Research Article

For reprint orders, please contact: reprints@future-science.com

Future Drug Discovery

Development of a novel chemoinformatic tool for natural product databases

Paulo Ricardo Viviurka do Carmo[1], Ricardo Marcacini[2], Marilia Valli[3], João Victor Silva-Silva[3], Leonardo Luiz Gomes Ferreira[3], Alan Cesar Pilon[4], Vanderlan da Silva Bolzani[4], Adriano D Andricopulo*[,3] & Edgard Marx**[,1]

[1]Agile Knowledge Engineering & Semantic Web (AKSW), Institute of Computer Science, Leipzig University, Leipzig, 04109, Germany

[2]Computer Science & Mathematics Institute, University of São Paulo, São Carlos, SP, 13566-590, Brazil

[3]Laboratory of Medicinal & Computational Chemistry, São Carlos Institute of Physics, University of São Paulo, São Carlos, SP, 13563-120, Brazil

[4]Nuclei of Bioassays, Biosynthesis & Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, São Paulo State University (UNESP), Araraquara, SP, 14800-901, Brazil

*Author for correspondence: aandrico@ifsc.usp.br

**Author for correspondence: marx@informatik.uni-leipzig.de

Aim: This study aimed to develop a chemoinformatic tool for extracting natural product information from academic literature. **Materials & methods:** Machine learning graph embeddings were used to extract knowledge from a knowledge graph, connecting properties, molecular data and BERTopic topics. **Results:** Metapath2Vec performed best in extracting compound names and showed improvement over evaluation stages. Embedding Propagation on Heterogeneous Networks achieved the best performance in extracting bioactivity information. Metapath2Vec excelled in extracting species information, while DeepWalk and Node2Vec performed well in one stage for species location extraction. Embedding Propagation on Heterogeneous Networks consistently improved performance and achieved the best overall scores. Unsupervised embeddings effectively extracted knowledge, with different methods excelling in different scenarios. **Conclusion:** This research establishes a foundation for frameworks in knowledge extraction, benefiting sustainable resource use.

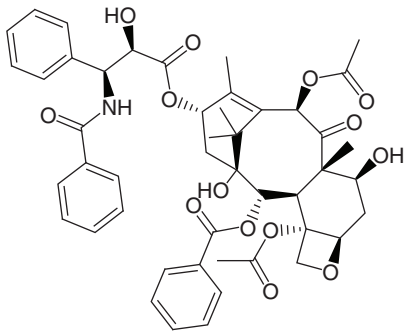
Plain language summary: In this study, a tool to extract relevant information on natural products from scientific papers was developed. Advanced machine learning techniques were used to create a knowledge graph by connecting different information sources. Several methods were tested, with some showing better performance in specific tasks such as the extraction of compound names and bioactivity information. The incorporation of additional data associated with the studied resources proved to improve the results of the models. This study provides a foundation for the development of future tools that can assist researchers in extracting valuable knowledge from scientific literature. Such tools have the potential to facilitate drug discovery efforts and promote the sustainable utilization of natural resources.

First draft submitted: 19 May 2023; Accepted for publication: 29 August 2023; Published online: 19 September 2023

Keywords: chemoinformatics • databases • drug discovery • knowledge extraction • knowledge graphs • machine learning • natural products • NuBBEDB

The biodiversity of tropical environments possesses an exceptionally valuable chemical diversity, being the most traditional source for the development of innovative drugs. The structural diversity of the small-molecule natural products continues to represent an important source of lead compounds for a wide variety of infectious and noncommunicable diseases. A detailed analysis of new drugs approved by the US regulatory agency (FDA) between 1981 and 2019 revealed that 23.4% of these molecules are natural products or derivatives [1]. Molecules such as

newlands press



Paclitaxel

Figure 1. Natural product used as a drug for the treatment of different types of neoplastic diseases.

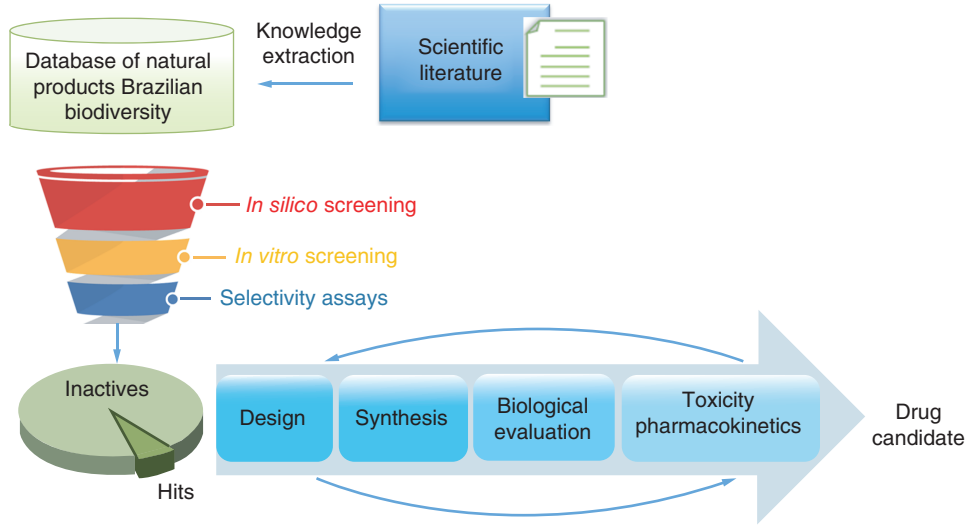


Figure 2. Natural products from Brazilian biodiversity as sources of hits with rich chemical diversity for drug discovery.

paclitaxel (1998), extracted from the bark of the Pacific yew tree (*Taxus brevifolia*) is a representative of a natural product that has been used in the treatment of a number of different cancer types (Figure 1) [1–3].

Computational drug discovery approaches that would allow research on natural products through data access, searches, exploration and organization are of great importance. The discovery of hits derived or inspired in natural products is a promising starting point for the design of drug candidates, as shown in Figure 2. This process starts with the *in silico* screening of libraries of compounds from suitable databases. The selected compounds are then tested *in vitro* for the evaluation of properties such as biological activity, selectivity and toxicity. These early steps lead to the selection of promising hits for further development. Pharmacokinetics (absorption, distribution, metabolism, and excretion) is also investigated *in silico* and *in vitro* to identify the best candidates for clinical trials [4]. The process of simultaneous optimization of multiple pharmacodynamic and pharmacokinetic properties, known as multiparameter optimization, is of special value in the early stages of drug discovery. The design and synthesis of series of compounds are usually assisted by ligand-based and structure-based drug design methods. Following several cycles of hit-to-lead and lead optimization, the most promising candidates can advance to clinical trials.

The Brazilian Biodiversity Natural Products Database (NuBBEDB) is a broadly known source of information on drug discovery research based on natural products from Brazilian biodiversity. Currently, this database contains 2223 compounds, providing valuable data such as chemical, spectral, biological, taxonomic, geographic and pharmacological information (https://nubbe.iq.unesp.br/portal/nubbe-search.html) [4,5]. The information is extracted (Figure 2) from scientific articles, including natural products identified or isolated from Brazilian species. The need to address major challenges in natural product drug discovery led the authors of the present study to investigate intelligent tools for natural product data extraction. Knowledge graphs (KGs) are crucial tools for the generation of

structured data for a number of scientific applications, which are often laborious and time consuming [6], requiring a variety of sophisticated reading and understanding natural language processing strategies. In addition, it is very difficult to keep the datasets up to date with the most recent information [7].

Knowledge extraction can be used to help keep datasets up to date – for example, the NED-EE method that combines Stanford NER with a conditional random fields classifier [8]. At the same time, JERL [9] uses a custom conditional random fields model, and ADEL [10] and USFD [11] use Stanford NER. Meanwhile, WAT [12] combines a maximum entropy model with OpenNLP's NER, and J-NERD [13] uses Stanford's dependency parse-tree as inputs for each sentence in a Gibbs sampling inference model.

In this paper, the authors have chosen a different approach based on unsupervised graph embedding methods. They compared four different unsupervised graph embedding methods in the task of knowledge extraction, using a technique called KG completion, presentend by Martinez-Rodriguez *et al.* [14]. The graph embedding models the authors benchmarked contain an unsupervised graph embedding technique, called DeepWalk [15]. It samples a training dataset for a skip-gram architecture using random walks. The DeepWalk technique is extended by Node2Vec [16] to provide the random walks additional control. Another DeepWalk modification, Metapath2Vec [17], converts random walks into meta-path-based walks. A regularization function is used to disseminate an initial embedding on a KG in the embedding propagation approach called Embedding Propagation on Heterogeneous Networks (EPHEN) [18]. As a result, in an unsupervised scenario, it considers both text and structured data. It propagates a Sentence BERT [19] multilingual model embedding, so that all nodes in a KG receive an embedding in the same vector space from the Sentence BERT regulated by all nodes. These approaches were selected because they produce embeddings for each node without requiring a complete KG (where all nodes are connected), predetermined weights or an ontology. This is significant because since only automatically extracted attributes will be used to connect papers when the model is applied in the actual world.

The development of advanced knowledge extraction techniques is therefore considered a benchmark for KG curation and maintenance. In this work, the authors investigated an approach for data extraction from a set of over 2000 papers in natural product sciences. The evaluation comprised three aspects: extraction of data from papers using a dataset of the natural product database NuBBEDB [4,5], acquisition of similarity distances from unsupervised graph embedding models by KG completion and evaluation of the data extraction of four different graph embedding models to compare their behavior in each property extraction. In addition, evaluations of different unsupervised incorporation generation methods and of extracting properties from natural products are presented.

Materials & methods

Dataset

Hundreds of peer-reviewed scientific articles with information on more than 2521 potential natural product extraction methods were used to create the dataset for testing and training. The NuBBEDB was originally designed by chemistry specialists, who compiled it manually by reading the articles and annotating the relevant information of each natural product [4,5]. As a starting point, the authors requested a list for the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico) containing scientific articles published between 1950 and 2015 by researchers in the area of "natural products" that were cataloged on the Lattes platform (http://lattes.cnpq.br/). A list of 32,524 papers resulted. From this list, the authors selected only articles with registered digital object identifiers reporting compounds that have been isolated/identified from species of Brazilian biodiversity. This work used 390 papers, which represented 5% of this list. The number of papers is representative for data extraction in the area of natural products. The models in this work were designed using five of the properties in NuBBEDB for training and prediction: compound name (*$NuBBE_{DB}$:common name*), bioactivity (*$NuBBE_{DB}$:biologicalActivity*), species (species where the natural product was identified), *$NuBBE_{DB}$:collectionSpecie*), species location (*$NuBBE_{DB}$:collectionSite*) and obtention method (*NuBBEDB:collectionType*). The authors used the ontology created for the Database of Natural Products from the Brazilian Biodiversity (https://github.com/AKSW/dinobbio/tree/main/ontology) for property extraction. The number of unique options for extraction in each information was #compound name = 446; #bioactivity = 34; #species = 116; #species location = 52; #obtention method = 6.

Experimental setup & evaluation criteria

The performance of several machine learning (ML) graph embeddings was examined for the unsupervised knowledge extraction challenge. Graph embeddings enable the extraction of data that have previously been extracted from

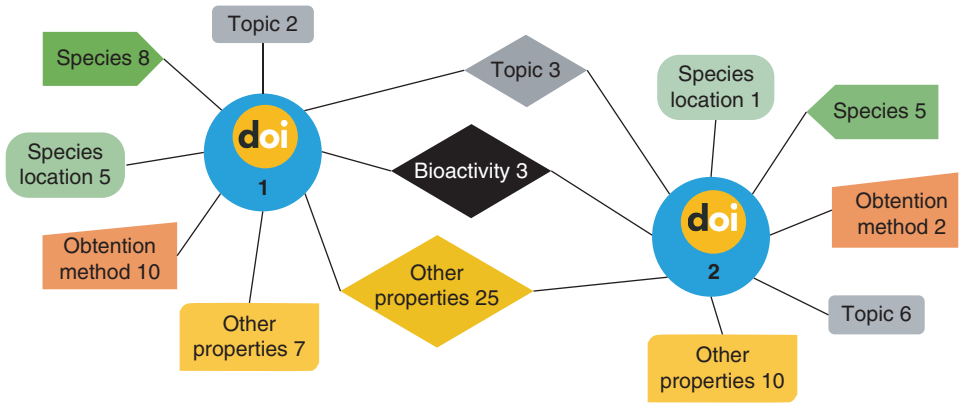


Figure 3. The proposed knowledge graph structure.

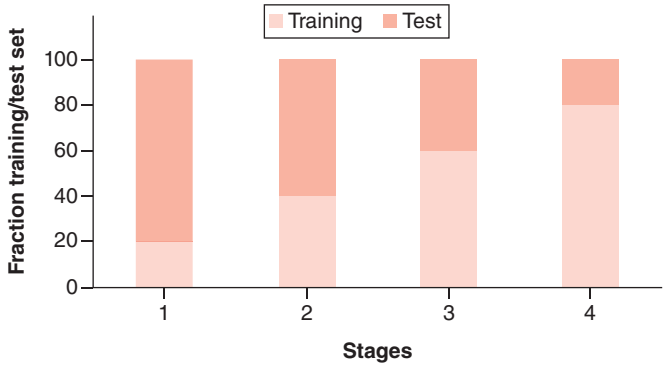


Figure 4. Dynamic stages for evaluation.

the graph and encoded in characteristic vectors. The authors modeled a KG with the digital object identifier of the publication acting as a central node in order to employ graph embedding methods (Figure 3). The previously extracted properties, associated molecular data properties and topics taken from BERTopic [20] are all connected by this. BERTopic is a topic modeling technique that uses a class-based term frequency–inverse document frequency and the bidirectional encoder representations from transformers (BERT) [21] method to construct dense clusters while preserving keywords from objects in these clusters to enable readily readable topics.

The BERTopic model prohibits the input of complete papers and demands a minimum number of tokens. In order to feed the BERTopic model the papers, they were divided into sentences. A multilingual pretrained BERTopic model was fitted to each sentence in order to extract nonduplicate topics. By enabling connections only when less than 80% of the papers had already been connected, the themes of the paper were filtered. This made it possible to weed out broad subjects that did not differentiate papers enough. When papers are chosen as testing data, they may be used to hide all ties to manually extracted data while keeping the nodes related to one another by their subjects.

The authors created a dynamic evaluation of four phases with varying quantities of training data in each step to replicate a situation where new training data are continuously added to the model. It is important to underline that "test data" means that all manually extracted information contained in the dataset is hidden so that it is represented as a new data point. Throughout the subsequent stages, the training split is raised by 20% until it achieves an 80/20 ratio. The initial stage consists of a training/test split with a 20/80 ratio (Figure 4). For all evaluation stages, the authors sorted a list by cosine similarity separately for each extraction scenario with the test data.

Models & framework models

Four different unsupervised graph embedding methods were used for the knowledge extraction task to obtain an embedding vector with 512 dimensions and the following parameters: DeepWalk [15] with 80 walks with the length of ten nodes; Node2Vec [16] with 80 walks with the length of ten nodes and a balance of 0.5 in p and 1 in q, meaning that the random walks explore the global network; Metapath2Vec [17] with one walk with the length of

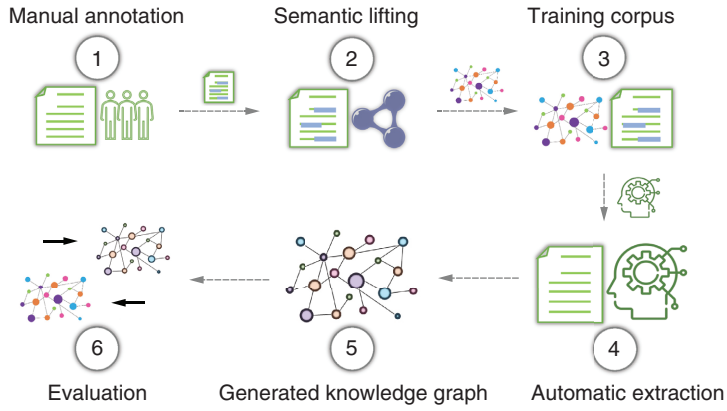


Figure 5. Representation of the pipeline for training the different machine learning models.

100, a context window size of 10 and a set of meta-paths; and EPHEN [18] propagating DistilBERT multilingual embedding in 30 iterations for the first evaluation stages and 20 upon the updates of the remaining stages.

Concept for quantified approach

The schematic representation of the pipeline for training the different ML models is depicted in Figure 5. First, a group of chemical experts crowdsource a paper and annotate its content manually. The annotations are then combined with a predefined semantic data model and used to instantiate a KG containing the natural products and their properties. Later, the resulting KG, as well as the annotated paper, are used for training ML models using different strategies. Those models are used to annotate automatically different samples of preannotated papers. The resulting generated natural product knowledge subgraph of each paper is then compared with the manually annotated one.

Different values of individuals per property were used and they were proportionally attributed to its extraction difficulty. For instance, it is significantly more challenging to extract the right compound name than it is to extract the obtaining method, because there are considerably fewer options in the training set for the compound name than there are for the obtaining method. For that, the authors evaluated with different k from 1 to 50, considering values multiples of 5. The final k value for each extraction was defined either when a score higher than 0.50 is achieved at any evaluation stage or the upper limit of k = 50. Using this rule, the final values of k were defined as follows: compound name, k = 50; biological activity, k = 5; species, k = 50; species location, k = 20; and obtention method, k = 1.

Results

The predictive ability of each approach to extract knowledge from various compound properties using hits@k was evaluated separately for each property. The statistic hits@k calculates the average number of forecasts that place in the top k results. The hits@k metric considers how many predictions achieve top k rankings and has been widely used to evaluate graph completion researches [22–24]. When there is just one right prediction, hits@k is a ranking metric that works in conjunction with the mean reciprocal rank. On the other hand, when a list of pertinent predictions is supplied as a ground truth, mean average precision, normalized discounted cumulative gain [25] and precision@k are developed for ranking. hits@k was chosen because, by adjusting the k value, it enables the evaluation of each data extraction with acceptable expectations and is frequently used in the evaluation of knowledge extraction.

All the results in the benchmark were executed in $NuBBE_{DB}$ and for each graph the bars represent the corresponding hits@k score and the lines represent the execution minutes for the entire process. The results for the compound name extraction are presented in Figure 6. In this scenario, it can be seen that Metapath2Vec provided the best results, being capable of increasing performance in the fourth evaluation stage, reaching 0.2 hits@50, whereas all the other evaluated methods had worse performance and declined with the increase in train data. Execution time dropped in all methods. This can be explained by the reduction of options in the ranking by cosine similarity. It can also be observed that in the third evaluation stage, EPHEN starts to have a better execution time than Metapath2Vec. This is due to EPHEN being able to reuse existing embeddings without recalculating the entire space every execution, like the other methods. This behavior is constant for all other scenarios.

10.4155/fdd-2023-0007

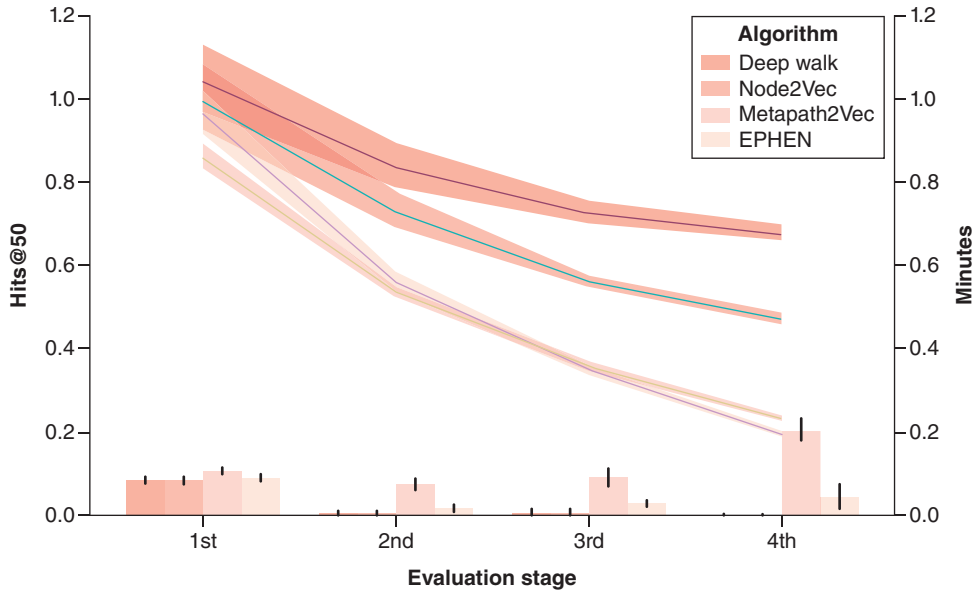


Figure 6. Compound name extraction with hits@50 results (bars). Minutes for execution are indicated in the lines.

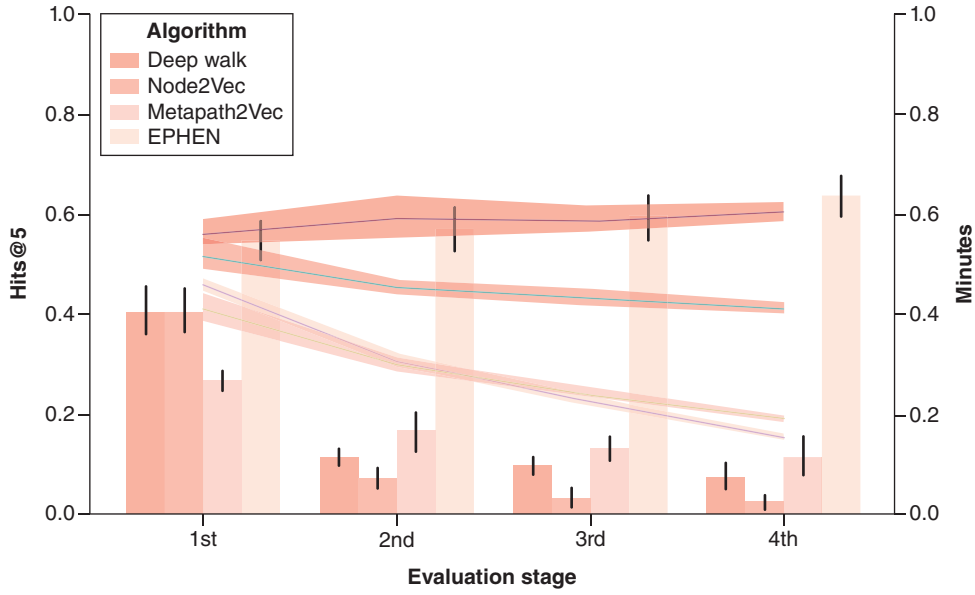


Figure 7. Bioactivity extraction with hits@5 results (bars). Minutes for execution are indicated in the lines.

The extraction of the bioactivity information was best achieved with EPHEN at more than 0.6 hits@5 in the fourth evaluation stage. Meanwhile, Metapath2Vec was the only method capable of increasing performance comparing the fourth with the first evaluation stage (Figure 7). It can also be seen that DeepWalk and Node2Vec achieved better performance than Metapath2Vec in the first evaluation stage, but they drastically lost performance in the following evaluation stages. Finally, in this scenario the performance achieved was much more reasonable than predicting the extraction of compound name that contains many other options. For execution times in this scenario, DeepWalk and Node2Vec did not reduce them in the more advanced evaluation stages. This can be explained by the existence of fewer options overall for bioactivity knowledge extraction.

The results for extracting species were similar to the compound name scenario, with Metapath2Vec being the best performer and able to increase performance (Figure 8). However, in this scenario the margins were much tighter, and the other methods were able to maintain their performance in other evaluation stages. Nevertheless, this shows that

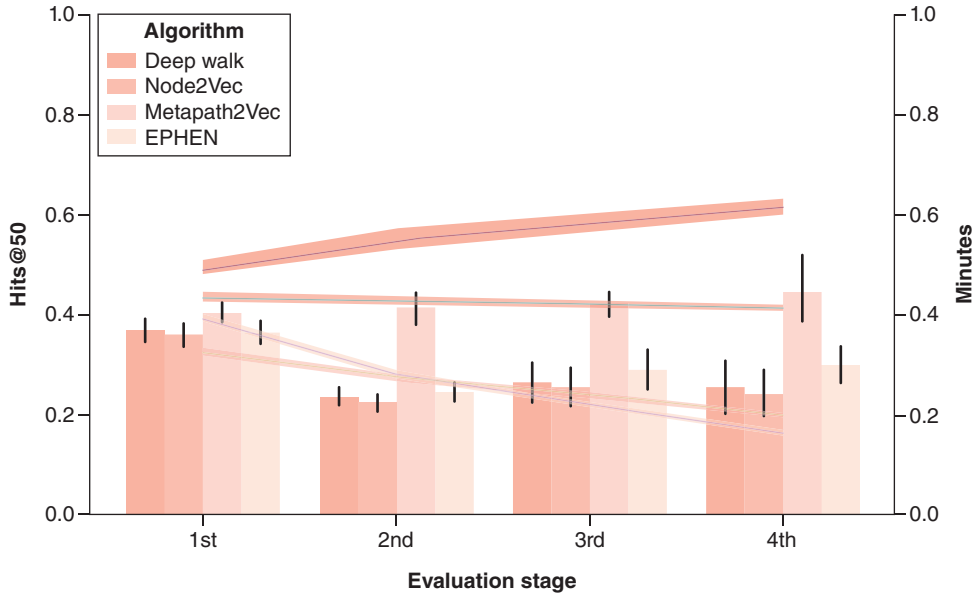


Figure 8. Species extraction with hits@50 (bars). Minutes for execution are indicated in the lines.

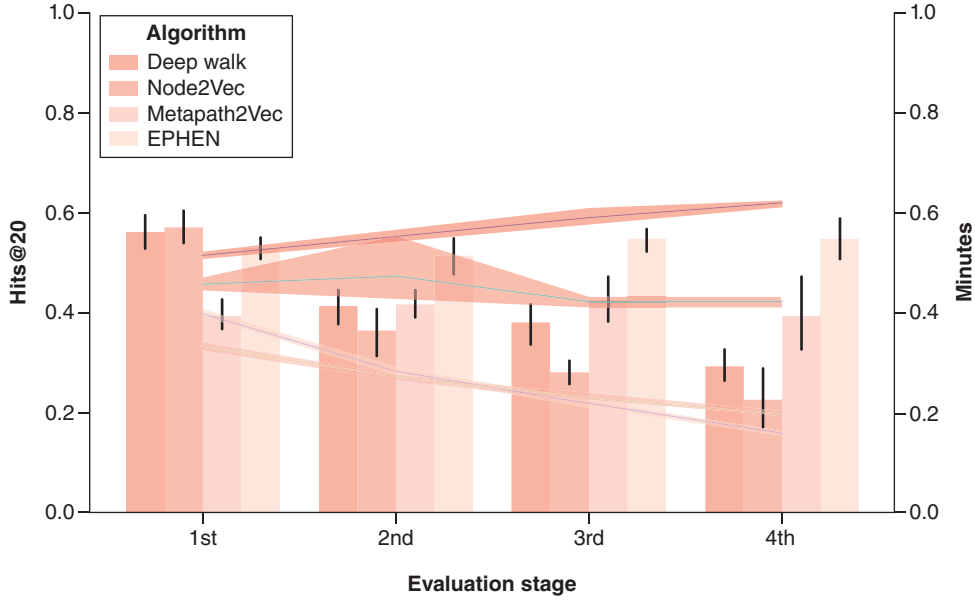


Figure 9. Species location extraction with hits@20 (bars). Minutes for execution are indicated in the lines.

Metapath2Vec's more robust pipeline to deal with different types of relations yields more performance in challenging knowledge extraction scenarios. As for execution times, DeepWalk and Node2Vec repeated their behavior from bioactivity but with a more accentuated increase in execution times for the third and fourth evaluation stages.

The species location extraction was the only scenario where DeepWalk and Node2Vec achieved the best performance in one evaluation stage (Figure 9). However, they were not able to increase performance in the following evaluation stages, unlike EPHEN, which was the only method capable of doing that. Once again, DeepWalk and Node2Vec execution times increased in the third and fourth evaluation stages.

The obtention method information extraction achieved the best scores overall, even though the results were measured with hits@1, which is basically accuracy (Figure 10). However, only EPHEN was able to achieve good performance in this scenario, with a significant margin from other methods. Finally, DeepWalk's execution had

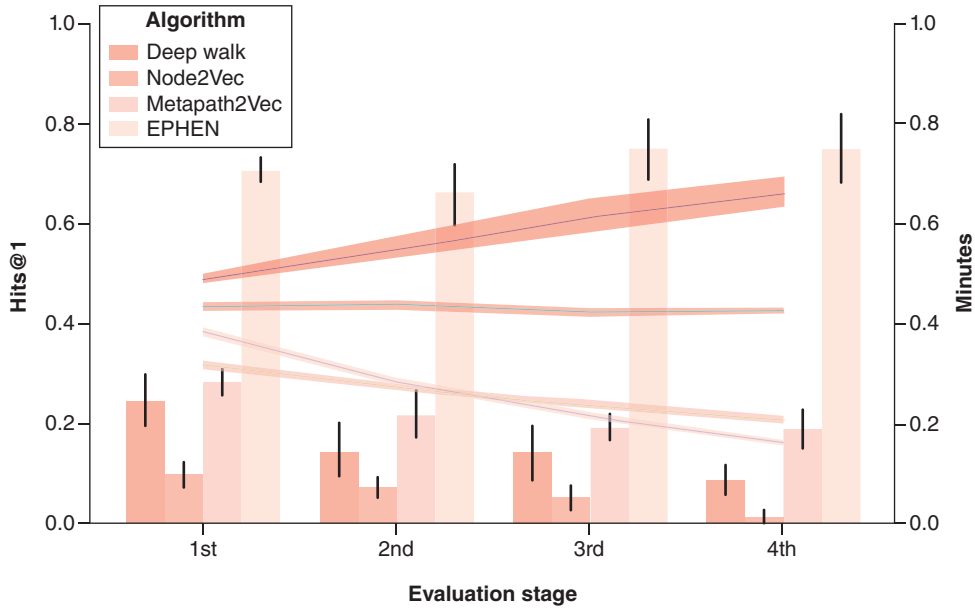


Figure 10. Obtention method extraction with hits@1 (bars). Minutes for execution are indicated in the lines.

the highest increase, starting in the second evaluation stage. This indicates that for the fewest options, DeepWalk started to be the bottleneck in execution time for this methodology of knowledge extraction.

Discussion

The quality of the database used for the drug discovery pipeline is extremely important. The NuBBEDB was originally a manually curated database. This means that the compound chemical structure and bioactivity, as well as the species from where the compound was isolated, the location where this species was collected and the isolation method were extracted by a human reading the paper and annotating the data. Extraction of the ever-increasing amount of data published in the natural product area is a very challenging task. Automation is essential to make work more structured, more efficient and faster. The main challenge of knowledge extraction from scientific literature is that it constitutes an unstructured data source, where authors write using different words and formats, sometimes describing the same compound, method or activity. Therefore, a robust ontology is proposed to create an in-depth extraction methodology. To achieve more stable and trustworthy results, this work used rule-based information extraction algorithms. A methodology was developed and evaluated in different ML embeddings for the task of unsupervised knowledge extraction. The evaluation was designed so that the performance of each approach was measured when inserting randomly selected portions of a crowdsourced training dataset.

The models were created adding new training data constantly, to ideally increase accuracy over the stages. All nodes that originated from the crowdsourced dataset out of the KG were removed, leaving the papers connected only to their topics. The first training/test consisted of a 20/80 ratio, and for other stages, the training ratio was increased by 20% until it reached 80/20. The KG was enriched with topics related to the papers using BERTopic [20]. Evaluation of this methodology was performed using hits@k, comparing the performance of the four approaches in extracting the knowledge of a different compound in the literature.

The first important remark considering all the results is the execution time duration to perform the algorithm. Metapath2Vec took the least amount of time to generate the embedding in all first evaluation stages. However, in the second evaluation stage EPHEN tied the amount of time it took and was the best performer after that. This can be explained by two basic characteristics of Metapath2Vec and EPHEN. Metapath2Vec is implemented as a parallel central processing unit method and EPHEN can dynamically update the embeddings instead of regenerating the entire vector space every execution.

Another important aspect is that Metapath2Vec achieved the best performance in the two most challenging scenarios. This indicates that the embeddings this method generates are more capable of discriminating more unique nodes. However, in all the remaining scenarios, EPHEN achieved the best performance, which can be

explained by the method characteristics of propagating BERT's embeddings to the graph's nodes, allowing the embeddings to be generated considering two different sources of data.

Conclusion

This study shows that it is possible to use unsupervised embedding approaches to extract natural product knowledge from academic literature – in particular, those with fewer unique options in our ontology (i.e., biological activity and obtention method). The quality of the extraction is generally improved by the incorporation of context aware data. The best outcomes were obtained by EPHEN, and Metapath2Vec performed well in more difficult situations (i.e., compound name and species location extraction). Last but not least, the random walks of DeepWalk and Node2Vec perform better with less training corpora. In order to improve EPHEN, resource similarity data could be used. In the future, this could lead to even better outcomes and the creation of a framework for extracting natural product knowledge with a human in the loop. The development of reliable extraction models to aid in the update of this enriched information will be important to drug discovery inspired by natural products as well as increase community awareness of natural resource values and their sustainable use.

Future perspective

This work opens new ways to gather relevant information on natural products that otherwise would be scattered in the literature. The algorithm herein reported will be applied to expand the number of articles to be assessed to extract data for $NuBBE_{DB}$. Over the next years, the number of papers from which the data on compounds are extracted will grow significantly faster using the algorithm, compared with doing data extraction manually. New cycles of training and test sets will be run and the algorithm will be supplied with new data/papers, therefore improving the relevant statistical indicators. The algorithm will become more robust, as more papers are analyzed. It is important the development of specific algorithms for each area – in this case, for natural products – given the lack of tools for data extraction in this area.

Summary points

- Machine learning graph embeddings, including DeepWalk, Node2Vec, Metapath2Vec and Embedding Propagation on Heterogeneous Networks (EPHEN), were used to extract knowledge from a knowledge graph.
- Metapath2Vec performed well in extracting compound names and showed improvement over evaluation stages.
- EPHEN achieved the best performance in extracting bioactivity information.
- Metapath2Vec excelled in extracting species information, while DeepWalk and Node2Vec performed well in one evaluation stage for species location extraction.
- EPHEN consistently improved performance across different scenarios and achieved the best overall scores in extracting obtention methods.
- Unsupervised embeddings effectively extracted natural product knowledge from academic literature.
- Resource similarity data can enhance EPHEN's performance.
- This research establishes a foundation for human-inclusive frameworks in knowledge extraction, benefiting drug discovery and sustainable resource use.

Author contributions

Conceptualization was done by E Marx, AC Pilon, VDS Bolzani and AD Andricopulo; methodology was developed by E Marx, AC Pilon, LLG Ferreira, PRVD Carmo, R Marcacini and M Valli; software was developed by PRVD Carmo and R Marcacini; validation was done by E Marx, PRVD Carmo, R Marcacini and M Valli; formal analysis was conducted by E Marx, PRVD Carmo and M Valli; investigation was carried out by E Marx, AC Pilon and PRVD Carmo; resources were provided by E Marx, M Valli, VDS Bolzani and AD Andricopulo; data curation was done by M Valli, VDS Bolzani and AD Andricopulo; PRVD Carmo and E Marx wrote the original draft; PRVD Carmo, E Marx, M Valli, JV Silva-Silva and LLG Ferreira reviewed and edited the draft; E Marx, VDS Bolzani and AD Andricopulo were responsible for supervision and project administration; E Marx, M Valli, JV Silva-Silva, VDS Bolzani and AD Andricopulo acquired funding.

Financial disclosure

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grants 2020/11967-3 (DFG/FAPESP), under the project DINOBBIO (DFG project 459288952) https://dinobbio.aksw.org, #2022/08333-8 (DAAD/FAPESP), #2013/07600-3 (CIBFar-CEPID), #2014/50926-0 (INCT BioNatCNPq/FAPESP), Conselho Nacional de Desenvolvimento Científico

e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for grant support and research fellowships. The authors acknowledge the scholarships conferred to M Valli (Fapesp #2019/05967-3) and JV Silva-Silva (CNPq #382364/2022-8). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending or royalties.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

Open access

References

Papers of special note have been highlighted as: •• of considerable interest

1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83(3), 770–803 (2020).
2. Gallego-Jara J, Lozano-Terol G, Sola-Martínez RA, Cánovas-Díaz M, de Diego Puente T. A compressive review about Taxol®: history and future challenges. *Molecules* 25(24), 5986 (2020).
3. Durand GA, Raoult D, Dubourg G. Antibiotic discovery: history, methods and perspectives. *Int. J. Antimicrob. Agents* 53(4), 371–382 (2019).
4. Valli M, dos Santos RN, Figueira LD *et al.* Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.* 76(3), 439–444 (2013).

•• **Article containing information for the creation of the first version of the Brazilian Biodiversity Natural Products Database.**

5. Pilon AC, Valli M, Dametto AC *et al.* NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci. Rep.* 7(1), 7215 (2017).

•• **Study on the profile of the compounds present in the Brazilian Biodiversity Natural Products Database.**

6. Zaveri A, Kontokostas D, Sherif MA *et al.* User-driven quality evaluation of DBpedia. Proceedings of: *The 9th International Conference on Semantic Systems.* NY, USA, 4–6 September 2013.
7. Hellmann S, Stadler C, Lehmann J, Auer S. DBpedia live extraction. Proceedings of: *In On the Move to Meaningful Internet Systems: OTM 2009: Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009, Part II* Springer Berlin Heidelberg, Vilamoura, Portugal, 1209–1223, 1–6 November 2009.
8. Hoffart J, Altun Y, Weikum G. Discovering emerging entities with ambiguous names. Proceedings of: *The 23rd International Conference on World Wide Web.* NY, USA, 7–11 April 2014.
9. Luo G, Huang X, Lin C-Y, Nie Z. Joint entity recognition and disambiguation. Proceedings of: *The 2015 Conference on Empirical Methods in Natural Language Processing.* Stroudsburg, PA, USA, 17–21 September 2015.
10. Plu J, Rizzo G, Troncy R. Enhancing entity linking by combining NER models. Proceeding of: *In Semantic Web Challenges: Third SemWebEval Challenge at ESWC 2016* Springer International Publishing, Heraklion, Crete, Greece, 17–32, 29 May–2 June 2016. (Revised Selected Papers 3).
11. Derczynski L, Augenstein I, Bontcheva K. USFD: Twitter NER with drift compensation and linked data. Proceedings of: *The Workshop on Noisy User-generated Text.* Stroudsburg, PA, USA, 31 July 2015.
12. Piccinno F, Ferragina P. From TagME to WAT. Proceedings of: *The First International Workshop on Entity Recognition & Disambiguation – ERD'14.* NY, USA, 11 July 2014.
13. Nguyen DB, Theobald M, Weikum G. J-NERD: Joint named entity recognition and disambiguation with rich linguistic features. *Trans. Assoc. Comput. Linguist* 4, 215–229 (2016).
14. Martinez-Rodriguez JL, Hogan A, Lopez-Arevalo I. Information extraction meets the semantic web: a survey. *Semant. Web* 11(2), 255–335 (2020).
15. Perozzi B, Al-Rfou R, Skiena S. DeepWalk. Proceedings of: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* NY, USA, 24–27 August 2014.

16. Grover A, Leskovec J. Node2vec. Proceedings of: *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* NY, USA, 13–17 August 2016.

17. Dong Y, Chawla NV, Swami A. Metapath2Vec. Proceedings of: *The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* NY, USA, 13–17 August 2017.

18. do Carmo P, Marcacini R. Embedding propagation over heterogeneous event networks for link prediction. Proceedings of: *The 2021 IEEE International Conference on Big Data (Big Data).* Virtual, 15–18 December 2021.

•• **The basic concepts for the development of the present work.**

19. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proceedings of: *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Stroudsburg, PA, USA, 3–7 November 2019.

20. Grootendorst M. BERTopic:Neural topic modeling with a class-based TF-IDF procedure. *arXiv* arXiv:2203.05794 (2022).

21. Devlin J, Chang M-W, Lee K, Toutanova KBERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* arXiv:1810.04805 (2018). https://github.com/tensorflow/tensor2tensor

22. Deng S, Rangwala H, Ning Y. Dynamic knowledge graph based multi-event forecasting. Proceedings of: *The 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* NY, USA, 6–10 July 2020.

23. Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems* 26 (2013).

24. Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling relational data with graph convolutional networks. Proceedings of: *In The Semantic Web: 15th International Conference, ESWC 2018.* Springer International Publishing, Heraklion, Crete, Greece, 593–607 3–7 June 2018.

25. Kishida K. Property of average precision and its generalization: an examination of evaluation indicator for information retrieval experiments. *Tokyo, Japan: National Institute of Informatics.* 19 (2005).