Latin American Natural Product Database (LANaPDB): an update

Alejandro Gómez-García,[1] Daniel A. Acuña Jiménez,[2] William J. Zamora,[2,3,4] Haruna L. Barazorda-Ccahuana,[5] Miguel Á. Chávez-Fumagalli,[5] Marilia Valli,[6] 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] José L. Medina-Franco[1,]*

[1] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico. alex.go.ga2121@gmail.com (A.G-G.)
[2] CBio3 Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José 11501-2060, Costa Rica; daniel.acunajimenez@ucr.ac.cr (D.A.A.J.); william.zamoraramirez@ucr.ac.cr (W.J.Z.)
[3] Laboratory of Computational Toxicology and Artificial Intelligence (LaToxCIA), Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José 11501-2060, Costa Rica
[4] Advanced Computing Lab (CNCA), National High Technology Center (CeNAT), Pavas, San José 1174-1200, Costa Rica
[5] Computational Biology and Chemistry Research Group, Vicerrectorado de Investigación, Universidad Católica de Santa María, Arequipa 04000, Peru; hbarazorda@ucsm.edu.pe (H.L.B.-C.); mchavezf@ucsm.edu.pe (M.Á.C.-F.)
[6] School of Pharmaceutical Sciences of Ribeirao Preto (FCFRP), University of São Paulo (USP), Avenida Professor Doutor Zeferino Vaz, s/n, Ribeirao Preto 14040-903, SP, Brazil
[7] 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), Av. João Dagnone, 1100, São Carlos 13563-120, SP, Brazil; marilia.valli@ifsc.usp.br (M.V.); aandrico@ifsc.usp.br (A.D.A.)
[8] Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, São Paulo State University (UNESP), Av. Prof. Francisco Degni, 55, Araraquara 14800-900, SP, Brazil; vanderlan.bolzani@unesp.br
[9] Center for Pharmacognostic Research on Panamanian Flora (CIFLORPAN), College of Pharmacy, University of Panama, Av. Manuel E. Batista and Jose De Fabrega, Panama City 3366, Panama; dionisio.olmedo@up.ac.pa (D.A.O.); pablo.solis@up.ac.pa (P.N.S.)
[10] Natural Product Research Laboratory, School of Chemistry and Pharmacy, University of El Salvador, Final Ave. Mártires Estudiantes del 30 de Julio, San Salvador 01101, El Salvador; marvin.nunez@ues.edu.sv
[11] GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia; johny.rodriguez@utp.edu.co (J.R.R.P.); hvalencia@utp.edu.co (H.A.V.S.); hfcortes@utp.edu.co (H.F.C.H.); omosquer@utp.edu.co (O.M.M.M.)
[12] GIEPRONAL Research Group, School of Basic Sciences, Technology and Engineering, Universidad Nacional Abierta y a Distancia, Dosquebradas 661001, Colombia
[13] GBPN Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira, Colombia
* Correspondence: medinajl@unam.mx; Tel.: +52-55-5622-3899

ABSTRACT

Natural product (NP) databases are crucial tools in computer-aided drug design (CADD). Over the last 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, to our knowledge, it 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, visualization of the chemical space, determination of the structural diversity, molecular complexity, synthetic feasibility, commercial availability, predicted and reported 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. The Latin American natural product collection LANaPDB is publicly available and can be downloaded at https://github.com/alexgoga21/LANaPDB-version-2/tree/main.

Keywords: chemoinformatics; chemical space; database; diversity; drug discovery; Latin America; open science

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 computer-aided drug design (CADD) NP-based. 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 CAAD NP-based 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 the CADD, since they provide access to thousands of molecules. In the last 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[9] with 449,048 NPs. The Collection of Open Natural Products (COCONUT)[10] contains 411,000 NPs and The Universal Natural Product Database[11] has 229,000 NPs. The Universal Natural Product Database is still accessible on the ISDB website.[11] NPASS database[12] has 94,413 NPs, of them 43,285 are annotated with the biological activity information. Hippo(crates)[13] database contains 45,300 NPs, NP derivatives and synthetic compounds, many of them 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, which is 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 AfroDb,[16] containing over 1000 molecules.

Latin America is a region with an extraordinary biodiversity and richness in endemic species. It is a region that may be home to at least a third of global biodiversity.[17] 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 mega-diverse 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.[18] 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.[19] Just recently, a NP database from Argentina[20] and Colombia[21] were published. In 2023 was published the first version of LANaPDB, a compendium that aims to gather and standardize the NP databases of Latin America[22] which was already included in COCONUT (https://coconut.naturalproducts.net/search?type=tags&q=Latin+America+dataset&tagType=dataSource).[10] In early 2024, and update was reported regarding the NP-likeness profile of the database.[23]

Herein, we report a major update of LANaPDB,[19] 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, visualization of the chemical space, quantification of the structural diversity, molecular complexity, synthetic feasibility, commercial availability, reported and predicted biological activity.

RESULTS AND DISCUSSION

Database update and data curation

The first version of LANaPDB comprised 12,959 compounds.[22] This reported update includes 619 new compounds, resulting a total of 13,578 compounds in its second version published in early 2024.[23] A new dataset 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, 1,707 compounds were considered for 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.

During the curation process, before the elimination of repeated molecules, a Venn diagram was constructed (Figure 1) to determine the number of unique molecules in every Latin American country of the databases that contain LANaPDB. It was found that the number of unique molecules is associated with the number of molecules in the country. Brazil is the country with more unique molecules (10,580), followed by Mexico (1360), Costa Rica (585), Panama (289), El Salvador (174), Peru (178), and Colombia (72).

Table 1. **Natural product databases in the updated version of LANaPDB**

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.	24,25
SistematX (Brazil)	9514	Plants	Database composed of secondary metabolites and developed at the Federal University of Paraiba.	26,27
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.	28
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.	21
NAPRORE-CR (Costa Rica)	~1600	Plants Microorganisms	Developed in the CBio3 and LaToxCIA Laboratories of the University of Costa Rica.	*
LAIPNUDELSAV (El Salvador)	214	Plants	Developed by the Research Laboratory in Natural Products of the University of El Salvador.	*
UNIIQUIM (Mexico)	1112	Plants	Natural products isolated and characterized at the Institute of Chemistry of the National Autonomous University of Mexico.	29
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.	30,31
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.	32,33

PeruNPDB (Peru)	280	Animals Plants	Natural products representative of Peruvian biodiversity. Created and curated at the Catholic University of Santa Maria.	[34]

*The database has not been published yet.

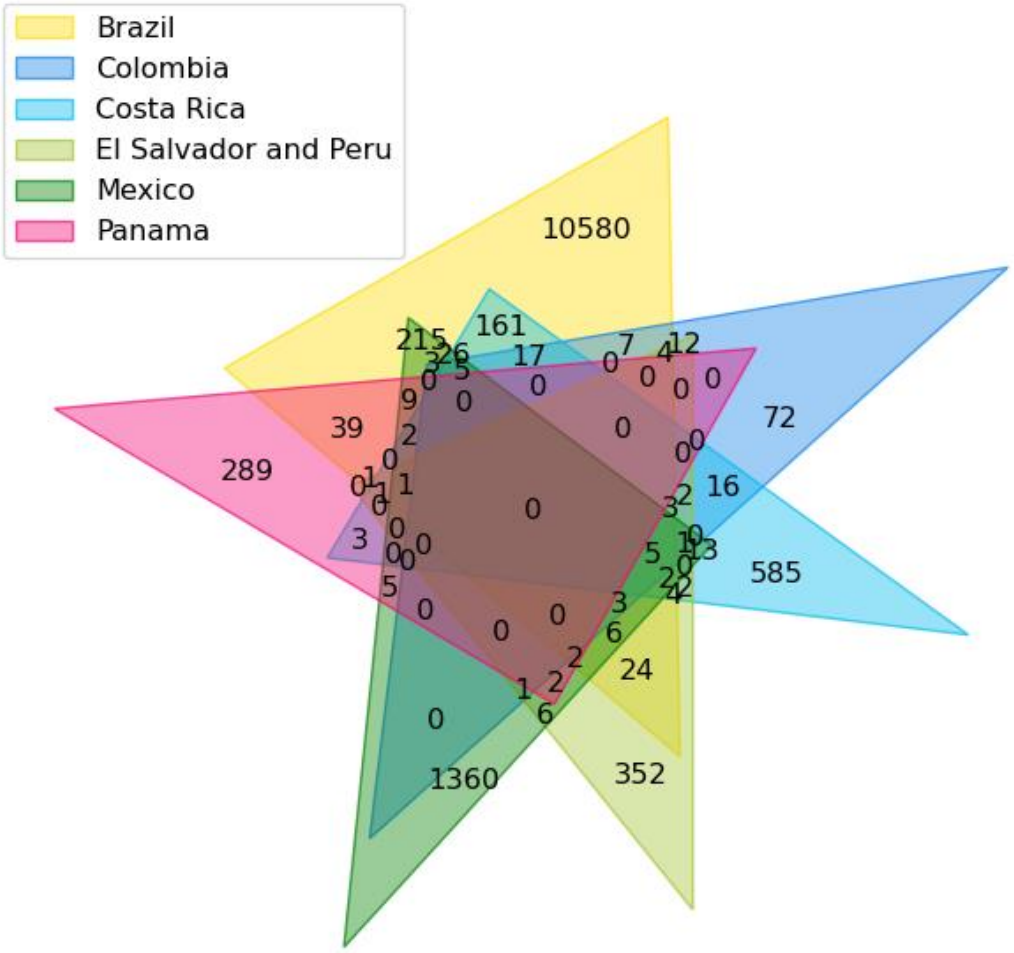

Figure 1. Venn diagram showing the number of unique and overlapped compounds of the Latin American natural product databases contained in LANaPDB. The compounds are grouped according to the country of origin of the database. The two countries with less compounds (El Salvador and Peru) were grouped. The number of unique molecules for El Salvador is 174 and Peru 178.

Structural classification

The compounds in LANaPDB were structurally classified according to a classification system based on the literature from the specialized metabolism of plants, marine organisms, fungi, and microorganisms. The classification system is divided in 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 the terpenoids (Figure 2). At the hierarchical level of pathway, the terpenoids, shikimates, phenylpropanoids, and alkaloids encompass more than 90% of the total

compounds. At the hierarchical level of superclass and class, the terpenoids and the flavonoids were the predominant compounds (Figure 2). The above was expected because the terpenoids are the predominant secondary metabolites produced by natural sources.[35] Compared to the previous version of LANaPDB the above tendencies have not changed.[22]

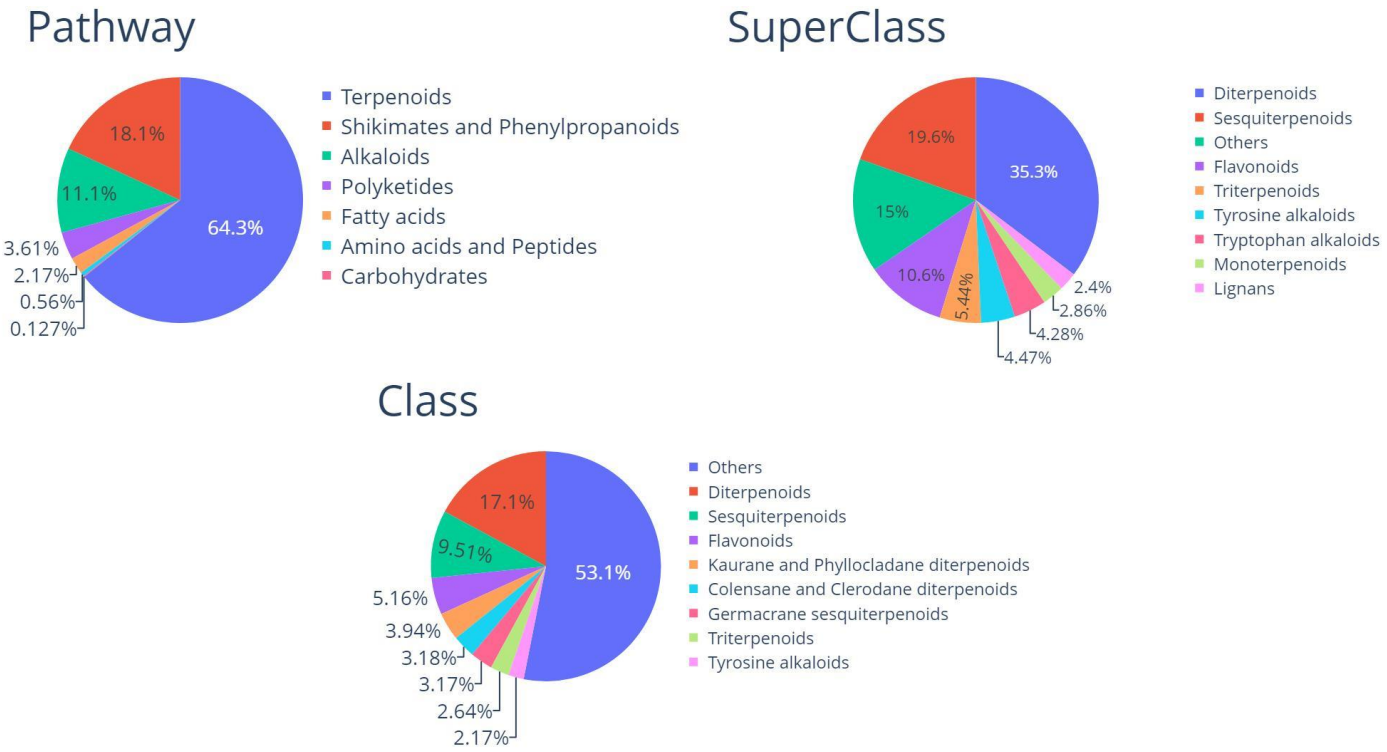

Figure 2. Pie charts showcasing the distribution of the LANaPDB compounds, according to a classification system[36] based on the literature from the specialized metabolism of the producing organisms.

Physicochemical properties

We calculated physicochemical properties of pharmaceutical interest for the LANaPDB compounds and compared them with two reference datasets: COCONUT[10] and FDA-approved small-molecule drugs.[37] Figures 3 and 4 show the distribution of the calculated physicochemical properties: SlogP,[38] molecular weight (MW), topological polar surface area (TPSA),[39] 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 the limits of some drug-likeness rules of thumb: Lipinski's rule of 5 (Ro5),[40,41] Veber's rules,[42] GlaxoSmithKline's (GSK) 4/400 rule,[43] and Pfizer 3/75 rule.[44] The improvement of some drug-like parameters is attributed to the fulfillment of these rules of thumb. In Figure 3 are not appreciated noticeable changes in the distribution of the

physicochemical properties of LANaPDB compared to the previous version.[22] The above can be attributed to the fact that the terpenoids remain as the prevalent compounds (Figure 2). Moreover, the LANaPDB and COCONUT distribution of the compounds (Figure 3) is very similar for all the physicochemical properties, except for the rotatable bonds whose distribution in LANaPDB is focused in a region with less rotatable bonds. The compound distribution (Figure 3) of LANaPDB and FDA-approved small-molecule drugs overlaps in all the cases, and in both datasets are the same regions where the greatest number of compounds are focused. Besides, the violin plots (Figure 3) shows that most of the LANaPDB compounds have physicochemical properties of pharmaceutical interest that fulfill the drug-likeness rules of thumb. 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 semi-synthetic drugs or pseudo-NP.

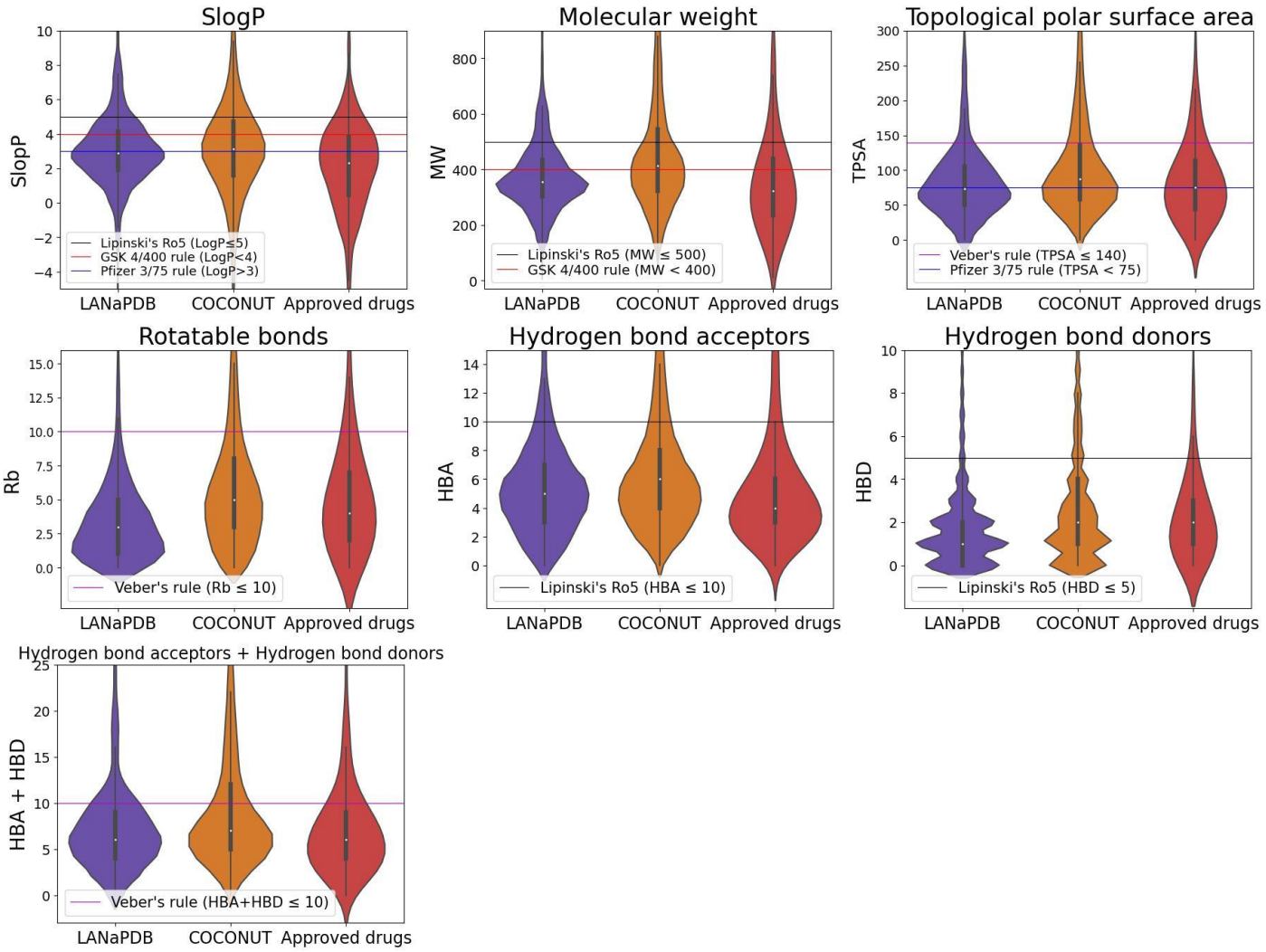

Figure 3. Violin plots summarizing the distribution of seven physicochemical properties of pharmaceutical interest of the compounds of three databases: LANaPDB, COCONUT and FDA-approved small-molecule drugs.

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 the 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 were added new compounds of Costa Rica and Mexico, nevertheless, the distribution of the physicochemical properties of the compounds of both countries compared to the previous version of LANaPDB[22] 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 in such a way that most of the compounds fulfill the drug-likeness rules of thumb.

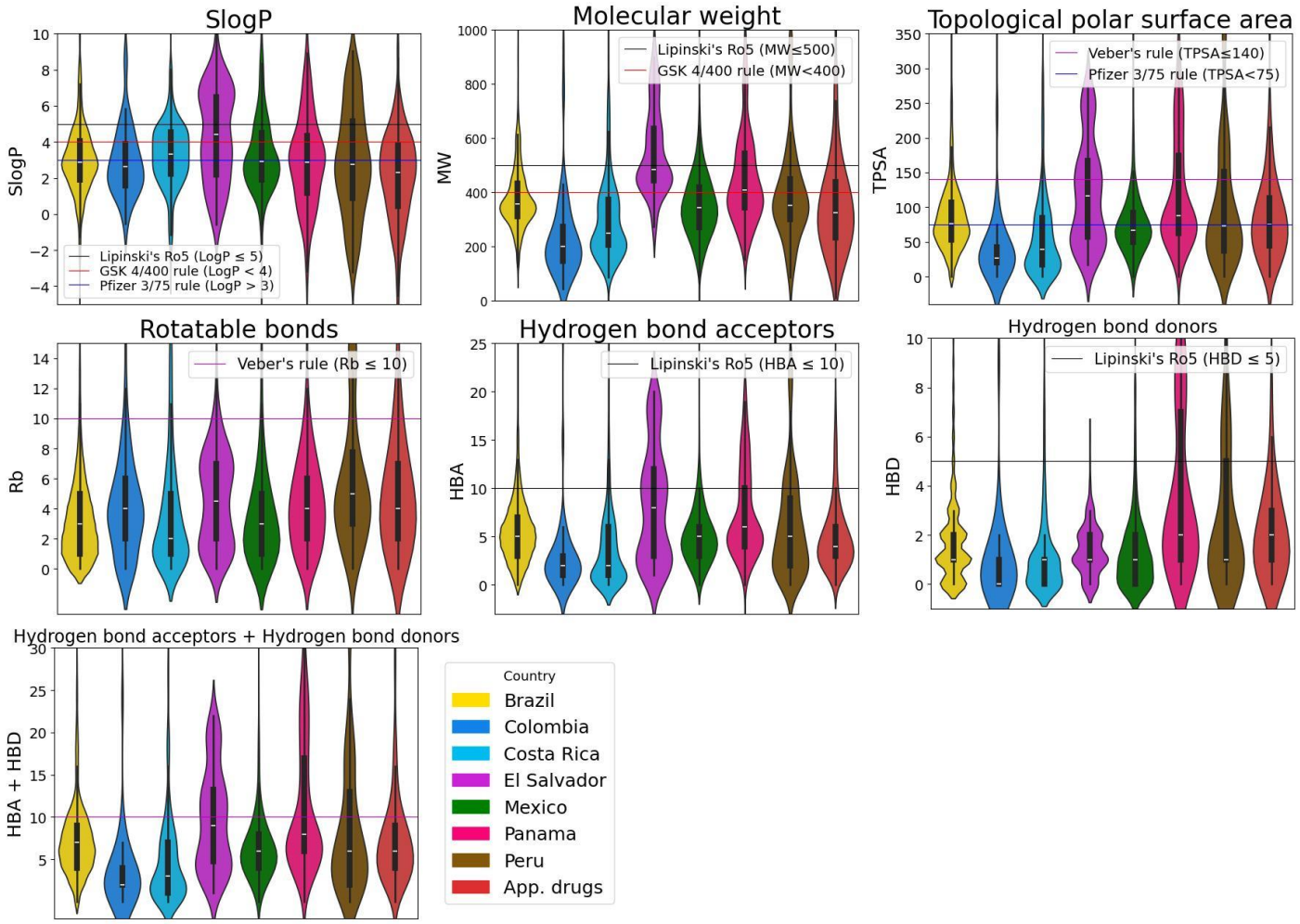

Figure 4. Violin plots summarizing the distribution of seven physicochemical properties of pharmaceutical interest of the compounds in LANaPDB and FDA-approved small-molecule drugs (App. drugs). The databases that encompass LANaPDB for every country: Brazil (NuBBEDB, SistematX and UEFS), Colombia (NPDB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

Chemical space visualization

Figure 5 shows the TMAP of LANAPDB generated from the MACCS keys (166-bit) fingerprint,[45] and their comparison with the FDA-approved small-molecule drugs.[37] The interactive version of the TMAP is freely available at https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP.html (to open the interactive map, download the file and open it in a web explorer, zoom in option available with the mouse scroll). In the interactive version of Figure 5, it can be appreciated that the TMAP accomplished effectively the clustering of structurally similar compounds in "branches." Figure 5 shows that all the countries and the approved drugs overlap with the Brazilian NPs. Therefore, Brazil is the country with the highest structural

diversity of NPs according to the TMAP. Moreover, Figure 5 shows that the compounds for each one of the seven Latin American countries 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. Figure 6 depicts the comparison of LANaPDB with COCONUT and the approved drugs. 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 5, it is appreciated that the approved drugs are distributed across the chemical space, overlapping LANaPDB and COCONUT in different regions.

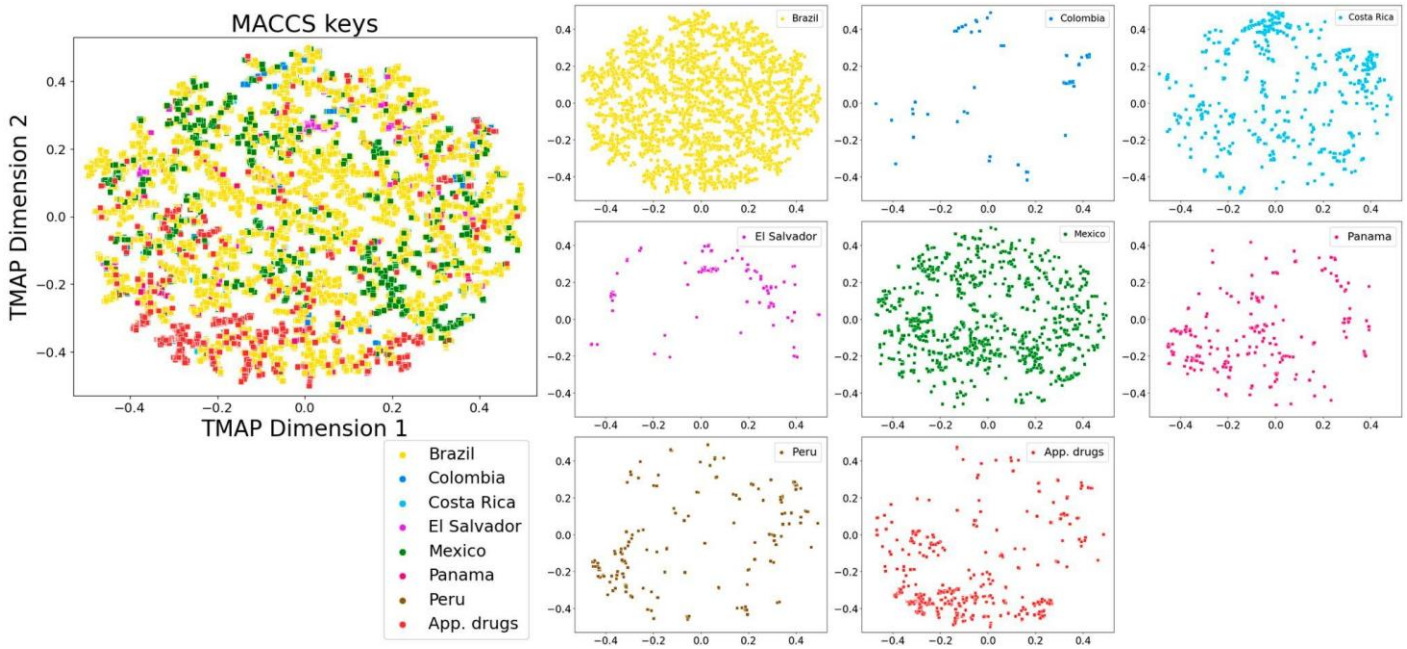

Figure 5. Tree MAP of LANaPDB and the comparison with FDA-approved small-molecule drugs, generated from MACCS keys (166-bits) fingerprint. An interactive version of the TMAP at https://github.com/alexgoga21/LANaPDB-version-2/blob/main/Interactive%20TMAP.html (to open the interactive map, download the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

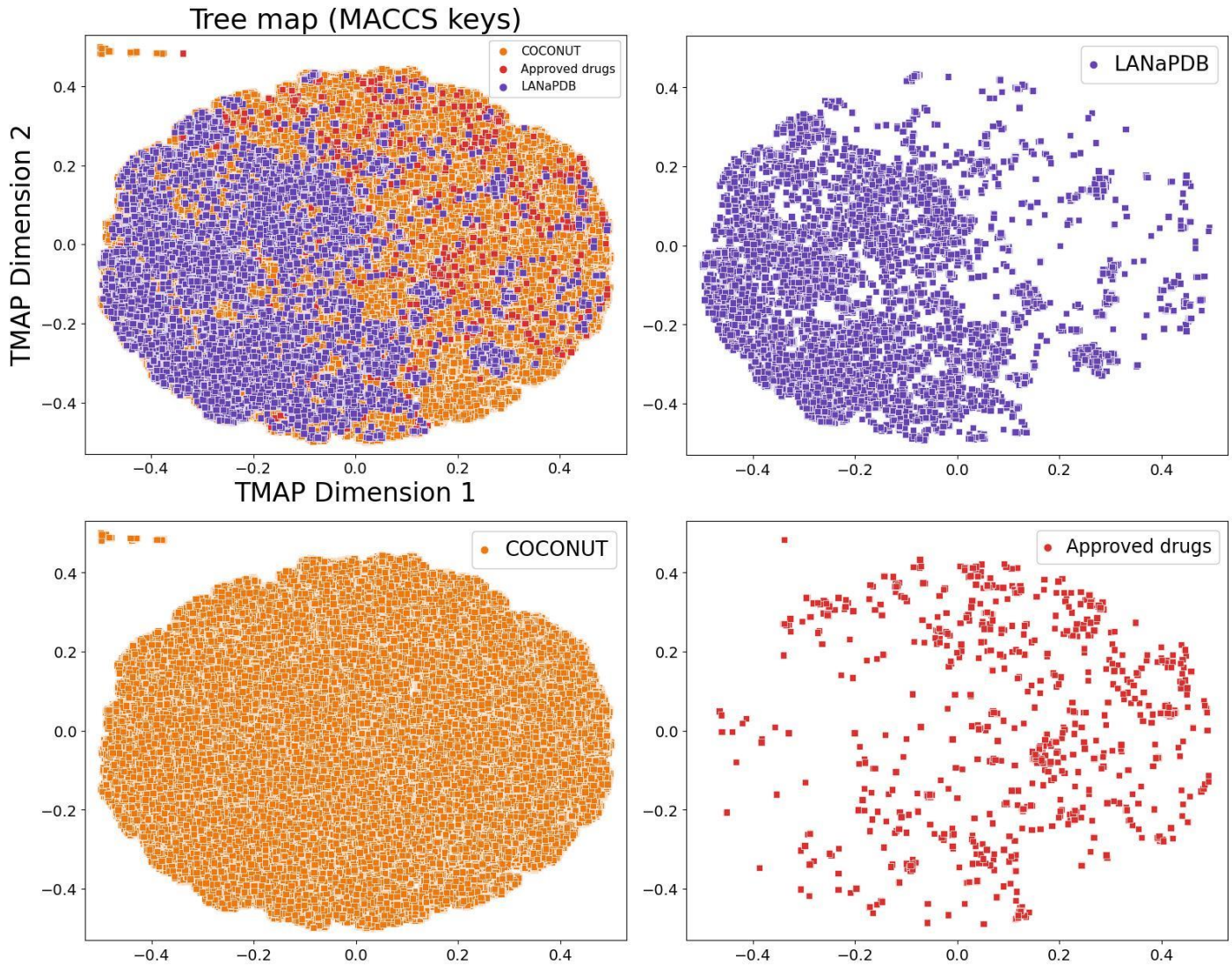

Figure 6. Tree MAP of LANaPDB and the comparison with COCONUT and FDA-approved small-molecule drugs, generated from MACCS keys (166-bits) fingerprint.

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[46] and ChEMBL, version 34.[47] From both databases was retrieved the ID (identification) code that allows to identify and differentiate every single compound. In the case of PubChem the ID code is known as CID (compound identification) and SID (substance identification). From all the LANaPDB compounds, were successfully retrieved 71.71% of the ID codes 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 correspondent ID code in the respective websites of both databases. The additional information that

can be checked in PubChem for the LANaPDB compounds includes spectral information, toxicity, and patents. ChEMBL contains information about the 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 in the PubChem website, from the PubChem ID codes added to LANaPDB. Moreover, all the molecules were classified in 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, version 34, employing 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 pharmaceutic aid (flavor) (4 compounds), pharmaceutic aid (solvent) (3 compounds), antifungal (3 compounds), pharmaceutic aid (antimicrobial agent) (2 compounds), pharmaceutic aid (emulsion adjunct) (2 compounds) and inhibitor (alpha-glucosidase) (2 compounds).

The second approach was based on a study of 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.[48] For LANaPDB, it was determined which compounds contain these bioactive ring systems reported by Ertl. It was found that 31.51% of the LANaPDB compounds (4,279 molecules) have bioactive ring systems. Chart 1 shows the twenty 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.[49] It is important to take in 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 reported Ertl.[48] 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 database of four million medicinal chemistry-relevant scaffolds that are not included in ChEMBL and PubChem.[50]

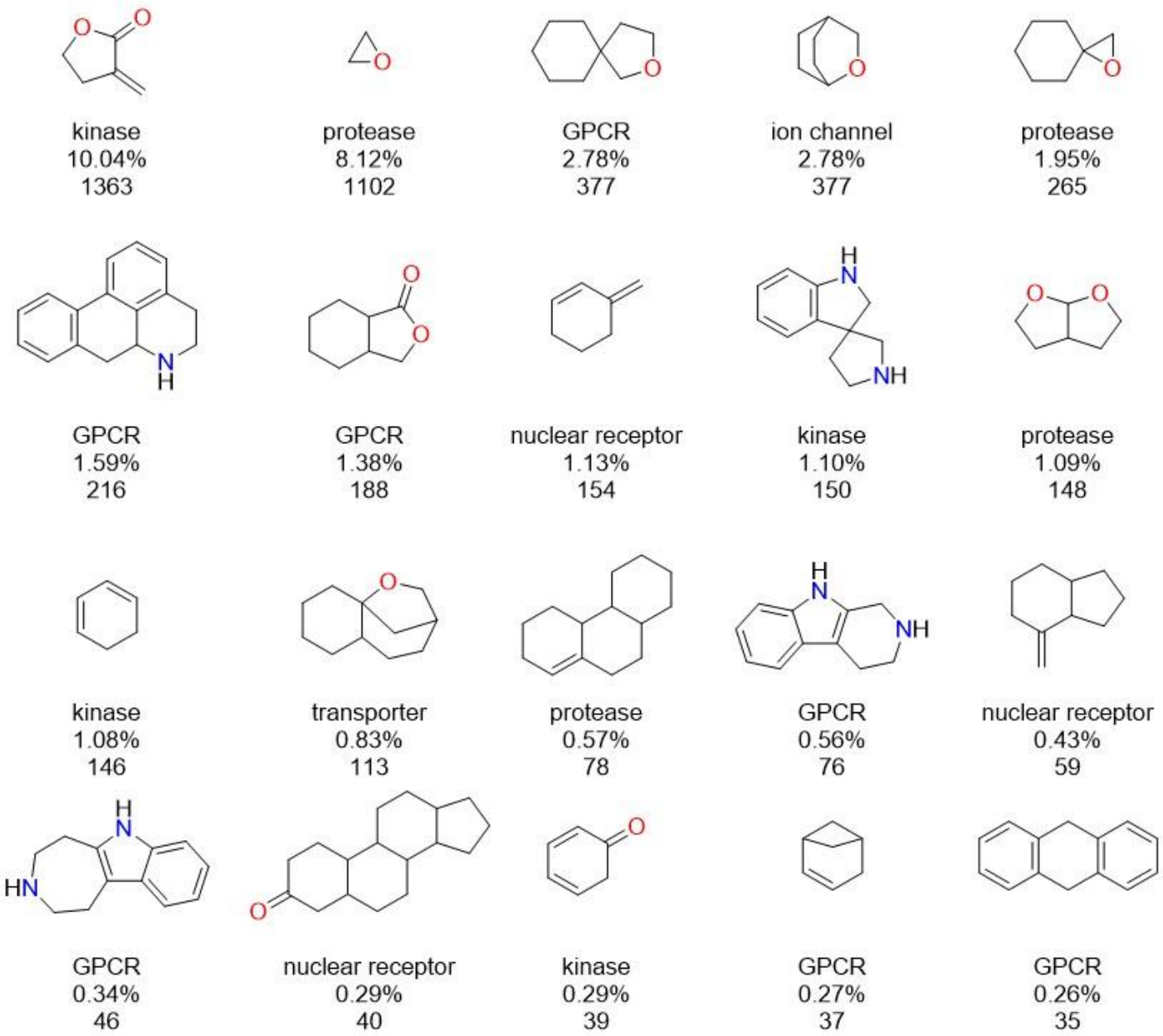

Chart 1. The twenty most abundant bioactive ring systems in LANaPDB, their biological target, percentage of occurrence and the total number of compounds that contain the ring system in LANaPDB.

Structural diversity

The structural diversity of LANaPDB was quantified with two types of molecular representations, molecular scaffolds and fingerprints. The diversity was compared with COCONUT and FDA-approved small-molecule drugs. The scaffold diversity of all data sets was measured with cumulative scaffold recovery (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 on the y-axis. Two metrics were obtained from the CSR curves: area under the curve (AUC) and the fraction of scaffolds to retrieve 50% of the compounds in the database (F_{50}) (i.e., if a dataset has F_{50}=0.43, 50% of the compounds in the dataset 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 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 where all the compounds have the same scaffold. In this case, the CSR function would be a vertical line with 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.[51,52]

In the consensus diversity plots (Figures 7A and 7B) it is shown that FDA-approved small-molecule drugs is the dataset 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 dataset has not just NPs, instead a significant proportion are NP-derivatives and purely synthetic molecules,[53] which increases the structural diversity. According to the MPTS metric, the side chain structural diversity of LANaPDB is lower than COCONUT. Nonetheless, considering the AUC and F_{50} metrics, LANaPDB has higher scaffold diversity than 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.

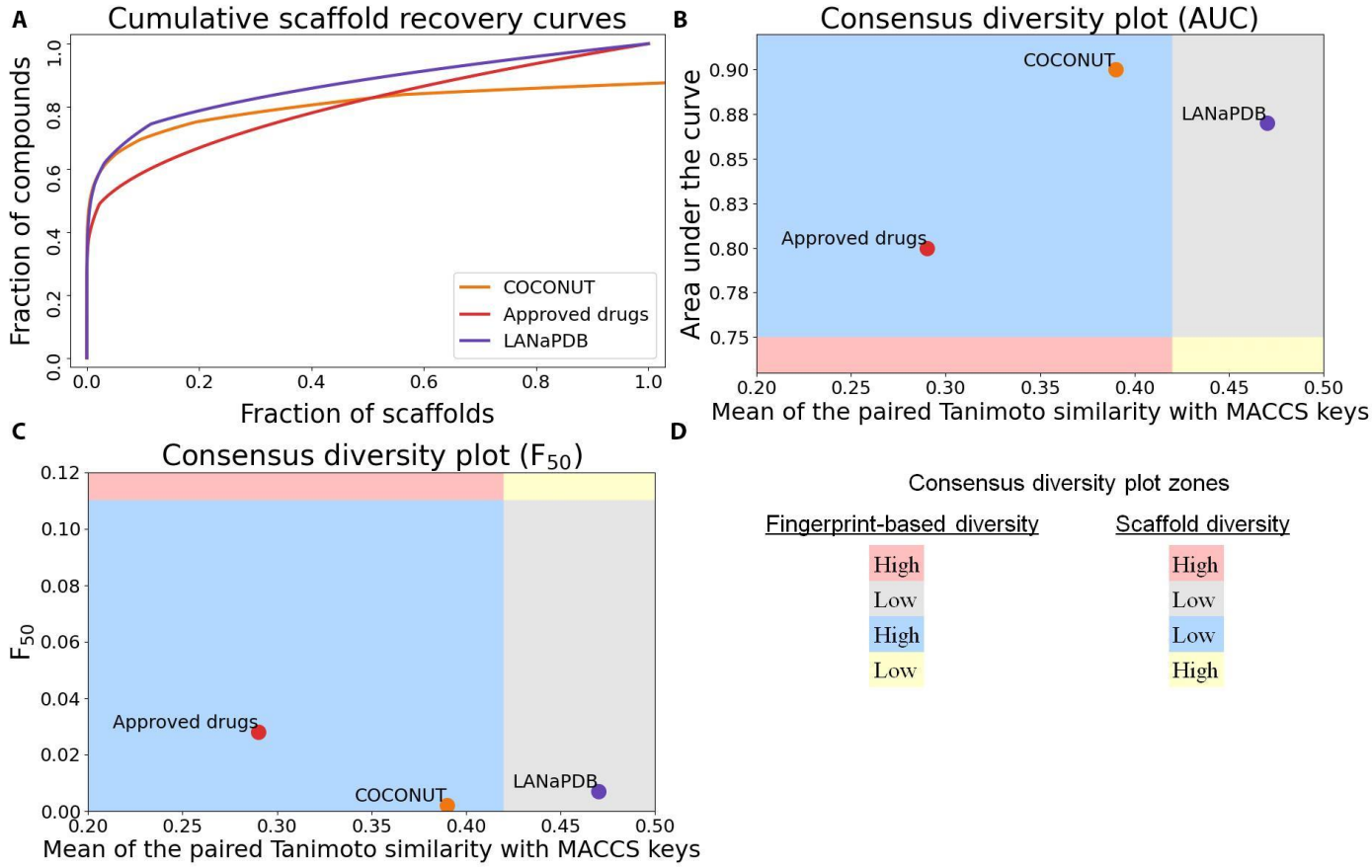

Figure 7. **A**) Cumulative scaffold recovery (CSR) curves of LANaPDB, COCONUT and FDA-approved small-molecule drugs. Consensus diversity plots of LANaPDB, COCONUT and FDA-approved small-molecule drugs, which describe the datasets diversity considering the MACCS keys (166-bit) fingerprint, **B**) area 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.

Molecular complexity and synthetic feasibility

Molecular complexity can be quantified using several different metrics.[54] In this work, as a quantitative measure of molecular complexity we employed the recently developed metric normalized spacial score (nSPS).[55] The synthetic feasibility was determined calculating the synthetic accessibility score (SAscore).[56] The distribution of both metrics was represented with kernel density estimate (KDE) plots which represent the data using continuous probability density curves (Figure 8). nSPS takes into account the atom hybridization, stereoisomerism, presence and complexity of aromatic or non-aromatic rings, and the number of heavy-atom neighbors.[55] As a reference, in an earlier study, it was found that the nSPS values of most of the approved drugs is between ten and twenty and this has remained without any significant changes in the last eight decades.[57] The nSPS values for the compounds of the three databases studied in this work are centered around

ten and twenty (Figure 8A). Thus, LANaPDB has a significant proportion of compounds with nSPS values between ten and twenty (39.78%) and those compounds are expected to have a similar pharmacokinetic profile to the approved drugs according to the molecular similarity principle.[57] Moreover, unlike the other two reference databases, the LANaPDB compounds presented mainly nSPS values around thirty and fifty (26.88%) (Figure 8A). Previously, it has been found that the ligand potency and target selectivity are maximized in compounds with nSPS values between twenty and forty.[55] Therefore, LANaPDB has a significant proportion of compounds with nSPS values between twenty and forty (37.95%) which are expected to have a good potency and target selectivity. The nSPS value for each compound in LANaPDB is indicated in the database publicly available.

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 is associated with a higher synthetic feasibility.[56] In this work, approved drugs and COCONUT presented mainly SAscores between two and three (Figure 8B). 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.[53] The LANaPDB compounds have mostly SAscores around five, which implies that a significant proportion of the LANaPDB compounds have a higher synthetic feasibility than the approved drugs.

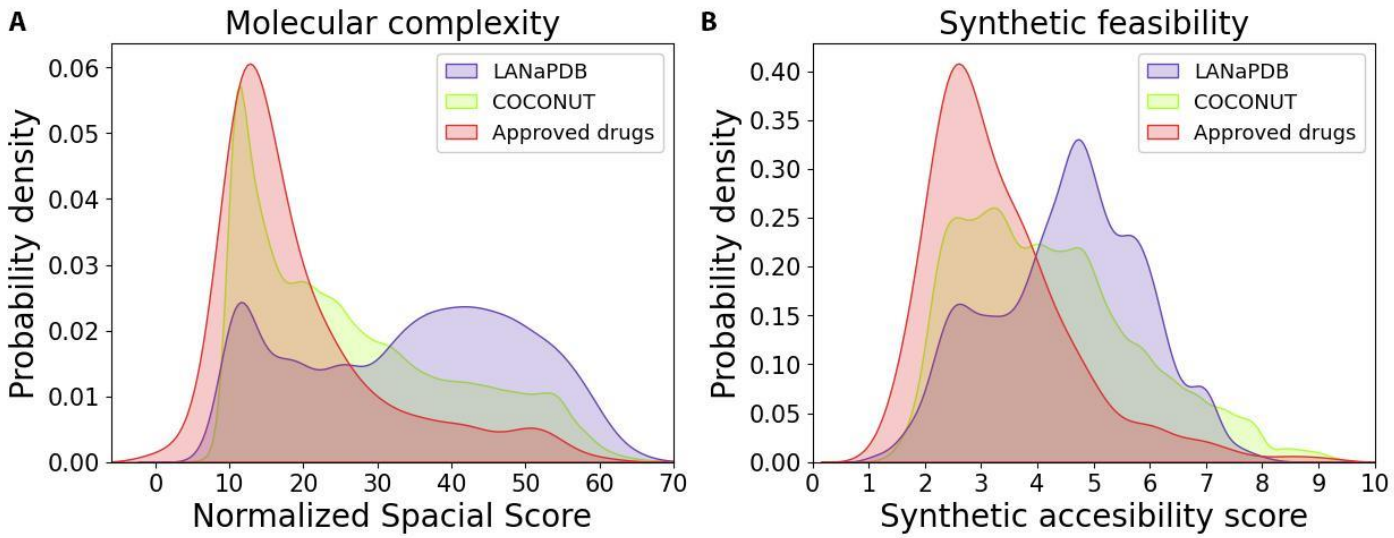

Figure 8. Kernel density estimate plots that represent the distribution of the **A)** Normalized Spacial Score and **B)** Synthetic accessibility score of LANaPDB, COCONUT, and FDA-approved small-molecule drugs.

EXPERIMENTAL SECTION

The version of the Python programming language that was used for all the analysis in this manuscript is 3.10.7. The version of the Python modules: RDKit (2022.03.5),[58] MolVS (0.1.1),[59] Venn (0.1.3),[60] Plotly express (0.4.1),[61] Scikit-learn (1.2.2),[62] NumPy (1.23.2)[63] and seaborn (0.12.2).[64]

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. To 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 is accessible through an open-data portal (www.npdbejecol.com). 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 of the second version of LANaPDB was carried out with the same workflow employed in the first version of the database.[22] The process was made in the Python programming language, employing the RDKit and MolVS modules. The standard curation process of MolVS was implemented through the standardize function included in this Python module: 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 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, just the largest fragment is kept and is tried to neutralize all the molecules of the database. The canonical tautomer was determined, and, from the InChIKey strings of the canonical tautomer, the duplicate compounds were removed. The same curation workflow was applied to two reference datasets

employed to compare LANaPDB: COCONUT[10] and FDA-approved small-molecule drugs, version 5.1.10 (released by DrugBank in January 2023).[37]

In LANaPDB, during the curation process, before the elimination of duplicate molecules, a Venn diagram (Figure 1) was constructed in the Python programming language employing the Venn module, from the InChIKey strings of the molecules.

Structural classification

The freely available online server NPClassifier[36] was employed to do 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 utilizing the Plotly express module.

Physicochemical properties

The following physicochemical properties of pharmaceutical interest were calculated in Python employing the RDKit module: SlogP,[38] molecular weight (MW), topological polar surface area (TPSA)[39], number of rotatable bonds (Rb), number of hydrogen bond acceptors (HBA), and number of hydrogen bond donors (HBD). The distribution of the physicochemical properties was depicted with violin plots, constructed in the Python programming language with the Scikit-learn module.

Chemical space visualization

The visualization of the chemical space of LANaPDB was made using the TMAP (Tree MAP) algorithm[65] from the MACCS keys fingerprint.[45] The determination of the MACCS keys fingerprint was made in the Python programming language with the RDKit module. The construction of the TMAP was made with Python following the reported protocol.[65] The results were compared with two reference datasets: COCONUT[10] 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 and ChEMBL (ID codes) were requested and retrieved from the respective websites of both databases. The request and retrieval of the ID codes was 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 module.

Commercial availability and chirality

The commercial availability of every compound of LANaPDB was obtained from the PubChem website.[46] 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 the chirality was made in Python, employing the function Chem.FindMolChiralCenters of the RDKit module.

Biological activity

In the first approach, with the Python programming language, employing the ChEMBL API, from the InChIKey strings was requested and retrieved from the ChEMBL database website the reported biological activity of the LANaPDB compounds. In the second approach, in the Python programming language employing the RDKit module, it was determined if the SMILES strings of the LANaPDB molecules contained the SMILES strings of the ChEMBL bioactive rings reported by Ertl.[48]

Structural diversity

The Bemis and Murcko scaffolds[66] were determined from the SMILES strings in the Python programming language with the RDKit module. The AUC was obtained from the CSR curves with the trapezoidal rule in the Python programming language with the trapz function of the numpy module. The F_{50} metric was obtained

from the CSR curves, 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 module. The MACCS keys (166-bit) fingerprint and the paired Tanimoto similarity were calculated in the Python programming language with the RDKit module. The paired Tanimoto similarity calculation for the COCONUT dataset was made with a random sample of the 10% (with more than 40,000 compounds) that represents the diversity of the whole database.[67]

Molecular complexity and synthetic feasibility

The nSPS and SAscore were determined in the Python programming language with the RDKit module, employing the SpacialScore and sascorer[68] functions. The KDE plots were constructed in the Python programming language with the seaborn module.

ASSOCIATED CONTENT

Supporting information

The Supporting Information is available free of charge at (Added by the editorial)

LANaPDB version 1 (WEB SERVER, XLSX)

LANaPDB version 2 (XLSX)

Interactive version of the TMAP (HTML)

AUTHOR INFORMATION

Corresponding Author

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;* https://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, Avenida Universidad 3000, Mexico City 04510, Mexico; https://orcid.org/0000-0003-4444-8221*; Email: alex.go.ga21@hotmail.com

Daniel A. Acuña Jiménez - *CBio3 Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José 11501-2060, Costa Rica;* Email:daniel.acunajimenez@ucr.ac.cr

William J. Zamora - *CBio3 Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José 11501-2060, Costa Rica;* Email: william.zamoraramirez@ucr.ac.cr

Haruna L. Barazorda-Ccahuana - *Computational Biology and Chemistry Research Group, Vicerrectorado de Investigación, Universidad Católica de Santa Maria, Arequipa 04000, Peru;* Email: hbarazorda@ucsm.edu.pe

Miguel Á. Chávez-Fumagalli - *Computational Biology and Chemistry Research Group, Vicerrectorado de Investigación, Universidad Católica de Santa Maria, Arequipa 04000, Peru;* Email: mchavezf@ucsm.edu.pe

Marilia Valli - *School of Pharmaceutical Sciences of Ribeirao Preto (FCFRP), University of São Paulo (USP), Avenida Professor Doutor Zeferino Vaz, s/n, Ribeirao Preto 14040-903, SP, Brazil;* Email: marilia.valli@usp.br

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), Av. João Dagnone, 1100, São Carlos 13563-120, SP, Brazil;* Email: aandrico@ifsc.usp.br

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), Av. Prof. Francisco Degni, 55, Araraquara 14800-900, SP, Brazil;* Email: vanderlan.bolzani@unesp.br

Dionisio A. Olmedo - *Center for Pharmacognostic Research on Panamanian Flora (CIFLORPAN), College of Pharmacy, University of Panama, Av. Manuel E. Batista and Jose De Fabrega, Panama City 3366, Panama;* Email: dionisio.olmedo@up.ac.pa

Pablo N. Solís - *Center for Pharmacognostic Research on Panamanian Flora (CIFLORPAN), College of Pharmacy, University of Panama, Av. Manuel E. Batista and Jose De Fabrega, Panama City 3366, Panama;* Email: pablonsolis@gmail.com

Marvin J. Núñez - *Natural Product Research Laboratory, School of Chemistry and Pharmacy, University of El Salvador, Final Ave. Mártires Estudiantes del 30 de Julio, San Salvador 01101, El Salvador;* Email: marvin.nunez@ues.edu.sv

Johny R. Rodríguez Pérez - *GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia;* Email: johny.rodriguez@utp.edu.co

Hoover A. Valencia Sánchez - *GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia;* Email: hvalencia@utp.edu.co

Héctor F. Cortés Hernández - *GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia;* Email: hfcortes@utp.edu.co

Oscar M. Mosquera Martinez - *GBPN Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira, Colombia;* Email: omosquer@utp.edu.co

Notes

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

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 #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), 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 thanks to the Vice Chancellor for Research of the University of Costa Rica for grant via the research project 115-C2-126.

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* (1), 20–32. DOI: 10.1039/d1np00039j.

(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 Discov.* **2023**, *22* (11), 895–916. DOI: 10.1038/s41573-023-00774-7.

(3) Medina-Franco, J. L.; Saldívar-González, F. I. Cheminformatics to Characterize Pharmacologically Active Natural Products. *Biomolecules* **2020**, *10* (11). DOI: 10.3390/biom10111566.

(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* (11), 4906–4920. DOI: 10.1021/acs.jcim.9b00489.

(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* (7), 1793–1802. DOI: 10.1021/acs.jpclett.0c03119.

(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 Medical Journal* **2022**, *34* (3), 276–286. DOI: 10.4103/tcmj.tcmj_318_21.

(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 (Basel)* **2022**, *12* (2). DOI: 10.3390/life12020315.

(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. DOI: 10.1016/j.biopha.2023.114247.

(9) Gallo, K.; Kemmler, E.; Goede, A.; Becker, F.; Dunkel, M.; Preissner, R.; Banerjee, P. SuperNatural 3.0-a Database of Natural Products and Natural Product-Based Derivatives. *Nucleic Acids Res.* **2023**, *51* (D1), D654–D659. DOI: 10.1093/nar/gkac1008.

(10) 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. DOI: 10.1186/s13321-020-00478-9.

(11) 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* (4), e62839. DOI: 10.1371/journal.pone.0062839.

(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* (D1), D621–D628. DOI: 10.1093/nar/gkac1069.

(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. DOI: 10.3892/wasj.2021.136.

(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* (1), e15939. DOI: 10.1371/journal.pone.0015939.

(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. DOI: 10.1038/s41598-018-22631-z.

(16) 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* (10), e78085. DOI: 10.1371/journal.pone.0078085.

(17) Raven, P. H.; Gereau, R. E.; Phillipson, P. B.; Chatelain, C.; Jenkins, C. N.; Ulloa Ulloa, C. The Distribution of Biodiversity Richness in the Tropics. *Sci. Adv.* **2020**, *6* (37). DOI: 10.1126/sciadv.abc6228.

(18) 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. DOI: 10.1007/978-3-642-20992-5_1.

(19) Gómez-García, A.; Medina-Franco, J. L. Progress and Impact of Latin American Natural Product Databases. *Biomolecules* **2022**, *12* (9). DOI: 10.3390/biom12091202.

(20) 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. **2024**. DOI: 10.26434/chemrxiv-2024-56rks.

(21) 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. **2024**. DOI: 10.26434/chemrxiv-2024-vp95j.

(22) 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. da 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* (10), 1388. DOI: 10.3390/ph16101388.

(23) 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. D.; da S Bolzani, V.; Olmedo, D. A.; Solís, P. N.; Núñez, M. J.; Rodríguez Pérez, J. R.; Sánchez, H. A. V.; Cortés Hernández, H. F.; Mosquera Martinez, O. M.; Koch, O.; Medina-Franco, J. L. Updating and Profiling the Natural Product-Likeness of Latin American Compound Libraries. *Mol. Inform.* **2024**, *43* (7), e202400052. DOI: 10.1002/minf.202400052.

(24) 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* (3), 439–444. DOI: 10.1021/np3006875.

(25) 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. DOI: 10.1038/s41598-017-07451-x.

(26) Scotti, M. T.; Herrera-Acevedo, C.; Oliveira, T. B.; Costa, R. P. O.; Santos, S. Y. K. de 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* (1). DOI: 10.3390/molecules23010103.

(27) 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* (6), 2516–2522. DOI: 10.1021/acs.jcim.1c00083.

(28) *UEFS Natural Products*. http://zinc12.docking.org/catalogs/uefsnp (accessed 2024-03-20).

(29) *UNIIQUIM*. https://uniiquim.iquimica.unam.mx/ (accessed 2024-03-20).

(30) 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* (1). DOI: 10.3390/biom9010031.

(31) Sánchez-Cruz, N.; Pilón-Jiménez, B. A.; Medina-Franco, J. L. Functional Group and Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database. *F1000Res.* **2020**, *8*, 2071. DOI: 10.12688/f1000research.21540.2.

(32) Olmedo, D. A.; González-Medina, M.; Gupta, M. P.; Medina-Franco, J. L. Cheminformatic Characterization of Natural Products from Panama. *Mol. Divers.* **2017**, *21* (4), 779–789. DOI: 10.1007/s11030-017-9781-4.

(33) A. Olmedo, D.; L. Medina-Franco, J. Chemoinformatic Approach: The Case of Natural Products of Panama. In *Cheminformatics and its applications [working title]*; IntechOpen, 2019. DOI: 10.5772/intechopen.87779.

(34) 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. DOI: 10.1038/s41598-023-34729-0.

(35) 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; Studies in natural products chemistry; Elsevier, 2018; Vol. 59, pp 227–250. DOI: 10.1016/B978-0-444-64179-3.00007-4.

(36) 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* (11), 2795–2807. DOI: 10.1021/acs.jnatprod.1c00399.

(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* (D1), D1265–D1275. DOI: 10.1093/nar/gkad976.

(38) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868–873. DOI: 10.1021/ci990307l.

(39) 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* (20), 3714–3717. DOI: 10.1021/jm000942e.

(40) 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 Deliv. Rev.* **2001**, *46* (1–3), 3–26. DOI: 10.1016/S0169-409X(00)00129-0.

(41) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discov. Today Technol.* **2004**, *1* (4), 337–341. DOI: 10.1016/j.ddtec.2004.11.007.

(42) 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* (12), 2615–2623. DOI: 10.1021/jm020017n.

(43) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51* (4), 817–834. DOI: 10.1021/jm701122q.

(44) 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* (17), 4872–4875. DOI: 10.1016/j.bmcl.2008.07.071.

(45) 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* (6), 1273–1280. DOI: 10.1021/ci010132r.

(46) 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* (D1), D1373–D1380. DOI: 10.1093/nar/gkac956.

(47) 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* (D1), D1180–D1192. DOI: 10.1093/nar/gkad1004.

(48) Ertl, P. Magic Rings: Navigation in the Ring Chemical Space Guided by the Bioactive Rings. *J. Chem. Inf. Model.* **2022**, *62* (9), 2164–2170. DOI: 10.1021/acs.jcim.1c00761.

(49) Ivanescu, B.; Miron, A.; Corciova, A. Sesquiterpene Lactones from Artemisia Genus: Biological Activities and Methods of Analysis. *J. Anal. Methods Chem.* **2015**, *2015*, 247685. DOI: 10.1155/2015/247685.

(50) Ertl, P. Database of 4 Million Medicinal Chemistry-Relevant Ring Systems. *J. Chem. Inf. Model.* **2024**, *64* (4), 1245–1250. DOI: 10.1021/acs.jcim.3c01812.

(51) 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* (5), 717–724. DOI: 10.1111/cbdd.12011.

(52) 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. DOI: 10.1186/s13321-016-0176-9.

(53) 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* (3), 770–803. DOI: 10.1021/acs.jnatprod.9b01285.

(54) 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. DOI: 10.1016/B978-0-12-818349-6.00003-0.

(55) Krzyzanowski, A.; Pahl, A.; Grigalunas, M.; Waldmann, H. Spacial Score─A Comprehensive Topological Indicator for Small-Molecule Complexity. *J. Med. Chem.* **2023**, *66* (18), 12739–12750. DOI: 10.1021/acs.jmedchem.3c00689.

(56) 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. DOI: 10.1186/1758-2946-1-8.

(57) Oprea, T. I.; Bologa, C. Molecular Complexity: You Know It When You See It. *J. Med. Chem.* **2023**, *66* (18), 12710–12714. DOI: 10.1021/acs.jmedchem.3c01507.

(58) Open-source chemoinformatics and machine learning. *RDKit: Open-Source Cheminformatics Software.* https://www.rdkit.org (accessed 2023-12-15).

(59) *MolVS. Molecule Validation and Standardization.* https://molvs.readthedocs.io/en/latest/index.html (accessed 2023-12-15).

(60) *Venn*. pypi. https://pypi.org/project/venn/ (accessed 2024-06-03).

(61) Plotly Technologies Inc. *Collaborative Data Science Publisher: Plotly Technologies Inc.*; Plotly Technologies Inc.: Montréal, QC, 2015.

(62) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *arXiv* **2012**. DOI: 10.48550/arxiv.1201.0490.

(63) 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* (7825), 357–362. DOI: 10.1038/s41586-020-2649-2.

(64) Waskom, M. Seaborn: Statistical Data Visualization. *JOSS* **2021**, *6* (60), 3021. DOI: 10.21105/joss.03021.

(65) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12* (1), 12. DOI: 10.1186/s13321-020-0416-x.

(66) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. DOI: 10.1021/jm9602928.

(67) Lohr, S. *Sampling:Design and Analysis*; Brooks/Cole: Boston, MA, United States, 2010.

(68) *Matter Modeling*. https://mattermodeling.stackexchange.com/questions/8541/how-to-compute-the-synthetic-accessibility-score-in-python (accessed 2024-08-07).

FOR TABLE OF CONTENTS ONLY