Brazilian Biodiversity Natural Products Database: BrNPDB

Fillipe A. Neves[1], Alice R. V. Carvalho[1], Lívia M. Campos[1], Thainan L. N. de Oliveira[1], Marilia Valli[1], Alan C. Pilon[2], Thomas Riechert[2], Vanderlan S. Bolzani[2], Edgard Marx[3], István J. Mócsy[3], Paulo R. V. do Carmo[3], Stefan Schmidt-Dichte[3]

[1]BioMedChem Research Group, School of Pharmaceutical Sciences of Ribeirão Preto University of São Paulo (USP) – Ribeirão Preto, SP – Brazil

[2]Institute of Chemistry, São Paulo State University (UNESP) – Araraquara, SP – Brazil

[3]Leipzig University of Applied Sciences (HTWK) – Leipzig – Germany

{fillipe.neves, alice.veloso, livia10mendes, marilia.valli}@usp.br
{alan.pilon, vanderlan.bolzani}@unesp.br, {edgard.marx, paulo.carmo}@htwk-leipzig.de

Abstract. *Released in 2013, the NuBBEDB is one of the best sources of natural products from Brazilian biodiversity. However, new possibilities have emerged with advances in scientific research and technologies. In this study, we demonstrate and discuss the development of the Brazilian Biodiversity Natural Products Database (BrNPDB), the relevance of Natural Products databases in Brazil, the materials and methods for data modeling, integration, standards and generated data from the cutting-edge sources, applications and technologies. The results show how BrNPDB upgrades NuBBEDB in many ways and contributes to the growing demand for platforms that enable more effective integration of chemical, biological, and ecological data.*

1. Introduction

In this study, we aim to demonstrate and discuss the development of a database containing secondary metabolites from the Brazilian biodiversity, the Brazilian Biodiversity Natural Products Database (BrNPDB). BrNPDB is an updated database offering a compilation of curated data done by specialized researchers, including the Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products Database (NuBBEDB), and data provided by Chemical Abstracts Service (CAS), part of American Chemical Society (ACS). Questions arose during the development of BrNPDB: Is there a best way to store data on Natural Products? Should we use relational or non-relational databases? How can we ensure integration and interoperability with other Natural Products data sources? How to ensure reliability and scalability as new data enters the database?

The following sections will present the importance of Natural Products databases in Brazil, the materials and methods used for data integration, standards and generate data, as well the technologies used and developed in this study, addressing the issues that arose during development. Finally, results, discussion and possibilities for future work will be presented.

2. Why Natural Products databases in Brazil are so relevant?

Brazilian biodiversity is widely recognized as one of the planet's natural treasures, harboring a diverse array of flora and fauna species. This diversity not only represents a unique natural heritage but also a source of bioactive compounds that can be exploited for the development of medicines, cosmetics, and other high-value-added products. It is estimated that only 8% of Brazilian plant species have been investigated for their pharmacological potential, highlighting the gap in the scientific exploitation of these resources [Calixto 2019]. In addition to its ecological significance, it offers untapped potential for the discovery of new bioactive compounds. Studies show that many essential medicines currently in use have their origins in natural products, reinforcing the need to preserve and study our biodiversity [Newman and Cragg 2020].

The relationship between science and biodiversity has historically been essential for the development of bioactive substances. Many of the pharmaceuticals currently used originate from natural products, isolated from plants, microorganisms, and marine organisms. In Brazil, the vast diversity of species represents a still largely untapped potential for medicinal chemistry and other areas of biotechnology. However, transforming this wealth into useful knowledge requires significant effort in cataloging, organizing, and accessing data. This need is exacerbated by the fact that information on Brazilian natural products is scattered across scientific articles not linked to unified platforms, hindering their practical application [Thessen and Patterson 2011]. In this context, specialized databases play a fundamental role in gathering, structuring, and making information available in a standardized format. Access to this data allows researchers and institutions to analyze chemical, biological, and pharmacological patterns, facilitating new discoveries and driving innovation in the natural products field. However, the lack of an efficient and up-to-date system can make this process fragmented and disorganized.

Natural product (NP) databases are important sources for the scientific community, as they provide access to thousands of compounds and exhibit a huge range of structural complexities, supporting relevant research in computer-aided drug discovery (CADD), chemical ecology and molecular biology. Some of these databases already established are continuously being updated and available in regional databases, for example, LANaPDB [Gomez-Garcia et al. 2024], SANCDB [Hatherley et al. 2015], TM-CM [Kim et al. 2015], TCM-Database@Taiwan [Chen 2011], NANPDB [Ntie-Kang et al. 2017] and TCMID [Xue et al. 2013], but also some researchers work on integrated this databases in a world-wide perspective, like in the COlleCtion of Open Natural ProdUcTs – COCONUT project [Steinbeck et al. 2025].

NuBBEDB was released in 2013, as a database focused on Brazilian biodiversity, consolidating data on natural compounds isolated from various species. NuBBEDB is the first library of natural products of Brazilian biodiversity, and it includes a large variety of compounds and structural types of secondary metabolites of plants, fungi, insects, marine organisms, and bacteria [Valli et al. 2013]. However, with the advancement of scientific research and database technologies, the need for a more robust and integrated system became evident, and BrNPDB emerged as an evolution of this initiative, incorporating new features and approaches to improve the organization and access to information.

3. Materials and Methods

The data sources for creating the BrNPDB have thousands of records arranged in different data models, so the first step was to create and define a new data model, taxonomy and ontology, followed by the integration of the databases, which included organization, checking for duplicates, inconsistencies, null values, and data cleaning to provide data standardization.

Applications and technologies were also researched and used to generate data of interest, like the compound identifiers Inchi, Inchikey (using OpenBabel [O'Boyle et al. 2011]) and Iupac Name (using STOUT v.2.0 [Rajan et al. 2022]), descriptors such as Topological Polar Surface Area (TPSA), Lipinski violations, H-bond donors, H-bond acceptors, Rotatable Bonds, mol2 and SDF files (using R-CDK [Guha 2007] and ChemmineR [Cao et al. 2008]), and identifiers for integration with other data sources, such as the Brazilian Institute of Geography and Statistics and Catalogue of Life [Bánki et al. 2025].

Python scripts were developed to use the Application Programming Interface – API from Crossref, for generated references details such as author, year, title, journal, and the API from NPClassifier [Kim et al. 2021] to generate data about metabolic class. Finally, a search engine was built using a Shiny application [Chang et al. 2025] based on programming language R [R Core Team 2024]. The next sections show details about the development of BrNPDB.

3.1. Data Taxonomy

The data taxonomy was defined for the BrNPDB, shown in figure 1. The most important and central information is the natural product compound structure in SMILES format (Mol2 and SDF file also available). Metadata includes the bibliographical reference with a DOI of the scientific paper, the species from where the compound was isolated and its geographical location, biological property, metabolic class, and chemical properties. This taxonomy was generated after years of research to provide linked data and offer transparency from where each of the metadata came from. Ontology was also researched for standardized terms, for interoperability with other databases.

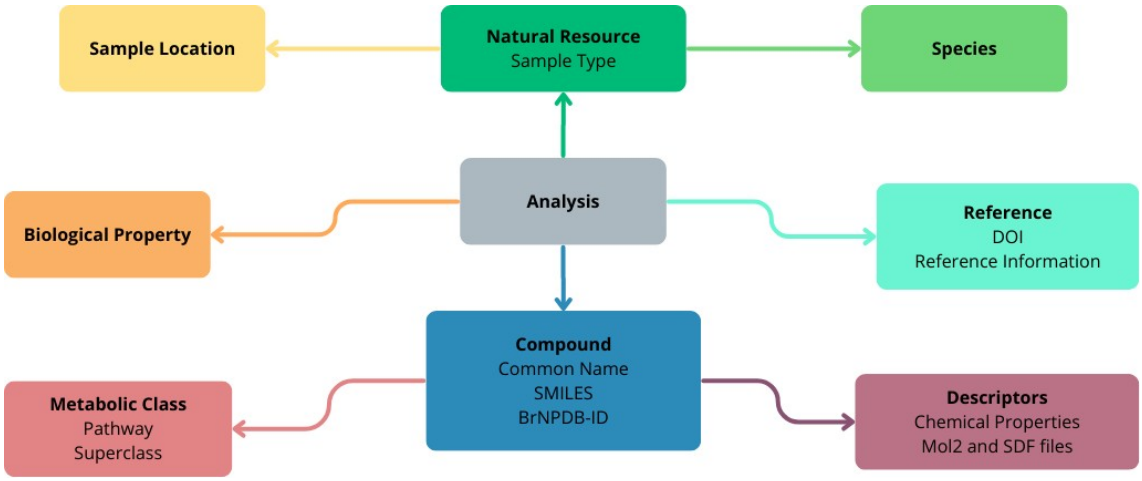

Figure 1. Data taxonomy for BrNPDB

3.2. Data Integration and Standards

Data integration was carried out to merge three sources of data, variables common to all three data sources were selected and described in table 1.

Table 1. Summary of variables used for BrNPDB data integration

Variable	Description
SMILES	Simplified Molecular-Input Line-Entry System
Reference	Digital Object Identifier - DOI
Natural Resource	Source type of the compound
Species	Species from where the compound was isolated
Location (State and City)	Sample location
Biological Property	Biological properties of the compound
Common Name	Common nomenclature of the compound

After selecting the variables considered in data integration and the desired data model for the BrNPDB, data cleaning and organization procedures were carried out to standardize the data and form the basis for generating the data that completes the data model, as shown in Figure 2.

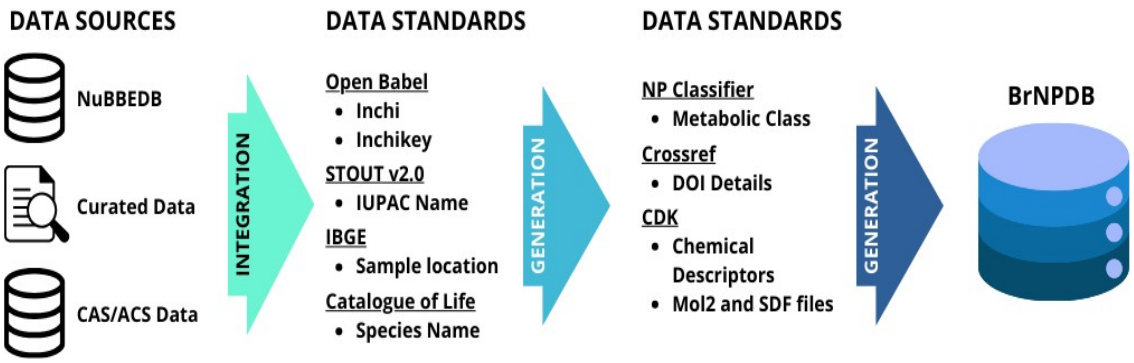

Figure 2. Data route to build BrNPDB

These choices for standards and generated data sources dialogue with the state-of-the-art in cheminformatics and bioinformatics, using Catalogue of Life, NP Classifier, Crossref, CDK and STOUT to improve the data sources and make easier to integrate BrNPDB with another databases. While the data from NuBBEDB is hard to manipulate nowadays, BrNPDB data was stored in a simple MySQL format that can easily be implemented in other data formats, such as PostgreSQL, SQLite and RDF versions, like ChEMBL [Zdrazil et al. 2023] and Pubchem [Kim et al. 2025] provides.

3.3. Search Engine

The search engine (Figure 3) was developed with seven filter blocks, covering general information, species, location, biological properties, source, reference, and chemical information. When performing a query, a list of compounds is displayed, based on the selected filters.

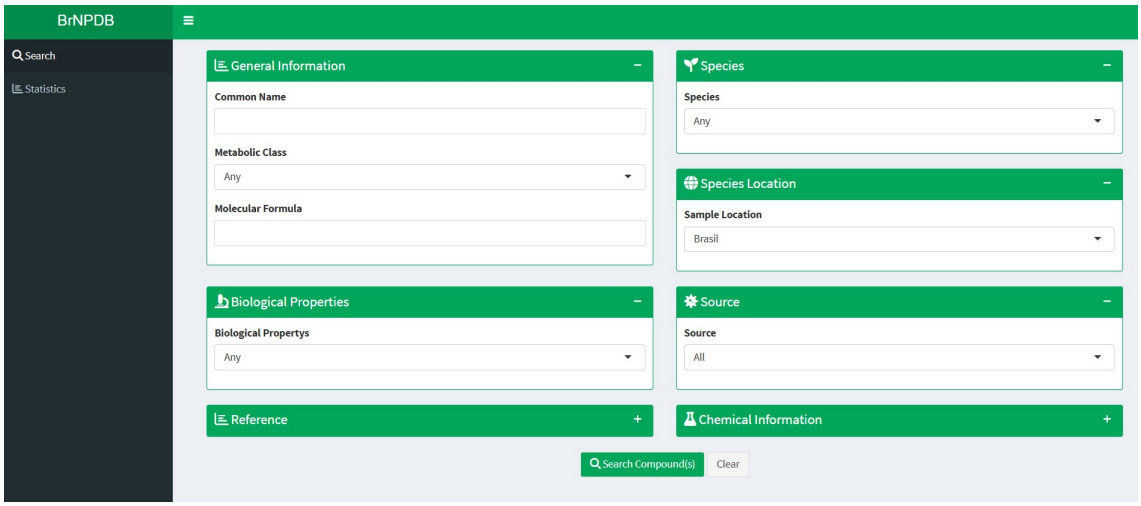

Figure 3. Search engine interface

When a compound from the result list is selected, a window with compound details will appear, as shown in Figure 4. Details from a compound include general information, chemical information, 2D visualization, and download options from Mol2 or SDF files. The BrNPDB search engine also can be used to identify relevant associations between species, biological properties and metabolic classification, besides providing useful chemical information for the scientific community.

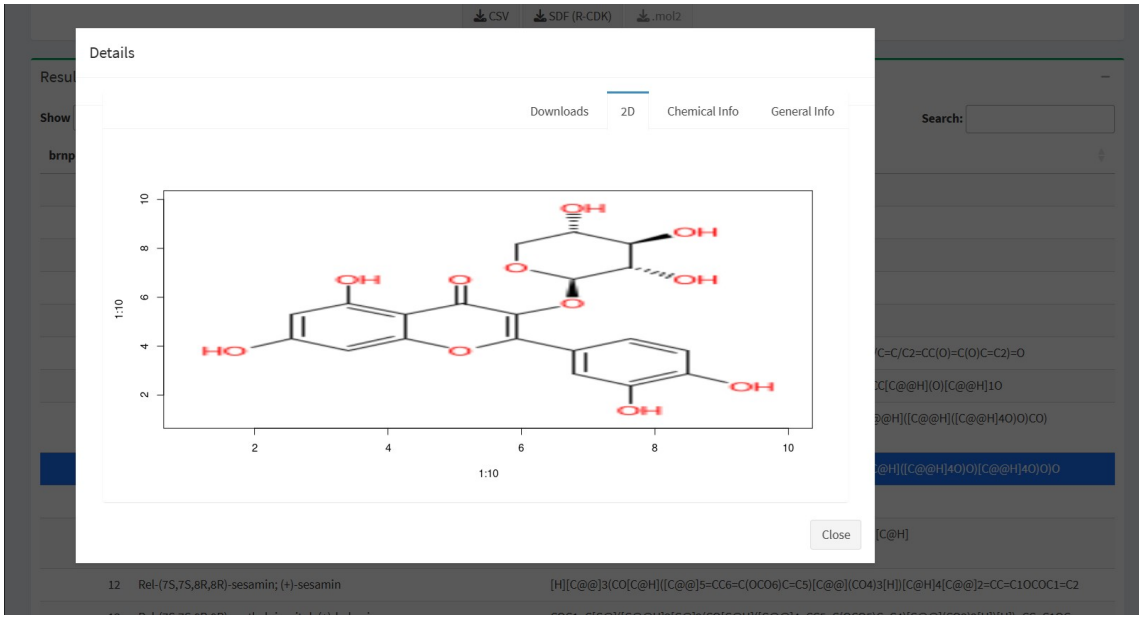

Figure 4. Search engine details

4. Results

The latest version of BrNPDB, on July 2025, records 9.962 compounds, 2.998 references, 2.229 species, 270 biological properties and 787 locations, organized into 47.522 analysis relationships. Regarding the source distribution of the compounds, the most frequent values are Plantae (82,6%), Animalia (7,58%) and Fungi (4,87%). All of this was classified as Natural Resources, while Semi-synthetic (2,65%) and Biotransformation products (0,98%) complete the sample types, as shown in figure 5.

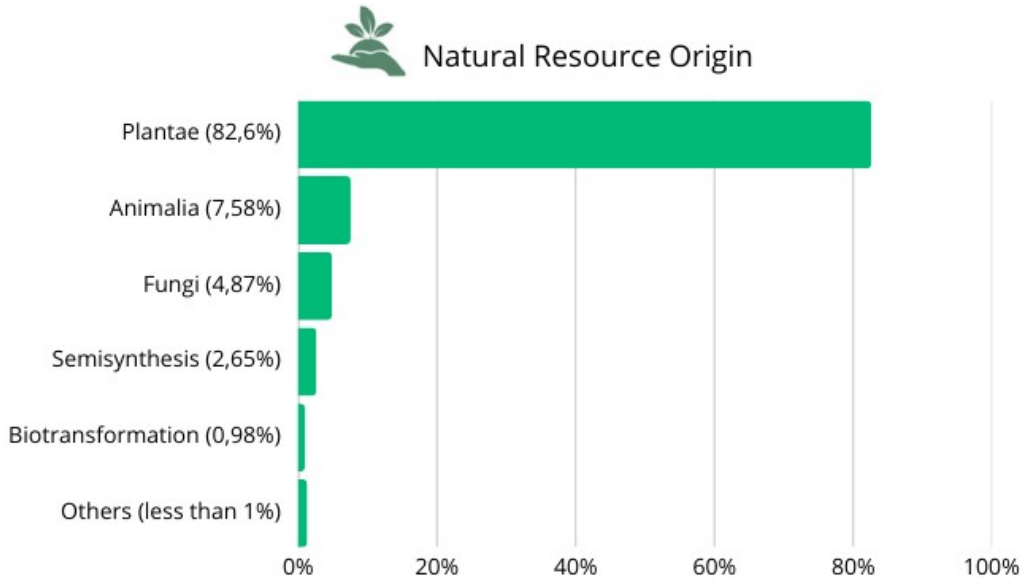

Figure 5. Distribution of sample's source types

The classification process in BrNPDB adopted the pattern suggested by the NP Classifier, which classifies the structure of a natural product at three levels into seven Pathways, 70 Superclasses, and 672 Classes, all generally recognized by the NP research community. The most frequently Pathways values are shown in figure 6.

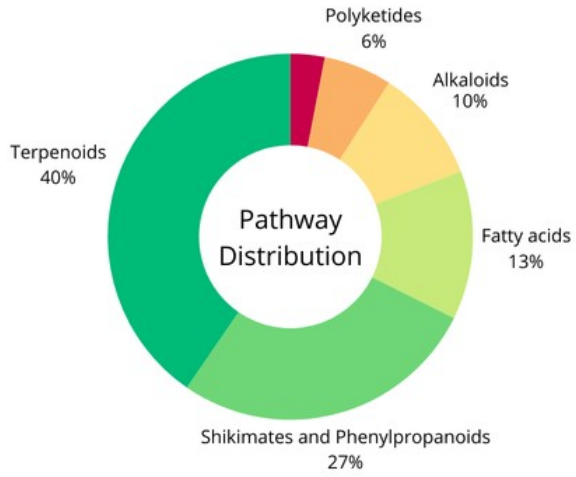

Figure 6. Distribution of compounds by metabolic class

According to the taxonomic distribution, the Asteraceae family is the most frequent in the BrNPDB (8,69%), followed by Fabaceae (8,54%), Myrtaceous (7,4%), Euphorbiaceous (3,67%), Lamiaceae (3,52%), Piperaceae (3,41%), Rubiaceous (2,59%), Lauraceous (2,02%), as shown in figure 7. The Rutaceae family, most frequently in NuBBE DB (14%) [Pilon et al. 2017], represents about 3% of the records in BrNPDB.

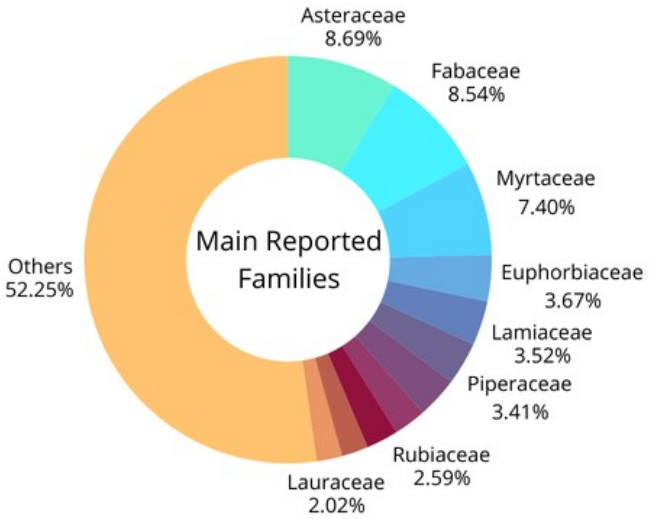

Figure 7. The families from which the most compounds were identified

Regarding biological activities frequency, the most common values in BrNPDB are Antioxidant (10%), Antimicrobial (8%), Cytoxic (5%), Antitumor (5%), Antibacterial (5%) and Fungicide (5%), as shown in figure 8. A different type of metric was shown in a study about the NuBBEDB update in 2017, associating Biological Properties with taxonomy distribution [Pilon et al. 2017].

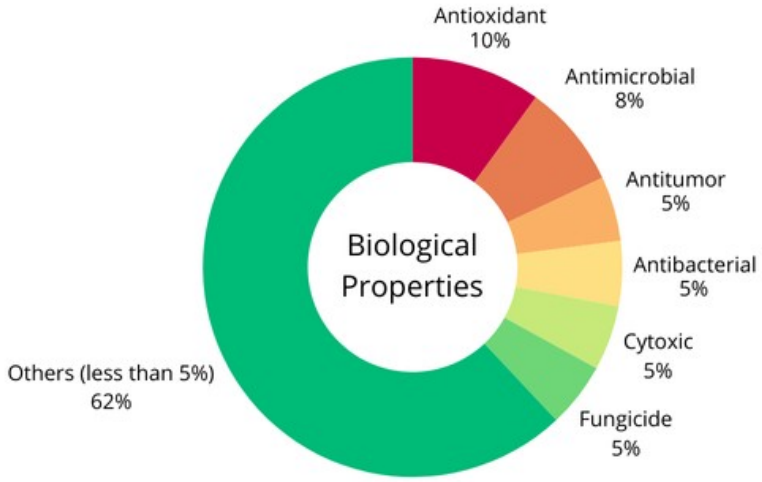

Figure 8. Biological properties distribution

About the sample location, Figure 9 shows what analysis was more frequent by regions of the country. This includes records about all the variables in data taxonomy and their multiples relationships, for example, some compound found from multiple sources, in distinct locations and study from distinct perspectives.

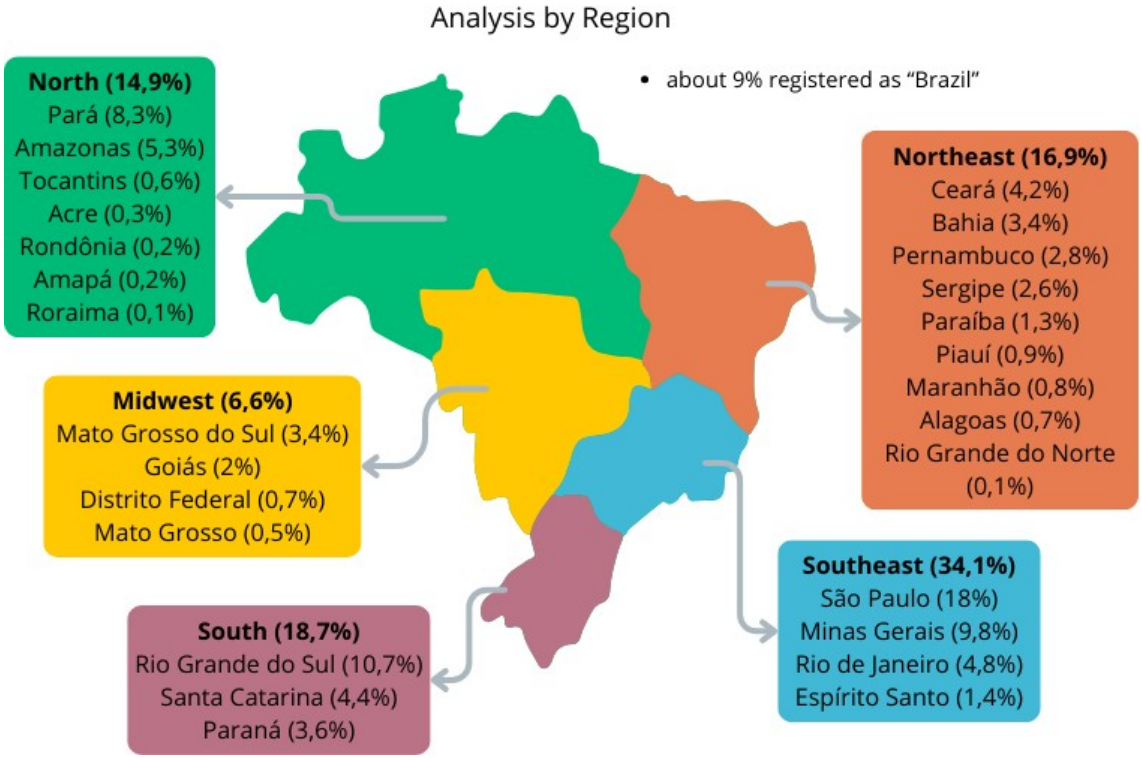

Figure 9. Region distribution analysis data from BrNPDB

Comparing all these values with those present in the latest version of NuBBEDB, we can verify an upgrade that may have occurred both from the addition of new information and from the standardization of values. The main upgrades from NuBBEDB latest version (obtained from a dump) to the BrNPDB are shown in Table 2.

Table 2. Comparing BrNPDB data with NuBBEDB data

Feature	***NuBBEDB (2017)***	***BrNPDB (2025)***
Compounds	2223	9962
References	715	2998
Species	441	2229
Main sample type source	Plants (78%)	Plants (82,6%)
Main reported family	Rutaceae (14%)	Asteraceae (8,69%)
Main metabolic class	Terpenoids (35%)	Terpenoids (40%)
Main biological property	Different evaluation	Antioxidant (10%)
Main location analysis	São Paulo (37%)	São Paulo (18%)

5. Discussion

Interesting dilemmas arose in the development of a natural products database, in a new context where computational technologies and applications using artificial intelligence (AI), machine learning (ML) and deep learning (DL) are on the rise. AI accelerates the way we conduct science, from folding proteins with AlphaFold [Jumper et al. 2021] and summarizing literature findings with large language models [Zhang et al. 2024] (LLMs), to annotating genomes and prioritizing newly generated molecules for screening using specialized software [Meijer et al. 2024]. The use of ML in natural product drug discovery has also been pivotal in identifying bioactive compounds and understanding their structural patterns for drug design [Saldívar-González et al. 2022].

This forces us to rethink methodologies and objectives in the face of so many possibilities. Unfortunately, in many cases, the integration between independent natural products resources is poor. This is due to both variations in compound content between databases and to challenges with standardization of compound structures and trivial names [Van Santen et al. 2019]. While available natural product data are multimodal, unbalanced, unstandardized, and scattered across multiple data repositories, a comprehensive graphical framework, incorporating all available data in natural product science, is the ideal framework for facilitating large-scale causal inference in the field of natural product science. Recently, a study showed with the Experimental Natural Products Knowledge Graph (ENPKG) how unpublished and unstructured data can be converted to public and connected data [Gaudry et al. 2024].

Initiatives like the BrNPDB play a crucial role in valuing and protecting this natural wealth. By documenting and making information available in a structured format, the database promotes the sustainable use of these resources and contributes to the development of innovative solutions in health, agriculture, and technology, strengthening Brazil's position in the international scientific landscape. However, challenges remain. Training researchers in the use of semantic technologies and expanding the database's coverage are essential aspects for consolidating the BrNPDB as a reference and improving the web interface and integrating it with other platforms remain priorities to increase the database's impact and usability.

6. Final Considerations

In this study, we demonstrate and discuss the development of a new Natural Product Database for secondary metabolites from the Brazilian biodiversity, the BrNPDB. The development of BrNPDB reflects the growing demand for platforms that enables more effective integration of chemical, biological, and ecological data. In addition to storing information on natural compounds, the database incorporates modern technologies, an approach that not only expands the reach of research on Brazilian Natural Products but also contributes to the preservation and appreciation of biodiversity in the global scientific landscape.

Finally, the BrNPDB represents not only a technological advancement in the management of information on natural products, but also a strategic initiative for the appreciation and preservation of Brazilian biodiversity, allowing researchers structured and reliable access to data essential for the country's scientific and technological development. The website and source code are available at http://brnpdb.fcfrp.usp.br, and updates will be implemented continuously.

References

Alho, C. (2008). “The value of biodiversity”. *In: Brazilian J. Biol.* 68, 1115–1118

Bánki, O., Roskov, Y. et al. (2025). “Catalogue of Life (Version 2025-07-10)”. *Catalogue of Life Foundation,* Amsterdam, Netherlands.

Calixto, J. B. (2019). “The role of natural products in modern drug discovery”. *Anais da Academia Brasileira de Ciências*, v. 91, n.3.

Cao Y., Charisi A., et al. (2008). “ChemmineR: a compound mining framework for R.” *Bioinformatics*, 24 (15), 1733–1734.

Chang W., Cheng J. et al. (2025). “Shiny: Web Application Framework for R”. *R package version* 1.11.1.9000

Chen, C.Y.-C (2011). "TCM Database@Taiwan: The world’s largest traditional Chinese medicine database for drug screening in silico". *PLoS ONE*, 6, e15939.

Gaudry, A., Pagni, M., et al (2024). “A Sample-Centric and Knowledge-Driven Computational Framework for Natural Products Drug Discovery”. *ACS Cent Sci.* 2024 Feb 20;10(3):494-510.

Gómez-García, A., Jiménez, D. A. A. et al. (2024). “Latin American Natural Product Database (LANaPDB): An Update”, *In: J. Chem. Inf. Model.* 64 (22), 8495-8509

Guha, R. (2007). “Chemical Informatics Functionality in R.” *J. of Stat. Software*, 18

Hatherley, R., Brown, D. K. et al (2015). "SANCDB: A South African natural compound database". J. *Cheminform.,* 7, 29.

Jumper, J. et al. (2021) "Highly accurate protein structure prediction with AlphaFold." *Nature* 596, 583–589

Kim S., Chen J., Cheng T., et al. (2025). “PubChem 2025 update”. *Nucleic Acids Res*. 2025;53(D1): D1516-D1525.

Kim, H.W., Wang, M. et al (2021). “NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products”. *J. Nat. Prod.*, 84, 2795–2807

Kim, S. K., Nam, S. et al (2015). “TM-MC: A database of medicinal materials and chemical compounds in Northeast Asian traditional medicine”. *BMC Complement. Altern. Med.,* 15, 218.

Marx E., Schmidt-Dichte S. and Mocsy I. J. (2023). “nubbe: The data model for nubbeDB Knowledge Graph”. https://nubbekg.aksw.org/ontology/index.html

Meijer, D., Beniddir, M. A. et al (2024). “Empowering natural product science with AI: leveraging multimodal data and knowledge graphs”, *Royal Society of Chemistry, Nat. Prod. Rep.*

Newman, D. J., Cragg, G. M. (2020). “Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019”. *Journal of Natural Products*, 27,83(3), 770-883.

Ntie-Kang, F., Telukunta, K.K. et al (2017). "NANPDB: A Resource for Natural Products from Northern African Sources". *J. Nat. Prod.*, 80, 2067–2076.

O'Boyle, N. M., Banck, M., et al (2011). “Open Babel: An open chemical toolbox”. *J Cheminform* 3, 33.

Pilon, A., Valli, M., et al. (2017). “NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity”, *In: Scientific Reports* | 7: 7215

R Core Team (2024). “R: A Language and Environment for Statistical Computing”. *R Foundation for Stat. Computing,* Vienna, Austria.

Rajan, K., Zielesny, A., and Steinbeck, C. (2022). “STOUT-V2 (2.0.0)”. *Zenodo*.

Saldívar-González, F. I., Aldas-Bulos, V. D., et al. (2022). “Natural product drug discovery in the artificial intelligence era”, *Chem. Sci.*, 13, 1526-1546.

Scotti, M.T., Herrera-Acevedo, C. et al (2018). “SistematX, an OnlineWeb-Based Cheminformatics Tool for Data Management of Secondary Metabolites”. *Molecules*, 23, 103

Sorokina, M., Steinbeck, C. (2020). “Review on natural products databases: where to find data in 2020”. J. *Cheminform* 12, 20.

Steinbeck C., Chandrasekhar V., Rajan K., et al. (2025). “COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database”. *Nucleic Acids Research*, Volume 53, Issue D1, 6 January 2025, Pages D634–D643

Thessen, A. E., Patterson, D. J. (2011). “Data issues in the life sciences”. *ZooKeys*, v. 150, p. 15-51.

Valli, M., dos Santos, R. N. et al (2013). "Development of a natural products database from the biodiversity of Brazil". *J. Nat. Prod.*, 76, 439–444.

Van Santen J. A., Jacob, G. et al. (2019). “The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery”. *ACS Central Science* 2019 5 (11), 1824-1833

Waagmeester, A.; Schriml, L., Su, A. (2019). “Wikidata as a linked-data hub for Biodiversity data”. *Biodiversity Information Science and Standards,* 3, e35206, 2019.

Xue, R., Fang, Z. et al (2013). "TCMID: Traditional Chinese medicine integrative database for herb molecular mechanism analysis". *Nucleic Acids Res.*, 41, D1089–D1095.

Zdrazil, B., Felix, E. et al (2024). “The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods”, *Nucleic Acids Research*, Volume 52, Issue D1, Pages D1180–D1192.

Zhang, Y., Jin, H., et al. (2024). “A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods.” *arXiv preprint,* arXiv:2403.02901.