Sarocladium strictum* `hasSynonymTaxon` *Cephalosporium acremonium*), ensures interoperability between sources of relations (`LotusNPR`, `TiabNPR`, `ChunkNPR`), and, differentiates evidence origins (`OL` vs. `CL`) and alert levels (`Strong`, `Medium`, `Weak`).

4 Results

4.1 Natural products literature: descriptive bibliometric analysis

Assessing the size and growth of the natural products and antibiotics literature is crucial to highlight the extensive effort required by reviewers. In 2024, it is more than 50,000 new articles that have been indexed in PubMed for the searches "natural products" and "antibiotics", reporting novel links between organisms, chemicals, and activities. While keeping up with new literature is crucial, Figure 4.A shows that a significant portion of annotated re-

[3] MeSH are standardized biomedical indexing terms in PubMed.

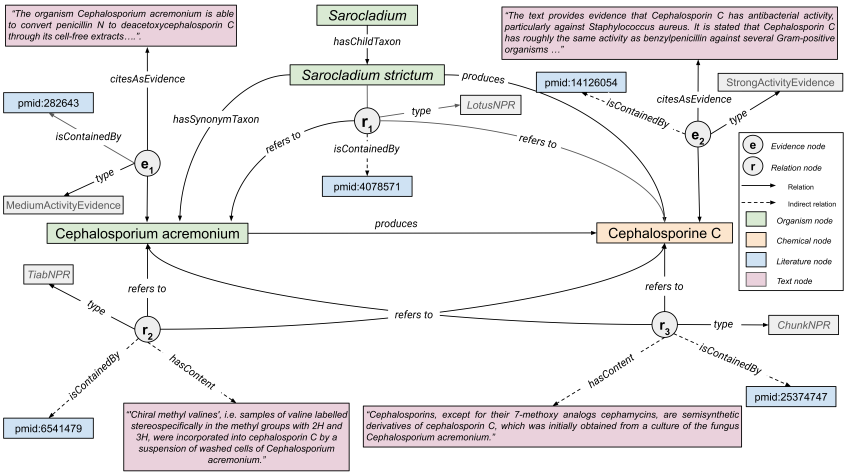

"The organism Cephalosporium acremonium is able to convert penicillin N to deacetoxycephalosporin C through its cell-free extracts….".
Sarocladium
hasChildTaxon
"The text provides evidence that Cephalosporin C has antibacterial activity, particularly against Staphylococcus aureus. It is stated that Cephalosporin C has roughly the same activity as benzylpenicillin against several Gram-positive organisms …"
citesAsEvidence
Sarocladium strictum
produces
pmid:14126054
StrongActivityEvidence
pmid:282643
hasSynonymTaxon
type
LotusNPR
isContainedBy
citesAsEvidence
isContainedBy
refers to
r1
refers to
e2
type
isContainedBy
e1
type
pmid:4078571
MediumActivityEvidence
e Evidence node
r Relation node
Relation
Indirect relation
Organism node
Chemical node
Literature node
Text node
Cephalosporium acremonium
produces
Cephalosporine C
TiabNPR
refers to
refers to
type
refers to
r2
r3
type
ChunkNPR
isContainedBy
hasContent
hasContent
isContainedBy
pmid:6541479
"'Chiral methyl valines', i.e. samples of valine labelled stereospecifically in the methyl groups with 2H and 3H, were incorporated into cephalosporin C by a suspension of washed cells of Cephalosporium acremonium."
"Cephalosporins, except for their 7-methoxy analogs cephamycins, are semisynthetic derivatives of cephalosporin C, which was initially obtained from a culture of the fungus Cephalosporium acremonium."
pmid:25374747

Figure 3: A snapshot of the built KG around the natural product relation between *Cephalosporium acremonium* and Cephalosporine C. Taxonomic and nomenclature relationships are represented between Organism nodes in green. Relation nodes (r_1, r_2, r_3) describe relations between organisms and the isolated natural product Cephalosprin C from different sources: LOTUS database (`LOTUSNPR`) and extracted from an abstract (`TiabNPR`) and a paragraph (`ChunkNPR`). Text nodes connected to relation nodes (r_2, r_3) refer to the text from which the relation was extracted. The evidence node e_1 is an example of `OL`-evidence associated with a `Medium` alert. The node e_2 is a `CL`-evidence associated with a `Strong` alert. Literature node connected to relation and evidence nodes allow for linked to the original reference in PubMed (or using the DOI if not available in the case of LOTUS annotations).

lations in the LOTUS database comes from older articles (pre-1970). Given the evolution of taxonomy and nomenclature over time, relying on original organism identifications from the text is unreliable, making synonym resolution essential for linking past and novel relations. Using the publicly available literature from PubMed as a reference for an alert system also requires evaluating its coverage. Although PubMed includes over 38 millions articles, Figure 4.B indicates that fewer than half of the annotated references in the LOTUS database are actually indexed in PubMed. This observation underscores a notable gap in PubMed's coverage. Nevertheless, given the extensive volume of literature within PubMed, it's also reasonable to expect that many relevant references may be missing from LOTUS. Also, while we observed that most articles are in English, this likely reflect a bias from the resources used in LOTUS, and, other corpora (eg. traditional Chinese medicine prescriptions) are also expected to be relevant.

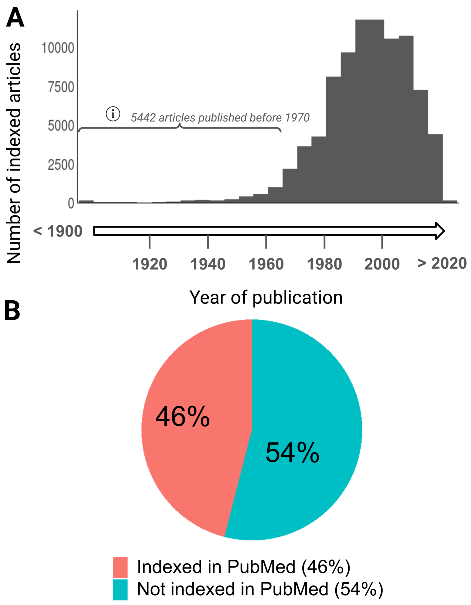


Figure 4: **A** describes the distribution of publication years for literature references annotated in the LOTUS database. Panels **B** represents the distribution of references indexed in PubMed for natural products relations annotated in LOTUS.

A notable example of the last points is Atranorin, an anti-inflammatory, analgesic, and antibacterial compound, isolated from *Gyrophora esculenta* (now named *Umbilicaria esculenta*), described in German by Hoppe (1958).

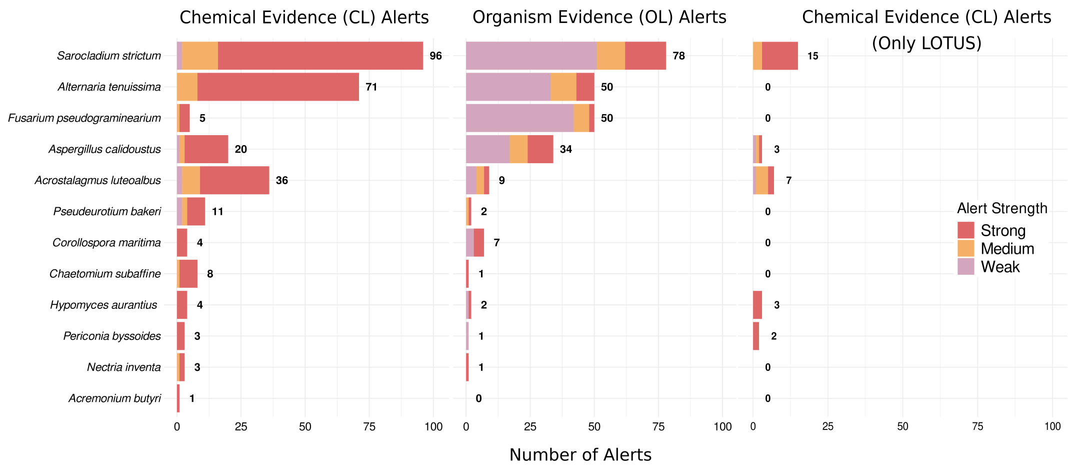


Figure 5: Distribution of all reported alerts per class (`Strong`, `Medium` and `Weak`) and categories for `CL` (left) and `OL` (center) evidence for the 12 discarded organisms. The right panel describes the reported evidence only using the LOTUS available natural products relations.

4.2 Pipeline execution

Starting from 73 initial identifications, the flow of extracted and processed literature is outlined in Figure 2. Over 50,000 paragraphs were processed, yielding 2,135 organism-chemical relations and 1,359 alerts directly from the literature of organisms. Expanding to the literature of identified chemicals, more than 2.7 million abstracts were processed, resulting in 33,724 alerts for potential antibacterial activity.[4]

4.3 Evaluation on Discarded Hits

Among the 73 initial identifications, 12 were discarded as negative hits after an extensive manual review. Figure 5 displays the distribution of alerts raised for each discarded organism from `CL` (left) and `OL` (center) evidence. While the number of alerts varies (max: 174, min: 1), each organism has at least one `Strong` alert. To assess the impact of the extraction pipeline, an ablation study (Figure 5 right) using only LOTUS database annotations showed that only 5 of the 12 negative hits could be identified, highlighting the added value of the RE step. For the 12 negative hits, the reviewers previously identified 27 evidence triples (*organism-chemical-activity*). Table 1 compares these with system-generated alerts from Figure 5, focusing on chemical-based alerts, as all evidence provided by the reviewers are linked to a chemical. An alert is considered missed if the chemical was not retrieved (via RE or LOTUS) or its activity was not reported[5]. Among the 27 reviewer-reported evidence, 6 were missed by the system, including 3 because of non-indexed references or unavailable texts in PubMed. Notably, 26 of the 27 chemicals were successfully retrieved, with 22 through the RE step. A detailed error analysis is provided in Appendix D. Except for *Acremonium butyri*, all negative hits were correctly discarded. Screenshots of the user interface, including an example for *Sarocladium strictum*, are shown in Appendix F.

5 Discussion

Most alert-associated chemicals were extracted from the public literature, suggesting an underestimation of PubMed's coverage in section 4.1, and, highlighting gaps in public databases, particularly for rarely mentioned organisms. However, given the nature of the task, and the cost of false negatives (e.g., *Acremonium butyri*), public resources alone are insufficient to prevent rediscoveries. Notably, half of the missing evidence could have been recovered by incorporating non-publicly accessible literature, beyond PubMed and LOTUS. From the initial set of 73 organisms, over 35,000 alerts were generated, which, paradoxically, could overwhelm the reviewers. To mitigate this, the prioritization system, categorizing evidence into `Strong`, `Medium`, and `Weak`, is essential for the reviewing process. Interestingly, in only 9 of the 27 evidence reported by the annotators, the activity of the chemical was reported in the same article as its isolation. This highlights the need for extending the search to the literature of individual

[4] Many alerts stem from genus-level identifications, which expand to numerous species.

[5] Neither `Strong`, `Medium`

Organisms	Chemicals	PubMed ID Isolation	PubMed ID Activity	RE / LOTUS	CL-evidence
A. butyri	Orbuticin	8982351	8982351	✓/✓	Missed
A. luteoalbus	Acrozine A-C	31226467	31226467	✓/✓	Strong
A. luteoalbus	T988 C	35621985	35621985	×/×	Missed
A. luteoalbus	Lasiodipline E	37627256	24529576	✓/×	Strong
A. luteoalbus	luteoalbusin A	23079524	35621985	✓/✓	Missed
A. tenuissima	Altertoxin I, II, III	25260957	37764307	✓/×	Strong
A. tenuissima	Tenuazonic acid	34575812	34575812	✓/×	Strong
A. tenuissima	Alternariol mono. ether	24071643	38470179	✓/×	Strong
A. calidoustus	Ophiobolin K	25812930	29375031	✓/×	Strong
A. calidoustus	Strobilactone A	8698631	*ext. ref(1)*	×/✓	Missed
S. strictum	Cephalosporin C	10397815	14126054	✓/✓	Strong
S. strictum	Isopenicillin N	575040	7107525	✓/✓	Strong
S. strictum *	Cytosporone E	29354097	22690142	✓/×	Strong
C. subaffine	Chrysophanol	35761187	25821480	✓/×	Strong
C. maritima	Corollosporine	16557326	16557326	✓/×	Strong
F. pseudograminearum	Deoxynivalenol	35878241	38408410	✓/×	Strong
F. pseudograminearum	Zearalenone	24291181	37929585	✓/×	Strong
H. aurantius *	Cladobotryal	9586194	12934912	×/✓	Strong
H. aurantius *	Furopyridine antibiotics	11918067	11918067	✓/×	Strong
H. aurantius	Hypomycetin	*ext. ref(2)*	*ext. ref(2)*	×/✓	Missed
N. inventa	Chaetocin	31569621	21140472	✓/×	Strong
N. inventa	Verticillin B	31569621	31569621	✓/×	Missed
P. byssoides	Pericosine A	18043803	26928999	✓/✓	Strong
P. byssoides	Macrosphelide A	15895526	19298513	×/✓	Strong
P. bakeri	Cytochalasin X	35841670	35841670	✓/×	Strong
P. bakeri	Chaetoglobosin B	36104717	26669098	✓/×	Strong
P. bakeri	Chaetoglobosin A	36104717	26669098	✓/×	Strong

Table 1: Comparison of reviewers extracted CL-evidence and system-extracted evidence for each discarded hits. When an organism is marked with a *, it indicates that the chemical has been retrieved for a synonym (eg. *Cladobotyryum varium* in the case of *Hypomyces aurantius*). "PubMed ID Isolation" and "PubMed ID Activity" list PubMed references for chemical isolation and antibiotic activity extracted by reviewers. The "RE/LOTUS" column uses a tick (✓) and a cross (×) to show whether the relationship organism-chemical is present or missing. The left symbol represents extraction from the Relation Extraction (RE) pipeline, while the right symbol indicates whether it is annotated in the LOTUS database. CL-evidence indicates the system's alert level (Strong, Medium, Weak, or Missed). Ext. ref(1) and ext. ref(2) are non-PubMed references: doi:10.1515/znb-2007-1218 and 10.3891/acta.chem.scand.51-0855.

chemicals, and reflects the 2-hop nature of the task. Moreover, accurate nomenclature resolution, inherently supported by the KG, remains critical. This is exemplified by the case of *Hypomyces aurantius*, where key evidence were retrieved under its synonym *Cladobotryum varium*. While a single (Strong) evidence is enough to discard an organism, comparing Table 1 and Figure 5 suggests that many pieces of evidence may have been overlooked by reviewers, considering the vast amount of literature to examine. Paradoxically, in the proposed scenario, a "positive" result is therefore an "empty" result, such that no external evidence was found to challenge the novelty. Finally, the versatility of LLMs has been instrumental in the development of the system, particularly for Zero-shot inference, reasoning-based activity extraction, and pseudo-labeling (see 3). This adaptability was crucial due to the lack of pre-existing models designed for such tasks. LLMs clearly open new opportunities for assisting large literature reviews in the pharmaceutical domain and, more broadly, across the biomedical domain. However, LLMs are also prone to hallucinations and can misinterpret evidence from the source text (*context inconsistency* (Huang et al., 2025)). While incorrect associations between organisms and natural products, or misidentified antibiotic activity evidence, can lead to false positives, it is the omission of such relations that is more detrimental for the alert system by introducing false negatives. Various strategies have been proposed to mitigate these errors in biomedical texts, such as adapting the decoding process (Xu et al., 2024) or incorporating a self-reflection mechanism (Ji et al., 2023).

6 Conclusion

Avoiding rediscoveries and dead-end paths is crucial in industrial antibiotic developments, saving time and resources. Yet, this process is itself resource-intensive, highlighting the need for semi-automatic reviewing. We present a practical application of LLMs to build an alert system that, given a list of organisms, flags evidence of previously reported activity from both the organism and chemical literature. We demonstrated the value of the system using 12 disclosed organisms and identified key factors: literature coverage, efficient natural products RE, synonym resolution and alert prioritization. The subset of the KG related to the negative hits, along with the code to reproduce the user interface and explore the results interactively, are available at `https://github.com/idiap/abroad-kg-store` and `https://github.com/idiap/abroad-demo-webapp`.

Acknowledgments

The authors are thankful to Vincent Mutel and his collaborators from Inflamalps, as well as to Joël Dumoulin, Joel Rossier, and Colombine Verzat for their help during the project.

References

Sirwan Khalid Ahmed, Safin Hussein, Karzan Qurbani, Radhwan Hussein Ibrahim, Abdulmalik Fareeq, Kochr Ali Mahmood, and Mona Gamal Mohamed. 2024. Antimicrobial resistance: Impacts, challenges, and future prospects. *Journal of Medicine, Surgery, and Public Health*, 2:100081.

I. E. Allen and I. Olkin. 1999. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA*, 282(7):634–635.

David Altarac, Michael Gutch, John Mueller, Matthew Ronsheim, Ruben Tommasi, and Manos Perros. 2021. Challenges and opportunities in the discovery, development, and commercialization of pathogen-targeted antibiotics. *Drug Discovery Today*, 26(9):2084–2089.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Peter G. Beninger and Thierry Backeljau. 2019. Understanding taxonomic and nomenclatural instability – a case study of the Manila clam. *Aquaculture*, 504:375–379.

Maxime Delmas, Magdalena Wysocka, and André Freitas. 2024. Relation extraction in underexplored biomedical domains: A diversity-optimized sampling and synthetic data generation approach. *Computational Linguistics*, 50(3):953–1000.

EClinicalMedicine. 2021. Antimicrobial resistance: a top ten global public health threat. *eClinicalMedicine*, 41:101221.

Stephen Walter Gabrielson. 2018. SciFinder. *Journal of the Medical Library Association : JMLA*, 106(4):588–590.

GBIF Secretariat. 2023. GBIF Backbone Taxonomy.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, 48(1):77–117.

Heinz A. Hoppe. 1958. Galanthus nivalis– Gyrophora esculenta. In *Drogenkunde*, pages 402–434. De Gruyter.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Wang, and Aakanksha Naik. 2024. CHIME: LLM-Assisted Hierarchical Organization of Scientific Studies for Literature Review Support. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 118–132, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Katia Iskandar, Jayaseelan Murugaiyan, Dalal Hammoudi Halat, Said El Hage, Vindana Chibabhai, Saranya Adukkadukkam, Christine Roques, Laurent Molinier, Pascale Salameh, and Maarten Van Dongen. 2022. Antibiotic Discovery and Resistance: The Chase and the Race. *Antibiotics*, 11(2):182.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *arXiv preprint*. ArXiv:2411.05025 [cs].

Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214. Publisher: Nature Publishing Group.

David J. Payne, Michael N. Gwynn, David J. Holmes, and David L. Pompliano. 2007. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Reviews Drug Discovery*, 6(1):29–40. Publisher: Nature Publishing Group.

Gerry A. Quinn and Paul J. Dyson. 2024. Going to extremes: progress in exploring new environments for novel antibiotics. *npj Antimicrobials and Resistance*, 2(1):1–9. Publisher: Nature Publishing Group.

Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, Christoph Steinbeck, Guido F Pauli, Jean-Luc Wolfender, Jonathan Bisson, and Pierre-Marie Allard. 2022. The LOTUS initiative for open knowledge management in natural products research. *eLife*, 11:e70780.

Selman A. Waksman. 1947. What Is an Antibiotic or an Antibiotic Substance? *Mycologia*, 39(5):565–569. Publisher: Mycological Society of America.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Olivier J. Wouters, Martin McKee, and Jeroen Luyten. 2020. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853.

Magdalena Wysocka, Oskar Wysocki, Maxime Delmas, Vincent Mutel, and André Freitas. 2024. Large Language Models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *Journal of Biomedical Informatics*, 158:104724.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7744–7757, Miami, Florida, USA. Association for Computational Linguistics.

Hye Yun, Iain Marshall, Thomas Trikalinos, and Byron Wallace. 2023. Appraising the potential uses and harms of LLMs for medical systematic reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10122–10139, Singapore. Association for Computational Linguistics.

Christine Årdal, Manica Balasegaram, Ramanan Laxminarayan, David McAdams, Kevin Outterson, John H. Rex, and Nithima Sumpradit. 2020. Antibiotic development — economic, regulatory and societal challenges. *Nature Reviews Microbiology*, 18(5):267–274.

Christine Årdal, Enrico Baraldi, Ursula Theuretzbacher, Kevin Outterson, Jens Plahte, Francesco Ciabuschi, and John-Arne Røttingen. 2018. Insights into early stage of antibiotic development in small- and medium-sized enterprises: a survey of targets, costs, and durations. *Journal of Pharmaceutical Policy and Practice*, 11:8.

A Manual review and evaluation

The review was conducted by a team of three experts (one biologist and two chemists) over several weeks (> 400 hours). In the process, they used PubMed, GBIF, CAS SciFinder (Gabrielson, 2018), and LOTUS (Rutz et al., 2022). CAS SciFinder, a proprietary tool, facilitating the retrieval of scientific literature and patents related to chemical names and structures.

In the initial phase, reviewers examined literature associated with the target organisms, focusing on `OL`-evidence and chemicals produced by the organisms (natural products). They also used GBIF to retrieve associated synonyms, and the LOTUS database to extend the search for natural products. As expected, few matches were found with the database, as the initial organism selection only involved weakly characterized organisms. No filters were applied to the original studies, but, only secondary metabolites were retained and primary metabolites (those involved in growth, development or other essential pathways) were automatically excluded.

For each organism, reviewers compiled a list of compounds and primarily relied on SciFinder to explore associated literature and patents. Any evidence of antibiotic activity (growth inhibition, organism elimination, etc.) was considered as a hit, even if quantitative measurements (e.g., IC50 values) were not specified. From these steps, they identified 27 *organism-chemical-activity* evidence triples corresponding to the 12 disclosed negative hits.

The reviewers emphasized that the first phase, identifying related natural products, is critical. Once compounds were identified, resources like SciFinder, alongside with expert knowledge, provide a detailed overview of the compounds' properties, literature, and associated patents. Nevertheless, the initial link between the organism and its chemical compounds remained often poorly documented. Finally, the goal is not to identify exhaustively all active molecules, rather, only identifying one or a few associated active compounds is sufficient to discard the organism.

B Activity Evidence Classification

Concrete examples of `Strong`, `Medium` and `Weak` antibiotic evidence alerts, extracted using the prompting strategy described in Figure 6.

Strong Activity Evidence: Cephalosporin C The following evidence text has been extracted and classified as `Strong` from `PMID:4078571`: *The text provides evidence that Cephalosporin C has antibacterial activity, particularly against Staphylococcus aureus. It is stated that Cephalosporin C has roughly the same activity as benzylpenicillin against several Gram-positive organisms and about one-eighth of the activity of benzylpenicillin against penicillin-sensitive strains of Staphylococcus aureus. Additionally, Cephalosporin C shows 4 to 8 times the activity of methicillin against penicillinase-producing staphylococcal strains. It also exhibits synergism in protection experiments in mice infected with a strong penicillinase-producing strain of Staphylococcus aureus when combined with benzylpenicillin.*". Here, the evidence of activity is supported by quantitative measurements.

Medium Activity Evidence: Cephalosporin C The following evidence text has been extracted and classified as `Medium` from `PMID:22136576`: *The evidence of the potential antibiotic activity of Acremostrictin is found in the statement "The new compound exhibited weak antibacterial activities." This suggests that Acremostrictin showed some level of antibacterial effect, although it was classified as weak.* Here, the article only reports weak antibacterial activity.

Weak Activity Evidence: Dipeptide delta-(L-alpha-aminoadipyl)-L-cysteine The following evidence text has been extracted and classified as `Weak` from `PMID:6684424`: *The text describes the biosynthesis of two compounds, the tripeptide delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine and the dipeptide delta-(L-alpha-aminoadipyl)-L-cysteine, using a cell-free extract of Cephalosporium acremonium. However, it does not provide any information about the potential antibiotic activity of the dipeptide delta-(L-alpha-aminoadipyl)-L-cysteine. Therefore, there is No evidence found in this text to support the potential antibiotic activity of this chemical compound.*

C Filtering Classifiers

Considering the massive amount of literature to be processed for both `NPR` and activity extraction, it is essential to integrate a pre-filtering step to exclude out-of-scope references. It is also particularly essential for the RE step, which uses a decoder-only architecture, where sending out-of-distribution abstracts (not mentioning any relations) lead to chaotic outputs.

C.1 `NPR` Filtering

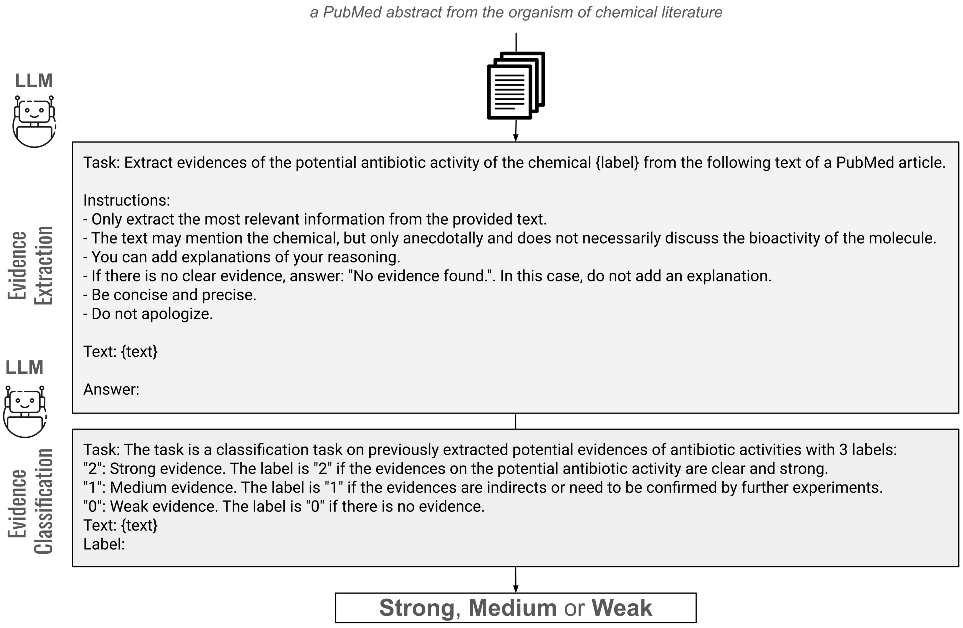
From the LOTUS database, we extracted the top-200 organism entities with the most associated relations and extracted 5k annotated abstracts, completed with 5k other abstracts not indexed in LOTUS. As LOTUS relations may not have been reported from the abstract (but from the full-text for instance) we annotated the dataset with LLM-generated pseudo-labels (prompt in Figure 7). We trained a simple lexical Naive Bayes classifier and compared the performance against more complex transformer architecture (BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019)) in Table 2.

Model	Recall	Precision	F1
Naive Bayes	96.8	77.9	86.4
BioBERT	89.8	91.6	90.6
SciBERT	91.1	88.3	89.7

Table 2: Performance comparison of different models on `NPR` classification.

C.2 Activity Filtering

While MeSH terms index articles in PubMed with relevant concepts such as *Anti-Bacterial Agents*,



Task: Extract evidences of the potential antibiotic activity of the chemical {label} from the following text of a PubMed article.

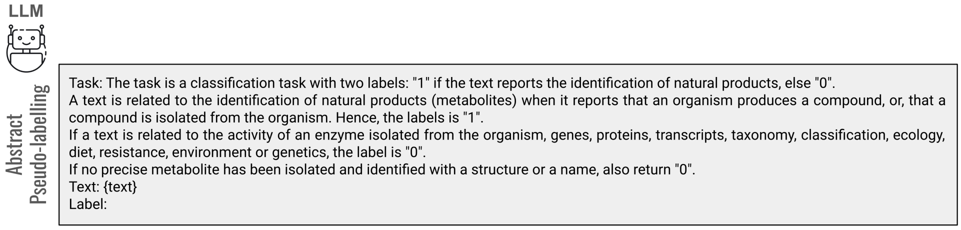
Instructions:
- Only extract the most relevant information from the provided text.
- The text may mention the chemical, but only anecdotally and does not necessarily discuss the bioactivity of the molecule.
- You can add explanations of your reasoning.
- If there is no clear evidence, answer: "No evidence found.". In this case, do not add an explanation.
- Be concise and precise.
- Do not apologize.

Text: {text}

Answer:

Task: The task is a classification task on previously extracted potential evidences of antibiotic activities with 3 labels:
"2": Strong evidence. The label is "2" if the evidences on the potential antibiotic activity are clear and strong.
"1": Medium evidence. The label is "1" if the evidences are indirects or need to be confirmed by further experiments.
"0": Weak evidence. The label is "0" if there is no evidence.
Text: {text}
Label:

Figure 6: Schema of the prompting strategy for the extraction and classification of antibiotic activity evidence from the literature of chemicals (the strategy is equivalent for the literature of organisms). When an organism has multiple synonyms, evidence extraction is performed independently for each synonym based on its associated literature. For chemicals, we rely on the labels provided by LOTUS or those extracted by the RE model. No synonym resolution is applied to chemicals.



Task: The task is a classification task with two labels: "1" if the text reports the identification of natural products, else "0".
A text is related to the identification of natural products (metabolites) when it reports that an organism produces a compound, or, that a compound is isolated from the organism. Hence, the labels is "1".
If a text is related to the activity of an enzyme isolated from the organism, genes, proteins, transcripts, taxonomy, classification, ecology, diet, resistance, environment or genetics, the label is "0".
If no precise metabolite has been isolated and identified with a structure or a name, also return "0".
Text: {text}
Label:

Figure 7: Prompt instructions for pseudo-labeling of natural products relationships.

most recent articles are not indexed, which can be critical for the alert system. Therefore, given the previously extracted top-200 organisms and their associated chemicals, we extracted their abstracts along with the MeSH annotations to build our dataset. We considered every article indexed with the concept *Anti-Bacterial Agents* (or narrower in the hierarchy) as positive examples and the rest as negatives. From the total set, we re-sampled 5k positives and negatives. Similarly to C.1 we trained a Naive Bayes classifier, BioBERT and SciBERT models (see Table 3)

As expected, simple lexical approaches compete in practice with more complex transformers architecture, given the simplicity of the task. Indeed, in both cases, a keyword matching strategy is sufficient to efficiently classify the abstracts. We logically decided to use the simpler Naive Bayes Classifer in both cases

Model	Recall	Precision	F1	F2
Naive Bayes	94.2	90.2	92.2	93.4
BioBERT	96.8	94.9	95.8	96.4
SciBERT	96.8	95.6	96.2	96.5

Table 3: Performance comparison of different models on AA classification.

D Error analysis

This section provides a detailed error analysis on the 6 evidence the system failed to retrieve.

Acremonium butyri - Orbuticin: While the chemical has been correctly extracted from the title of PMID:8982351 the abstract and full-text of the article are not publicly available on PubMed, hence the system failed to extract the activity. The reported

Strong evidence *Acremonium butyri* in Figure 5 actually refers to "Isoprenoids", which is a chemical family and not a single molecule. The Strong evidence is erroneously linked to articles reporting that the biosynthesis pathway for Isoprenoids is a target for many antibiotics.

Acrostalagmus luteoalbus - T988 C: The RE model failed to extract the natural product from PMID:35621985. This relation is also not annotated in LOTUS.

Acrostalagmus luteoalbus - Luteoalbusin A: The chemical has been correctly extracted from PMID:35621985 but the activity information from PMID:35621985 have not been extracted as only the abstract was processed.

Aspergillus calidoustus - Strobilactone A: The article reporting the relation in LOTUS is not publicly available (DOI:A10.7164/antibiotics.49.505)

Hypomyces aurantius - Hypomycetin: The reference article identified by the reviewers (DOI:10.3891/acta.chem.scand.51-0855) is indexed in LOTUS. This article also describes the antifungal activity of Hypomycetin. However, since the article is not indexed in PubMed, the evidence of its activity has not been extracted.

Nectria inventa - Verticillin B: The relation has correctly been identified in PMID:31569621, but the activity information from PMID:31569621 have not been extracted as only the abstracts are processed.

E Ontology schema

Figure 8 presents the main classes and properties of the proposed ontology used in the KG.

F Screenshots of the User Interface

Figures 9 and 10 present screenshots of the user interface.

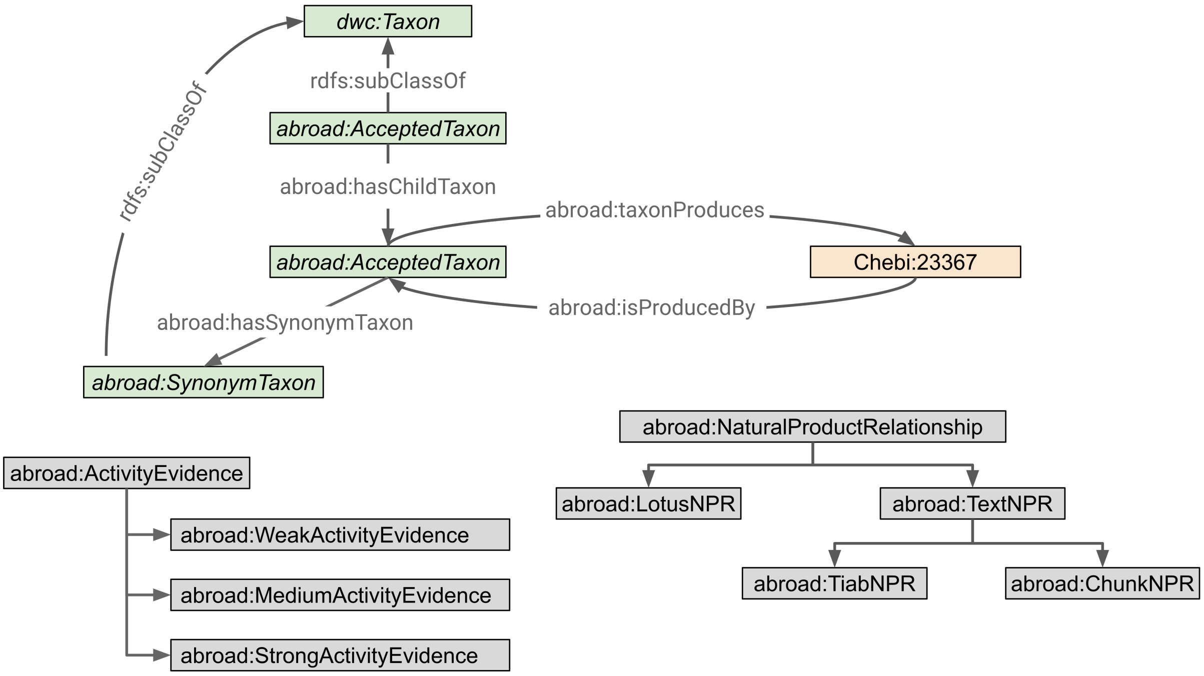
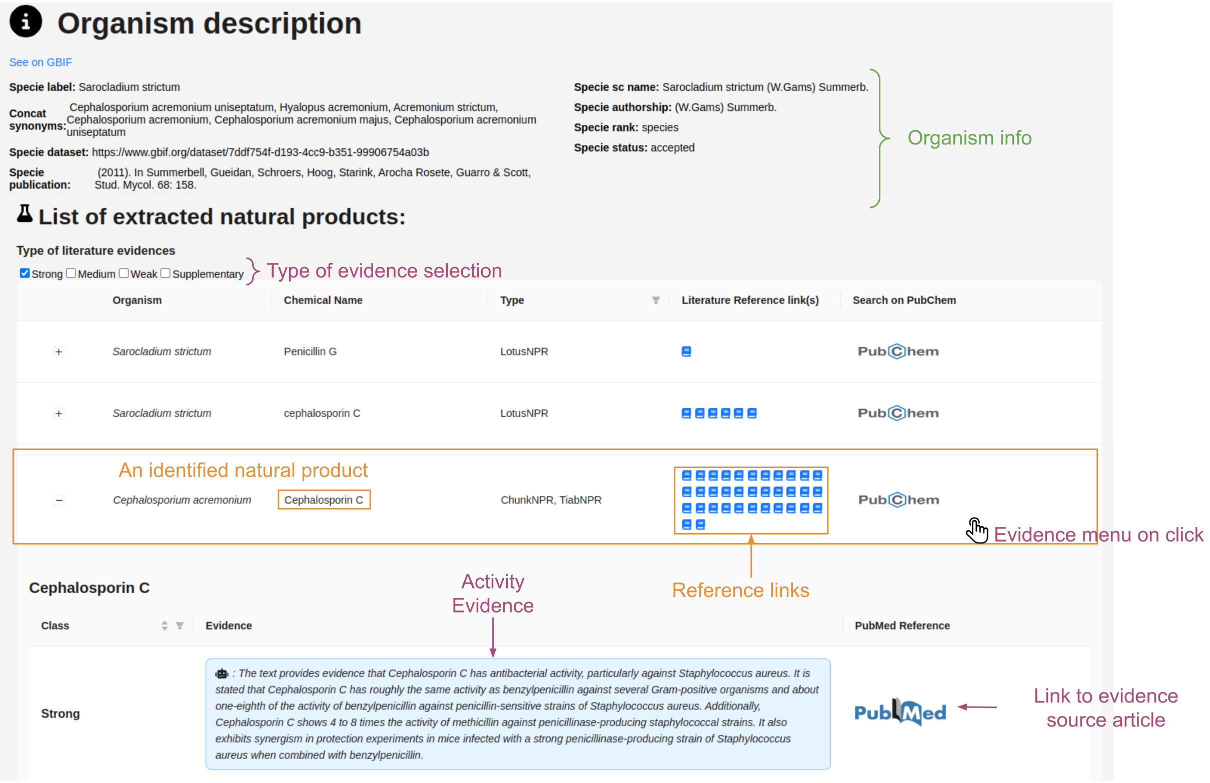


Figure 8: The core structure of the proposed ontology, forming the backbone of the KG. Taxonomic relationships between `abroad:AcceptedTaxon` instances are defined via the transitive property `abroad:hasChildTaxon`. Synonyms are linked using the similarly symmetric and transitive `abroad:hasSynonymTaxon` property. Both `abroad:AcceptedTaxon` and `abroad:SynonymTaxon` are subclasses of `dwc:Taxon` (Darwin Core). Organisms are connected to chemical entities (`chebi:23367`, *molecular entity*) using `abroad:taxonProduces`. The `abroad:NaturalProductRelationship` class defines a hierarchy of extracted relationships, integrating data from LOTUS and from the RE pipeline. Antibiotic evidence is categorized into disjoint classes: `abroad:WeakActivityEvidence`, `abroad:MediumActivityEvidence`, and `abroad:StrongActivityEvidence`



Organism description

See on GBIF

Specie label: Sarocladium strictum

Concat synonyms: Cephalosporium acremonium uniseptatum, Hyalopus acremonium, Acremonium strictum, Cephalosporium acremonium, Cephalosporium acremonium majus, Cephalosporium acremonium uniseptatum

Specie dataset: https://www.gbif.org/dataset/7ddf754f-d193-4cc9-b351-99906754a03b

Specie publication: (2011). In Summerbell, Gueidan, Schroers, Hoog, Starink, Arocha Rosete, Guarro & Scott, Stud. Mycol. 68: 158.

Specie sc name: Sarocladium strictum (W.Gams) Summerb.

Specie authorship: (W.Gams) Summerb.

Specie rank: species

Specie status: accepted

Organism info

List of extracted natural products:

Type of literature evidences

☑ Strong ☐ Medium ☐ Weak ☐ Supplementary

Type of evidence selection

	Organism	Chemical Name	Type	Literature Reference link(s)	Search on PubChem
+	*Sarocladium strictum*	Penicillin G	LotusNPR	▤	PubChem
+	*Sarocladium strictum*	cephalosporin C	LotusNPR	▤▤▤▤▤▤	PubChem
−	*Cephalosporium acremonium*	Cephalosporin C	ChunkNPR, TiabNPR	▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤▤	PubChem

An identified natural product

Evidence menu on click

Reference links

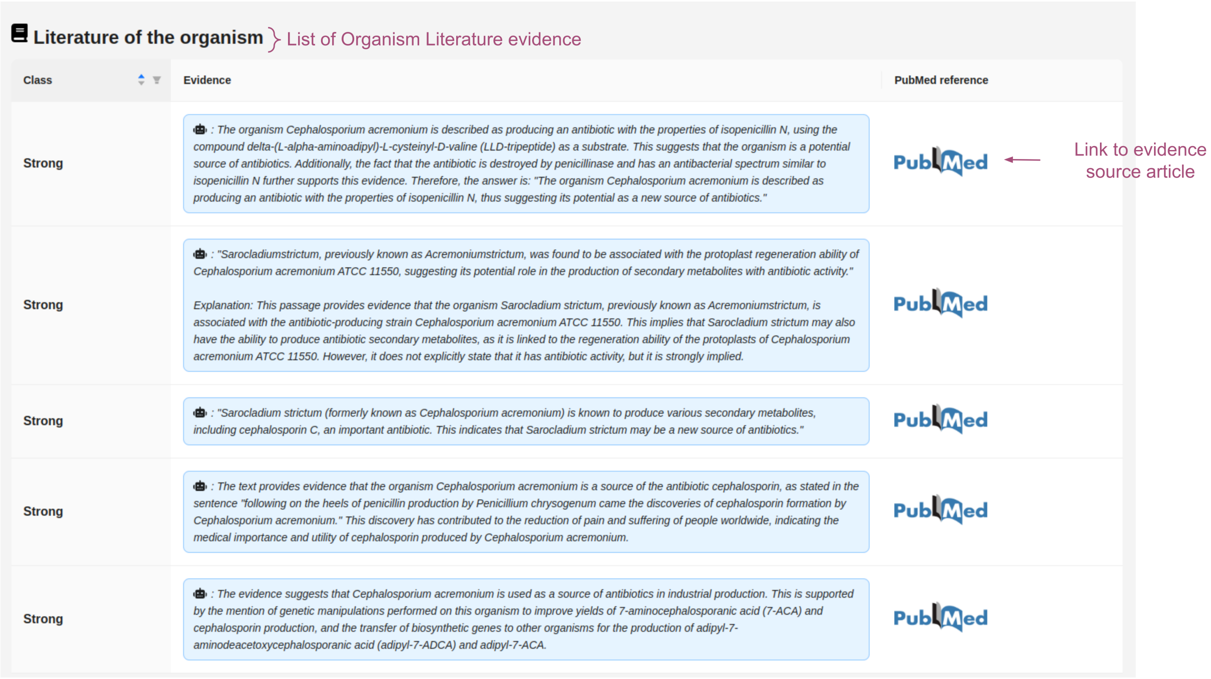
Cephalosporin C

Activity Evidence

Class	Evidence	PubMed Reference
Strong	*: The text provides evidence that Cephalosporin C has antibacterial activity, particularly against Staphylococcus aureus. It is stated that Cephalosporin C has roughly the same activity as benzylpenicillin against several Gram-positive organisms and about one-eighth of the activity of benzylpenicillin against penicillin-sensitive strains of Staphylococcus aureus. Additionally, Cephalosporin C shows 4 to 8 times the activity of methicillin against penicillinase-producing staphylococcal strains. It also exhibits synergism in protection experiments in mice infected with a strong penicillinase-producing strain of Staphylococcus aureus when combined with benzylpenicillin.*	PubMed

Link to evidence source article

Figure 9: Screenshot of the `CL`-evidence alert panel for *S. strictum*



Literature of the organism

List of Organism Literature evidence

Class	Evidence	PubMed reference
Strong	🤖 : *The organism Cephalosporium acremonium is described as producing an antibiotic with the properties of isopenicillin N, using the compound delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine (LLD-tripeptide) as a substrate. This suggests that the organism is a potential source of antibiotics. Additionally, the fact that the antibiotic is destroyed by penicillinase and has an antibacterial spectrum similar to isopenicillin N further supports this evidence. Therefore, the answer is: "The organism Cephalosporium acremonium is described as producing an antibiotic with the properties of isopenicillin N, thus suggesting its potential as a new source of antibiotics."*	PubMed
Strong	🤖 : *"Sarocladiumstrictum, previously known as Acremoniumstrictum, was found to be associated with the protoplast regeneration ability of Cephalosporium acremonium ATCC 11550, suggesting its potential role in the production of secondary metabolites with antibiotic activity."* *Explanation: This passage provides evidence that the organism Sarocladium strictum, previously known as Acremoniumstrictum, is associated with the antibiotic-producing strain Cephalosporium acremonium ATCC 11550. This implies that Sarocladium strictum may also have the ability to produce antibiotic secondary metabolites, as it is linked to the regeneration ability of the protoplasts of Cephalosporium acremonium ATCC 11550. However, it does not explicitly state that it has antibiotic activity, but it is strongly implied.*	PubMed
Strong	🤖 : *"Sarocladium strictum (formerly known as Cephalosporium acremonium) is known to produce various secondary metabolites, including cephalosporin C, an important antibiotic. This indicates that Sarocladium strictum may be a new source of antibiotics."*	PubMed
Strong	🤖 : *The text provides evidence that the organism Cephalosporium acremonium is a source of the antibiotic cephalosporin, as stated in the sentence "following on the heels of penicillin production by Penicillium chrysogenum came the discoveries of cephalosporin formation by Cephalosporium acremonium." This discovery has contributed to the reduction of pain and suffering of people worldwide, indicating the medical importance and utility of cephalosporin produced by Cephalosporium acremonium.*	PubMed
Strong	🤖 : *The evidence suggests that Cephalosporium acremonium is used as a source of antibiotics in industrial production. This is supported by the mention of genetic manipulations performed on this organism to improve yields of 7-aminocephalosporanic acid (7-ACA) and cephalosporin production, and the transfer of biosynthetic genes to other organisms for the production of adipyl-7-aminodeacetoxycephalosporanic acid (adipyl-7-ADCA) and adipyl-7-ACA.*	PubMed

Link to evidence source article

Figure 10: Screenshot of the OL-evidence alert panel for *S. strictum*