A zero-shot NLP-based pipeline for automated processing of antimicrobial-related scientific texts

Oscar A. Bustos-Brinez[1,2], *Fabio A.* González[2], and *Daniel* Restrepo-Montoya[1*]

[1]PhytoMicrOmics - Max Planck Tandem Group, Faculty of Engineering, Universidad Nacional de Colombia, Bogotá 111321, Colombia
[2]MindLab Research Group, Faculty of Engineering, Universidad Nacional de Colombia, Bogotá 111321, Colombia

Abstract. Information extraction from literature is a fundamental process in the construction of knowledge in the life sciences. However, it is also a process that often requires time and effort to obtain accurate results. This work proposes a fast and adaptable scheme for the automatic processing of article texts (abstracts) based on the use of NLP models, specifically designed to identify publications related to the evaluation of antimicrobial compounds. The proposed mechanism receives an abstract as input and determines whether the article meets a series of criteria, also generating a list of the chemical compounds present in the text. The NLP models applied to the texts are executed without additional training (zero-shot learning), and as many filtering criteria as necessary can be used. The quality of this proposal is determined by its use in 368 abstracts of articles, employing three acceptance criteria. The results indicate a high precision of the proposed mechanism for both classifying texts in the area of antimicrobial prospecting and recognizing chemical entities.

1 Introduction

The proliferation of diverse sources of biochemical and biomedical information, including multiple databases and ontologies, scientific articles, patents, and clinical reports, represents a significant opportunity for the application of artificial intelligence and machine learning models, specifically, models designed for natural language processing (NLP). The retrieval of useful information in this area has become an increasingly complex and time-consuming task, particularly when considering multidisciplinary problems such as antimicrobial resistance (AMR) [1], in which sources can also include chemical databases, phytochemical studies, and microbiology records. In biology, automated information extraction has played an important role in understanding the roles of species and their interactions within ecosystems based on morphological and phylogenetic information [2], and has been used in ethnobotany to integrate disparate sources such as traditional medicine literature, taxonomic

* Corresponding author: drestrepom@unal.edu.co

information, and phytochemical databases to provide a comprehensive overview in the search for plants with medicinal properties [3].

Common approaches to automating information extraction in biological sciences have been developed with NLP models at their core, from neural networks that are specifically designed for text processing tasks such as BERT [4, 5], to large generative language models (LLMs) trained with millions of texts, such as the well-known GPT models [6]. BERT is a fast neural network capable of encoding rich contexts, is relatively simple, and is flexible enough to be adapted to multiple tasks with few changes. A well-known proposal based on BERT is BioBERT [7], a model specifically trained in biomedical texts to tackle multiple issues in medicine and biochemistry, including abstract reviews of clinical texts [8] or question answering (extracting portions of text that answer a given question) [9]. Although more complex models are more capable to address difficult issues that involve millions of texts. In this regard, PubTator 3.0 [10], an automated system that classifies documents and annotates key concepts in the texts of all articles available in PubMed, has become the de-facto standard for automatic scientific article processing in medicine and biology. PubTator combines annotation models that identify key concepts with NLP tools that focus on identifying relationships between them. Thanks to its advantages, PubTator has been used, among multiple cases, to boost research into specific diseases such as tuberculosis [11] and to contribute to the development of precision medicines [12].

This work is part of a multidisciplinary project called "Evaluation of Native Plants of Colombia and Their Antimicrobial Activity", which focuses on bioprospecting bioactive compounds from various species of plants collected in different areas of Colombia to identify metabolites with antimicrobial properties against bacteria or fungi. A significant aspect of the overall project pipeline focuses on screening the literature for metabolites found during the chemical analysis stages, to determine whether their antimicrobial potential has been previously established and against which microorganisms. This work focuses on scientific articles, with the main purpose of facilitating (or automating, if possible) the processing of references to determine their usefulness to the project and the extraction of the most relevant words from these articles, i.e., the chemical compounds tested in each research.

To perform the automated processing of the articles, a two-phase pipeline is proposed. The expected input of this pipeline is the plain text of abstracts from scientific articles. The first phase of processing aims to determine if the article meets the project requirements by checking if it corresponds to antimicrobial research centered around the testing of chemical compounds. The second phase applies only to articles that pass the filter and seeks to analyze the text to detect names of chemical compounds, whether associated with specific molecules or broader categories of compounds. With this output from the proposed pipeline (the compounds clearly identified), further refinements can be made, including determining whether the compounds are listed as natural products or plant metabolites or associating them with known interactions between plants and their endophytes. As an important advantage of the proposed pipeline, the models used in the proposed methodology do not require specific retraining for the task, saving time and computational resources. Thus, the pipeline can significantly accelerate the extraction of information from scientific texts and effectively contribute to the understanding of biochemical and bioprospecting processes.

2 Methodology

The proposed methodology focuses on automatic review of scientific articles by processing their abstracts. Abstracts are commonly easier to retrieve than complete texts, particularly when searching in large indexed databases, and tend to concentrate the most useful insights from the text. For the model, it is important to determine which articles are relevant to the project's topic of interest, i.e., the analysis of chemical compounds with antimicrobial

properties. Once a useful article has been identified, the next step is to isolate the names of the compounds present in its abstract, since it is very likely that the abstract mentions the main compound or compounds of the research. Given an abstract as input, the expected output of the methodology consists of the acceptance or rejection of the text and, if accepted, a list of the chemical compounds within the text. Although there are multiple proposals that address abstract reviews or recognition of chemical compounds, the combination of both processes oriented to the specific area of antimicrobial prospection is thoroughly unexplored. Figure 1 illustrates a graphical description of the proposed processing structure.

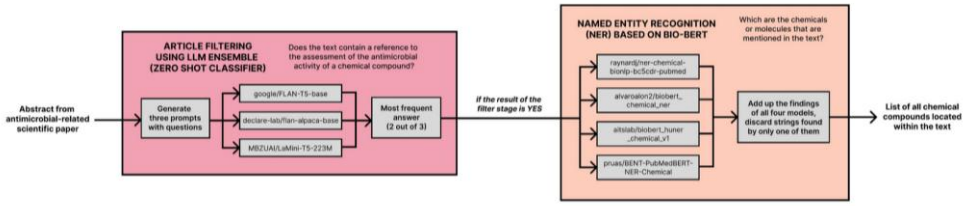

Fig. 1. Description of the proposed methodology. First stage: filtering the text to determine whether there are references to compounds with antimicrobial properties using three text generation models. Second stage: if the text passes the filter, a list is constructed of all the portions of the text that are identified by four named entity recognition (NER) models as names of chemical compounds.

2.1 First stage: filtering using LLM

The first stage of the process (left box in Figure 1) involves filtering the texts using large generative language models (LLMs). Since LLMs are commonly trained on general tasks and fine-tuning is highly costly in terms of time and computational resources, a zero-shot approach is chosen: instead of retraining the model, it works through a natural language description of the task using an input prompt. This approach has demonstrated competitive performance in multiple scenarios, given a sufficiently powerful language model [3, 6].

In this research, three text generation models are used. Google's FLAN-T5 [13], a model based on Alpaca/Llama [14], and a model developed by MBZUAI called LaMini [15]. Since abstracts are relatively short, the proposed approach is to construct an input prompt with the abstract as the context and a question for the models to answer based on that context. The models act as classifiers so that the generated response can only be interpreted as "yes" or "no". The three models receive their input prompts independently from each other, so they may generate the same or different responses. Therefore, the definitive response is found by aggregating the three responses in an ensemble mechanism, such that a text is approved only if "yes" is the output of at least two of the three models.

2.2 Second stage: recognition of chemical compounds

The second stage (right box in Figure 1) applies only to texts that have been accepted by the filter and have been identified as appropriate based on their subject. This stage involves using Named Entity Recognition (NER) models specifically designed to recognize chemical entities to identify the names of compounds in the text. Four models based on BERT are used instead of a single model, and a similar ensemble mechanism is used to combine their answers. There are slight differences between these four models in their architecture and the data with which they have been trained. One of these four models is a version of RoBERTa [16] trained on a chemical dataset, another model is a modified BERT model with an optimized embedding mechanism that favors the extraction of relationships [17], and the

remaining two are models based directly on BioBERT and trained on various corpora [18, 19]. Each of these models receives the abstract as input, independently of the others, and their objective is to identify all portions of the text corresponding to chemical compound names. The four outputs are integrated by considering names that have been identified in at least two responses and discarding those returned by only one model, which are interpreted as artifacts.

3 Results and Discussion

3.1 Dataset construction

In order to determine the quality of the proposed process for isolating items of interest (as a proof of concept), two sets of texts are constructed:

• **Positive Dataset.** This first dataset contains 190 articles that have been extracted from the AntibioticDB database [20]. This hand-curated database records hundreds of research articles on antimicrobial compounds at various stages of development, with a focus on drug-resistant bacteria. Articles like these in AntibioticDB are the target of the proposed pipeline, since they contain information about chemical compounds and their potential to treat bacterial infections. For each compound, the database includes a reference to the URL of the article in which it is mentioned or tested. The dataset for the experimentation is constructed by selecting a subset of the list of compounds, selecting those whose associated articles are included in the PubMed library. This allows for easy access to their abstracts through the Eutils API provided by the National Center for Biotechnology Information (NCBI). Additionally, since these articles are in PubMed, annotations of chemical compounds are available through PubTator 3.0 and can be obtained automatically via its own API. The list of annotations identified by PubTator in each text will serve as the ground truth when measuring the performance of compound recognition models.

• **Negative Dataset**. The 178 articles in this second dataset were found using Google Scholar and Semantic Scholar as search engines, using various keywords aimed at selecting articles that contain references to the antimicrobial field without being associated with compound screening or testing. Most of the articles are related to research on biologically inspired algorithms, AI algorithms applied to medical problems, and a variety of applications in genomics, nanotechnology, and materials science. By examining these articles, which share some keywords or concepts with the desired scope of the project, we hope to verify whether the proposed process can understand the underlying contexts in which these antimicrobial concepts are used and recognize that they are not useful.

The first step to obtaining an overview of the two data sets and their semantic similarities and differences was preprocessing and normalizing the texts, which included removing special characters and stop words. Only normalized texts were used as input for the process. Figure 2 shows a two-dimensional scatter plot visualization of the texts, generated by applying the TF-IDF vectorization, the LDA semantic modeling, and the PCA dimensional reduction to the texts. The negative data (blue dots) and positive data (yellow dots) are clearly distinguishable, indicating thematic differences between the two datasets. However, there is a region where some of these dots overlap because of the use of "antimicrobial" and other similar keywords in both scenarios.

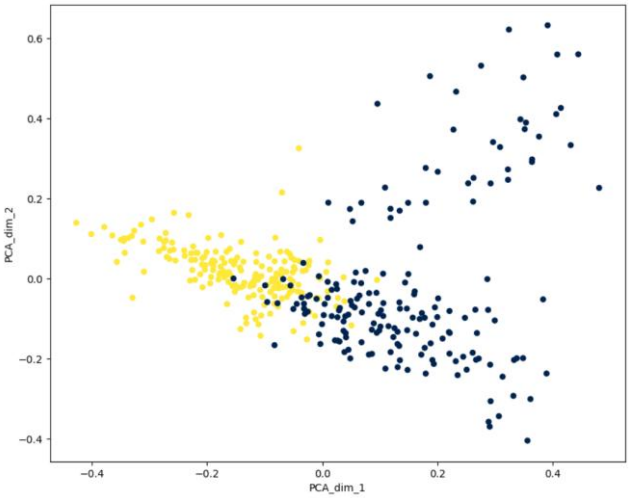

Fig. 2. Distribution of texts in the positive (yellow) and negative (dark blue) datasets.

3.2 Implementation and metrics

The implementation of the proposed methodology, as well as any additional processing, was carried out using the Python programming language, and all code was executed on Google Colab machines. To access the models, the Transformers library from HuggingFace was used. The source code for the implementation of the proposed methodology can be found in the Github repository: https://github.com/oabustosb/nlp-pipe-phytomicromics.

A total of seven NLP models were chosen to perform the tasks defined in the two stages of the proposed process. These models were selected based on two criteria: a good trade-off between the complexity of the model and its speed for the given task, and the accessibility of information about its development, whether through a published article or a code repository. The three models selected for the filtering process (first stage) can be identified in HuggingFace as "google/FLAN-T5-base", "declare-lab/flan-alpaca-base" and "MBZUAI/LaMini-T5-223M". The four models for the compound recognition process (second stage) can be identified in HuggingFace as "alvaroalon2/biobert_chemical_ner", "aitslab/biobert_huner_chemical_v1", "raynardj/ner-chemical-bionlp-bc5cdr-pubmed", and "pruas/BENT-PubMedBERT-NER-Chemical".

The models for the first stage were designed to filter articles according to three criteria: mention of antimicrobial activity, presence of chemical compounds used as drugs, and relationship between chemical compounds and use as anti-infective agents. To achieve this, three different input prompts were constructed and processed independently, each representing one of these three criteria. All prompts follow the pattern "**question:** *question*, **context:** *text_abstract*" where the question may change depending on the specified criterion. The questions used in the prompts to address the three criteria are as follows:

1. There is a reference to antimicrobial, antibacterial, or antiviral activity? Say yes or no.
2. There is a reference to the use of chemical compounds as medicines in the context? Say yes or no.
3. Is any antibiotic, anti-infective, antimicrobial compound used against biofilm, parasites, protozoans, virus, fungi or bacteria? Say yes or no.

A given text can meet none, one, two, or all the criteria. These answers can be combined into a single output, so the filter only accepts texts that meet all the criteria.

3.3 Results

3.3.1 Data filter

The results of the first stage are presented in Table 1, which shows each question associated with the acceptance criteria individually and the combined results of the three questions. Since the filter behaves like a binary classifier, its quality is measured using common metrics for this scenario: accuracy, precision, recall, F1-Score, and AUC-ROC. The first question, the antimicrobial mention criterion, has the lowest metrics of the three questions, due to a high number of false positives that appear because of the antimicrobial theme appearing in almost all texts in the negative dataset. The third question (relationship between compounds and diseases) best discriminates the negative dataset, showing the best F1-Score, but tends to leave out texts in the positive dataset where this relationship is not explicit. The second question (the presence of compounds) has the highest precision score thanks to its quality in the positive dataset, where all texts meet the criteria of showcasing one or more chemicals. Finally, combining the three criteria yields the best values for all other metrics, clearly surpassing the individual questions while balancing their quality between both datasets (i.e., there is a small rise in false negatives, but an important reduction in false positives). This indicates that having more acceptance criteria can improve quality by allowing a more specific delimitation of the topic of interest.

3.3.2 Chemical compound recognition

Each text in the positive dataset is accompanied by its respective PubTator annotations, to compare the performance of the second stage with respect to these annotations. This comparison is performed individually for each text that passed the filter in the previous step, taking into account only the joint output of the four models. Quality is measured using metrics similar to those in the previous case, particularly recall, precision, and F1-Score. In one of the texts, neither PubTator nor our proposal was able to find an annotation, so this text is not considered. Only in five texts was the output of the process completely different from the PubTator annotations; for those cases, the majority of the PubTator annotations corresponded to composite words that the BERT-based models struggled to detect. In one of these cases, PubTator did not find an annotation, but our process detected four words.

Table 1. Results of the first stage of the methodology on the dataset. The metrics obtained for each of the three questions are presented separately, along with the general result of the filter. TP, FN, TN, and FP show the number of true positives, false negatives, true negatives, and false positives, respectively. Precision, Recall, and F1-Score are calculated using a macro average.

	TP	FN	TN	FP	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Question 1	178	12	30	148	0.5652	0.6301	0.5527	0.4813	0.5527
Question 2	190	0	73	105	0.7141	0.8220	0.7051	0.6826	0.7051
Question 3	170	20	91	87	0.7092	0.7406	0.7030	0.6952	0.7030
General Result	166	24	118	60	0.7717	0.7827	0.7683	0.7678	0.7683

Figure 3 shows the distribution of the three metrics, separating F1-Score (plot in the left) and Precision and Recall (plot in the right). F1-Score showed a value of 1.0 in 62 of the 165 filtered abstracts, implying that PubTator's output and the process's output were the same for these texts. These 62 texts are clearly distinguishable on the right side of the red histogram, and most of the remaining values are concentrated between 0.65 and 0.9. Precision and Recall distributions show an even stronger skewness towards high values, and their joint distribution shows a positive correlation between the two metrics. Most texts tend to show both high values for the two metrics, with an important number of texts in which one metric was 1.0 at the cost of the other being comparatively low, and only a handful of texts with low values for both metrics. Table 2 summarizes the distribution plots by presenting the average values of the metrics, also including their standard deviation and the percentage of texts that had a score of 1.0 in each case. The slight difference between average Recall and average Precision shows that the method tends to be pretty good at recovering the same words as PubTator at the cost of generating some more words that do not coincide with the expected output.

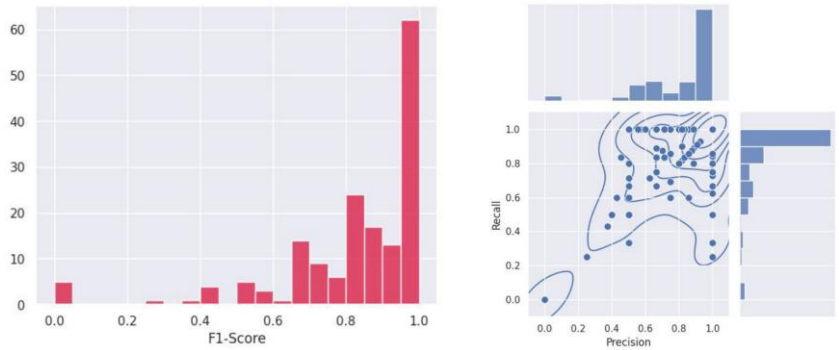

Fig. 3. Results of the second stage of the methodology. The metrics are calculated based on how many words the proposed model correctly extracted compared to the PubTator annotations. Left, the distribution of F1-Scores among the 165 texts. Right, the distribution of Precision and Recall scores, showing both their joint and marginal distributions.

Table 2. Results of the second stage of the methodology on the dataset. Three metrics of performance are presented: average score, standard deviation, and how many of the texts had a perfect score of 1.0.

	Average	Std. deviation	Texts with perfect score
Precision	0.8288	0.2398	53.9%
Recall	0.8495	0.2332	56.4%
F1-Score	0.8232	0.2181	37.6%

4 Conclusions

This paper presents an NLP-based methodology for automatic review of the literature in the context of analyzing compounds with antimicrobial properties. The method applies to

abstracts of scientific articles and consists of two stages. First, zero-shot filtering uses generative language models to determine if an article meets criteria that measure its relevance to the project. Second, NER models trained for chemical entity recognition identify the names of chemical compounds in the text. These two stages can be utilized as a preprocessing step that extracts useful information from unstructured texts to explore further connections between the compounds and other properties, such as their biological origins or chemical structures. The capabilities of the proposed model are tested using 368 abstract texts, including 190 articles related to antimicrobial prospection and 178 articles that may include references to antimicrobial topics without direct association with compound testing. Using three acceptance criteria, the first stage of the methodology correctly classified 77.2% of the texts with a F1-Score of 0.7678. The second stage was applied to 165 articles that passed the filter, and the NER models were tested against annotations provided by PubTator 3.0, PubMed's state-of-the-art annotation mechanism. The models achieved highly accurate detection of compounds in the vast majority of texts, with an average F1-Score of 0.8232.

The proposed methodology demonstrates relatively good performance in both text classification (determining whether texts fit a specific category or topic) and the detection of chemical compounds within texts. It achieves capabilities comparable to PubTator's without the need for fine-tuning the models that are utilized throughout the process, thus avoiding the costs that retraining may imply. The experimentation presented here is intended to demonstrate the feasibility of using this process, and further improvements may include adapting the method for use with longer texts, such as complete papers when available (either texts in PubMed or outside of it), or applying it to larger datasets. Filtering criteria, which in this case focused on antimicrobial compounds, can be modified to adapt this strategy to other scenarios or objectives. An interesting application scenario for the proposed method could focus on finding mentions of endophytic relationships between plants and microorganisms, particularly where these microorganisms are identified as the source of novel chemical compounds.

Funding Statement. The funding grant was assigned by Ministerio de Ciencias de Colombia and the Max Planck Tandem Group-Universidad Nacional de Colombia (project code 91382).

Acknowledgements. The authors thank the members of PhytoMicrOmics-Max Planck Tandem Group (Janice Valencia-Duarte, Lorena Miranda, Anyi Lamprea, Betty Herrera, and Ailyn Villamil) for their continued dedication and valuable feedback.

References

1. A A. Cesaro, S.C. Hoffman, P. Das, C. de la Fuente-Nunez, Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance, npj Antimicrobials and Resistance **3**, 2 (2025).
2. V. Domazetoski, H. Kreft, H. Bestova, P. Wieder, R. Koynov, A. Zarei, P. Weigelt, Using large language models to extract plant functional traits from unstructured text, Applications in Plant Sciences **13**, 70011 (2025). https://doi.org/10.1002/aps3.70011
3. D. Domingo-Fernández, Y. Gadiya, S. Mubeen, T.J. Bollerman, M.D. Healy, S. Chanana, R.G. Sadovsky, D. Healey, V. Colluru, Modern drug discovery using ethnobotany: A large-scale cross-cultural analysis of traditional medicine reveals common therapeutic uses, iScience **26**, 107729 (2023).
4. R.E. Turner, An introduction to transformers, arXiv preprint arXiv:2304.10557 (2023).
5. J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in Proceedings of the 2019 conference of the North

American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (2019), pp. 4171–4186

6. T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language Models are Few-Shot Learners, in Advances in Neural Information Processing Systems, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Curran Associates, Inc., 2020), Vol. 33, pp. 1877–1901
7. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics **36**, 1234 (2020).
8. S. Masoumi, H. Amirkhani, N. Sadeghian, S. Shahraz, Natural language processing (nlp) to facilitate abstract review in medical research: the application of biobert to exploring the 20-year use of nlp in medical research, Systematic Reviews **13**, 107 (2024).
9. M. Guo, M. Guo, E.T. Dougherty, F. Jin, MSQ-BioBERT: Ambiguity Resolution to Enhance BioBERT Medical Question-Answering, in Proceedings of the ACM Web Conference 2023 (Association for Computing Machinery, New York, NY, USA, 2023), WWW'23, p. 4020–4028, ISBN 9781450394161
10. C.H. Wei, A. Allot, P.T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, Z. Lu, Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge, Nucleic Acids Research **52**, W540 (2024). https://doi.org/10.1093/nar/gkae235
11. V. Kumar, G. Shankar, Y. Akhter, Deciphering drug discovery and microbial pathogenesis research in tuberculosis during the two decades of postgenomic era using entity mining approach, Archives of Microbiology **206**, 46 (2024).
12. T. He, K. Kreimeyer, M. Najjar, J. Spiker, M. Fatteh, V. Anagnostou, T. Botsis, Artificial Intelligence-assisted Biomedical Literature Knowledge Synthesis to Support Decision-making in Precision Oncology, in AMIA Annual Symposium Proceedings (2025), Vol. 2024, p. 513
13. H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma et al., Scaling instruction-finetuned language models (2022), https://arxiv.org/abs/2210.11416
14. R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T.B. Hashimoto, Stanford alpaca: An instruction-following llama model (2023), https://github.com/tatsu-lab/stanford_alpaca
15. M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, A.F. Aji, Lamini-lm: A diverse herd of distilled models from large-scale instructions, CoRR (2023), https://doi.org/2304.14402.
16. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
17. P. Ruas, F.M. Couto, Nilinker: Attention-based approach to nil entity linking, Journal of Biomedical Informatics **132**, 104137 (2022). https://doi.org/10.1016/j.jbi.2022.104137
18. Á. Alonso Casero, Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature, Ph.D. thesis, ETSI_Informatica (2021), https://oa.upm.es/67933/
19. R. Ahmed, P. Berntsson, A. Skafte, S.K. Rashed, M. Klang, A. Barvesten, O. Olde, W. Lindholm, A.L. Arrizabalaga, P. Nugues et al., Easyner: A customizable easy-to-use pipeline for deep learning-and dictionary-based named entity recognition from medical text, arXiv preprint arXiv:2304.07805 (2023).
20. L.J. Farrell, R. Lo, J.J. Wanford, A. Jenkins, A. Maxwell, L.J.V. Piddock, Revitalizing the drug pipeline: Antibioticdb, an open access database to aid antibacterial research and development, Journal of Antimicrobial Chemotherapy **73**, 2284 (2018). https://doi.org/10.1093/jac/dky208