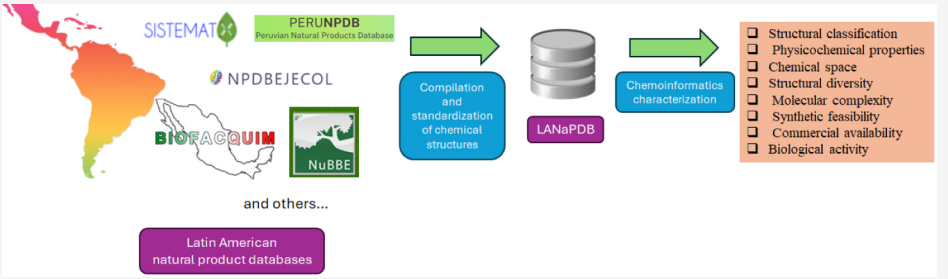
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

Latin American Natural Product Database (LANaPDB): An Update

Alejandro Gómez-García, Daniel A Acuña Jiménez, William J Zamora, Haruna L Barazorda-Ccahuana, Miguel Á. Chávez-Fumagalli, Marilia Valli, Adriano D Andricopulo, Vanderlan da S Bolzani, Dionisio A Olmedo, Pablo N Solís, Marvin J Núñez, Johny R Rodríguez Pérez, Hoover A Valencia Sánchez, Héctor F Cortés Hernández, Oscar M Mosquera Martinez, and José L Medina-Franco*

Cite This: J. Chem. Inf. Model. 2024, 64, 8495−8509



ABSTRACT: Natural product (NP) databases are crucial tools in computer-aided drug design (CADD). Over the past decade, there has been a worldwide effort to assemble information regarding natural products (NPs) isolated and characterized in certain geographical regions. In 2023, it was published LANaPDB, and to our knowledge, this is the first attempt to gather and standardize all the NP databases of Latin America. Herein, we present and analyze in detail the contents of an updated version of LANaPDB, which includes 619 newly added compounds from Colombia, Costa Rica, and Mexico. The present version of LANaPDB has a total of 13 578 compounds, coming from ten databases of seven Latin American countries. A chemoinformatic characterization of LANaPDB was carried out, which includes the structural classification of the compounds, calculation of six physicochemical properties of pharmaceutical interest, and visualization of the chemical space by employing and comparing two different fingerprints (MACCS keys (166-bit) and Morgan2 (2048-bit)). Furthermore, additional analyses were made, and valuable information not included in the first version of LANaPDB was added, which includes structural diversity, molecular complexity, synthetic feasibility, commercial availability, and reported and predicted biological activity. In addition, the LANaPDB compounds were cross-referenced to two of the largest public chemical compound databases annotated with biological activity: ChEMBL and PubChem.

INTRODUCTION

Historically, natural products (NPs) have been the largest source of inspiration for the design of new drugs. In recent years (2018 compared to 2006), there has been a significant increase in the number of NP-based drugs.[1] The recent technological advances, especially in the artificial intelligence (AI)[2] and chemoinformatics[3] areas, have boosted the NP-based computer-aided drug design (CADD). Among the recent progress in AI, the development of machine learning models to predict the target proteins of natural products stands out.[4] During the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic outbreak, the NP-based CAAD represented a main approach in the design and identification of lead compounds against the virus.[5−8] Natural product (NP) databases are crucial tools for CADD since they provide access to thousands of molecules. In the past few years, the number of NP databases has grown, and some of the databases already established are continuously being updated. Among the largest freely available NP databases is Supernatural 3.0, with 449 048 NPs.[9] The collection of open natural products (COCONUT 1.0)[9] contains 411 000 NPs, and the universal natural product database[10] has 229 000 NPs. The universal natural product database is still accessible on another online repository.[11] The NP activity and species source (NPASS) database[12] has 94 413 NPs, of which 43 285 are annotated with biological activity information. The Hippo(crates)[13] database contains 45 300 NPs, NP derivatives, and synthetic compounds, many of which

Received: August 29, 2024
Revised: October 19, 2024
Accepted: October 29, 2024
Published: November 6, 2024

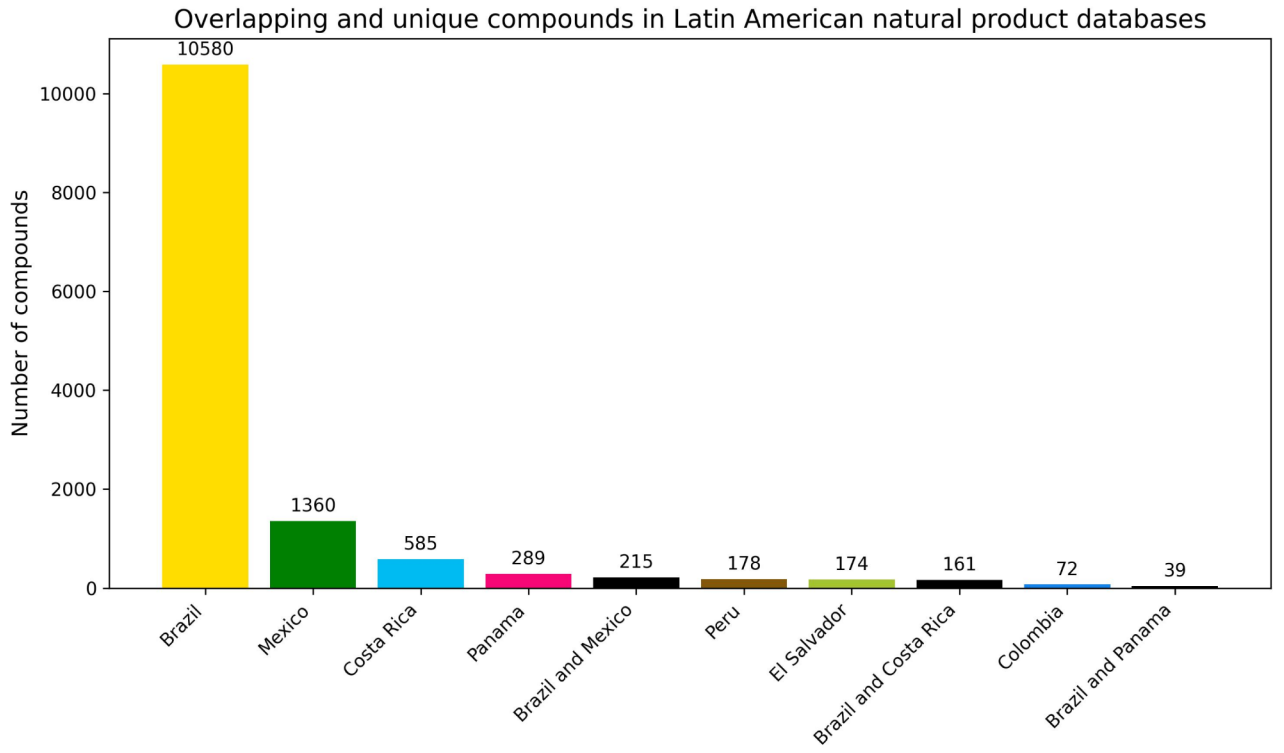


Figure 1. Bar chart showing the countries with the highest number of unique and overlapping compounds from the Latin American natural product databases contained in LANaPDB. The compounds are grouped according to the country of origin of the database.

are annotated with their biological targets. There are NP databases that contain NPs isolated and characterized in certain geographical areas. TCM@Taiwan[14] is the largest database of NPs from China, many of them employed in traditional Chinese medicine and contains 58 000 compounds. IMPPAT 2.0[15] is the largest compilation of NPs from India with 17 967 phytochemicals, employed in traditional Indian medicine. The largest collection of NPs from Africa is NANPDB,[16] containing over 4500 molecules from North Africa; nonetheless, there are other minor African NP databases.[17−20]

Latin America is a region with extraordinary biodiversity and richness in endemic species. It is a region that may be home to at least a third of global biodiversity.[21] Brazil, for example, is considered to host the earth's richest flora, with at least 50 000 species or one-sixth of the planetary total. Another example is Ecuador, with its megadiverse flora comprising more than 25 000 plant species (and thus twice the number of plant species found in Europe). Ecuador also has the highest vertebrate species density worldwide.[22] Therefore, Latin America is a major source of bioactive compounds. Moreover, it has been reported that several databases contain NPs isolated and characterized in Latin American countries. More than 92 molecules with therapeutic effects have been identified from Latin American NP databases.[23] Just recently, an NP database from Argentina[24] and Colombia[25] was published. In 2023, the first version of LANaPDB was published, a compendium that aims to gather and standardize the NP databases of Latin America[26] which was already included in COCONUT (https://coconut.naturalproducts.net/search?type=tags&q=Latin+America+dataset&tagType=dataSource).[9] In early 2024, an update was reported regarding the NP-likeness profile of the database.[27]

Herein, we report a major update of LANaPDB,[23] a compound collection that aims to gather and standardize all the Latin American NP databases. The analysis of the database includes the structural classification of the compounds, calculation of six physicochemical properties of pharmaceutical interest, and visualization of the chemical space by employing and comparing two different fingerprints (MACCS keys (166-bit) and Morgan2 (2048-bit)). Furthermore, additional analyses were made, and valuable information not included in the first version of LANaPDB was added, which includes structural diversity, molecular complexity, synthetic feasibility, commercial availability, and reported and predicted biological activity. Moreover, the database was cross-referenced to two of the largest public chemical compound databases annotated with biological activity: ChEMBL[28] and PubChem.[29]

METHODS

The version of the Python programming language that was used for all of the analyses in this article is 3.10.7. The versions of the Python packages are RDKit (2022.03.5),[30] MolVS (0.1.1),[31] Venn (0.1.3),[32] Plotly Express (0.4.1),[33] Scikit-learn (1.2.2),[34] NumPy (1.23.2),[35] and Seaborn (0.12.2).[36]

Database Update and Data Curation. The first version of LANaPDB had 12 959 NPs coming from nine different databases of six different Latin American countries.[23] To the first version of LANaPDB, a new database was added: NPDB EjeCol, which is a compilation of NPs isolated and characterized in Colombia, specifically from the region known as the Coffee Region.[25] This database is set to be published in 2024 and is accessible through an open-data portal (www.npdbejecol.com). Furthermore, the LANaPDB was updated with new NPs from Costa Rica (NAPRORE-CR) and Mexico (BIOFACQUIM). In total, 619 new compounds were added to LANaPDB, resulting in a total of 13 578 NPs in

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

the second version of the database. The curation of the second version of LANaPDB was carried out with the same workflow employed in the first version of the database.[26] The process was performed in the Python programming language, employing the RDKit and MolVS packages. The standard curation process of MolVS was implemented through the standardize_smiles function included in this Python package, which includes and implements some functions from RDKit (SanitizeMol, RemoveHs, and AssignStereochemistry) and MolVS (disconnect, normalize, and reionize): verify and correct valencies, aromaticity, and hybridization, removal of explicit hydrogens, disconnection of covalent bonds between metals and organic atoms (the disconnected metal is removed later), 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 the preservation of the original stereochemistry). From the molecules that are fragmented, i.e., the molecules that used to be connected with metals or other salts, only the largest fragment is kept (choose function from MolVS) and an attempt is made to neutralize all the molecules of the database (uncharge function from MolVS). The canonical SMILES strings were retrieved (MolToSmiles function of RDKit), and these are the SMILES strings included for every LANaPDB compound. The canonical tautomer was determined (canonicalize function from MolVS), and from the InChIKey strings of the canonical tautomer, the duplicate compounds were removed. The canonical tautomers were used only as part of the duplicate compound removal process; thus, the reported structures of the LANaPDB compounds correspond to the canonical SMILES strings retrieved before the elimination of the repeated molecules and not to the structure of the canonical tautomer. The same curation workflow was applied to two reference data sets employed to compare LANaPDB: COCONUT 1.0[9] and FDA-approved small-molecule drugs, version 5.1.10 (released by DrugBank in January 2023).[37]

To determine the number of unique and overlapping molecules in the different Latin American countries, the databases that encompass this version of LANaPDB were subjected to the above-described curation process; nonetheless, the duplicate removal step was omitted. Finally, from the molecule structures in the Python programming language employing the Venn package, the unique and overlapping molecules were determined (Figure 1).

Structural Classification. The freely available online server NPClassifier[38] was employed to perform the structural classification of the LANaPDB compounds. NPClassifier is a deep neural network-based structural classification tool for NPs. The distribution of the classified compounds was represented with pie plots constructed in Python using the Plotly Express package.

Physicochemical Properties. The following physicochemical properties of pharmaceutical interest were calculated in Python employing the RDKit package: SlogP,[39] molecular weight (MW), topological polar surface area (TPSA),[40] rotatable bonds (Rb), hydrogen bond acceptors (HBA), and hydrogen bond donors (HBD). The distribution of the physicochemical properties was depicted with violin plots,[41] constructed in the Python programming language with the Scikit-learn package.

Chemical Space Visualization. The visualization of the chemical space of LANaPDB was made using the TMAP (Tree MAP) algorithm[42] from the MACCS keys[43] and Morgan2[44] fingerprints. The determination of both fingerprints was made using the Python programming language with the RDKit package. The construction of the TMAP was made with Python, following the reported protocol.[42] The results were compared with two reference data sets: COCONUT[9] and FDA-approved small-molecule drugs, version 5.1.10 (released by DrugBank in January 2023).[37]

Cross-References to Other Databases. The cross-references to PubChem[29] and ChEMBL[28] identification (ID) codes were requested and retrieved from the respective websites of both databases. The request and retrieval of the ID codes were made in the Python programming language, employing the corresponding application programming interface (API) for PubChem and ChEMBL. The InChIKey strings of the LANaPDB compounds were utilized to make the requests with the PubChem and ChEMBL application programming interfaces (APIs). The InChIKey strings were calculated in the Python programming language, employing the RDKit package.

Commercial Availability and Chirality. The commercial availability of every compound of LANaPDB was obtained from the PubChem website.[29] It is not information that can be retrieved with the PubChem API. Therefore, the Python programming language was used to retrieve the commercial availability, but without using the PubChem API. The classification of every compound based on chirality was made in Python, employing the function Chem.FindMolChiralCenters of the RDKit package.

Biological Activity. The biological activity of the LANaPDB compounds was retrieved from ChEMBL, version 34, employing two different approaches. In the first approach, with the Python programming language, employing the ChEMBL API, from the InChIKey strings the reported biological activity of the LANaPDB compounds was requested and retrieved from the ChEMBL database website. In the second approach, in the Python programming language employing the RDKit package, it was determined if the SMILES strings of the LANaPDB molecules contained the SMILES strings of the ChEMBL bioactive rings reported by Ertl.[45]

Structural Diversity. The Bemis and Murcko scaffolds[46] were determined from the SMILES strings in the Python programming language with the RDKit package. The area under the curve (AUC) was obtained from the cumulative scaffold recovery (CSR) curves with the trapezoidal rule in the Python programming language with the trapz function of the numpy package. The fraction of scaffolds to retrieve 50% of the compounds in the database (F_{50}) metric was obtained from the CSR curves by interpolating the x-axis value of 0.5 to find the corresponding y-axis value, in the Python programming language with the interp function of the numpy package. The MACCS keys (166-bit) fingerprint and the paired Tanimoto similarity[47] were calculated in the Python programming language with the RDKit package. The paired Tanimoto similarity calculation for the COCONUT 1.0 data set was made with a random sample of 10% (with more than 40 000 compounds) that represents the diversity of the whole database.[48]

Molecular Complexity and Synthetic Feasibility. The normalized spacial score (nSPS)[49] and synthetic accessibility

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

Table 1. Natural Product Databases in the Updated Version of LANaPDB

Database	Number of compounds	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.	53,54
SistematX (Brazil)	9514	plants	Database composed of secondary metabolites and developed at the Federal University of Paraiba.	55,56
UEFS (Brazil)	503	plants	Natural products that have been separately published, but there is no common publication or public database for it. Developed at the State University of Feira de Santana.	57
NPDB EjeCol (Colombia)	236	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.	25
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.	58
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.	59,60
CIFPMA (Panama)	363	plants	Natural products that have been tested in over 25 in vitro and in vivo bioassays for different therapeutic targets, developed at the University of Panama.	61,62
PeruNPDB (Peru)	280	animals, plants	Natural products representative of Peruvian biodiversity. Created and curated at the Catholic University of Santa Maria.	63

[a]The database has not been published yet.

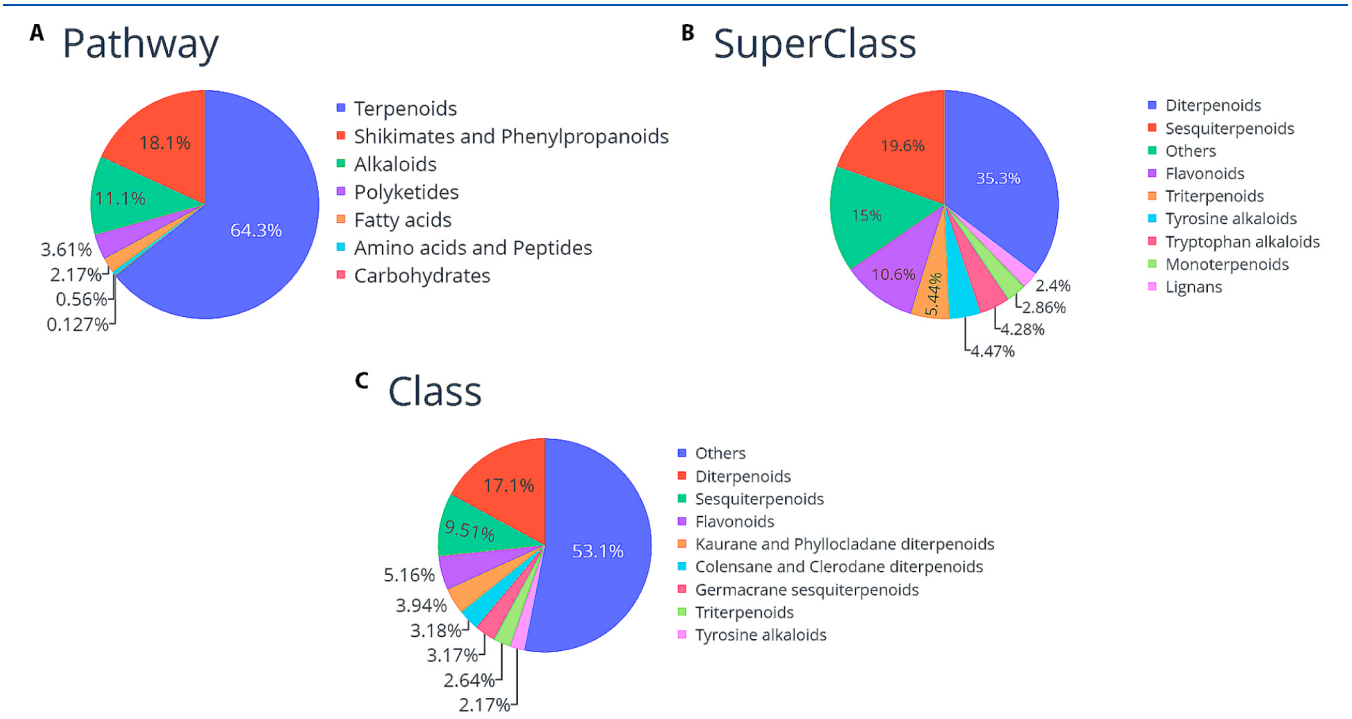


Figure 2. Pie charts showcasing the distribution of the LANaPDB compounds, according to a classification system[38] based on the literature from the specialized metabolism of the producing organisms. **A**) Pathway: related to the nature of the biosynthetic pathway. **B**) SuperClass: associated with chemical properties or chemotaxonomic information, and **C**) Class: correlated to structural details.

score (SAscore)[50] were determined in the Python programming language with the RDKit package, employing the SpacialScore and sascorer[51] functions. The kernel density estimate (KDE) plots[52] were constructed in the Python programming language with the Seaborn package.

RESULTS AND DISCUSSION

Database Update and Data Curation. The first version of LANaPDB comprised 12 959 compounds.[26] This reported update includes 619 new compounds, resulting in a total of 13 578 compounds in its second version published in early 2024.[27] A new data set was included: NPDB EjeCol, which contains NPs from foods and plants isolated and characterized in Colombia, from the Coffee Region (Eje Cafetero). Moreover, the database was updated with new NPs from Costa Rica (NAPRORE-CR) and Mexico (BIOFACQUIM). Table 1 shows the ten Latin American NP databases currently contained in LANaPDB. Initially, 1707 compounds were considered for the update of LANaPDB from the two updated databases, BIOFACQUIM and NAPRORE-CR, and the new

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

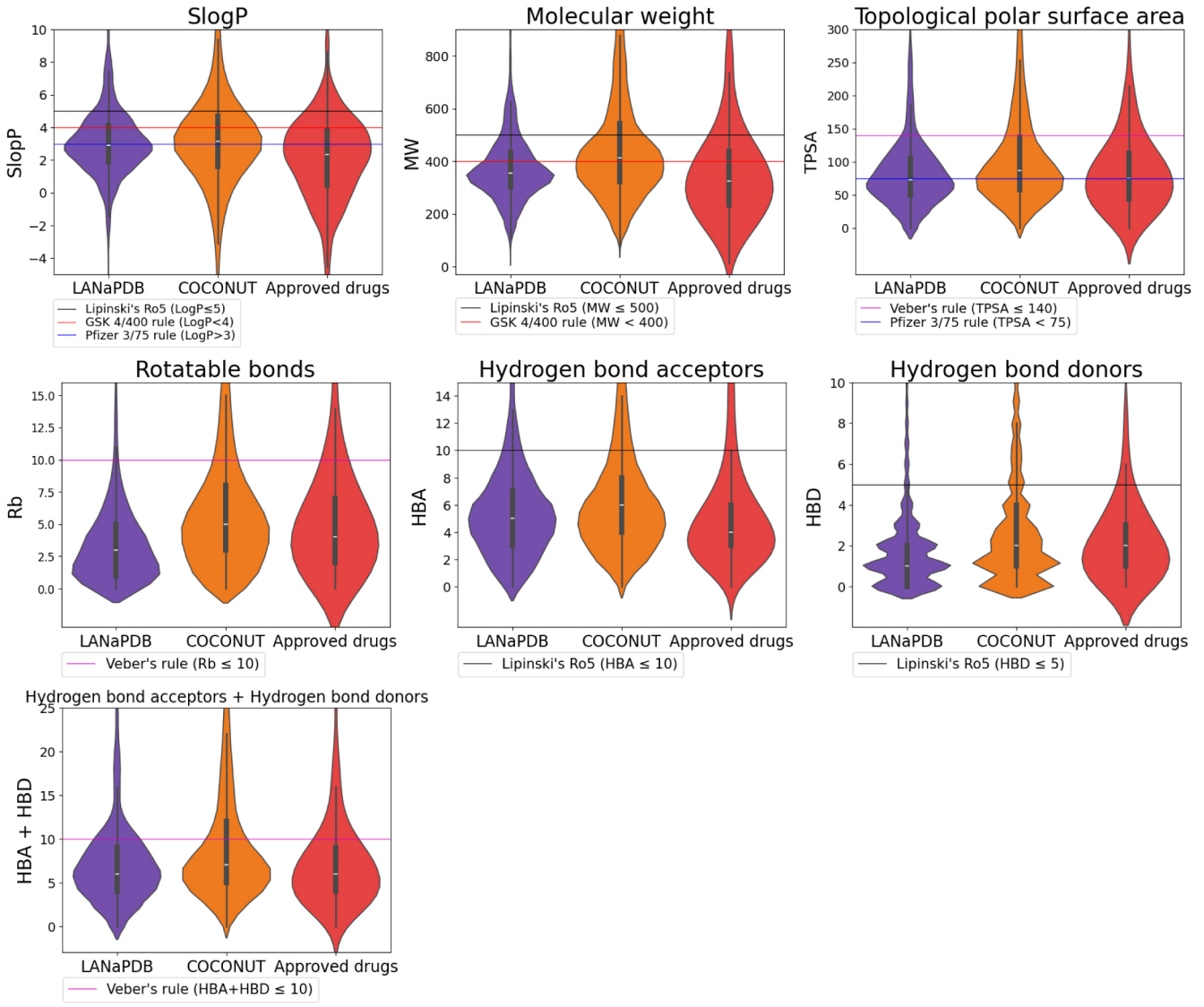


Figure 3. Violin plots summarizing the distribution of seven physicochemical properties of pharmaceutical interest in the compounds from three databases: LANaPDB, COCONUT, and FDA-approved small-molecule drugs.

database NPDB EjeCol. Nevertheless, from the initial 1707 compounds, 1088 molecules were duplicates and were no longer included. The remaining 619 molecules were added to LANaPDB.

The number of unique and overlapping molecules in every Latin American country was determined from the databases that contain LANaPDB, and Figure 1 shows the countries with the highest number of unique and overlapping compounds. It was found that the number of unique molecules is associated with the number of molecules in the country. Brazil is the country with the most unique molecules (10 580), followed by Mexico (1360), Costa Rica (585), Panama (289), Peru (178), El Salvador (174), and Colombia (72). Furthermore, it was found that Brazil has the highest number of overlapping compounds with other countries (Mexico: 215, Costa Rica: 161, and Panama: 39), which can be attributed to the fact that Mexico, Costa Rica, and Panama have the largest number of reported compounds after Brazil (Table 1). Nonetheless, it can also imply that Brazil shares flora and fauna with these three countries, with Mexico being the country with the highest number of shared compounds. There is a very small number of overlapping compounds among the other countries, with almost zero overlapping compounds in most cases. A possible explanation is that the remaining countries (Colombia, El Salvador, Panama, and Peru) have much fewer reported compounds than Brazil, Costa Rica, and Mexico.

Structural Classification. The compounds in LANaPDB were structurally classified according to a classification system based on the literature on the specialized metabolism of plants, marine organisms, fungi, and microorganisms. The classification system is divided into three hierarchical levels: pathway (nature of the biosynthetic pathway), superclass (chemical properties or chemotaxonomic information), and class (structural details). At the three hierarchical levels, the predominant compounds are terpenoids (Figure 2). At the hierarchical level of the pathway, terpenoids, shikimates, phenylpropanoids, and alkaloids encompass more than 90% of the total compounds. At the hierarchical level of superclass and class, terpenoids and flavonoids were the predominant compounds (Figure 2). The above was expected because terpenoids are the predominant secondary metabolites produced by natural sources.[64] Compared to the previous

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

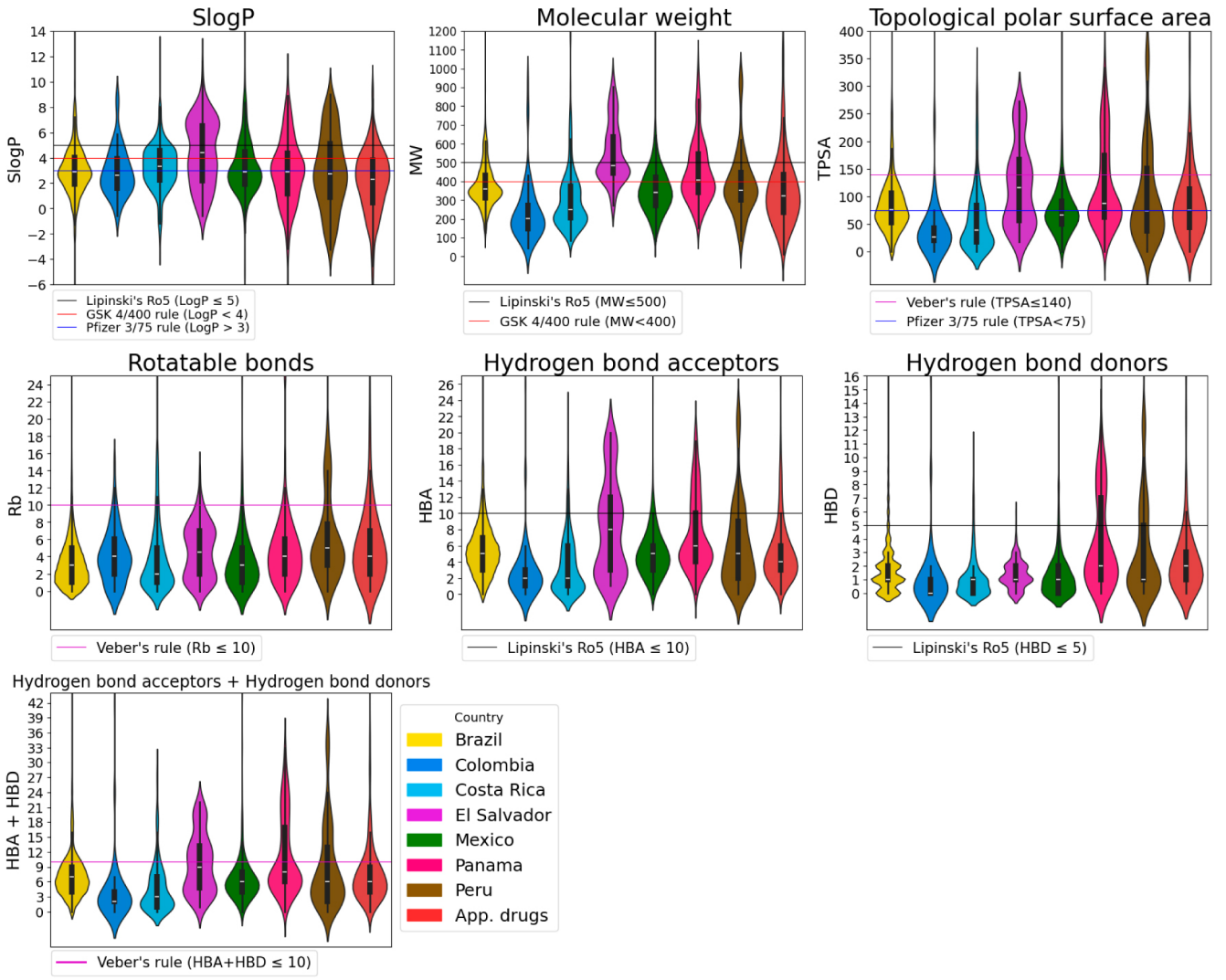


Figure 4. Violin plots summarizing the distribution of seven physicochemical properties of pharmaceutical interest of the compounds in LANaPDB and FDA-approved small-molecule drugs (App. drugs). The databases that encompass LANaPDB for every country: Brazil (NuBBEDB, SistematX, and UEFS), Colombia (NPDB EjeCol), Costa Rica (NAPRORE-CR), El Salvador (LAIPNUDELSAV), Mexico (UNIIQUIM and BIOFACQUIM), Panama (CIFPMA), and Peru (PeruNPDB).

version of LANaPDB, the above tendencies have not changed.[26]

Physicochemical Properties. We calculated the physicochemical properties of pharmaceutical interest for the LANaPDB compounds and compared them with two reference data sets: COCONUT[9] and FDA-approved small-molecule drugs.[37] Figures 3 and 4 show the distribution of the calculated physicochemical properties: SlogP,[39] molecular weight (MW), topological polar surface area (TPSA),[40] number of rotatable bonds (Rb), hydrogen bond acceptors (HBA), and hydrogen bond donors (HBD). The violin plots (Figures 3 and 4) are marked with a horizontal line indicating the limits of some drug-likeness rules of thumb: Lipinski's rule of 5 (Ro5),[65,66] Veber's rules,[67] GlaxoSmithKline's (GSK) 4/400 rule,[68] and Pfizer 3/75 rule.[69] Physicochemical properties within the limits of either Lipinski's, Veber's, or GSK rules are usually related to good oral bioavailability. The fulfillment of these rules of thumb is associated with the improvement of the following parameters: aqueous solubility and intestinal permeability (Lipinski's Ro5); passive membrane permeation (Veber's rules); absorption, distribution, metabolism, excretion, and toxicity (ADMET) profile (GlaxoSmithKline's 4/400 rule); and toxicity (Pfizer 3/75 rule). In Figure 3, noticeable changes in the distribution of the physicochemical properties of LANaPDB are not appreciated compared to the previous version.[26] This can be attributed to the fact that the terpenoids remain as the prevalent compounds (Figure 2). The physicochemical properties related to the Ro5 (SlogP, MW, HBA, and HBD), Veber's rules (HBA + HBD, TPSA, and Rb), and the GlaxoSmithKline's 4/400 rule (SlogP and MW) are within the limits of these three rules of thumb for most of the compounds in the three databases (Figure 3). Therefore, the aqueous solubility, intestinal permeability, oral bioavailability, and in general the ADMET profile are desirable for the three databases. Moreover, the three databases have a similar distribution for these physicochemical properties. Nevertheless, regarding the Pfizer 3/75 rule, which is related to toxicity, just approximately half of the compounds in the three databases satisfy the requirements of SlogP > 3 and TPSA < 75 (Figure 3). Therefore, according to the obtained values of SlogP and TPSA, considering the Pfizer 3/75 rule, half of the compounds in the three databases have a desirable toxicity

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

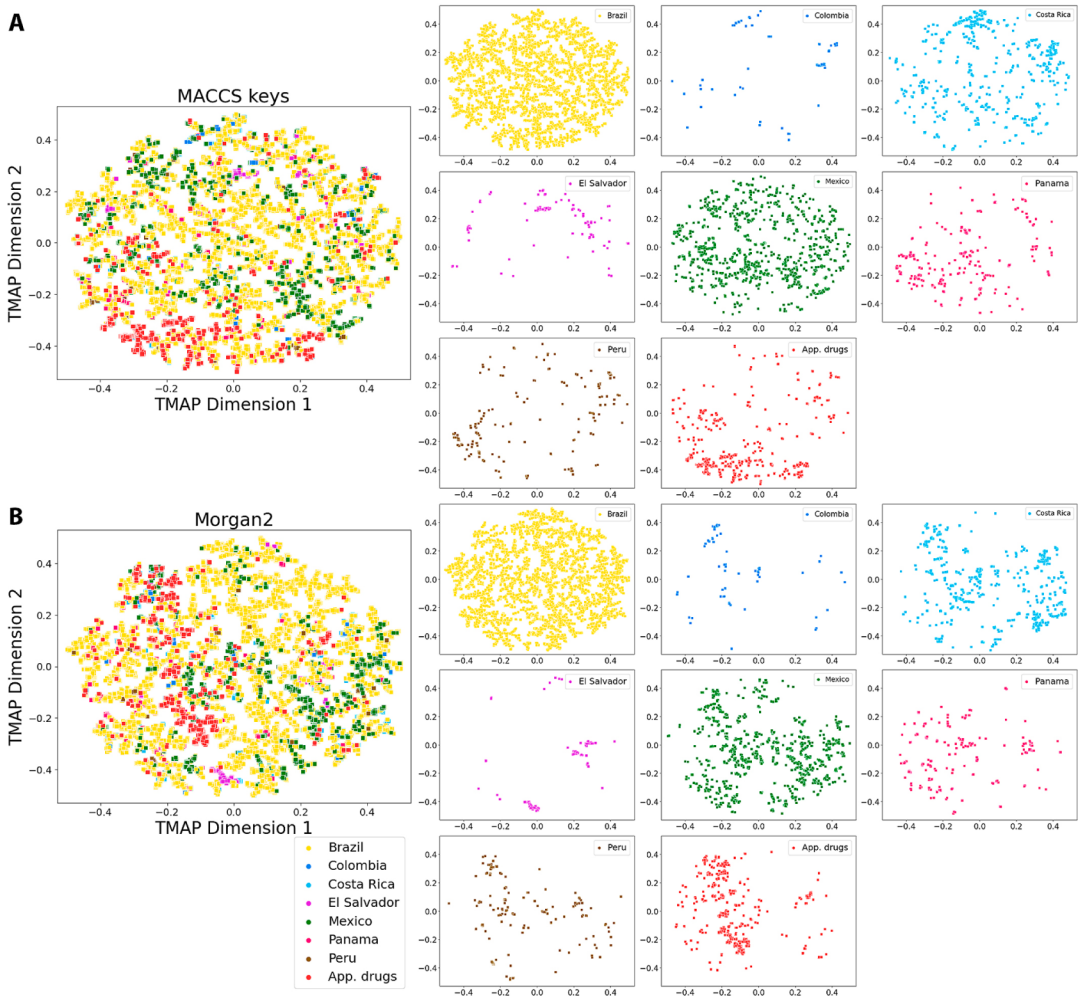


Figure 5. Tree MAP of LANaPDB and the comparison with FDA-approved small-molecule drugs, generated from **A**) MACCS keys (166-bit) and **B**) the Morgan2 (2048-bit) fingerprint. An interactive version of the TMAP is available for free download (MACCS keys (166-bit): https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP%20MACCS%20keys.html and Morgan2: https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP%20Morgan2.html) (to open the interactive map, download the file and open it in a web explorer; zoom in option is available with the mouse scroll).

profile. Regardless, half of the FDA-approved small-molecule drugs satisfy the Pfizer 3/75 rule; therefore, the compounds that do not satisfy this rule are still worth consideration in drug design because the toxicity is not just related to the SlogP and TPSA.

The LANaPDB compounds presented fewer rotatable bonds compared to COCONUT. This result can be attributed to the fact that the number of compounds in COCONUT is much larger compared to LANaPDB and as a consequence, the diversity in the structures contributes to a wider distribution in the rotatable bonds. Nonetheless, comparing this version of LANaPDB to the prior version of the database, the distribution of rotatable bonds is the same, which can be attributed to the fact that in both versions of the database, the terpenoid compounds are predominant and on average, they have fewer than four rotatable bonds. Regardless, the rotatable bonds in LANaPDB fulfill Veber's rules (Rb < 10). Therefore, it is expected to have good passive membrane permeation.

In the case of the individual countries that encompass LANaPDB, a similar behavior was observed for the physicochemical properties, where most of the compounds satisfy the Ro5, Veber's rules, and GlaxoSmithKline's 4/400 rule, but a lower proportion satisfies the Pfizer 3/75 rule. Nevertheless, Panama, compared to the other countries, shows a higher proportion of compounds with higher SlogP and MW, which can be detrimental to intestinal permeability and, in general, to the ADMET profile; therefore, other routes of administration should be considered for these compounds, for instance, the nasal delivery route.[70] Therefore, most of the LANaPDB compounds have a desirable physicochemical profile that allows them to be employed in the design of new drugs, either as potential drug candidates or as a starting point to design semisynthetic drugs or pseudo-NP.

Figure 4 shows the distribution of the physicochemical properties of pharmaceutical interest of LANaPDB, considering the seven countries individually. For comparison, the distribution of the compounds in the FDA-approved drugs is included. In general, it is observed that the distribution of compounds is mainly focused on regions that fulfill the drug-likeness rules of thumb. Nonetheless, El Salvador is a country with many compounds outside of the drug-likeness parameters considering the SlogP and MW. In the current version of LANaPDB, new compounds from Costa Rica and Mexico were added; nevertheless, the distribution of the physicochemical

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

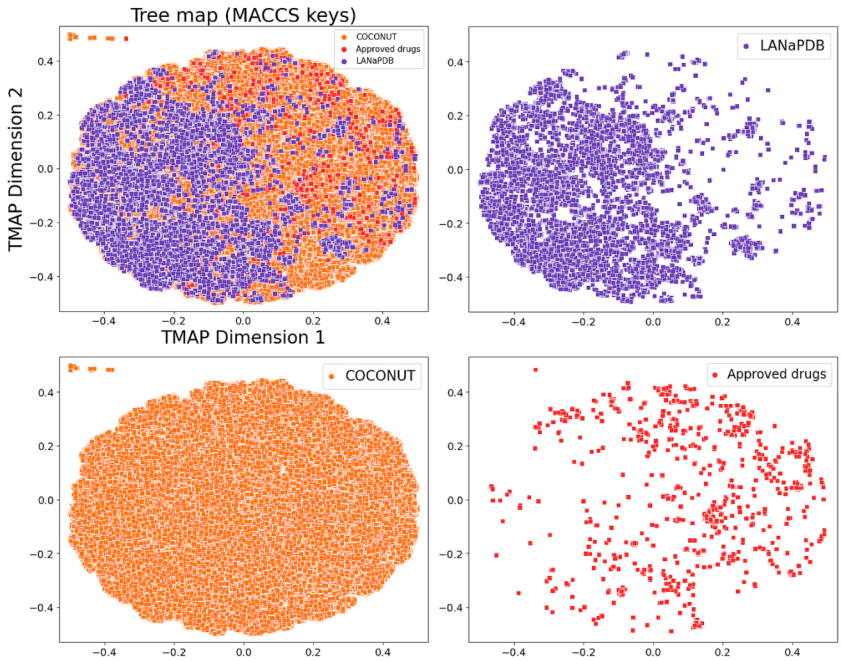


Figure 6. Tree MAP of LANaPDB and the comparison with COCONUT and FDA-approved small-molecule drugs, generated from the MACCS keys (166-bit) fingerprint.

properties of the compounds of both countries compared to the previous version of LANaPDB[26] remained without significant changes. The distribution of the physicochemical properties of the compounds of the new country added to the current version of LANaPDB, Colombia, is such that most of the compounds fulfill the drug-likeness rules of thumb.

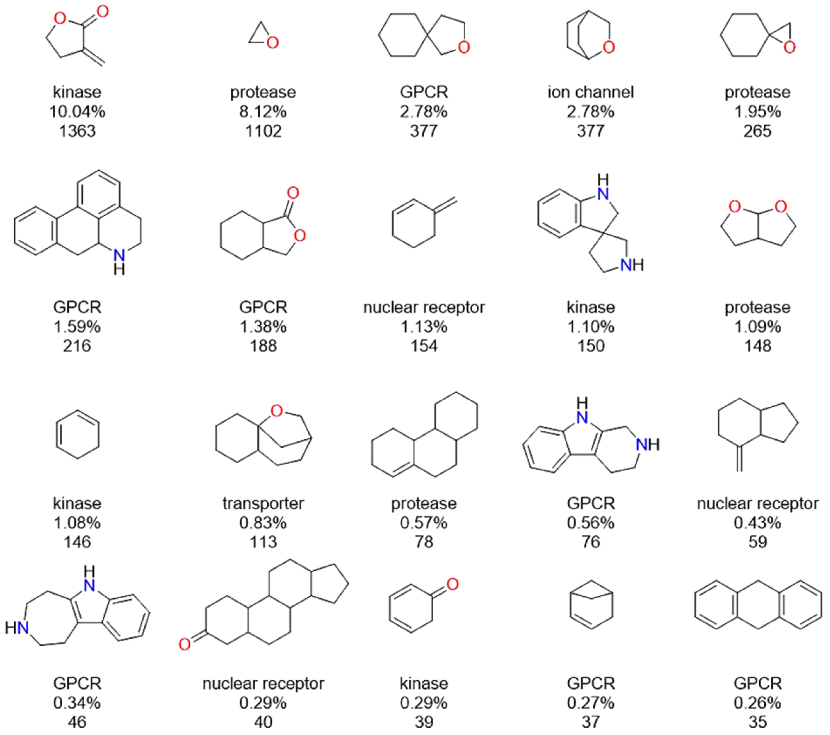
Chemical Space Visualization. Figure 5 shows the TMAP of LANaPDB generated from the MACCS keys (166-bit),[43] Morgan2[44] fingerprints, and their comparison with the FDA-approved small-molecule drugs.[37] The structural features of the compounds are not necessarily correlated to the numerical values of the x and y axes. Therefore, an interactive version of the TMAP is also freely available for download (MACCS keys (166-bit): https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP%20MACCS%20keys.html and Morgan2: https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP%20Morgan2.html) (to open the interactive map, download the file and open it in a web explorer; zoom in option is available with the mouse scroll). MACCS keys (166-bit) were chosen for their capacity to capture structural features from well-known predefined fragments and Morgan2 (2048-bit) for their efficiency in capturing detailed structural features. In the interactive version of Figure 5, it can be appreciated that the TMAP effectively accomplished the clustering of structurally similar compounds in "branches" for both fingerprints. Therefore, both fingerprints showed similar and very good capacities to capture the structural features of NPs. Neither MACCS keys (166-bit) nor Morgan2 (2048-bit) fingerprints appear to outperform the other in capturing structural features according to both interactive plots of Figure 5. Figure 5 shows that all of the countries and the approved drugs with both fingerprints overlap with the Brazilian NPs. Therefore, Brazil is the country with the highest structural diversity of NPs according to the TMAP. Moreover, the compounds for each of the seven Latin American countries with both fingerprints are in general not focused on a certain region of the chemical space. Instead, they are distributed across the chemical space and, in many cases, clustered, forming branches of structurally similar compounds. Besides, all the Latin American countries partially overlap with the approved drugs in specific regions for both fingerprints. Figure 6 depicts the comparison of LANaPDB with COCONUT and the approved drugs with the MACCS keys (166-bit) fingerprint. LANaPDB totally overlaps with COCONUT. Interestingly, the overlap of LANaPDB with COCONUT is mostly in a well-defined area (left side of the TMAP), which shows that COCONUT covers a huge area (right side of the TMAP) of the chemical space not covered by LANaPDB. It is important to consider that COCONUT has more than 400 000 compounds and LANaPDB 13 578. In Figure 6, it is appreciated that the approved drugs are distributed across the chemical space, overlapping LANaPDB and COCONUT in different regions.

Cross-References to Other Databases. The LANaPDB compounds were cross-referenced to two of the biggest publicly available chemical compound databases annotated with biological activity: PubChem, version 2024[29] and ChEMBL, version 34.[28] From both databases, the ID code was retrieved. The ID code allowed to identify and differentiate every single compound. In the case of PubChem, the ID codes are known as CID (compound identification) and SID (substance identification). From all the LANaPDB compounds, 71.71% of the ID codes were successfully retrieved from PubChem and 23.69% from ChEMBL.

Therefore, most of the LANaPDB compounds can be found in PubChem, and just a minority in ChEMBL. To consult additional information for the LANaPDB compounds in PubChem and ChEMBL, it is just needed to type the corresponding ID code in the respective websites of both databases. The SMILES strings contained in ChEMBL and the ones determined for the compounds of LANaPDB versions 1 and 2 were obtained with RDKit, which uses its own canonicalization method; thus, they are comparable to each

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

Chart 1. 20 Most Abundant Bioactive Ring Systems (with Reported Biological Activity in ChEMBL) in LANaPDB, their Biological Targets, Percentage of Occurrence, and the Total Number of Compounds that Contain the Ring System in LANaPDB



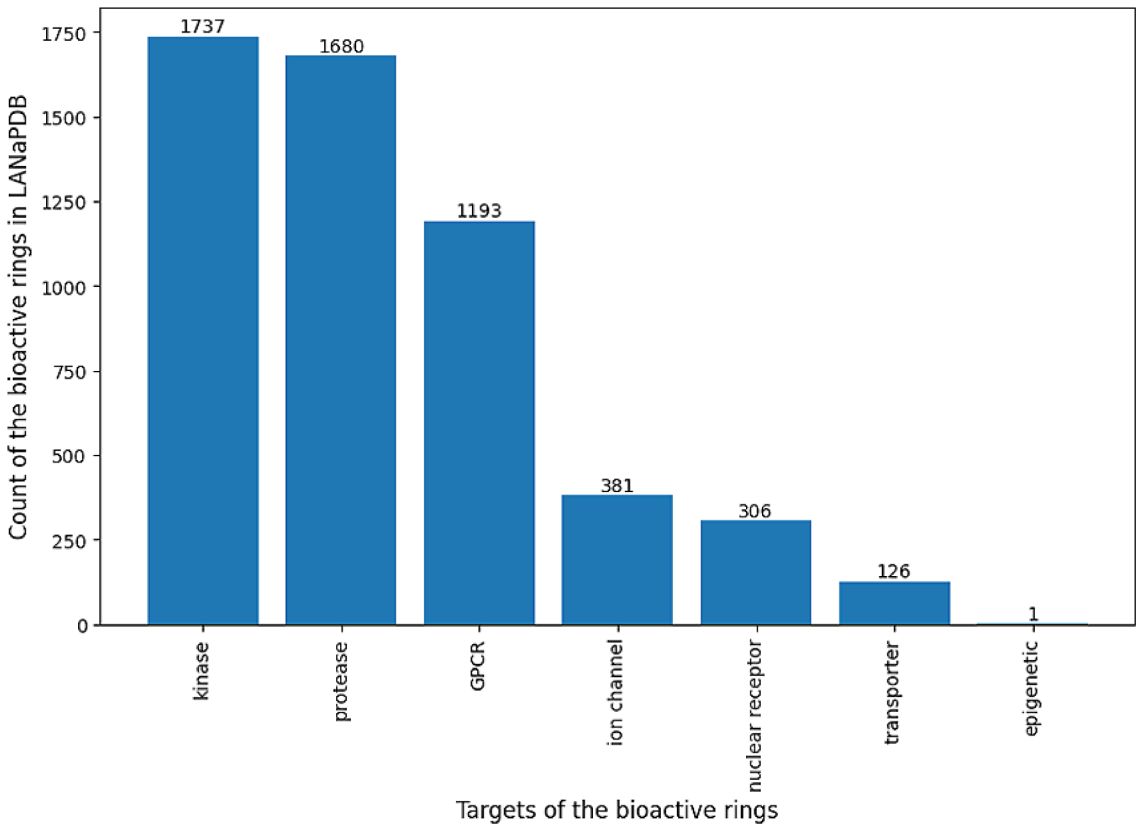


Figure 7. Histogram that shows the occurrence of bioactive rings (with reported bioactivity in ChEMBL) in LANaPDB and their biological target. Consider that every molecule can have more than one bioactive ring in its structure.

other. The additional information that can be checked in PubChem for the LANaPDB compounds includes spectral

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

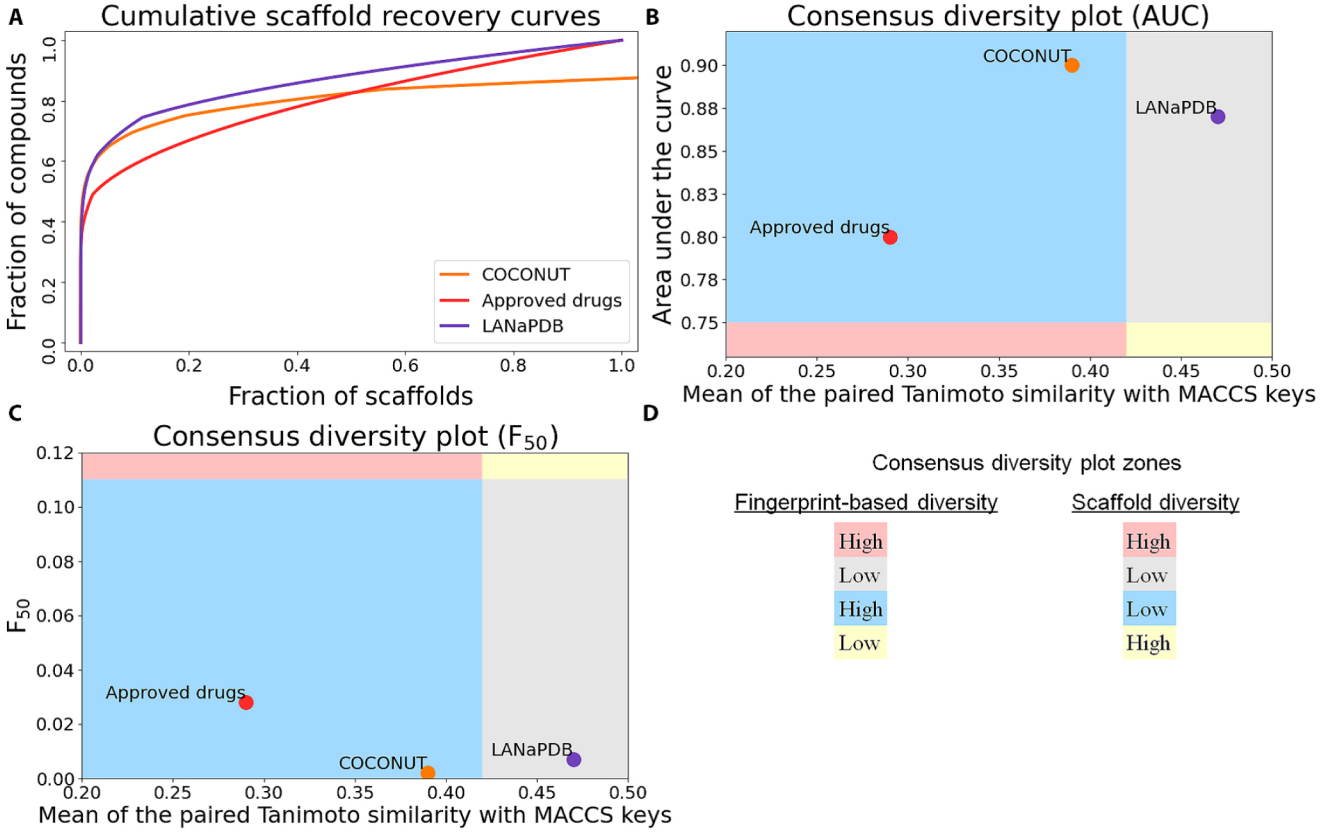


Figure 8. **A**) Cumulative scaffold recovery (CSR) curves of LANaPDB, COCONUT, and FDA-approved small-molecule drugs. Consensus diversity plots of LANaPDB, COCONUT, and FDA-approved small-molecule drugs, which describe the data set diversity considering the MACCS keys (166-bit) fingerprint, **B**) area under the curve, and **C**) the fraction of scaffolds to retrieve 50% of the database (F_{50}). **D**) Degree of scaffold and fingerprint-based diversity in the consensus diversity plots' quadrants.

information, toxicity, and patents. ChEMBL contains information about metabolism, target predictions, drug indications, and mechanism of action.

Commercial Availability and Chirality. It was found that 70.5% of the LANaPDB compounds are commercially available, as annotated on the PubChem website. The information about the companies that sell the individual molecules can be consulted on the PubChem website, from the PubChem ID codes added to LANaPDB. Moreover, all the molecules were classified into three categories: achiral (16.16%) and chiral with chirality annotated (55.53%) or not annotated (28.31%).

Biological Activity. The biological activity of the LANaPDB compounds was retrieved from ChEMBL, with two different approaches. In the first one, the biological activity was retrieved from the ChEMBL website with the ChEMBL API. It was found that only 0.29% of the LANaPDB compounds (39 molecules) have a reported biological activity that can be retrieved with the ChEMBL API. These compounds have up to three biological activities reported. The most common biological activities are pharmaceutical aid (flavor) (4 compounds), pharmaceutical aid (solvent) (3 compounds), antifungal (3 compounds), pharmaceutical aid (antimicrobial agent) (2 compounds), pharmaceutical aid (emulsion adjunct) (2 compounds), and inhibitor (alpha-glucosidase) (2 compounds).

The second approach was based on a study by Peter Ertl who previously extracted the ring systems from the molecules in ChEMBL (version not specified) and associated them with their reported bioactivity in ChEMBL against the following biological target families: G protein-coupled receptor (GPCR), kinase, protease, nuclear receptor, ion channel, transporters, and epigenetic targets.[45] For LANaPDB, it was determined which compounds contain the bioactive ring systems reported by Ertl. It was found that 31.51% of the LANaPDB compounds (4279 molecules) have bioactive ring systems. Chart 1 shows the 20 most abundant ring systems found in the LANaPDB compounds; the most abundant ring system agrees with the most abundant pathway found in LANaPDB (Figure 2) as it pertains to bioactive sesquiterpenic lactones.[71] Figure 7 shows that the bioactive rings in LANaPDB mainly target kinases, proteases, and G protein-coupled receptors (GPCRs), which is related to the fact that they are among the most extensively studied drug targets. GPCRs are the most studied drug targets,[72] and approximately up to 2018, 35% of the approved drugs (~700) target GPCRs.[73] Kinases are the second most therapeutically targeted group of proteins, after GPCRs, and up to 2023, 98 kinase inhibitors were approved.[74] Proteases are another extensively studied therapeutic target; up to 2011, 12 drugs that target proteases had been approved.[75]

It is important to take into account that the remaining percentage of compounds (68.49%) without bioactive ring systems are not necessarily inactive compounds; they may be active but against other biological targets different from the ones that were reported by Ertl.[45] Take into account that the currently known scaffold space is far from being fully explored. This is exemplified by the fact that in 2024, Ertl published a

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

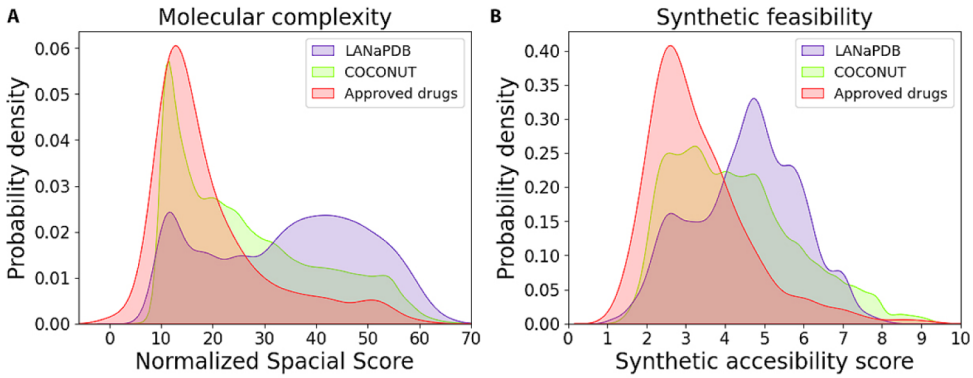


Figure 9. Kernel density estimate plots that represent the distribution of the **A**) normalized spacial score and **B**) synthetic accessibility score of LANaPDB, COCONUT, and FDA-approved small-molecule drugs.

database of four million medicinal chemistry-relevant scaffolds that are not included in ChEMBL and PubChem.[76]

Structural Diversity. The structural diversity of LANaPDB was quantified with two types of molecular representations: molecular scaffolds and fingerprints. The diversity was compared to those of COCONUT and FDA-approved small-molecule drugs. The scaffold diversity of all data sets was measured with CSR curves that represent the fraction of molecules in the data set contained in a fraction of scaffolds. To generate the CSR curves, the scaffolds are ordered by their frequency of occurrence (most to least common). Then, the fraction of scaffolds is plotted on the *x*-axis, and the fraction of compounds that contain those scaffolds is plotted on the *y*-axis. Two metrics were obtained from the CSR curves: AUC and F_{50} (i.e., if a data set has F_{50} = 0.43, 50% of the compounds in the data set are distributed in 43% of the scaffolds). A data set with maximum diversity would contain a different scaffold for each molecule in the library, and the curve would be a diagonal with an AUC of 0.5. As the scaffold diversity decreases, the curve will move away from the diagonal. The minimum diversity would be a data set in which all of the compounds have the same scaffold. In this case, the CSR function would be a vertical line with an AUC equal to 1.0. The fingerprint-based diversity was assessed with the mean of the paired Tanimoto similarity (MPTS), using the MACCS keys (166-bit) fingerprint (mainly quantifies the side chain structural diversity).[77,78]

In the consensus diversity plots (Figures 8B,C) it is shown that FDA-approved small-molecule drugs is the data set with the highest scaffold and fingerprint-based diversity (AUC = 0.80, F_{50} = 0.028, and MPTS = 0.29), followed by LANaPDB (AUC = 0.87, F_{50} = 0.007, and MPTS = 0.47) and COCONUT (AUC = 0.90, F_{50} = 0.002, and MPTS = 0.39). This result can be attributed to the fact that this data set contains not just NPs; instead, a significant proportion are NP derivatives and purely synthetic molecules,[79] which increases the structural diversity. According to the MPTS metric, the side chain structural diversity of LANaPDB is lower than that of COCONUT. Nonetheless, considering the AUC and F_{50} metrics, LANaPDB has higher scaffold diversity than that of COCONUT; nevertheless, the difference between both databases considering these two metrics is small (ΔAUC = 0.03 and ΔF_{50} = 0.005). Therefore, the structural diversity of LANaPDB is very similar to COCONUT, with less side chain diversity and a little more scaffold diversity.

Molecular Complexity and Synthetic Feasibility. Molecular complexity can be quantified using different metrics.[80] In this work, as a quantitative measure of molecular complexity, we employed the recently developed metric nSPS.[49] The synthetic feasibility was determined by calculating the SAscore.[50] The distribution of both metrics was represented with KDE plots, which represent the data using continuous probability density curves (Figure 9). nSPS takes into account the atom hybridization, stereoisomerism, presence and complexity of aromatic or nonaromatic rings, and the number of heavy-atom neighbors.[49] As a reference, in an earlier study, it was found that the nSPS values of most of the approved drugs are between 10 and 20, and this has remained without any significant changes in the last eight decades.[81] The nSPS values for the compounds of the three databases studied in this work are centered around 10 and 20 (Figure 9A). Thus, LANaPDB has a significant proportion of compounds with nSPS values between 10 and 20 (39.78%), and those compounds are expected to have a similar pharmacokinetic profile to the approved drugs according to the molecular similarity principle.[81] Moreover, unlike the other two reference databases, the LANaPDB compounds presented mainly nSPS values around 30 and 50 (26.88%) (Figure 9A). Previously, it has been found that the ligand potency and target selectivity are maximized in compounds with nSPS values between 20 and 40.[49] Therefore, LANaPDB has a significant proportion of compounds with nSPS values between 20 and 40 (37.95%), which are expected to have good potency and target selectivity. The nSPS value for each compound in LANaPDB is indicated in the publicly available database.

The synthetic feasibility was estimated with the SAscore, which considers the complexity of the molecular fragments, stereocomplexity, and molecule size. The synthetic feasibility is positively correlated with the SAscore, i.e., highest SAscores are associated with higher synthetic feasibility.[50] In this work, approved drugs and COCONUT presented mainly SAscores between two and three (Figure 9B). The accumulation of SAscores of approved drugs and COCONUT in the same zone can be attributed to the fact that a large proportion of the approved drugs are NPs or NP-based molecules.[79] The LANaPDB compounds have mostly SAscores around five, which implies that a significant proportion of the LANaPDB compounds have a synthetic feasibility higher than that of the approved drugs.

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

CONCLUSIONS

LANaPDB was updated with 619 new molecules from Colombia, Costa Rica, and Mexico, resulting in a total of 13 578 compounds. It is highlighted that the addition of a new database of NPs from Colombia, NPDB EjeCol, is the first database that gathers NPs from Colombia. In the structural classification of the compounds, it was found that terpenoids are still the dominant compounds in the database. According to the calculated physicochemical properties of pharmaceutical interest, most of the LANaPDB compounds have a desirable physicochemical profile, which allows them to be employed in the design of new drugs. In the chemical space visualization, it was found that LANaPDB totally overlaps with COCONUT and partially overlaps with FDA-approved small-molecule drugs. Furthermore, MACCS keys (166-bit) and Morgan2 (2048-bit) showed similar and good capacities to capture structural features from the LANaPDB compounds. Moreover, the LANaPDB compounds were cross-referenced to ChEMBL and PubChem. It was found that 70.5% of the database molecules are commercially available, and the information regarding the vendors can be consulted on the PubChem website, employing the PubChem IDs that were added to the LANaPDB compounds. Only 39 molecules of LANaPDB have reported biological activity on ChEMBL; nonetheless, 4279 molecules have bioactive ring systems. From the structural diversity analysis, it was found that LANaPDB has less scaffold and fingerprint-based diversity than FDA-approved small-molecule drugs; nevertheless, compared to COCONUT, LANaPDB has less side chain diversity and a little more scaffold diversity. According to the molecular complexity of the molecules in the database, they are expected to have a similar pharmacokinetic profile to the approved drugs, and most of the compounds have high synthetic feasibility. LANaPDB is an ongoing project and is planned to keep updating with more compounds and adding more information, such as spectroscopic data and the ADMET profile.

ASSOCIATED CONTENT

Data Availability Statement

The LANaPDB database is publicly available at https://github.com/alexgoga21/LANaPDB-version-2. The whole database can be downloaded as an xlsx file at https://github.com/alexgoga21/LANaPDB-version-2/blob/main/LANaPDB%20version%202.xlsx. The interactive tree MAP can be downloaded as an html file at https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP%20MACCS%20keys.html and https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP%20Morgan2.html. To open the interactive map, download the file and open it in a web explorer; zoom in option is available with the mouse scroll. The first version of LANaPDB can be consulted on the COCONUT web server at https://coconut.naturalproducts.net/search?type=tags&q=Latin+America+dataset&tagType=dataSource.

AUTHOR INFORMATION

Corresponding Author

José L Medina-Franco − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0003-4940-1107; Phone: +52-55-5622-3899; Email: medinajl@unam.mx

Authors

Alejandro Gómez-García − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0003-4444-8221

Daniel A Acuña Jiménez − *CBio3 Laboratory, School of Chemistry, University of Costa Rica, San José 11501-2060, Costa Rica*

William J Zamora − *CBio3 Laboratory, School of Chemistry, University of Costa Rica, San José 11501-2060, Costa Rica; Laboratory of Computational Toxicology and Artificial Intelligence (LaToxCIA), Biological Testing Laboratory (LEBi), University of Costa Rica, San José 11501-2060, Costa Rica; Advanced Computing Lab (CNCA), National High Technology Center (CeNAT), San José 1174-1200, Costa Rica;* orcid.org/0000-0003-4029-4528

Haruna L Barazorda-Ccahuana − *Computational Biology and Chemistry Research Group, Vicerrectorado de Investigación, Universidad Católica de Santa María, Arequipa 04000, Peru*

Miguel A. Chávez-Fumagalli − *Computational Biology and Chemistry Research Group, Vicerrectorado de Investigación, Universidad Católica de Santa María, Arequipa 04000, Peru*

Marilia Valli − *School of Pharmaceutical Sciences of Ribeirao Preto (FCFRP), University of São Paulo (USP), Ribeirao Preto 14040-903 SP, Brazil*

Adriano D Andricopulo − *Laboratory of Medicinal and Computational Chemistry (LQMC), Centre for Research and Innovation in Biodiversity and Drug Discovery (CIBFar), São Carlos Institute of Physics (IFSC), University of São Paulo (USP), São Carlos 13563-120 SP, Brazil;* orcid.org/0000-0002-0457-818X

Vanderlan da S Bolzani − *Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, São Paulo State University (UNESP), Araraquara 14800-900 SP, Brazil*

Dionisio A Olmedo − *Center for Pharmacognostic Research on Panamanian Flora (CIFLORPAN), College of Pharmacy, University of Panama, Panama City 3366, Panama*

Pablo N Solís − *Center for Pharmacognostic Research on Panamanian Flora (CIFLORPAN), College of Pharmacy, University of Panama, Panama City 3366, Panama*

Marvin J Núñez − *Natural Product Research Laboratory, School of Chemistry and Pharmacy, University of El Salvador, San Salvador 01101, El Salvador*

Johny R Rodríguez Pérez − *GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia; GIEPRONAL Research Group, School of Basic Sciences, Technology and Engineering, Universidad Nacional Abierta y a Distancia, Dosquebradas 661001, Colombia;* orcid.org/0000-0002-5216-220X

Hoover A Valencia Sánchez − *GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia*

Héctor F Cortés Hernández − *GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia;* orcid.org/0000-0001-7242-0192

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509

Oscar M Mosquera Martinez − *GBPN Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c01560

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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 thank the Sao Paulo Research Foundation (FAPESP) grants #2020/11967-3 (DFG/FAPESP), #2022/08333-8 (DAAD/FAPESP), #2013/07600-3 (CIBFar-CEPID), #2014/50926-0, #465637/2014-0 (INCT BioNat CNPq/FAPESP), the National Council for Scientific and Technological Development (CNPq), and Coordination for the Improvement of Higher Education Personnel (CAPES). 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 thank the Vice Chancellor for Research of the University of Costa Rica for the grant via the research project 115-C2-126. DAO thanks the Vice-rectory of Research and Postgraduate Studies of the University of Panama for University Research Funds CUFI-VIP-01-14-2019-05 and SNI sponsor 2022 to 2024.

■ REFERENCES

(1) Stone, S.; Newman, D. J.; Colletti, S. L.; Tan, D. S. Cheminformatic Analysis of Natural Product-Based Drugs and Chemical Probes. *Nat. Prod. Rep.* **2022**, *39*, 20−32.

(2) Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. Artificial Intelligence for Natural Product Drug Discovery. *Nat. Rev. Drug Discovery* **2023**, *22*, 895−916.

(3) Medina-Franco, J. L.; Saldívar-González, F. I. Cheminformatics to Characterize Pharmacologically Active Natural Products. *Biomolecules* **2020**, *10*, 1566.

(4) Cockroft, N. T.; Cheng, X.; Fuchs, J. R. Starfish: A Stacked Ensemble Target Fishing Approach and Its Application to Natural Products. *J. Chem. Inf. Model.* **2019**, *59*, 4906−4920.

(5) Gangadevi, S.; Badavath, V. N.; Thakur, A.; Yin, N.; De Jonghe, S.; Acevedo, O.; Jochmans, D.; Leyssen, P.; Wang, K.; Neyts, J.; Yujie, T.; Blum, G. Kobophenol A Inhibits Binding of Host ACE2 Receptor with Spike RBD Domain of SARS-CoV-2, a Lead Compound for Blocking COVID-19. *J. Phys. Chem. Lett.* **2021**, *12*, 1793−1802.

(6) Chang, C.-C.; Hsu, H.-J.; Wu, T.-Y.; Liou, J.-W. Computer-Aided Discovery, Design, and Investigation of COVID-19 Therapeutics. *Tzu Chi Med. J.* **2022**, *34*, 276−286.

(7) Siva Kumar, B.; Anuragh, S.; Kammala, A. K.; Ilango, K. Computer Aided Drug Design Approach to Screen Phytoconstituents of Adhatoda Vasica as Potential Inhibitors of SARS-CoV-2 Main Protease Enzyme. *Life* **2022**, *12*, 315.

(8) Gao, H.; Dai, R.; Su, R. Computer-Aided Drug Design for the Pain-like Protease (PLpro) Inhibitors against SARS-CoV-2. *Biomed. Pharmacother.* **2023**, *159*, 114247.

(9) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminform.* **2021**, *13* (1), 2.

(10) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS One* **2013**, *8*, No. e62839.

(11) ISDB. *A database of In-Silico predicted MS/MS spectrum of Natural Products*, http://oolonek.github.io/ISDB/. (accessed 23 September 2024).

(12) Zhao, H.; Yang, Y.; Wang, S.; Yang, X.; Zhou, K.; Xu, C.; Zhang, X.; Fan, J.; Hou, D.; Li, X.; Lin, H.; Tan, Y.; Wang, S.; Chu, X.-Y.; Zhuoma, D.; Zhang, F.; Ju, D.; Zeng, X.; Chen, Y. Z. NPASS Database Update 2023: Quantitative Natural Product Activity and Species Source Database for Biomedical Research. *Nucleic Acids Res.* **2023**, *51*, D621−D628.

(13) Papageorgiou, L.; Andreou, A.; Christoforides, E.; Bethanis, K.; Vlachakis, D.; Thireou, T.; Eliopoulos, E. Hippo(crates): An integrated atlas for natural product exploration through a state-of-the art pipeline in chemoinformatics. *Wrld Acd Sci.* **2021**, *4* (1), 1.

(14) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in Silico. *PLoS One* **2011**, *6*, No. e15939.

(15) Mohanraj, K.; Karthikeyan, B. S.; Vivek-Ananth, R. P.; Chand, R. P. B.; Aparna, S. R.; Mangalapandi, P.; Samal, A. IMPPAT: A Curated Database of Indian Medicinal Plants, Phytochemistry And Therapeutics. *Sci. Rep.* **2018**, *8* (1), 4329.

(16) Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; Moumbock, A. F. A.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S. NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **2017**, *80* (7), 2067−2076.

(17) Ntie-Kang, F.; Zofou, D.; Babiaka, S. B.; Meudom, R.; Scharfe, M.; Lifongo, L. L.; Mbah, J. A.; Mbaze, L. M.; Sippl, W.; Efange, S. M. N. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS One* **2013**, *8*, No. e78085.

(18) Diallo, B. N.; Glenister, M.; Musyoka, T. M.; Lobb, K.; Bishop, Ö. T. SANCDB: An Update on South African Natural Compounds and Their Readily Available Analogs. *J. Cheminform.* **2021**, *13* (1), 37.

(19) Ntie-Kang, F.; Amoa Onguéné, P.; Fotso, G. W.; Andrae-Marobela, K.; Bezabih, M.; Ndom, J. C.; Ngadjui, B. T.; Ogundaini, A. O.; Abegaz, B. M.; Meva'a, L. M. Virtualizing the P-ANAPL Library: A Step towards Drug Discovery from African Medicinal Plants. *PLoS One* **2014**, *9*, No. e90655.

(20) Simoben, C. V.; Qaseem, A.; Moumbock, A. F. A.; Telukunta, K. K.; Günther, S.; Sippl, W.; Ntie-Kang, F. Pharmacoinformatic Investigation of Medicinal Plants from East Africa. *Mol. Inform.* **2020**, *39* (11), 2000163.

(21) Raven, P. H.; Gereau, R. E.; Phillipson, P. B.; Chatelain, C.; Jenkins, C. N.; Ulloa, C. U. The Distribution of Biodiversity Richness in the Tropics. *Sci. Adv.* **2020**, *6* (37), No. eabc6228.

(22) Mittermeier, R. A.; Turner, W. R.; Larsen, F. W.; Brooks, T. M.; Gascon, C. Global Biodiversity Conservation: The Critical Role of Hotspots. In *Biodiversity Hotspots*, Zachos, F. E.; Habel, J. C., Eds.; Springer: Berlin Heidelberg: Berlin, Heidelberg, 2011; pp. 3−22.

(23) Gómez-García, A.; Medina-Franco, J. L. Progress and Impact of Latin American Natural Product Databases. *Biomolecules* **2022**, *12*, 1202.

(24) Martínez-Heredia, L.; Quispe, P.; Fernández, J.; Lavecchia, M. NaturAr, a Collaborative, Open Source, Database of Natural Products

from Argentinian Biodiversity for Drug Discovery and Bioprospecting. *ChemRxiv*, **2024**.
(25) Rodríguez-Pérez, J. R.; Valencia-Sanchez, H. A.; Mosquera-Martinez, O. M.; Gómez-García, A.; Medina-Franco, J. L.; Cortes-Hernandez, H. F. NPDBEjeCol: A Natural Products Database from Colombia. *ChemRxiv*, **2024**.
(26) Gómez-García, A.; Jiménez, D. A. A.; Zamora, W. J.; Barazorda-Ccahuana, H. L.; Chávez-Fumagalli, M. Á.; Valli, M.; Andricopulo, A. D.; Bolzani, V. D. S.; Olmedo, D. A.; Solís, P. N.; Núñez, M. J.; Rodríguez Pérez, J. R.; Valencia Sánchez, H. A.; Cortés Hernández, H. F.; Medina-Franco, J. L. Navigating the Chemical Space and Chemical Multiverse of a Unified Latin American Natural Product Database: Lanapdb. *Pharmaceuticals* **2023**, *16*, 1388.
(27) Gómez-García, A.; Prinz, A.-K.; Jiménez, D. A. A.; Zamora, W. J.; Barazorda-Ccahuana, H. L.; Chávez-Fumagalli, M. Á.; Valli, M.; Andricopulo, A.; Bolzani, V. D. S.; Olmedo, D. A.; Solís, P. N.; Núñez, M. J.; Pérez, J. R. R.; Sánchez, H. A. V.; Hernández, H. F. C.; Martinez, O. M. M.; Koch, O.; Medina-Franco, J. L. Updating and Profiling the Natural Product-Likeness of Latin American Compound Libraries. *Mol. Inform.* **2024**, *43*, No. e202400052.
(28) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, *52*, D1180−D1192.
(29) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51*, D1373−D1380.
(30) Open-source chemoinformatics and machine learning, *RDKit: Open-Source Cheminformatics Software*. https://www.rdkit.org. (accessed 15December 2023).
(31) MolVS, *Molecule Validation and Standardization*. https://molvs.readthedocs.io/en/latest/index.html. (accessed 15 December 2023).
(32) *Venn, pyvenn: Venn diagrams for 2, 3, 4, 5, 6 sets*. https://pypi.org/project/venn/. (accessed 3 June 2024).
(33) Plotly Technologies Inc. *Collaborative Data Science Publisher: plotly Technologies Inc*; Plotly Technologies Inc.: Montréal, QC, 2015.
(34) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(35) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; Del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585*, 357−362.
(36) Waskom, M. L. Statistical Data Visualization. *JOSS* **2021**, *6* (60), 3021.
(37) Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E. L.; Strawbridge, S. A.; Garcia-Patino, M.; Kruger, R.; Sivakumaran, A.; Sanford, S.; Doshi, R.; Khetarpal, N.; Fatokun, O.; Doucet, D.; Zubkowski, A.; Rayat, D. Y.; Jackson, H.; Harford, K.; Anjum, A.; Zakir, M.; Wang, F.; Tian, S.; Lee, B.; Liigand, J.; Peters, H.; Wang, R. Q. R.; Nguyen, T.; So, D.; Sharp, M.; da Silva, R.; Gabriel, C.; Scantlebury, J.; Jasinski, M.; Ackerman, D.; Jewison, T.; Sajed, T.; Gautam, V.; Wishart, D. S. Drugbank 6.0: The Drugbank Knowledgebase for 2024. *Nucleic Acids Res.* **2024**, *52*, D1265−D1275.
(38) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **2021**, *84*, 2795−2807.
(39) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(40) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(41) Tanious, R.; Manolov, R. Violin Plots as Visual Tools in the Meta-Analysis of Single-Case Experimental Designs. *Methodology* **2022**, *18*, 221−238.
(42) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12* (1), 12.
(43) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(44) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(45) Ertl, P. Magic Rings: Navigation in the Ring Chemical Space Guided by the Bioactive Rings. *J. Chem. Inf. Model.* **2022**, *62*, 2164−2170.
(46) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(47) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7* (1), 20.
(48) Lohr, S. *Sampling:design and Analysis*; Brooks/Cole: Boston, MA, United States, 2010.
(49) Krzyzanowski, A.; Pahl, A.; Grigalunas, M.; Waldmann, H. Spacial Score—A Comprehensive Topological Indicator for Small-Molecule Complexity. *J. Med. Chem.* **2023**, *66*, 12739−12750.
(50) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1* (1), 8.
(51) *Matter Modeling*, https://mattermodeling.stackexchange.com/questions/8541/how-to-compute-the-synthetic-accessibility-score-in-python. (accessed 7 Auguest 2024).
(52) Węglarczyk, S. Kernel Density Estimation and Its Application. *ITM Web Of Conferences* **2018**, *23*, 00037.
(53) Valli, M.; dos Santos, R. N.; Figueira, L. D.; Nakajima, C. H.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. Development of a Natural Products Database from the Biodiversity of Brazil. *J. Nat. Prod.* **2013**, *76*, 439−444.
(54) Pilon, A. C.; Valli, M.; Dametto, A. C.; Pinto, M. E. F.; Freire, R. T.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. NuBBEDB: An Updated Database to Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep.* **2017**, *7* (1), 7215.
(55) Scotti, M. T.; Herrera-Acevedo, C.; Oliveira, T. B.; Costa, R. P. O.; Santos, S. Y. K. D. O.; Rodrigues, R. P.; Scotti, L.; Da-Costa, F. B. SistematX, an Online Web-Based Cheminformatics Tool for Data Management of Secondary Metabolites. *Molecules* **2018**, *23*, 103.
(56) Costa, R. P. O.; Lucena, L. F.; Silva, L. M. A.; Zocolo, G. J.; Herrera-Acevedo, C.; Scotti, L.; Da-Costa, F. B.; Ionov, N.; Poroikov, V.; Muratov, E. N.; Scotti, M. T. The Sistematx Web Portal of Natural Products: An Update. *J. Chem. Inf. Model.* **2021**, *61*, 2516−2522.
(57) UEFS Natural Products, http://zinc12.docking.org/catalogs/uefsnp. (accessed 20 March 2024).
(58) UNIIQUIM, https://uniiquim.iquimica.unam.mx/. (accessed 20 March 2024).
(59) Pilón-Jiménez, B. A.; Saldívar-González, F. I.; Díaz-Eufracio, B. I.; Medina-Franco, J. L. BIOFACQUIM: A Mexican Compound Database of Natural Products. *Biomolecules* **2019**, *9*, 31.
(60) Sánchez-Cruz, N.; Pilón-Jiménez, B. A.; Medina-Franco, J. L. Functional Group and Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database. *F1000Research*, **2020**, *8*, 2071.
(61) Olmedo, D. A.; González-Medina, M.; Gupta, M. P.; Medina-Franco, J. L. Cheminformatic Characterization of Natural Products from Panama. *Mol. Divers.* **2017**, *21*, 779−789.

(62) Olmedo, D. A.; Medina-Franco, J. L. Chemoinformatic Approach: The Case of Natural Products of Panama. In *Cheminformatics and its applications*; IntechOpen, 2019.
(63) Barazorda-Ccahuana, H. L.; Ranilla, L. G.; Candia-Puma, M. A.; Cárcamo-Rodriguez, E. G.; Centeno-Lopez, A. E.; Davila-Del-Carpio, G.; Medina-Franco, J. L.; Chávez-Fumagalli, M. A. PeruNPDB: The Peruvian Natural Products Database for in Silico Drug Screening. *Sci. Rep.* **2023**, *13* (1), 7577.
(64) Isah, M. B.; Tajuddeen, N.; Umar, M. I.; Alhafiz, Z. A.; Mohammed, A.; Ibrahim, M. A. Terpenoids as Emerging Therapeutic Agents: Cellular Targets and Mechanisms of Action against Protozoan Parasites. *Stud. Nat. Prod. Chem.* **2018**, *59*, 227−250.
(65) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(66) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technol.* **2004**, *1*, 337−341.
(67) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(68) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51*, 817−834.
(69) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; Decrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical Drug Properties Associated with in Vivo Toxicological Outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872−4875.
(70) Ozsoy, Y.; Gungor, S.; Cevher, E. Nasal Delivery of High Molecular Weight Drugs. *Molecules* **2009**, *14*, 3754−3779.
(71) Ivanescu, B.; Miron, A.; Corciova, A. Sesquiterpene Lactones from Artemisia Genus: Biological Activities and Methods of Analysis. *J. Anal. Methods Chem.* **2015**, *2015*, 1−21.
(72) Zhang, M.; Chen, T.; Lu, X.; Lan, X.; Chen, Z.; Lu, S. G Protein-Coupled Receptors (GPCRs): Advances in Structures, Mechanisms, and Drug Discovery. *Signal Transduct. Targeted Ther.* **2024**, *9* (1), 88.
(73) Sriram, K.; Insel, P. A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol. Pharmacol.* **2018**, *93*, 251−258.
(74) Silnitsky, S.; Rubin, S. J. S.; Zerihun, M.; Qyit, N. An Update on Protein Kinases as Therapeutic Targets-Part I: Protein Kinase C Activation and Its Role in Cancer and Cardiovascular Diseases. *Int. J. Mol. Sci.* **2023**, *24*, 17600.
(75) Craik, C. S.; Page, M. J.; Madison, E. L. Proteases as Therapeutics. *Biochem. J.* **2011**, *435*, 1−16.
(76) Ertl, P. Database of 4 Million Medicinal Chemistry-Relevant Ring Systems. *J. Chem. Inf. Model.* **2024**, *64*, 1245−1250.
(77) Yongye, A. B.; Waddell, J.; Medina-Franco, J. L. Molecular Scaffold Analysis of Natural Products Databases in the Public Domain. *Chem. Biol. Drug Des.* **2012**, *80*, 717−724.
(78) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus Diversity Plots: A Global Diversity Analysis of Chemical Libraries. *J. Cheminform.* **2016**, *8*, 63.
(79) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770−803.
(80) Saldívar-González, F. I.; Medina-Franco, J. L. Chemoinformatics Approaches to Assess Chemical Diversity and Complexity of Small Molecules. In *Small molecule drug discovery*; Elsevier, 2020; pp. 83−102.
(81) Oprea, T. I.; Bologa, C. Molecular Complexity: You Know It When You See It. *J. Med. Chem.* **2023**, *66*, 12710−12714.

https://doi.org/10.1021/acs.jcim.4c01560
J. Chem. Inf. Model. 2024, 64, 8495−8509