Received: 8 February 2024 | Revised: 3 April 2024 | Accepted: 13 April 2024

DOI: 10.1002/minf.202400052

RESEARCH ARTICLE

molecular informatics

Updating and profiling the natural product-likeness of Latin American compound libraries

Alejandro Gómez-García[1] | Ann-Kathrin Prinz[2] | Daniel A. Acuña Jiménez[3] | William J. Zamora[3, 4, 5] | Haruna L. Barazorda-Ccahuana[6] | Miguel Á. Chávez-Fumagalli[6] | Marilia Valli[7] | Adriano D. Andricopulo[7] | Vanderlan da S. Bolzani[8] | Dionisio A. Olmedo[9] | Pablo N. Solís[9] | Marvin J. Núñez[10] | Johny R. Rodríguez Pérez[11, 12] | Hoover A. Valencia Sánchez[11] | Héctor F. Cortés Hernández[11] | Oscar M. Mosquera Martinez[13] | Oliver Koch[2] | José L. Medina-Franco[1]

[1]DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico

[2]Institute of Pharmaceutical and Medicinal Chemistry, Westfälische Wilhelms-Universität Münster, Münster, Germany

[3]CBio3 Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José, Costa Rica

[4]Laboratory of Computational Toxicology and Artificial Intelligence (LaToxCIA), Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Costa Rica

[5]Advanced Computing Lab (CNCA), National High Technology Center (CeNAT), Pavas, San José, Costa Rica

[6]Computational Biology and Chemistry Research Group, Vicerrectorado de Investigación, Universidad Católica de Santa Maria, Arequipa, Peru

[7]Laboratory of Medicinal and Computational Chemistry (LQMC), Centre for Research and Innovation in Biodiversity and Drug Discovery

Abstract

Compound databases of natural products play a crucial role in drug discovery and development projects and have implications in other areas, such as food chemical research, ecology and metabolomics. Recently, we put together the first version of the Latin American Natural Product database (LANaPDB) as a collective effort of researchers from six countries to ensemble a public and representative library of natural products in a geographical region with a large biodiversity. The present work aims to conduct a comparative and extensive profiling of the natural product-likeness of an updated version of LANaPDB and the individual ten compound databases that form part of LANaPDB. The natural product-likeness profile of the Latin American compound databases is contrasted with the profile of other major natural product databases in the public domain and a set of small-molecule drugs approved for clinical use. As part of the extensive characterization, we employed several chemoinformatics metrics of natural product likeness. The results of this study will capture the attention of the global community engaged in natural product databases, not only in Latin America but across the world.

KEYWORDS

chemical space, chemoinformatics, databases, LANaPDB, natural products

COCONUT, Collection of Open Natural Products; FDA, Food and Drug Administration; HTS, high throughput screening; KDE, kernel density estimate; LANaPDB, Latin American Natural Product Database; ML, machine learning; NP, natural product; NPs, natural products; N3PL, neural networks natural product-likeness; NPLC, natural product-likeness calculator; SM, synthetic molecule; SMs, synthetic molecules.

(CIBFar), São Carlos Institute of Physics (IFSC), University of São Paulo (USP), São Carlos, SP, Brazil

[8]Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, São Paulo State University (UNESP), Araraquara, SP, Brazil

[9]Center for Pharmacognostic Research on Panamanian Flora (CIFLORPAN), College of Pharmacy, University of Panama, Panama City, Panama

[10]Natural Product Research Laboratory, School of Chemistry and Pharmacy, University of El Salvador, San Salvador, El Salvador

[11]GIFAMOL Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira, Colombia

[12]GIEPRONAL Research Group, School of Basic Sciences, Technology and Engineering, Universidad Nacional Abierta y a Distancia, Dosquebradas, Colombia

[13]GBPN Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira, Colombia

Correspondence
Alejandro Gómez-García and José L. Medina-Franco, DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico.
Email: alex.go.ga21@hotmail.com and medinajl@unam.mx

Oliver Koch, Institute of Pharmaceutical and Medicinal Chemistry, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany.
Email: oliver.koch@uni-muenster.de

1 | INTRODUCTION

In addition to studies on biological diversity and ecology, natural products (NPs) have a unique and favorable property profile that is particularly useful in drug discovery [1]. It is therefore of most importance to collect available NPs and make them available for the identification of new bioactive molecules. The first chemoinformatics analysis that involved a compendium of natural product (NP) collections in the public domain was published over ten years ago [2] while the number of public databases has increased significantly in the past few years. As extensively reviewed elsewhere [3–5], more than twenty NP collections are organized by, for example, source (e.g., plants, fungi metabolites, marine) and geographical region (e.g., country or continent). In this context, Latin America is one of the largest biodiverse regions in the world. Hence, several countries have been developing NP collections and making them public [6]. In a collective effort to join NP databases in a single public repository, six countries so far have joined efforts and have put together a unified database termed the Latin

American Natural Product Database (LANaPDB) that currently holds more than 13,000 compounds [6,7]. The database is freely available at https://github.com/alexgoga21/LaNaPDB. Such efforts are contributing to further advancing the progress of chemoinformatics in Latin America [8].

Quantifying the NP-likeness of compound libraries is a relevant feature to fully characterize the contents of compound databases from their natural origin. For instance, to prioritize the most or least suitable libraries for a drug discovery project involving virtual screening and designing NP-like libraries [9]. To this end, several chemoinformatics metrics have been developed and used [9–13]. For instance, a new NP-likeness score based on a trained neural network was recently developed by some of the authors which can differentiate very well between NPs and synthetic molecules (SMs) [13]. Concerning the profiling of Latin American databases, the authors previously conducted an NP-likeness study of the Mexican database BIOFACQUIM [15] and other NP collections [16]. However, most of the compound libraries in LANaPDB have not been analyzed in terms of NP-likeness.

The goal of this work is to conduct a comprehensive profiling of NP-likeness of an updated version of LANaPDB that now has ten compound collections. The natural product-likeness profile of LANaPDB is discussed with two other major reference libraries, namely the Collection of Open Natural Products (COCONUT) [4], and a set of small-molecule drugs approved by the United States Food and Drug Administration (FDA) for clinical use [19].

2 | MATERIALS AND METHODS

2.1 | Data sets and data curation

In this study we profiled the updated, second version of LANaPDB. The first version of the database had 12,959 NPs coming from nine different databases of six different Latin American countries. In the first version of LANaPDB was added a new database: NPDB EjeCol which is a compilation of NPs isolated and characterized in Colombia, specifically from the region known as the Coffee Region. This database is set to be published in 2024 and will be accessible through an open-data portal. Furthermore, LANaPDB was updated with new NPs from Costa Rica (NAPRORE-CR) and Mexico (BIOFACQUIM). In total, 619 new compounds were added to LANaPDB, to have a total of 13,578 NPs in the second version of the database. The curation process of the new NPs added to LANaPDB was the same as that implemented in the first version of the database [7]. The process was carried out in the Python programming language (version 3.10.7), employing the RDKit (version 2022.03.5) [17] and MolVS (version 0.1.1) [18] modules. The standard curation process of MolVS was implemented: removal of explicit hydrogens, disconnection of covalent bonds between metals and organic atoms (the disconnected metal is not preserved), application of normalization rules (transformations to correct common drawing errors and standardization of functional groups), reionization (ensure the strongest acid groups protonate first in partially ionized molecules), and recalculation of the stereochemistry (ensures preservation of the original stereochemistry). The salts were removed, keeping the largest fragment, which was neutralized, and the remaining partially ionized fragments were reionized. The canonical tautomer was determined, and, from the InChIKey strings of the canonical tautomer, the duplicate compounds were removed.

The NP-likeness profile of the ten individual compound databases in LANaPDB and the entire collection was compared to the profile of COCONUT, one of the largest collections of NPs in the public domain. A set of drugs approved for clinical use was also used as a reference. The same curation process described above was applied to the two reference databases. Table 1 summarizes the data sets analyzed in this work and Table S1 in the Supplementary information summarizes the number of compounds analyzed with each approach.

The NP-likeness profile of LANaPDB and the two reference datasets was depicted with kernel density estimate (KDE) plots which represent the data using continuous probability density curves. The KDE plots were

TABLE 1 Compound databases that are analyzed in this work.

Database	Description	References
LANaPDB (version 2)	A database aimed to gather and standardize the natural product databases of Latin America. Composed of 13,578 natural products isolated and characterized in Latin America.	[7][a]
COCONUT	One of the most comprehensive, freely accessible natural product databases with more than 411,000 compounds from 50 open access natural product databases.	[4]
Approved drugs	FDA-approved small-molecule drugs, version 5.1.10 (released by DrugBank in January 2023).	[19]

[a] The cited reference [6] alludes to the first version of LANaPDB. This manuscript reports the second and updated version of LANaPDB.

created in the Python programming language (version 3.10.7), employing the seaborn module (version 0.12.2).

2.2 | Quantification of natural product-likeness

We used three well-known and validated approaches, summarized in Table 2.

3 | RESULTS AND DISCUSSION

3.1 | LANaPDB update

Table 3 summarizes the contents of the most current and updated version of LANaPDB. As commented in section 2.1 the previous version of LANaPDB was updated with 619 compounds adding a new data set (NPs from Colombia), and updated databases from Costa Rica and Mexico. Initially, 1,707 compounds were considered for

TABLE 2 Approaches used in this work to profile the NP-likeness of compound databases.

Approach	Basis of the method and accessibility	References
NP-likeness calculator (NPLC)	http://sourceforge.net/projects/np-likeness/	[10]
	Machine learning approach. NP-scout also generates similarity maps, highlighting atoms contributing significantly to the classification of small molecules as a NP or SM. Accesible at https://nerdd.univie.ac.at/	[12]
Neural networks NP-likeness (N3PL) score	Scorer based on a multi-layer perceptron network and a training database of natural products and synthetic compounds. The code is freely available at https://github.com/kochgroup/neural_npfp; natural product likeness score	[13]

TABLE 3 Natural product databases in the updated version of LANaPBD.

Database	Size	Source	General description	References
$NuBBE_{DB}$ (Brazil)	2223	Plants Microorganisms Terrestrial and marine animals	Natural products of Brazilian biodiversity. Developed by the São Paulo State University and the University of São Paulo.	[20,21]
SistematX (Brazil)	9514	Plants	Database composed of secondary metabolites and developed at the Federal University of Paraiba.	[22,23]
UEFS (Brazil)	503	Plants	Natural products that have been separately published, but there is no common publication nor public database for it. Developed at the State University of Feira de Santana.	[24]
NPDB EjeCol (Colombia)	200	Plants Plants-derived food	Natural products and foods derived from plants present in the Eje Cafetero Región of Colombia, database created and curated at the Technological University of Pereira.	[a]
NAPRORE-CR (Costa Rica)	~1600	Plants Microorganisms	Developed in the CBio3 and LaToxCIA Laboratories of the University of Costa Rica.	[a]
LAIPNUDELSAV (El Salvador)	214	Plants	Developed by the Research Laboratory in Natural Products of the University of El Salvador.	[a]
UNIIQUIM (Mexico)	1112	Plants	Natural products isolated and characterized at the Institute of Chemistry of the National Autonomous University of Mexico.	[25]
BIOFACQUIM (Mexico)	750	Plants Fungus Propolis Marine animals	Natural products isolated and characterized in Mexico at the School of Chemistry of the National Autonomous University of Mexico and other Mexican institutions.	[15,26]
CIFPMA (Panama)	363	Plants	Natural products that have been tested in over twenty-five in vitro and in vivo bioassays for different therapeutic targets. Developed at the University of Panama.	[27,28]
PeruNPDB (Peru)	280	Animals Plants	Natural products representative of Peruvian biodiversity. Created and curated at the Catholic University of Santa Maria.	[29]

[a] The database has not been published yet.

the update of LANaPDB from the two updated databases BIOFACQUIM, NAPRORE-CR, and the new database NPDB EjeCol. Nevertheless, from the initial 1,707 compounds, 1,088 molecules were duplicates and were no longer included. The remaining 619 molecules were added to LANaPDB. The NP-likeness scores of LANaPDB, COCONUT and approved drugs are freely available at https://doi.org/10.17879/77968651865.

3.2 | Profiling with NP-likeness calculator

The NP-likeness calculator (NPLC) was reported in the year 2012 [10] and is an open source re-implementation of the well-known Ertl NP-likeness score created at Novartis in 2008 [9]. The NP-likeness score of Novartis [9] was incorporated into several standard processes including virtual screening, selection of compound samples for purchasing, prioritizing hits from high-throughput screening (HTS), and library design. The NPLC's approach divides a molecule into structural fragments and every fragment has a different contribution to the NP-likeness. The contribution of a structural fragment to the NP-likeness depends on their frequency of appearance in the reference libraries (training sets) used by the developers of NPLC, namely, NPs and SMs. The contribution of a structural fragment to the NP-likeness is positive if the fragment is present in the NPs training dataset and negative if the fragment is present in the SMs training dataset. The positive or negative contribution of the structural fragment will be greater according to their frequency of appearance in the NPs or SMs datasets.

The NP-likeness of the LANaPDB compounds was calculated with the NPLC and compared with two reference datasets: COCONUT and a set of drugs approved for clinical use by the FDA. Figure 1A shows that most of the numeric values of the NP-likeness of LANaPDB are positive, in fact, only 1% of the compounds have negative values. Therefore, the proportion of the structural fragments of the molecules in the database comes predominantly from NPs and just a little proportion from SMs. In the reference dataset COCONUT, it was found that 74.8% of the compounds have a NP-likeness positive score, nonetheless, 25.2% of the compounds have a negative NP-likeness positive score. Hence, a quarter of the COCONUT compounds have predominantly overlapping structural fragments with synthetic drugs. In the reference dataset of FDA-approved small-molecule drugs 51.2% of the compounds have negative NP-likeness values. Thus, these molecules have mainly a synthetic origin. Nonetheless, 48.6% of the FDA-approved small-molecule drugs have positive scores, which mean that they are NP-derivatives or NP-like. It is important to consider that a high proportion of the small-molecule approved drugs are, of course, NPs and NP-derivatives. For instance, from 1946 to 1980, 53% of the small molecules approved were NPs or NP-derivatives and from 1981 to 2019 the proportion increased to 64.9% [30].

The NP-likeness score of the individual Latin American countries is mostly positive (Figures 2A and A1). In most of the countries the distribution of the score is around the numbers one and two. The presence of at least one peak around the number one and two is especially noticeable in the case of Brazil, Costa Rica, El Salvador, Mexico, and Peru. A possible explanation for the peaks around the NP-likeness one and two is that a certain combination of NPs fragments is mainly present in LANaPDB.

3.3 | Profiling with NP-Scout

NP-Scout [12], reported in the year 2019, is a practical machine learning (ML) approach based on random

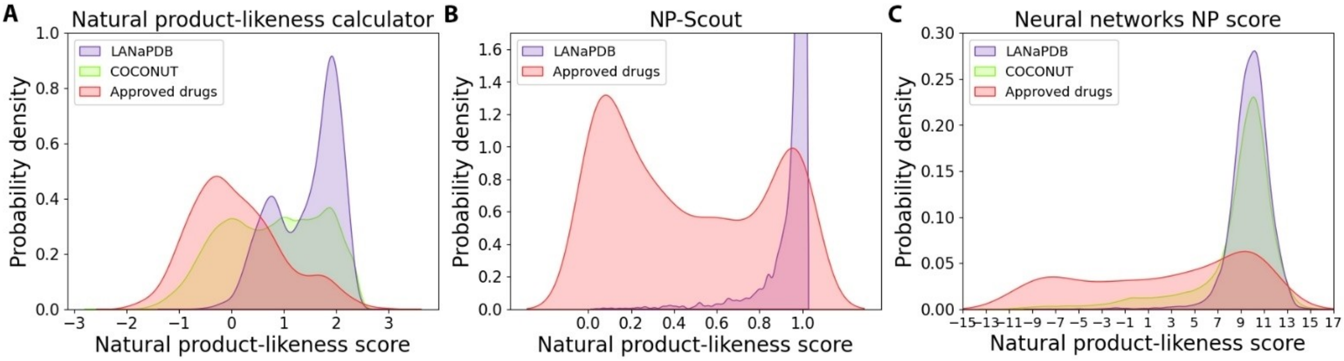

FIGURE 1 Kernel density estimate plots that represent the distribution of the natural product-likeness scores of LANaPDB, COCONUT and approved drugs calculated with three different algorithms: A) Natural product-likeness calculator [10] B) NP-Scout [12] and C) Neural networks NP score [13].

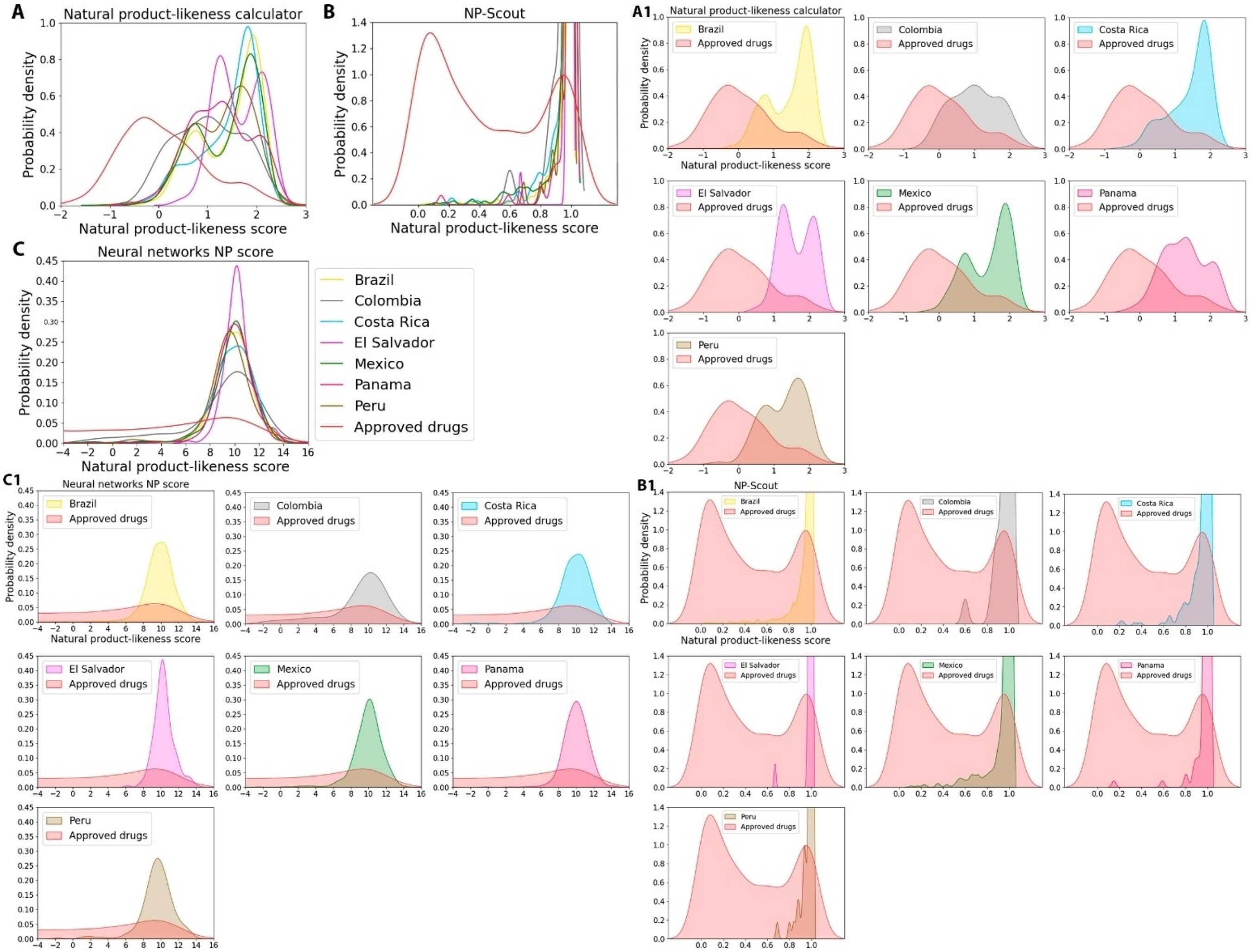

FIGURE 2 Kernel density estimate plots that represent the distribution of the natural product-likeness scores of LANaPDB and approved drugs databases calculated with three different algorithms: **A,A1)** Natural product-likeness calculator [10] **B,B1)** NP-Scout [12] **C,C1)** Neural networks NP score [13]. The distribution of the natural product-likeness score of the LANaPDB compounds is depicted with different colors according to the Latin American country of origin.

forest classifiers for the differentiation of NPs and SMs and for the quantification of NP-likeness. An additional value of NP-scout is the implementation of similarity maps to visualize atoms in molecules making decisive contributions to the assignment of compounds to NP or synthetic molecule (SM). The similarity maps highlight the atoms that contribute significantly to the classification of small molecules as NPs or SMs. The range of the NP class probability is from zero to one: one represents a probability of 100% of a compound to be a NP and zero represents a probability of 0% to be a NP. The SMs have values of zero or close to zero. Figure 1B shows that the LANaPDB compounds have a NP class probability of one or very near to one. Therefore, according to NP-Scout, the probability of the LANaPDB compounds to be labeled as NPs is very close to 100%. In the approved drugs set, 56.5% of the compounds have a NP-class probability between 0 and 0.5 and 43.5% among 0.51 and 1. Therefore, 56.5% of the compounds have a high probability of being SMs and 43.5% a high probability of being NPs. In the approved drugs, the prevalence of SMs over NPs aligns with the findings of the NP-likeness calculator, which indicates that SMs have a higher proportion than NPs. The above is evident in the Figure 1B, the highest peak is around zero and the peak around one is shorter. According to the NP class probabilities of the individual Latin American countries (Figures 2B and B1) the values are around one in every country. Thus, in all the Latin American countries the compounds have a probability of 100% or near to 100%, to be labeled as NPs.

3.4 | Profiling with neural networks NP score

The most recent methodology to calculate the NP-likeness [13] was reported in the year 2021. It is based on the training of neural networks to produce molecular representations which are better suited for NPs. The NP-likeness score is extracted from the trained neural networks. On this approach, the neural network NP-likeness (N3PL) score values for the NPs are around ten and minus ten for SMs.

For the LANaPDB database, the NP-likeness scores are around ten (Figure 1C), which means that the compounds are labeled as NP-like. The N3PL scores of the COCONUT compounds are also around ten, consequently, these compounds are labeled as NP-like. Nonetheless, in COCONUT the distribution of the NP-likeness scores is extended to the negative values to the LANaPDB distribution. The compounds with NP-likeness scores extended more to the left can be considered as less NP-like. The presence of less NP-like compounds in COCONUT is consistent with the NP-likeness calculator results, which identified compounds containing predominantly structural fragments found in small molecule drugs (Figure 1A).

The distribution of the N3PL scores of the approved drugs encompasses all the range from the NP-like molecules to the small molecule drugs. The distribution shows that the NP-like molecules have a higher abundance than the small molecule drugs. NP-likeness calculator and NP-Scout results showed the opposite trend: a higher abundance of synthetic-like molecules (Figure 1A). This can be explained by the neural network training, the underlying training dataset and the properties of drug-like molecules. The training dataset was compiled out of the COCONUT database and SMs from the ZINC database with an Ertl score of less than zero. Interestingly, the ZINC database shows only a small number of compounds with an Ertl score greater than zero [13], which is in contrast to the analysis of approved drugs. This shows that approved drugs have different properties and are more NP-like than the space of synthetic available compounds. In contrast to the Ertl score, the N3PL score was not trained on specific fragments but on NP-likeness and 48 additional surface descriptors to cover the physicochemical space of NPs. Since NPs show a property profile that is particularly useful for drug discovery, it is not a surprise that developed drugs also show a similar property profile which is reflected in the N3PL score. To summarize, the NPLC score analyzes typical NP fragments and the N3PL likeness score also includes NP physicochemical properties besides structural features. This leads to a higher NP-likeness in the approved drugs. This also explains the rather smooth distribution around a score of ten in contrast to NPLC scores with often two peaks. Regarding the N3PL scores of the individual Latin American countries (Figures 2C and C1), they showed a nearly normal distribution form centered in the number ten. Thus, these compounds are labeled as NP-like. The NP-likeness score calculated with the neural networks approach has the least variation among the three methodologies in the case of LANaPDB, showing more variation in the previous two methodologies (Figures 1A and B) and a normal distribution form with the neural networks (Figure 1C). Regarding COCONUT, there is also less variation in the N3PL score (Figure 1C) compared to the NP-likeness calculator (Figure 1A).

4 | CONCLUSIONS

LANaPDB was updated with new NPs from Costa Rica (NAPRORE-CR) and Mexico (BIOFACQUIM). In total, 619 new compounds were added to LANaPDB. The three methodologies employed for the calculation of the NP-likeness scores have a good performance distinguishing NPs or identifying NP-likeness, but have different definitions of NP likeness. The NPLC score, which reimplements the well-known Ertl score, is based on fragments that are typical for NPs with the goal to create a score that evaluates the NP likeness. In contrast, the N3PL network was trained to identify natural products and thus it evaluates how likely a molecule is a natural product. Menke et al. describe a good example of a furan which is ranked high by the network score but low by Ertl's score. It is part of the COCONUT database but have many specific NP-like features [13]. In addition, the neural network based N3PL score does not only take structural features into account, but also surface descriptors. This means that this definition is not only a structural NP-like score but also a NP-property like score. This leads to slightly different outcomes in the NP-likeness analysis. From the pure structural point of view, the NPLC score revealed a large variability between the analyzed NP collections and distinct combinations of NP fragments present in LANaPDB. The NP-likeness profile determined with NP-Scout and N3PL indicates that the NPs of LANaPDB have a probability of 100 % or near to 100 %, to be labeled as NPs. According to the results of NP-likeness calculator and NP-Scout, the FDA-approved small-molecule drugs reference dataset is composed of more synthetic-like molecules. The N3PL shows a distribution that is slightly shifted to more NP likeness. This is due to the fact that desired drug properties are of course similar to NP properties.

AUTHOR CONTRIBUTION STATEMENT

Conceptualization, J.L.M.F., O.K.; methodology, J.L.M.F., A.G.G., A-K.P.; software, J.L.M.F.; validation, W.J.Z., M.A.C.F., D.A.O.; formal analysis, J.L.M.F., A.G.G., A-K.p.; investigation, A.G.G., W.J.Z., H.L.B.C, M.A.C.F., M.V., A.D.A., V.S.B., D.A.O., P.N.S, M.J.N., J.R.R.P, H.A.V.S, H.F.C.H, O.M.M., O.K., J.L.M.F.; resources, J.L.M.F., O.K.; data curation, A.G.G.; writing—original draft preparation, J.L.M.F., A.G.G.; writing—review and editing, A.G.G., W.J.Z., H.L.B.C, M.A.C.F., M.V., A.D.A., V.S.B., D.A.O., P.N.S, M.J.N., J.R.R.P, H.A.V.S, H.F.C.H, O.M.M., O.K., J.L.M.F.; visualization, A.G.G.; supervision, O.K., J.L.M.F.; project administration, O.K., J.L.M.F.; funding acquisition, J.L.M.F., O.K., W.J.Z., M.A.C.F., M.V., A.D.A., V.S.B., D.A.O., M.J.N., O.M.M. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

We thank the critical reading and insightful comments of Samuel Homberg. The project was funded by DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), Grant No. IG200124. A.G.-G. thanks the Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCyT) for the PhD scholarship 912137. V.S.B, M.V. and A.D.A thanks the Sao Paulo Research Foundation (FAPESP) grants #2013/07600-3 (CIBFar-CEPID), #2014/50926-0/ #465637/2014-0 (INCT BioNat CNPq/FAPESP), and National Council for Scientific and Technological Development (CNPq) grant #382044/2023-1. The authors also thank the Technological University of Pereira (UTP) through the Vicerrectoria de investigaciones, Innovación y Extensión for the development of the funded project: "Development of a library of isolated and characterized natural products from plant species studied in the Coffee Axis region, Colombia," code E3-23-1. WZR and DAJ thanks to the Vice Chancellor for Research of the University of Costa Rica for grant via the research project 115-C2-126. OK and AKP gratefully acknowledge the financial support by the German Research Foundation (DFG) in the context of the Heisenberg-Program (KO 4689/5-1) and a Research Grant (KO 4689/6-1).

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The Latin American Natural Product Database (LANaPDB) is freely available at https://github.com/alexgoga21/LaNaPDB. The NP-likeness scores of LANaPDB, COCONUT, and approved drugs can be found in https://doi.org/10.17879/77968651865. The software used to calculate the NP-likness scores can be found in: http://sourceforge.net/projects/np-likeness/, https://nerdd.univie.ac.at/, and https://github.com/kochgroup/neural_npfp; natural product likeness score.

ORCID

Alejandro Gómez-García http://orcid.org/0000-0003-4444-8221

REFERENCES

1. A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, *Nat. Rev. Drug Discov.* **2021**, *20*, 200–216.
2. A. B. Yongye, J. Waddell, J. L. Medina-Franco, *Chem. Biol. Drug Des.* **2012**, *80*, 717–724.
3. Y. Chen, M. Garcia de Lomana, N.-O. Friedrich, J. Kirchmair, *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532.
4. M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, C. Steinbeck, *J. Cheminf.* **2021**, *13*, 2.
5. A. Gómez-García, J. L. Medina-Franco, *Biomolecules* **2022**, *12*, 1202.
6. J. L. Medina-Franco, *Future Science OA* **2020**, *6*, FSO468, https://doi.org/10.2144/fsoa-2020-0068.
7. A. Gómez-García, D. A. A. Jiménez, W. J. Zamora, H. L. Barazorda-Ccahuana, M. Chávez-Fumagalli, M. Valli, A. D. Andricopulo, V. S. Bolzani, D. A. Olmedo, P. N. Solís, M. J. Núñez, J. R. Rodríguez Pérez, H. A. Valencia Sánchez, H. F. Cortés Hernández, J. L. Medina-Franco, *Pharmaceuticals* **2023**, *16*, 1388.
8. J. Miranda-Salas, C. Peña-Varas, I. Valenzuela Martínez, D. A. Olmedo, W. J. Zamora, M. A. Chávez-Fumagalli, D. Q. Azevedo, R. O. Castilho, V. G. Maltarollo, D. Ramírez, J. L. Medina-Franco, *Artif. Intell. Life Sci.* **2023**, *3*, 100077.
9. P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
10. K. V. Jayaseelan, P. Moreno, A. Truszkowski, P. Ertl, C. Steinbeck, *BMC Bioinf.* **2012**, *13*, 106.
11. M. Sorokina, C. Steinbeck, *J. Cheminf.* **2019**, *11*, 55.
12. Y. Chen, C. Stork, S. Hirte, J. Kirchmair, *Biomolecules* **2019**, *9*, 43.
13. J. Menke, J. Massa, O. Koch, *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4593–4602.
14. D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
15. N. Sánchez-Cruz, B. A. Pilón-Jiménez, J. L. Medina-Franco, *F1000Res.* **2020**, *8*, 2071.
16. J. L. Medina-Franco, R. Gutiérrez-Nieto, H. Gómez-Velasco, in *Drug Target Selection and Validation* (Eds.: M.T. Scotti, C.L. Bellera), Springer International Publishing, Cham **2022**, 227–249.
17. Open-source chemoinformatics and machine learning., "RDKit: Open-Source Cheminformatics Software.," can be found under https://www.rdkit.org, 2022.
18. "MolVS. Molecule Validation and Standardization.," can be found under https://molvs.readthedocs.io/en/latest/index.html.
19. C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. L. Chin, S. A. Strawbridge, M. Garcia-Patino, R. Kruger, A. Sivakumaran, S. Sanford, R.

Doshi, N. Khetarpal, O. Fatokun, D. Doucet, A. Zubkowski, D. Y. Rayat, D. S. Wishart, *Nucleic Acids Res.* **2023**, *52*, 1265–1275.

20. M. Valli, R. N. dos Santos, L. D. Figueira, C. H. Nakajima, I. Castro-Gamboa, A. D. Andricopulo, V. S. Bolzani, *J. Nat. Prod.* **2013**, *76*, 439–444.
21. A. C. Pilon, M. Valli, A. C. Dametto, M. E. F. Pinto, R. T. Freire, I. Castro-Gamboa, A. D. Andricopulo, V. S. Bolzani, *Sci. Rep.* **2017**, *7*, 7215.
22. M. T. Scotti, C. Herrera-Acevedo, T. B. Oliveira, R. P. O. Costa, S. Y. K. O. Santos, R. P. Rodrigues, L. Scotti, F. B. Da-Costa, *Molecules* **2018**, *23*, 103.
23. R. P. O. Costa, L. F. Lucena, L. M. A. Silva, G. J. Zocolo, C. Herrera-Acevedo, L. Scotti, F. B. Da-Costa, N. Ionov, V. Poroikov, E. N. Muratov, M. T. Scotti, *J. Chem. Inf. Model.* **2021**, *61*, 2516–2522.
24. "UEFS Natural Products," can be found under http://zinc12.docking.org/catalogs/uefsnp.
25. "UNIIQUIM," can be found under https://uniiquim.iquimica.unam.mx/.
26. B. A. Pilón-Jiménez, F. I. Saldívar-González, B. I. Díaz-Eufracio, J. L. Medina-Franco, *Biomolecules* **2019**, *9*, 31.
27. D. A. Olmedo, M. González-Medina, M. P. Gupta, J. L. Medina-Franco, *Mol. Diversity* **2017**, *21*, 779–789.
28. D. A. Olmedo, J. L. Medina-Franco, *Cheminformatics and its Applications.* IntechOpen; **2020**. Available from: https://doi.org/10.5772/intechopen.87779.
29. H. L. Barazorda-Ccahuana, L. G. Ranilla, M. A. Candia-Puma, E. G. Cárcamo-Rodriguez, A. E. Centeno-Lopez, G. Davila-Del-Carpio, J. L. Medina-Franco, M. A. Chávez-Fumagalli, *Sci. Rep.* **2023**, *13*, 7577.
30. D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2020**, *83*, 770–803.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: A. Gómez-García, A.-K. Prinz, D. A. A. Jiménez, W. J. Zamora, H. L. Barazorda-Ccahuana, M. Á. Chávez-Fumagalli, M. Valli, A. D. Andricopulo, V. d. S. Bolzani, D. A. Olmedo, P. N. Solís, M. J. Núñez, J. R. Rodríguez Pérez, H. A. V. Sánchez, H. F. Cortés Hernández, O. M. Mosquera Martinez, O. Koch, J. L. Medina-Franco, *Molecular Informatics* **2024**, *43*, e202400052. https://doi.org/10.1002/minf.202400052