*Article*

# Navigating the Chemical Space and Chemical Multiverse of a Unified Latin American Natural Product Database: LANaPDB

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 [6],
Vanderlan da S. Bolzani [7], Dionisio A. Olmedo [8], Pablo N. Solís [8], Marvin J. Núñez [9],
Johny R. Rodríguez Pérez [10,11], Hoover A. Valencia Sánchez [10], Héctor F. Cortés Hernández [10]
and 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.ga21@hotmail.com
[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 Maria, Arequipa 04000, Peru; hbarazorda@ucsm.edu.pe (H.L.B.-C.);
    mchavezf@ucsm.edu.pe (M.Á.C.-F.)
[6] 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.)
[7] 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
[8] 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.); pablonsolis@gmail.com (P.N.S.)
[9] 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
[10] GIFES 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.)
[11] GIEPRONAL Research Group, School of Basic Sciences, Technology and Engineering, Universidad Nacional
    Abierta y a Distancia, Dosquebradas 661001, Colombia
*  Correspondence: medinajl@unam.mx; Tel.: +52-55-5622-3899

**Abstract:** The number of databases of natural products (NPs) has increased substantially. Latin America is extraordinarily rich in biodiversity, enabling the identification of novel NPs, which has encouraged both the development of databases and the implementation of those that are being created or are under development. In a collective effort from several Latin American countries, herein we introduce the first version of the Latin American Natural Products Database (LANaPDB), a public compound collection that gathers the chemical information of NPs contained in diverse databases from this geographical region. The current version of LANaPDB unifies the information from six countries and contains 12,959 chemical structures. The structural classification showed that the most abundant compounds are the terpenoids (63.2%), phenylpropanoids (18%) and alkaloids (11.8%). From the analysis of the distribution of properties of pharmaceutical interest, it was observed that many LANaPDB compounds satisfy some drug-like rules of thumb for physicochemical properties. The concept of the chemical multiverse was employed to generate multiple chemical spaces from two different fingerprints and two dimensionality reduction techniques. Comparing LANaPDB with FDA-approved drugs and the major open-access repository of NPs, COCONUT, it was concluded that the chemical space covered by LANaPDB completely overlaps with COCONUT and, in some

regions, with FDA-approved drugs. LANaPDB will be updated, adding more compounds from each database, plus the addition of databases from other Latin American countries.

## 1. Introduction

Historically, natural products (NPs) have been the biggest source of bioactive compounds for medicinal chemistry. For instance, in cancer research, in the lapse of time from 1946 to 1980, seventy-five small molecules were approved worldwide, of which 53% were unaltered NPs or natural product (NP) derivatives. Moreover, from 1981 to 2019, of the 185 small molecules approved to treat cancer, 64.9% were unaltered NPs and synthetic drugs with a NP pharmacophore [1]. Another example is the actual development of new promising antibiotics against drug-resistant bacteria from NPs [2]. Furthermore, in a recent review, it was shown that 697 natural steroidal alkaloids were isolated and characterized with various biological activities, from 1926 to 2021 [3]. The bioactive compounds encompass marine [4,5], fungal [6,7], bacteria [8], plants [9] and endogenous substances produced by human and animal sources [10], including venoms and poisons produced by different animals [11]. Even, as recently reviewed, the fruit peels are a source of bioactive compounds, which, in many instances, display better biological and pharmacological applications than the compounds of other sections of the fruit [12].

There are several approaches to the drug discovery process for NPs. The information of the therapeutic effects or even the side effects can serve as a starting point [10]. The stress-driven growth of plants and microorganisms stimulates the production of secondary metabolites with biological activities different to the primary metabolites produced under normal conditions [13]. Drug repositioning of NPs is another option, which offers lower development costs and shorter time frames [14]. Moreover, NPs are a rich source of "privileged scaffolds", structures capable of providing useful ligands for more than one receptor [15]. Privileged scaffolds are useful because they can serve as a core structure to construct compound libraries around them [1]. Some examples of privileged scaffolds that are actually used for drug design purposes are the alkaloid, terpenoid polyketide and phenylpropanoid structures [16]. Furthermore, the preparation of biologically relevant small-molecule libraries through unprecedented combinations of NP fragments to afford novel scaffolds that do not occur in nature is an approach that involves the preparation of molecules named "pseudo-natural products" (pseudo-NPs) [17]. The pseudo-NPs retain the biological relevance of NPs, yet exhibit structures and bioactivities not accessible to nature or through the use of existing design strategies. Pseudo-NPs may display unexpected bioactivities that differ from the activities of the NPs from which their fragments are derived. That is why their bioactivity should be monitored in a wide biological space through different biochemical and biological assays. Most of the pseudo-NP collections fall within the "Lipinski rule of 5" (Ro5) space, showing advantageous physicochemical "drug-like" properties. For the design of pseudo-NP libraries, it is important to consider that the combination of biosynthetically unrelated NP fragments may be beneficial for novel bioactivity, maximizing the biological relevance of the resulting pseudo-NP scaffold. Chromopynones, indotropanes, pyrrotropanes and pyrroquinolinones are part of pseudo-NP collections that have been developed for the first time, an unprecedented combination of these scaffolds, resulting in totally new chemical entities [18,19].

Over time, NPs have been a source of compounds with therapeutical effects and many of them, later, end up converting into approved drugs. Some of them have been approved as drugs without suffering a structural modification. In other cases, they serve as starting points that, later, with further structural modifications, are approved as drugs. Sometimes, bioactive molecules from NPs lack suitable physicochemical properties, and

their synthetic complexity may hinder their direct use as therapeutics. In this case, to be developed as drug candidates, NPs need to go through an optimization process that usually involves structural modifications to improve one or more of the following characteristics: potency, selectivity, solubility, metabolic and chemical stability; and the removal (or at least significant reduction) of toxicity [20]. This is usually done by decreasing the molecular size, eliminating the unnecessary functional groups and chiral centers and introducing nitrogen atoms if they are needed, because in the NPs the presence of nitrogen is limited.

To date, the discovery process of more than seventy commercialized drugs has included the rational use of at least a computational method [21]. Computer-aided drug design (CADD) has the potential to reduce the cost and decrease the time required for the drug design process, e.g., the hit identification rate for high-throughput screening (HTS) to discover novel inhibitors for the enzyme protein tyrosine phosphatase-1B is only 0.021% and the one for molecular docking is 34.8% [22]. Some crucial resources in CADD are the databases of chemical compounds, including NP databases. From the compound databases, it is possible to identify potential hit molecules through several virtual screening (VS) techniques [23,24], including the training of artificial intelligence (AI) algorithms [25]. When the compound databases are annotated with biological activity (or other property of relevance), it is possible to use the data to measure structure–activity (property) relationships and develop predictive models.

VS techniques are usually classified into two major categories: structure-based (SBVS) and ligand-based (LBVS). In general, SBVS is more suitable for finding structurally novel ligands, and is the preferred method when the three-dimensional (3D) structure of the target protein has been characterized experimentally [23]. When the structure of the target is unknown, or its prediction by structure-based methods is challenging, LBVS is the choice [24]. LBVS is based on the assumption that molecules with similar structures exhibit similar behavior. Among the LBVS techniques are the quantitative structure–activity relationship (QSAR) [23] and quantitative structure–property relationship (QSRP) [26] studies. QSAR/QSPR studies aim to find a mathematical association between the molecule structure and a given property, such as biological activity [24]. In this sense, the bioactivity and chemical information (i.e., chemogenomic) databases are crucial to allow the creation of QSAR/QSPR models that predict certain pharmacological activity or properties of pharmaceutical interest for a determined molecule or set of analog molecules.

Another important application of the databases in the drug discovery process is the training of AI algorithms. AI encompasses a set of computational algorithms that allow computers to simulate human cognitive abilities, such as learning from experience and solving problems [27]. Among the LBVS is the AI-based QSAR, the creation and training of these models relies on the data found in the bioactivity databases. AI can be applied to the SBVS; specifically, to the docking of the protein-ligand complexes [25]. AI-based scoring functions have shown better performance in benchmark studies [28,29]. The creation of AI-based scoring functions depends on the availability of the required data in the database to train the model. AI algorithms have already been applied in the drug discovery process from NPs. To name a few: data-mining into traditional medicines and peer-reviewed articles, prediction of chemical structures from microbial genomes, automation of NPs dereplication process, encoding NPs into molecular representations, vectorization of NPs with molecular descriptors, mapping of NPs in the chemical space, engineering likeness scores, deorphanization and de novo generation of natural product-inspired compounds [30]. The report on using AI to create models that allow for the prediction of biological effects from NPs has been rising in recent years. The application of AI models to predict biological effects of molecules, toxicity, drug–target and drug–drug interactions has been reviewed [31].

From 2003 to 2018, 104 research articles reported the identification of potential drug candidates from NP databases by using computational tools [32]. Moreover, during the current pandemic outbreak, NPs have been a rich source for discovering potential lead compounds against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [33–36].

Nonetheless, the computational methods, such as the VS techniques, are valuable tools and should be implemented together with the in vivo or in vitro assays to increase the success rate in the identification of bioactive molecules. Recently (August 2022), the outperformance of combining the computational methods with the biological assays was shown [37]. There are still many challenges in the identification of bioactive compounds from the screening of large collections of compounds. In the HTS approach, a common problem is the presence of frequent hitters: compounds that would form aggregates, react with proteins or interfere in screening assays, leading to false positives [38]. Combining the HTS approach with computational methods can lead to the identification and elimination of the frequent hitters prior to the start of the HTS study.

Between 2000 and 2019, one hundred and twenty-three commercial and public NP databases have been published. Among them, ninety-eight are still somehow accessible (online or under request access), ninety-two are free access and only fifty contain molecular structures that can be retrieved for a chemoinformatic analysis [39]. Examples of the most representative open-access NP databases include: The Collection of Open Natural Products (COCONUT) [40], which is a major repository containing more than 411,000 NPs collected from 50 open access NP databases. The Universal Natural Product Database [41] is a compilation that tries to gather all the known NPs—it has more than 229,000 NPs. It is not yet accessible through the link in the original publication, nevertheless, it is contained and maintained on the ISDB website [42]. The SuperNatural 3.0 [43] database contains over 449,048 NPs—it provides a bulk download and has information on pathways, mechanism of action, toxicity and vendor information if available. The ZINC [44] database has over 80,000 NPs, approximately 48,000 purchasable. Moreover, it contains some NP databases that are no longer accessible through the link provided in the original publication, e.g., the Herbal Ingredient Targets [45] and the Herbal Ingredients In Vivo Metabolism database [46], which mostly contain NPs from Chinese plants. A practical application of the public molecular databases that exemplifies their utility in the drug discovery area is the recent identification of inhibitors of the human immunodeficiency virus-1 from 1.6 million commercially available drug-like compounds from the ZINC database. Moreover, there are NP databases that contain compounds isolated and characterized in certain geographical areas. That is the case of China, where multiple compound databases containing only NPs from this country have been published [47–53], of which, TCM@Taiwan [54] is the largest, containing 58,000 compounds. There are two databases of NPs from India, IMPPAT [55], composed of approximately 10,000 phytochemicals extracted from 1700 medicinal plants, and MedPServer [56], containing 1124 NPs. Regarding NPs from Africa, there are several NP databases [57–62], of which, AfroDB [63] is the most extensive, containing over one thousand NPs. Recently, Phyto4Health was published [64], an NP database with 3128 NPs isolated from medicinal plants of Russia.

Latin America contains at least a third of the global biodiversity [65]; in fact, half of the countries have been classified as megadiverse: Bolivia, Brazil, Colombia, Costa Rica, Ecuador, Mexico, Peru and Venezuela) [66]. Therefore, Latin America represents a large source of bioactive molecules and potential drug candidates (Figure 1). There have been published databases containing NPs from some Latin American countries, such as NaturAr [67] (Argentina), NuBBE$_{DB}$ [68,69], SistematX [70,71], UEFS [72] (Brazil), CIF-PMA [73,74] (Panama), PeruNPDB [75], (Peru), UNIIQUIM [76] and BIOFACQUIM [77,78] (Mexico). Recently, the present state of the art in developing Latin American NP databases and their practical applications to the drug discovery area were reviewed [79]. Multiple drug candidates have been identified from the Latin American NP databases as therapeutic agents for diseases caused by infectious agents (Chagas disease [80,81], tuberculosis [82], Leishmaniasis [83,84], schistosomiasis [85], coronavirus disease [86], human immunodeficiency virus infection and acquired immunodeficiency syndrome, hepatitis B and C) [87], pain [88], obesity, diabetes, hyperlipoproteinemia, cancer and age-related diseases [89,90].
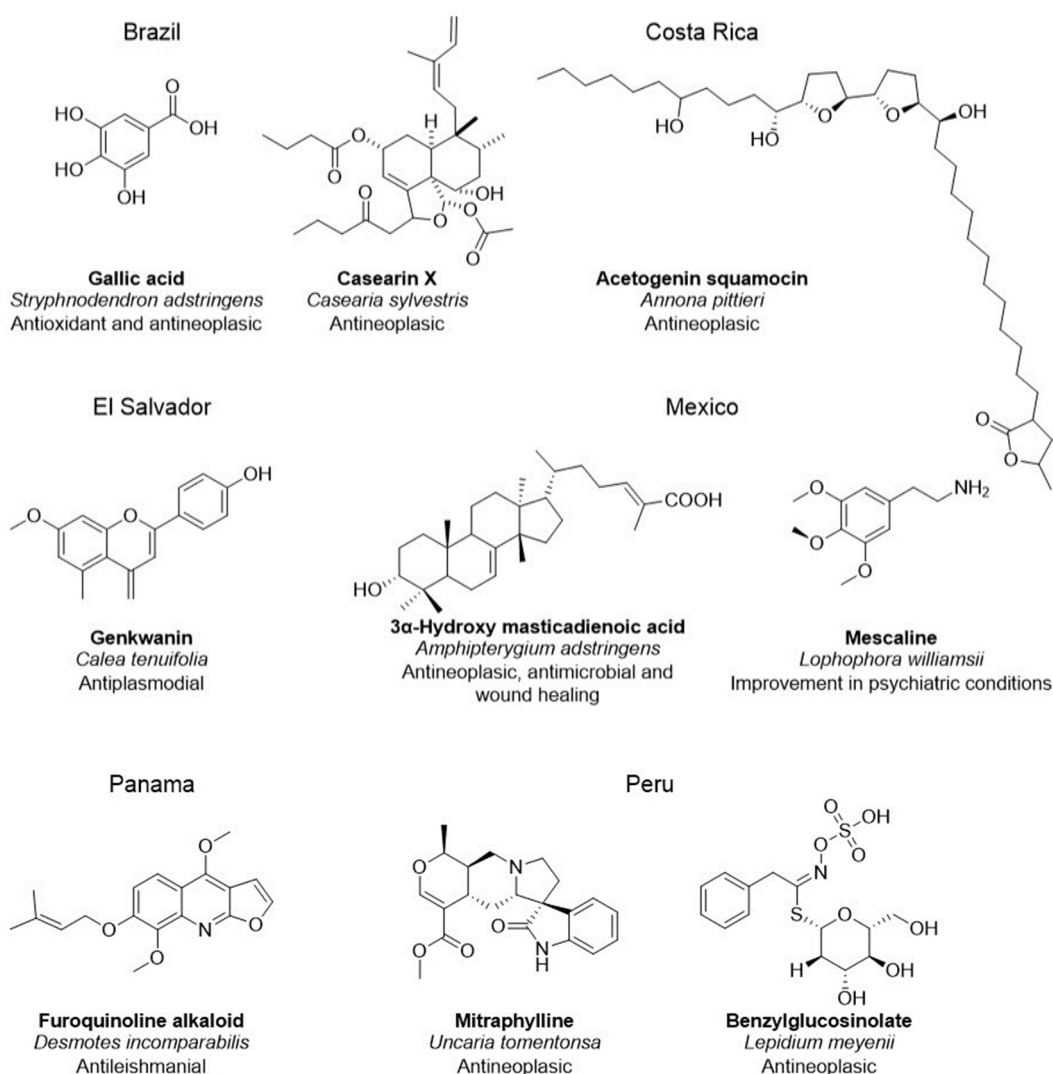
**Figure 1.** Active compounds of representative medicinal plants from Latin American countries and some of their therapeutic effects described in the literature. Brazil (gallic acid [91] and casearin x [92]), Costa Rica (acetogenin squamocin [93]), El Salvador (Genkwanin [94]), Mexico (3α-hydroxy masticadienoic acid [95] and mescaline [96]), Panama (furoquinoline alkaloid [97]) and Peru (mitraphylline [98] and benzylglucosinolate [99]).

The long-term goal of the project is to collect, unify and standardize the Latin American NP collections available in the public domain into one public database. In this study, we report significant advances towards this goal through the assembly of the first version of the unified database, herein called the Latin American Natural Products Database (LANaPD). We report its curation, standardization and a comprehensive analysis of nine compound databases, totaling 12,959 unique molecules. As part of this study, we analyzed the structural content and determined some physicochemical properties of pharmaceutical interest of the compounds in LANaPDB. We also represent coverage in the chemical space of compounds in LANaPDB using the concept of a chemical multiverse [100]. The database is freely available at https://github.com/alexgoga21/LaNaPDB (accessed on 28 September 2023).

## 2. Results and Discussion

### 2.1. Bioactive Compounds from Latin American Natural Product Databases

Bioactive compounds have been identified from Latin American NP databases. Ten compounds against *Trypanosoma cruzi* [80] have been identified from the NuBBE_DB database.

Moreover, 13 compounds against *Mycobacterium tuberculosis* were identified in another study from NuBBE_DB [82].

Several bioactive compounds have been identified from five VS studies from the SistematX database. The bioactive compounds found include 1306 sesquiterpene lactones with potential activity against Trypanosoma cruzi [81]. In another VS study, 13 promising antileishmanial compounds were identified [83]. In the third VS study, the researchers looked for compounds against Schistosoma mansoni; from this, five compounds were identified with potential schistosomicidal activity [85]. In the fourth VS study, 19 compounds were identified as potential SARS-CoV-2 inhibitors [86]. In the last VS study, two promising compounds were identified for the treatment of Alzheimer's disease [90].

The compounds of CIFPMA have been tested in over 25 in vitro and in vivo bioassays for different therapeutic targets, including anti-HIV (human immunodeficiency virus), antioxidants and anticancer [73].

In the UNIIQUIM database, molecules were found with potential analgesic activity [88].

In the BIOFACQUIM database, eight beta-glucosidase inhibitors were identified. The pharmacological applications of these compounds include obesity, diabetes, hyperlipoproteinemia, cancer, HIV and hepatitis B and C [87]. In another study, three compounds were identified to prevent and improve multiple adverse outcomes related to age [89]. The identified compounds from Latin American NP databases are in the Table 1.

**Table 1.** Bioactive compounds found from the Latin American Natural Products Databases.

| Database Name | Disease or Symptom | Number of Identified Compounds | Reference |
|---|---|---|---|
| NuBBE_DB | Chagas disease | 10 | [80] |
| | Tuberculosis | 13 | [82] |
| SistematX | Chagas disease | 13 | [81] |
| | Leishmaniasis | 13 | [83] |
| | Schistosomiasis | 5 | [85] |
| | Coronavirus disease 2019 | 19 | [86] |
| | Alzheimer's disease | 2 | [90] |
| UNIIQUIM | Pain | 6 | [88] |
| BIOFACQUIM | Obesity | 8 | [87] |
| | Diabetes | | |
| | Hyperlipoproteinemia Cancer | | |
| | HIV/AIDS * | | |
| | Hepatitis B and C. | | |
| | Age-related diseases | 3 | [89] |

* Human immunodeficiency virus infection and acquired immunodeficiency syndrome (HIV/AIDS).

### 2.2. Dataset Curation

From nine Latin American NP databases of six different countries (Table 2), the first version of LANaPDB, which currently contains 12959 compounds in total, was constructed. The number of unique and overlapping compounds is shown in the Figure 2. The number of unique compounds is proportionally similar to the number of compounds contained in the databases that comprise every country.

**Table 2.** Latin American Natural Product Databases analyzed in this work.

| Database Name (Country) | Number of Compounds [a] | 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. | [68,69] |
| SistematX (Brazil) | 9514 | Plants | Database composed of secondary metabolites and developed at the Federal University of Paraiba. | [70,71] |
| 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. | [72] |
| NAPRORE-CR (Costa Rica) | 359 | 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. | [76] |
| BIOFACQUIM (Mexico) | 553 | 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. | [77,78] |
| 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. | [73,74] |
| PeruNPDB (Peru) | 280 | Animals Plants | Created and curated at the Catholic University of Santa Maria. | [75] |

The URL of the websites where the natural product databases of Latin America are allocated is in the Supplementary Material (Table S1). [a] Number of compounds contained in each database previous to the curation process. * Actually, there is not a publication associated with the database.
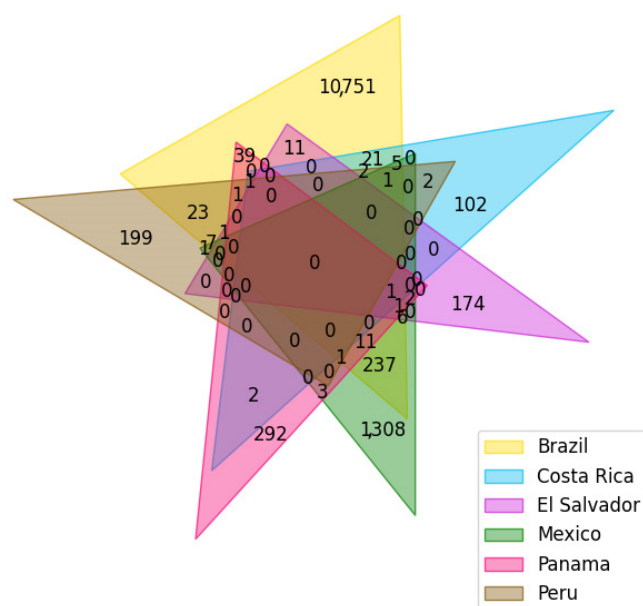


**Figure 2.** Venn diagram showing the number of unique and overlapping compounds contained in the nine Latin American databases.

## 2.3. Structural Classification

The compounds were classified in a total of seven different pathways, fifty-three superclasses and three hundred and thirty-six classes (Figure 3). The three predominant pathways are terpenoids (63.2%), shikimates and phenylpropanoids (18%) and alkaloids (11.8%). The main superclasses are diterpenoids (34.3%), sesquiterpenoids (17.6%) and flavonoids (10.3%). The prevalent classes are kaurane and phyllocladane diterpenoids (6.99%), colensane and clerodane diterpenoids (5.91%) and germacrane sesquiterpenoids (5.36%). The results are in accordance with expectations because the terpenoids are the most diverse group of secondary metabolites derived from natural sources [101].
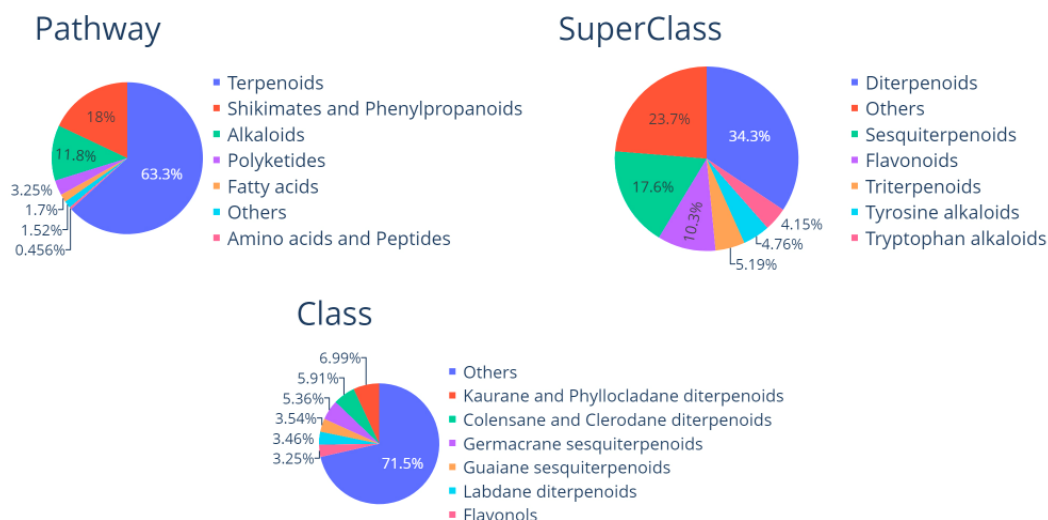


**Figure 3.** Structural classification of the compounds in the current (first) version of LANaPDB.

## 2.4. Physicochemical Properties

The violin plots show the distribution of six physicochemical properties of pharmaceutical interest: SlogP [102], molecular weight (MW), topological polar surface area (TPSA) [103], rotatable bonds (Rb), hydrogen bond acceptors (HBA) and hydrogen bond donors (HBD) (Figures 4 and 5). In the violin plots, the limits of the following rules of thumb of drug-likeness are marked with a horizontal line: Lipinski's rule of 5 (Ro5) [104,105], Veber's rules [106], GlaxoSmithKline's (GSK) 4/400 rule [107] and Pfizer 3/75 rule [108] (Table S2). Physicochemical properties in the limits of either Lipinski's, Veber's or GSK rules is usually related with a 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).

NPs contain complex structures and are large and diverse; therefore, compared with synthetic drugs, it is not easy for them to satisfy most of the criteria of Lipinski's Ro5 [109] or the other drug-likeness parameters mentioned above. Nevertheless, it is shown in the violin plots that a broad range of the LANaPDB compounds satisfy most of the rules of thumb of Table S2 for the physicochemical properties of pharmaceutical interest. The LANaPDB and COCONUT compound distributions of the physicochemical properties are, in general, similar (Figure 4). Additionally, as expected, COCONUT covers the broadest area of the chemical space, because it is the largest database (411,000 compounds) (Figure 6). Many compounds of LANaPDB fulfill the rules of thumb associated with drug-likeness (Figure 4) and part of the LANaPDB chemical space overlaps with the chemical space comprised by the approved drugs (Figure 6).
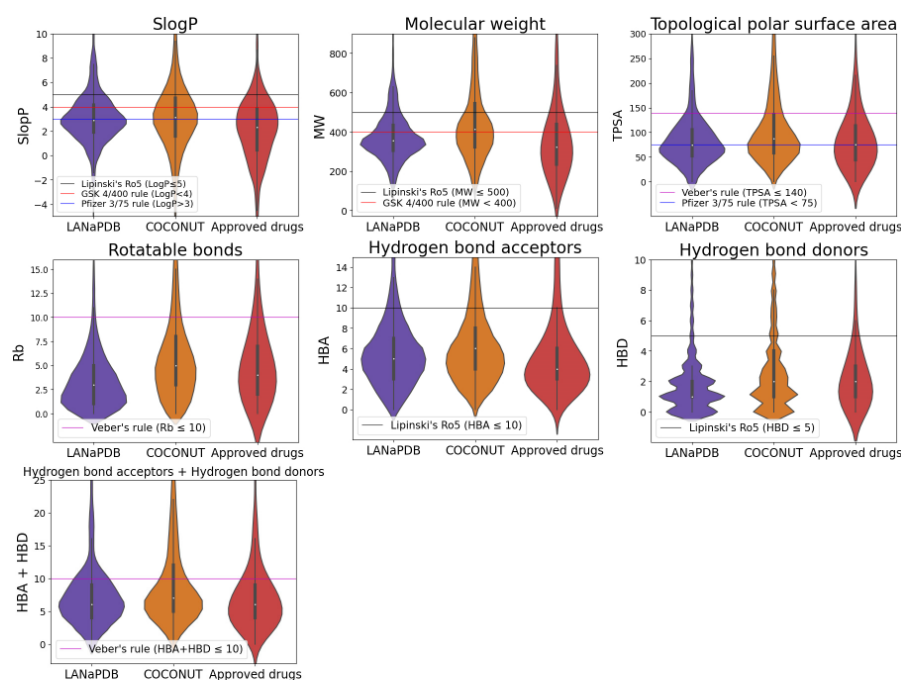
**Figure 4.** Violin plots summarizing the distribution of the representing physicochemical properties of pharmaceutical interest of the compounds of three databases LANaPDB, COCONUT and FDA-approved small molecule drugs.
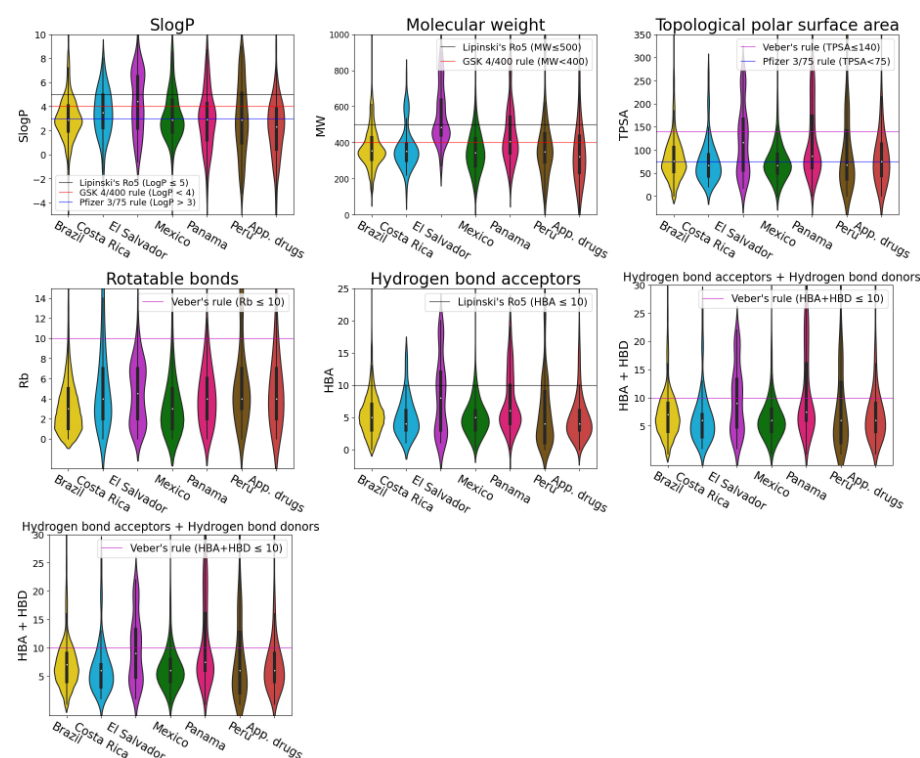


**Figure 5.** Violin plots summarizing the distribution of the physicochemical properties of pharmaceutical interest of the compounds in LANaPDB and FDA-approved small molecule drugs (App. drugs). The databases that encompass LANaPDB for every country: Brazil (NuBBE_DB, SistematX and UEFS), Costa Rica (NAPRORE-CR), El Salvador (LAIPNUDELSAV), Mexico (UNIIQUIM and BIOFACQUIM), Panama (CIFPMA) and Peru (PeruNPDB).
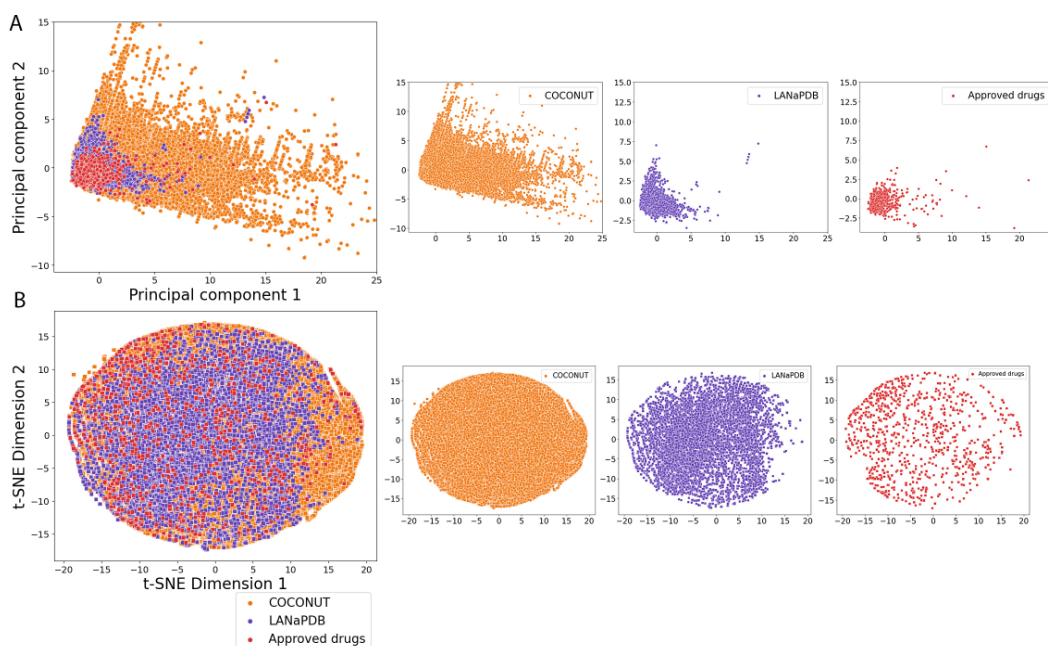
**Figure 6.** Visual representation of the chemical space based on six physicochemical properties of pharmaceutical interest of LANaPDB and its comparison with COCONUT and approved drugs. The chemical space was generated with (**A**) principal component analysis (PCA), the first two principal components capture 89.3% of the total variance; (**B**) t-distributed stochastic neighbor embedding (t-SNE).

The distribution of the physicochemical properties of the NPs in the countries with more compounds (Brazil and Mexico) is, in general, more focused in certain regions, compared with the NPs from countries with less compounds (Costa Rica, El Salvador, Panama and Peru) which are more broadly distributed (e.g., SlogP, Brazil vs Peru from Figure 5). Panama and Peru show similar distributions in all the physicochemical properties, which may be due to the similar distribution of the dominant structural features in both datasets (Panama: 42% shikimates and phenylpropanoids and 34.3% terpenoids; Peru: 29.6% shikimates and phenylpropanoids and 39.2% terpenoids). In general, the distributions of physicochemical properties of the Latin American countries and the approved drugs are focused in the same regions (Figure 5). The chemical space represented by the six physicochemical properties is overlapped among the NPs from the six Latin American countries (Figure 7). In the principal component analysis (PCA), the first two principal components are enough to represent most of the explained variance percentage: 89.3% in the LANaPDB, COCONUT and approved drugs comparison (Figure 6A) and 84.6% in the Latin American countries comparison (Figure 7A). Moreover, TPSA, MW, HBD and HBA are the descriptors with greater contributions to principal component 1. The descriptors with greater contributions to principal component 2 are SlogP and Rb (Table S3).
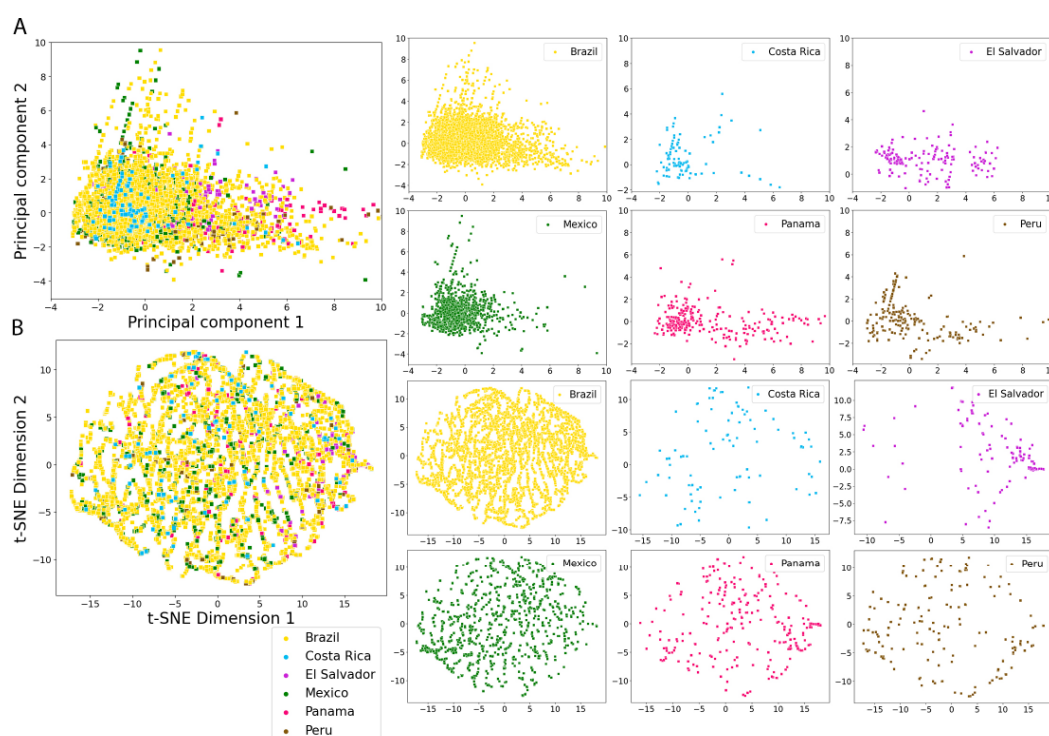
**Figure 7.** Visual representation of the chemical space based on six physicochemical properties of pharmaceutical interest of LANaPDB and individual Latin American natural product databases. The chemical space was generated with (**A**) principal component analysis (PCA), with the first two principal components capturing 84.6% of the total variance; (**B**) t-distributed stochastic neighbor embedding (t-SNE).

### 2.5. Molecular Fingerprints

Figures 8 and 9 show the visual representation of the chemical multiverse of LANaPDB generated with t-distributed stochastic neighbor embedding (t-SNE) and tree MAP (TMAP) [110] and two fingerprints of different designs: MACCS keys (166-bits) (Figures 8A and 9A) and MAP4 (Figures 8B and 9B). As discussed recently, the chemical multiverse can be defined as a group of chemical spaces, each generated with a diverse set of descriptors [100]. A chemical multiverse is a natural extension of the concept of chemical space and its advantage is that it provides a more complete description of the chemical space of a set of compounds as opposed to using only one representation. Moreover, substructure fingerprints perform best for small molecules, such as drugs, while atom–pair fingerprints are preferable for large molecules. Given that it is common to see large molecules among the NPs, MAP4 fingerprint was chosen because it is suitable for both small and large molecules by combining substructure and atom–pair concepts [111]. The MACCS keys (166-bits) [112] fingerprint was employed to compare the results obtained with the MAP4 fingerprint with a well-known substructure fingerprint.
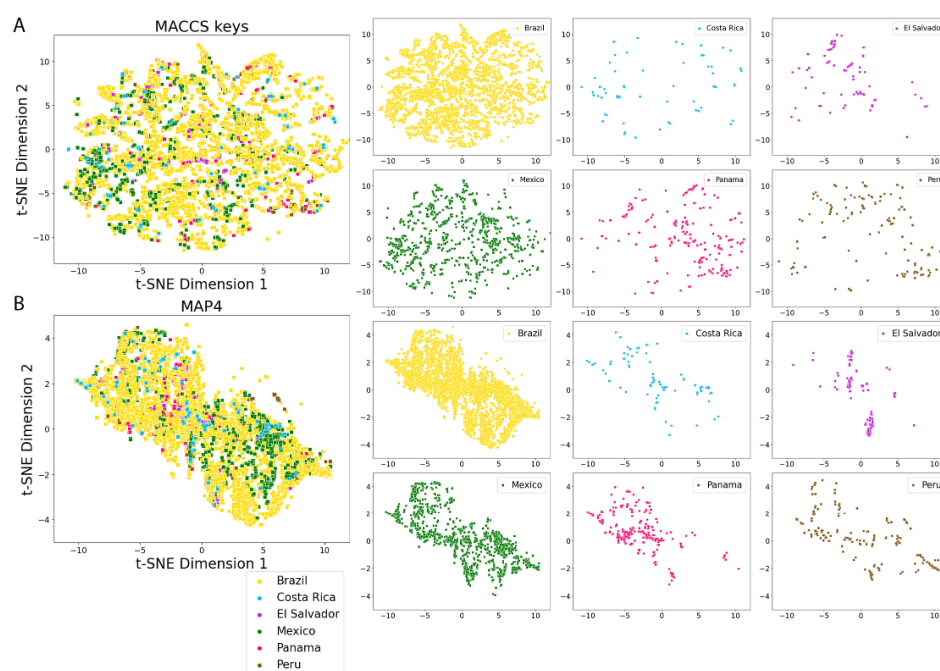
**Figure 8.** Visual representation of the chemical multiverse of LANaPDB and individual Latin American natural product databases. The chemical multiverse is a group of chemical spaces, each generated with a different set of descriptors. Chemical space comprised by (**A**) (t-SNE)-MACCS keys (166-bits) fingerprint; (**B**) (t-SNE)-MAP4 fingerprint.
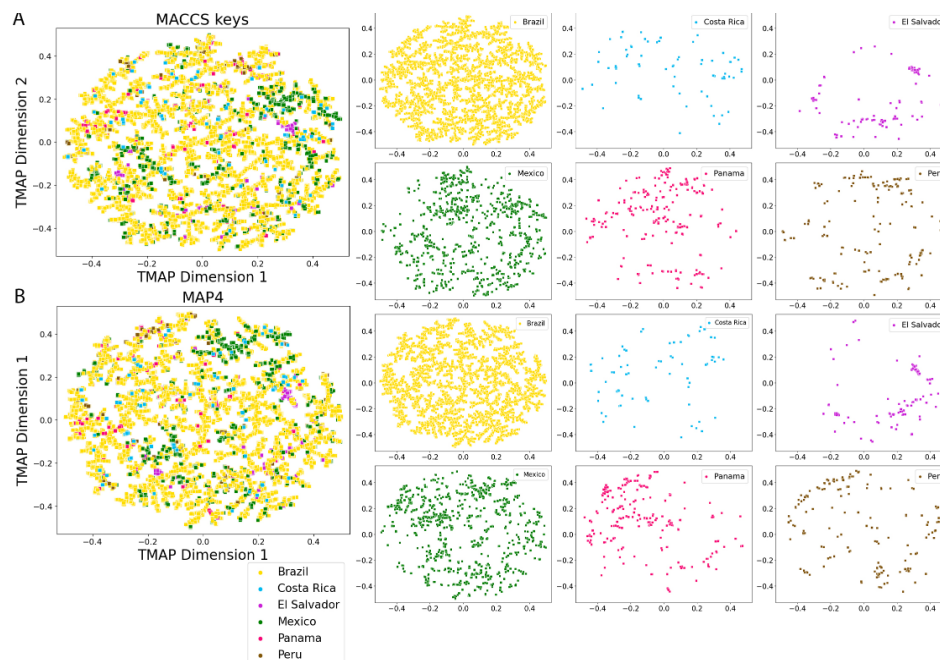


**Figure 9.** Visual representation of the chemical multiverse of LANaPDB and individual Latin American natural product databases. The chemical multiverse is a group of chemical spaces, each generated with a different set of descriptors. Chemical space comprised by (**A**) (TMAP)-MACCS keys (166-bits) fingerprint; (**B**) (TMAP)-MAP4 fingerprint (interactive version: https://github.com/alexgoga21/LaNaPDB/blob/main/Interactive%20TMAP_MAP4.html) (accessed on 28 September 2023).

Based on the visual representation of the chemical multiverse, it is concluded that t-SNE has a better performance with MACCS keys (166-bits) fingerprint over MAP4 fingerprint, separating the NPs on clusters according to the structural features (Figure 8). The efficacy of TMAP to separate compounds in clusters from MACCS keys (166-bits) and MAP4 fingerprints is similar with both fingerprints (Figure 9). Moreover, TMAP performed better than t-SNE in the NPs cluster creation with both fingerprints. An interactive version of the scatter plot created with TMAP from MAP4 fingerprints (Figure 9B) is freely available at https://github.com/alexgoga21/LaNaPDB/blob/main/Interactive%20TMAP_MAP4 .html (accessed on 28 September 2023). To open the interactive map, download the file and open it in a web explorer. Given that TMAP performed better than t-SNE, and MACCS keys (166-bits) and MAP4 fingerprints showed a similar efficacy in the TMAP, the comparison of LANaPDB with the reference databases was made with TMAP and MACCS keys (166-bits) fingerprint (Figure 10). It can be observed that LANaPDB overlaps with COCONUT in well-defined areas; nevertheless, the approved drugs are more dispersed and some of them overlap with the compounds in LANaPDB (Figure 10).
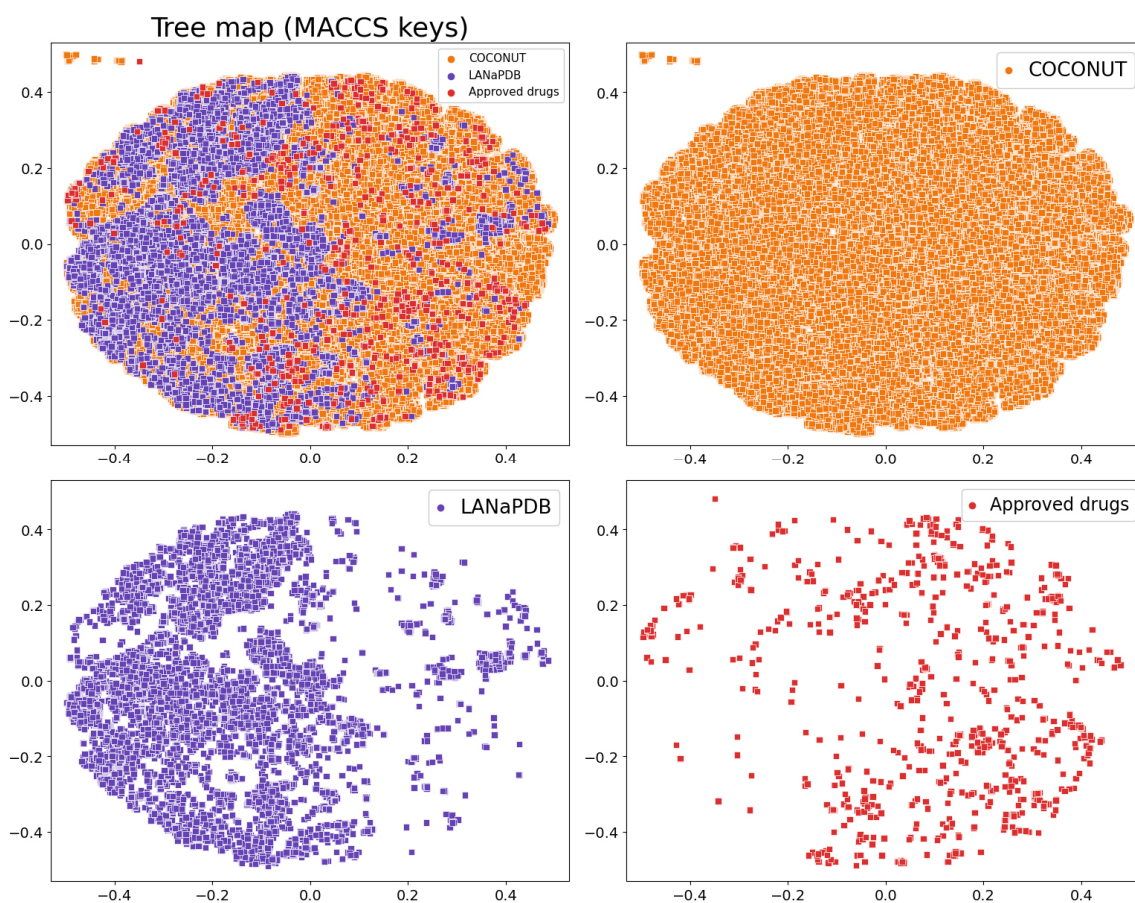


**Figure 10.** Tree map from MACCS keys (166-bits) of LANaPDB and the comparison with COCONUT and approved drugs.

## 3. Materials and Methods

The visual representations of the different chemical spaces that consider either physico-chemical properties or molecular fingerprints are illustrated with scatter plots (Figures 6–10). Every point in the scatter plots represents a unique compound. The scatter plots were created in the python programming language (version 3.10.7), employing the seaborn module (0.12.2) [113].

### 3.1. Dataset Curation

The Latin American NP databases of Table 2 were used to construct the unified NP database LANaPDB. The process was carried out in the python programming language (version 3.10.7), employing the RDKit (version 2022.03.5) [114] and MolVS (version 0.1.1) [115] modules. The standardization process of MolVS was applied, which consisted of the remotion of explicit hydrogens, disconnection of covalent bonds between metals and organic atoms (the disconnected metal is not preserved), application of normalization rules (transformations to correct common drawing errors and standardization of functional groups), reionization (ensure the strongest acid groups protonate first in partially ionized molecules) and recalculation of the stereochemistry (ensures preservation of the original stereochemistry). The salts were removed, keeping the largest fragment, which was neutralized, and the remaining partially ionized fragments were reionized. The canonical tautomer was determined, and, from the InChIKey strings of the canonical tautomer, the duplicate compounds were removed.

A Venn diagram was constructed with python programming language, employing the Venn module (version 0.1.3), to see the number of unique and overlapping compounds along the nine Latin American databases.

### 3.2. Structural Classification

Compounds in LANaPDB were classified with NPClassifier [116], which is a freely available deep neural network-based structural classification tool for NPs. NPClassifier establishes a classification system based on the literature from the specialized metabolism of plants, marine organisms, fungi and microorganisms. The categories used in NPClassifier were defined at three hierarchical levels: pathway (nature of the biosynthetic pathway), superclass (chemical properties or chemotaxonomic information) and class (structural details). Pie charts were constructed with the python programming language (version 3.10.7), employing the Plotly express module (version 0.4.1) [117], to represent the distribution of the dominant structures for each of the three categories.

### 3.3. Physicochemical Properties

Employing the software KNIME [118] version 4.7.1, with the RDKit nodes, six physicochemical properties of pharmaceutical interest were calculated: SlogP [102], MW, TPSA [103], Rb, HBA, HBD. Violin plots were constructed to summarize the distribution of each property individually. In each violin plot, we highlighted the limit of drug-like rules of thumb (Table S2). To generate a visual representation of the chemical space of the compound libraries based on the six properties, we reduced the data dimensionality to two dimensions, employing PCA and t-SNE with the python module Scikit-learn version 1.2.2 [119]. PCA: principal component one and principal component two to represent the six physicochemical properties. t-SNE hyperparameters: perplexity = 40 and number of iterations = 300. The distribution of the individual properties and the two-dimensional representation of the chemical space were conducted to analyze and compare the properties of the NPs among the six Latin American countries and with two other reference datasets, COCONUT [40] and FDA-approved small-molecule drugs, version 5.1.10 (released by DrugBank in January 2023) [120].

### 3.4. Molecular Fingerprints

A fingerprint encodes the structural information of a molecule in a vector [121]. Two different fingerprints were determined for each molecule: MACCS keys (166-bits) fingerprint and MAP4 fingerprint. MACCS keys (166-bits) fingerprints were calculated with KNIME [118] version 4.7.1, employing the Chemistry Development Kit (CDK) nodes [122]. MAP4 fingerprints were determined with the python programming language, following the instructions of the creators of this fingerprint [111]. To allow a 2D representation of the molecules, two different techniques for dimensionality reduction were employed: t-SNE and TMAP [110]. For t-SNE, the same hyperparameters of Section 3.3. were used. Em-

ploying the TMAP with MACCS keys (166-bits), LANaPDB was compared with the two reference datasets used in Section 3.3. From the MAP4 fingerprints, an interactive TMAP was created in the python programming language (version 3.9.17) with the faerun module (version 0.4.2).

## 4. Conclusions

Here we report progress towards the assembly of the first version of a unified Latin American Natural Products Database. The current version has 12,959 compounds from nine compound databases of six different Latin American countries. The database is freely available and the information of each compound in this first version includes the structures in SMILES format, the structural classification and six physicochemical properties of pharmaceutical interest. The LANaPDB compounds are produced by plants, terrestrial and marine animals, fungi and bacteria. Moreover, the most abundant NPs were the terpenoids (63.2%), followed by the shikimates and phenylpropanoids (18%) and the alkaloids (11.8%). Although it is not easy for NPs to fulfill most of the drug-likeness parameters compared with synthetic drugs, many LANaPDB compounds satisfy some drug-like rules of thumb for physicochemical properties. Moreover, the chemical space covered by LANaPDB completely overlaps with COCONUT and, in some regions, with the FDA-approved drugs. The concept of the chemical multiverse was used to generate multiple chemical spaces from two different dimensionality reduction techniques (t-SNE and TMAP) and two fingerprints (MACCS keys (166-bits) and MAP4). MAP4 performed better than t-SNE in separating the compounds into clusters according to their structural features. All the resources used for the assembly, curation, analysis and graphics creation are freely available.

LANaPDB is part of one of the strategic actions to contribute to the further development of chemoinformatics and related disciplines in Latin America and strengthen the interactions between Latin America and other geographical regions [123]. We encourage the community to visit the websites where the individual NP databases of the different Latin American countries are available (Table S1).

We anticipate that LANaPDB will continue growing and evolving with the update of more compounds from each existing database, along with the addition of databases from other Latin American countries. One of the first steps in this direction is the integration of a larger set of NAPRORE-CR and the incorporation of natural products database NPDB-EjeCol from Colombia. Another perspective is the implementation of the database in a free web server. Likewise, LANaPDB could be integrated with other large public databases of natural products, such as COCONUT or LOTUS.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ph16101388/s1, Table S1: Websites of the natural product databases of Latin America; Table S2: Rules of thumb guides associated with drug-likeness; Table S3: Analysis metrics of the principal component analysis.

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

**Funding:** We thank the innovation space UNAM-HUAWEI the computational resources to use their supercomputer under project No. 7 "Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer". We also thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grants #2019/05967-3 (Scholarship MV), #2020/11967-3 (DFG/FAPESP), under the project DINOBBIO (DFG Project #459288952)

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data are contained within the article, Supplementary Material and the following github repository https://github.com/alexgoga21/LaNaPDB (accessed on 28 September 2023).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1.  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. [CrossRef] [PubMed]
2.  Porras-Alcalá, C.; Moya-Utrera, F.; García-Castro, M.; Sánchez-Ruiz, A.; López-Romero, J.M.; Pino-González, M.S.; Díaz-Morilla, A.; Kitamura, S.; Wolan, D.W.; Prados, J.; et al. The development of the bengamides as new antibiotics against drug-resistant bacteria. *Mar. Drugs* **2022**, *20*, 373. [CrossRef] [PubMed]
3.  Xiang, M.-L.; Hu, B.-Y.; Qi, Z.-H.; Wang, X.-N.; Xie, T.-Z.; Wang, Z.-J.; Ma, D.-Y.; Zeng, Q.; Luo, X.-D. Chemistry and bioactivities of natural steroidal alkaloids. *Nat. Prod. Bioprospect.* **2022**, *12*, 23. [CrossRef] [PubMed]
4.  Li, X.-W. Chemical ecology-driven discovery of bioactive marine natural products as potential drug leads. *Chin. J. Nat. Med.* **2020**, *18*, 837–838. [CrossRef]
5.  Banerjee, P.; Mandhare, A.; Bagalkote, V. Marine natural products as source of new drugs: An updated patent review (July 2018–July 2021). *Expert Opin. Ther. Pat.* **2022**, *32*, 317–363. [CrossRef]
6.  Singh, A.; Singh, D.K.; Kharwar, R.N.; White, J.F.; Gond, S.K. Fungal endophytes as efficient sources of plant-derived bioactive compounds and their prospective applications in natural product drug discovery: Insights, avenues, and challenges. *Microorganisms* **2021**, *9*, 197. [CrossRef] [PubMed]
7.  Tiwari, P.; Bae, H. Endophytic fungi: Key insights, emerging prospects, and challenges in natural product drug discovery. *Microorganisms* **2022**, *10*, 360. [CrossRef]
8.  Foxfire, A.; Buhrow, A.R.; Orugunty, R.S.; Smith, L. Drug discovery through the isolation of natural products from Burkholderia. *Expert Opin. Drug Discov.* **2021**, *16*, 807–822. [CrossRef]
9.  Porras, G.; Chassagne, F.; Lyles, J.T.; Marquez, L.; Dettweiler, M.; Salam, A.M.; Samarakoon, T.; Shabih, S.; Farrokhi, D.R.; Quave, C.L. Ethnobotany and the role of plant natural products in antibiotic drug discovery. *Chem. Rev.* **2021**, *121*, 3495–3560. [CrossRef]
10. Zhang, L.; Song, J.; Kong, L.; Yuan, T.; Li, W.; Zhang, W.; Hou, B.; Lu, Y.; Du, G. The strategies and techniques of drug discovery from natural products. *Pharmacol. Ther.* **2020**, *216*, 107686. [CrossRef]
11. Bordon, K.; de, C.F.; Cologna, C.T.; Fornari-Baldo, E.C.; Pinheiro-Júnior, E.L.; Cerni, F.A.; Amorim, F.G.; Anjolette, F.A.P.; Cordeiro, F.A.; Wiezel, G.A.; et al. From animal poisons and venoms to medicines: Achievements, challenges and perspectives in drug discovery. *Front. Pharmacol.* **2020**, *11*, 1132. [CrossRef] [PubMed]
12. Hussain, H.; Mamadalieva, N.Z.; Hussain, A.; Hassan, U.; Rabnawaz, A.; Ahmed, I.; Green, I.R. Fruit peels: Food waste as a valuable source of bioactive natural products for drug discovery. *Curr. Issues Mol. Biol.* **2022**, *44*, 1960–1994. [CrossRef] [PubMed]
13. Shams, U.L.; Hassan, S.; Jin, H.-Z.; Abu-Izneid, T.; Rauf, A.; Ishaq, M.; Suleria, H.A.R. Stress-driven discovery in the natural products: A gateway towards new drugs. *Biomed. Pharmacother.* **2019**, *109*, 459–467. [CrossRef] [PubMed]
14. Huang, B.; Zhang, Y. Teaching an old dog new tricks: Drug discovery by repositioning natural products and their derivatives. *Drug Discov. Today* **2022**, *27*, 1936–1944. [CrossRef]
15. Evans, B.E.; Rittle, K.E.; Bock, M.G.; DiPardo, R.M.; Freidinger, R.M.; Whitter, W.L.; Lundell, G.F.; Veber, D.F.; Anderson, P.S.; Chang, R.S. Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246. [CrossRef]
16. Davison, E.K.; Brimble, M.A. Natural product derived privileged scaffolds in drug discovery. *Curr. Opin. Chem. Biol.* **2019**, *52*, 1–8. [CrossRef]

17. Karageorgis, G.; Foley, D.J.; Laraia, L.; Brakmann, S.; Waldmann, H. Pseudo natural products-chemical evolution of natural product structure. *Angew. Chem. Int. Ed.* **2021**, *60*, 15705–15723. [CrossRef]

18. Karageorgis, G.; Foley, D.J.; Laraia, L.; Waldmann, H. Principle and design of pseudo-natural products. *Nat. Chem.* **2020**, *12*, 227–235. [CrossRef]

19. Cremosnik, G.S.; Liu, J.; Waldmann, H. Guided by evolution: From biology oriented synthesis to pseudo natural products. *Nat. Prod. Rep.* **2020**, *37*, 1497–1510. [CrossRef]

20. Guo, Z. The modification of natural products for medical use. *Acta Pharm. Sin. B* **2017**, *7*, 119–136. [CrossRef]

21. Sabe, V.T.; Ntombela, T.; Jhamba, L.A.; Maguire, G.E.M.; Govender, T.; Naicker, T.; Kruger, H.G. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **2021**, *224*, 113705. [CrossRef]

22. Doman, T.N.; McGovern, S.L.; Witherbee, B.J.; Kasten, T.P.; Kurumbail, R.; Stallings, W.C.; Connolly, D.T.; Shoichet, B.K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213–2221. [CrossRef] [PubMed]

23. Kar, S.; Roy, K. How far can virtual screening take us in drug discovery? *Expert Opin. Drug Discov.* **2013**, *8*, 245–261. [CrossRef] [PubMed]

24. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334–395. [CrossRef]

25. Vijayan, R.S.K.; Kihlberg, J.; Cross, J.B.; Poongavanam, V. Enhancing preclinical drug discovery with artificial intelligence. *Drug Discov. Today* **2022**, *27*, 967–984. [CrossRef]

26. Grover, I.; Singh, I.; Bakshi, I. Quantitative structure-property relationships in pharmaceutical research-Part 1. *Pharm. Sci. Technol. Today* **2000**, *3*, 28–35. [CrossRef] [PubMed]

27. Cavasotto, C.N.; Di Filippo, J.I. Artificial intelligence in the early stages of drug discovery. *Arch. Biochem. Biophys.* **2021**, *698*, 108730. [CrossRef]

28. Shen, C.; Hu, Y.; Wang, Z.; Zhang, X.; Zhong, H.; Wang, G.; Yao, X.; Xu, L.; Cao, D.; Hou, T. Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions. *Brief. Bioinform.* **2021**, *22*, 497–514. [CrossRef]

29. Ain, Q.U.; Aleksandrova, A.; Roessler, F.D.; Ballester, P.J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2015**, *5*, 405–424. [CrossRef]

30. Saldívar-González, F.I.; Aldas-Bulos, V.D.; Medina-Franco, J.L.; Plisson, F. Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* **2022**, *13*, 1526–1546. [CrossRef]

31. Jeon, J.; Kang, S.; Kim, H.U. Predicting biochemical and physiological effects of natural products from molecular structures using machine learning. *Nat. Prod. Rep.* **2021**, *38*, 1954–1966. [CrossRef] [PubMed]

32. de Sousa Luis, J.A.; Barros, R.P.C.; de Sousa, N.F.; Muratov, E.; Scotti, L.; Scotti, M.T. Virtual screening of natural products database. *Mini Rev. Med. Chem.* **2020**. [CrossRef]

33. Gangadevi, S.; Badavath, V.N.; Thakur, A.; Yin, N.; De Jonghe, S.; Acevedo, O.; Jochmans, D.; Leyssen, P.; Wang, K.; Neyts, J.; et al. 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. [CrossRef] [PubMed]

34. 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. [CrossRef]

35. 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. [CrossRef] [PubMed]

36. 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. [CrossRef]

37. Cerón-Carrasco, J.P. When virtual screening yields inactive drugs: Dealing with false theoretical friends. *ChemMedChem* **2022**, *17*, e202200278. [CrossRef]

38. Dantas, R.F.; Evangelista, T.C.S.; Neves, B.J.; Senger, M.R.; Andrade, C.H.; Ferreira, S.B.; Silva-Junior, F.P. Dealing with frequent hitters in drug discovery: A multidisciplinary view on the issue of filtering compounds on biological screenings. *Expert Opin. Drug Discov.* **2019**, *14*, 1269–1282. [CrossRef]

39. Sorokina, M.; Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminf.* **2020**, *12*, 20. [CrossRef]

40. Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M.A.; Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminf.* **2021**, *13*, 2. [CrossRef]

41. 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*, e62839. [CrossRef] [PubMed]

42. ISDB. A database of In-Silico predicted MS/MS spectrum of Natural Products. Available online: http://oolonek.github.io/ISDB/ (accessed on 12 June 2023).

43. 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*, D654–D659. [CrossRef] [PubMed]

44. Sterling, T.; Irwin, J.J. ZINC 15--ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef] [PubMed]

45. Ye, H.; Ye, L.; Kang, H.; Zhang, D.; Tao, L.; Tang, K.; Liu, X.; Zhu, R.; Liu, Q.; Chen, Y.Z.; et al. HIT: Linking herbal active ingredients to targets. *Nucleic Acids Res.* **2011**, *39*, D1055–D1059. [CrossRef]

46. Kang, H.; Tang, K.; Liu, Q.; Sun, Y.; Huang, Q.; Zhu, R.; Gao, J.; Zhang, D.; Huang, C.; Cao, Z. HIM-herbal ingredients in-vivo metabolism database. *J. Cheminf.* **2013**, *5*, 28. [CrossRef] [PubMed]

47. Li, B.; Ma, C.; Zhao, X.; Hu, Z.; Du, T.; Xu, X.; Wang, Z.; Lin, J. YaTCM: Yet another traditional chinese medicine database for drug discovery. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 600–610. [CrossRef] [PubMed]

48. Ru, J.; Li, P.; Wang, J.; Zhou, W.; Li, B.; Huang, C.; Li, P.; Guo, Z.; Tao, W.; Yang, Y.; et al. TCMSP: A database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminf.* **2014**, *6*, 13. [CrossRef]

49. Kim, S.-K.; Nam, S.; Jang, H.; Kim, A.; Lee, J.-J. TM-MC: A database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. *BMC Complement. Altern. Med.* **2015**, *15*, 218. [CrossRef]

50. Xu, H.-Y.; Zhang, Y.-Q.; Liu, Z.-M.; Chen, T.; Lv, C.-Y.; Tang, S.-H.; Zhang, X.-B.; Zhang, W.; Li, Z.-Y.; Zhou, R.-R.; et al. ETCM: An encyclopaedia of traditional Chinese medicine. *Nucleic Acids Res.* **2019**, *47*, D976–D982. [CrossRef]

51. Fang, X.; Shao, L.; Zhang, H.; Wang, S. CHMIS-C: A comprehensive herbal medicine information system for cancer. *J. Med. Chem.* **2005**, *48*, 1481–1488. [CrossRef]

52. Qiao, X.; Hou, T.; Zhang, W.; Guo, S.; Xu, X. A 3D structure database of components from Chinese traditional medicinal herbs. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 481–489. [CrossRef] [PubMed]

53. Huang, J.; Zheng, Y.; Wu, W.; Xie, T.; Yao, H.; Pang, X.; Sun, F.; Ouyang, L.; Wang, J. CEMTDD: The database for elucidating the relationships among herbs, compounds, targets and related diseases for Chinese ethnic minority traditional drugs. *Oncotarget* **2015**, *6*, 17675–17684. [CrossRef] [PubMed]

54. Chen, C.Y.-C. TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening in silico. *PLoS ONE* **2011**, *6*, e15939. [CrossRef]

55. 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*, 4329. [CrossRef] [PubMed]

56. Potshangbam, A.M.; Polavarapu, R.; Rathore, R.S.; Naresh, D.; Prabhu, N.P.; Potshangbam, N.; Kumar, P.; Vindal, V. MedPServer: A database for identification of therapeutic targets and novel leads pertaining to natural products. *Chem. Biol. Drug Des.* **2019**, *93*, 438–446. [CrossRef] [PubMed]

57. Bultum, L.E.; Woyessa, A.M.; Lee, D. ETM-DB: Integrated Ethiopian traditional herbal medicine and phytochemicals database. *BMC Complement. Altern. Med.* **2019**, *19*, 212. [CrossRef] [PubMed]

58. Ntie-Kang, F.; Onguéné, P.A.; Scharfe, M.; Owono Owono, L.C.; Megnassan, E.; Mbaze, L.M.; Sippl, W.; Efange, S.M.N. ConMedNP: A natural product library from Central African medicinal plants for drug discovery. *RSC Adv.* **2014**, *4*, 409–419. [CrossRef]

59. Ibezim, A.; Debnath, B.; Ntie-Kang, F.; Mbah, C.J.; Nwodo, N.J. Binding of anti-Trypanosoma natural products from African flora against selected drug targets: A docking study. *Med. Chem. Res.* **2017**, *26*, 562–579. [CrossRef]

60. Onguéné, P.A.; Ntie-Kang, F.; Mbah, J.A.; Lifongo, L.L.; Ndom, J.C.; Sippl, W.; Mbaze, L.M. The potential of anti-malarial compounds derived from African medicinal plants, part III: An in silico evaluation of drug metabolism and pharmacokinetics profiling. *Org. Med. Chem. Lett.* **2014**, *4*, 6. [CrossRef]

61. Ntie-Kang, F.; Nwodo, J.N.; Ibezim, A.; Simoben, C.V.; Karaman, B.; Ngwa, V.F.; Sippl, W.; Adikwu, M.U.; Mbaze, L.M. Molecular modeling of potential anticancer agents from African medicinal plants. *J. Chem. Inf. Model.* **2014**, *54*, 2433–2450. [CrossRef]

62. 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*, e90655. [CrossRef] [PubMed]

63. 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*, e78085. [CrossRef]

64. Ionov, N.; Druzhilovskiy, D.; Filimonov, D.; Poroikov, V. Phyto4Health: Database of phytocomponents from russian pharmacopoeia plants. *J. Chem. Inf. Model.* **2023**, *63*, 1847–1851. [CrossRef] [PubMed]

65. 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*, eabc6228. [CrossRef] [PubMed]

66. 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, 2011; pp. 3–22, ISBN 978-3-642-20991-8.

67. NaturAr. Available online: https://naturar.quimica.unlp.edu.ar/es/ (accessed on 9 December 2022).

68. 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. [CrossRef]

69. 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*, 7215. [CrossRef]

70. 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*, 103. [CrossRef]

71. 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.; et al. The Sistematx web portal of natural products: An update. *J. Chem. Inf. Model.* **2021**, *61*, 2516–2522. [CrossRef]

72. UEFS Natural Products. Available online: http://zinc12.docking.org/catalogs/uefsnp (accessed on 2 December 2022).

73. 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. [CrossRef]

74. Olmedo, A.D.; Medina-Franco, L.J. Chemoinformatic approach: The case of natural products of Panama. In *Cheminformatics and Its Applications*; IntechOpen: London, UK, 2019. [CrossRef]

75. 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*, 7577. [CrossRef]

76. UNIIQUIM. Available online: https://uniiquim.iquimica.unam.mx/ (accessed on 6 December 2022).

77. 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. [CrossRef] [PubMed]

78. Sánchez-Cruz, N.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database. *F1000Research* **2019**, *8*, 2071. [CrossRef]

79. Gómez-García, A.; Medina-Franco, J.L. Progress and impact of latin american natural product databases. *Biomolecules* **2022**, *12*, 1202. [CrossRef] [PubMed]

80. do Carmo Santos, N.; da Paixão, V.G.; da Rocha Pita, S.S. New Trypanosoma cruzi Trypanothione Reductase Inhibitors Identification using the Virtual Screening in Database of Nucleus Bioassay, Biosynthesis and Ecophysiology (NuBBE). *Anti-Infect. Agents* **2019**, *17*, 138–149. [CrossRef]

81. Acevedo, C.H.; Scotti, L.; Scotti, M.T. In Silico studies designed to select sesquiterpene lactones with potential antichagasic activity from an in-house asteraceae database. *ChemMedChem* **2018**, *13*, 634–645. [CrossRef]

82. Antunes, S.S.; Won-Held Rabelo, V.; Romeiro, N.C. Natural products from Brazilian biodiversity identified as potential inhibitors of PknA and PknB of M. tuberculosis using molecular modeling tools. *Comput. Biol. Med.* **2021**, *136*, 104694. [CrossRef]

83. Herrera-Acevedo, C.; Dos Santos Maia, M.; Cavalcanti, É.B.V.S.; Coy-Barrera, E.; Scotti, L.; Scotti, M.T. Selection of antileishmanial sesquiterpene lactones from SistematX database using a combined ligand-/structure-based virtual screening approach. *Mol. Divers.* **2021**, *25*, 2411–2427. [CrossRef]

84. Barazorda-Ccahuana, H.L.; Goyzueta-Mamani, L.D.; Candia Puma, M.A.; Simões de Freitas, C.; de Sousa Vieria Tavares, G.; Pagliara Lage, D.; Ferraz Coelho, E.A.; Chávez-Fumagalli, M.A. Computer-aided drug design approaches applied to screen natural product's structural analogs targeting arginase in *Leishmania* spp. *F1000Research* **2023**, *12*, 93. [CrossRef]

85. Menezes, R.P.B.; de Viana, J.; de, O.; Muratov, E.; Scotti, L.; Scotti, M.T. Computer-assisted discovery of alkaloids with schistosomicidal activity. *Curr. Issues Mol. Biol.* **2022**, *44*, 383–408. [CrossRef]

86. Rodrigues, G.C.S.; Dos Santos Maia, M.; de Menezes, R.P.B.; Cavalcanti, A.B.S.; de Sousa, N.F.; de Moura, É.P.; Monteiro, A.F.M.; Scotti, L.; Scotti, M.T. Ligand and structure-based virtual screening of lamiaceae diterpenes with potential activity against a novel coronavirus (2019-nCoV). *Curr. Top. Med. Chem.* **2020**, *20*, 2126–2145. [CrossRef]

87. Przybyłek, M. Application 2D descriptors and artificial neural networks for beta-glucosidase inhibitors screening. *Molecules* **2020**, *25*, 5942. [CrossRef]

88. Martinez-Mayorga, K.; Marmolejo-Valencia, A.F.; Cortes-Guzman, F.; García-Ramos, J.C.; Sánchez-Flores, E.I.; Barroso-Flores, J.; Medina-Franco, J.L.; Esquivel-Rodriguez, B. Toxicity assessment of structurally relevant natural products from Mexican plants with antinociceptive activity. *J. Mex. Chem. Soc.* **2017**, *61*, 186–196. [CrossRef]

89. Barrera-Vázquez, O.S.; Gómez-Verjan, J.C.; Magos-Guerrero, G.A. Chemoinformatic screening for the selection of potential senolytic compounds from natural products. *Biomolecules* **2021**, *11*, 467. [CrossRef] [PubMed]

90. Herrera-Acevedo, C.; Perdomo-Madrigal, C.; Herrera-Acevedo, K.; Coy-Barrera, E.; Scotti, L.; Scotti, M.T. Machine learning models to select potential inhibitors of acetylcholinesterase activity from SistematX: A natural products database. *Mol. Divers.* **2021**, *25*, 1553–1568. [CrossRef] [PubMed]

91. de Souza Ribeiro, M.M.; dos Santos, L.C.; de Novais, N.S.; Viganó, J.; Veggi, P.C. An evaluative review on Stryphnodendron adstringens extract composition: Current and future perspectives on extraction and application. *Ind. Crops Prod.* **2022**, *187*, 115325. [CrossRef]

92. Li, R.; Morris-Natschke, S.L.; Lee, K.-H. Clerodane diterpenes: Sources, structures, and biological activities. *Nat. Prod. Rep.* **2016**, *33*, 1166–1226. [CrossRef]

93. Parra, J.; Ford, C.D.; Murillo, R. Phytochemical study of endemic Costa rican annonaceae species annona pittieri and Cymbopetalum costaricense. *J. Chil. Chem. Soc.* **2021**, *66*, 5047–5050. [CrossRef]

94. Köhler, I.; Jenett-Siems, K.; Siems, K.; Hernández, M.A.; Ibarra, R.A.; Berendsohn, W.G.; Bienzle, U.; Eich, E. In vitro antiplasmodial investigation of medicinal plants from El Salvador. *Z. Naturforsch. C. J. Biosci.* **2002**, *57*, 277–281. [CrossRef]

95. Sotelo-Barrera, M.; Cília-García, M.; Luna-Cavazos, M.; Díaz-Núñez, J.L.; Romero-Manzanares, A.; Soto-Hernández, R.M.; Castillo-Juárez, I. Amphipterygium adstringens (Schltdl.) Schiede ex Standl (Anacardiaceae): An Endemic Plant with Relevant Pharmacological Properties. *Plants* **2022**, *11*, 1766. [CrossRef]

96. Agin-Liebes, G.; Haas, T.F.; Lancelotta, R.; Uthaug, M.V.; Ramaekers, J.G.; Davis, A.K. Naturalistic use of mescaline is associated with self-reported psychiatric improvements and enduring positive life changes. *ACS Pharmacol. Transl. Sci.* **2021**, *4*, 543–552. [CrossRef]

97. Cubilla-Rios, L.; Chérigo, L.; Ríos, C.; Togna, G.D.; Gerwick, W.H. Phytochemical analysis and antileishmanial activity of Desmotes incomparabilis, an endemic plant from Panama. *Planta Med.* **2008**, *74*, PA98. [CrossRef]

98. García Giménez, D.; García Prado, E.; Sáenz Rodríguez, T.; Fernández Arche, A.; De la Puerta, R. Cytotoxic effect of the pentacyclic oxindole alkaloid mitraphylline isolated from Uncaria tomentosa bark on human Ewing's sarcoma and breast cancer cell lines. *Planta Med.* **2010**, *76*, 133–136. [CrossRef]

99. Gonzales, G.F.; Valerio, L.G. Medicinal plants from Peru: A review of plants as potential agents against cancer. *Anticancer. Agents Med. Chem.* **2006**, *6*, 429–444. [CrossRef] [PubMed]

100. Medina-Franco, J.L.; Chávez-Hernández, A.L.; López-López, E.; Saldívar-González, F.I. Chemical multiverse: An expanded view of chemical space. *Mol. Inf.* **2022**, *41*, e2200116. [CrossRef] [PubMed]

101. 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. In *Studies in Natural Products Chemistry*; Elsevier: Amsterdam, The Netherlands, 2018; ISBN 9780444641793.

102. Wildman, S.A.; Crippen, G.M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873. [CrossRef]

103. 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. [CrossRef] [PubMed]

104. 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*, 3–26. [CrossRef]

105. Lipinski, C.A. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337–341. [CrossRef]

106. 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. [CrossRef]

107. Gleeson, M.P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51*, 817–834. [CrossRef]

108. 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.; et al. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872–4875. [CrossRef] [PubMed]

109. Ntie-Kang, F.; Nyongbela, K.D.; Ayimele, G.A.; Shekfeh, S. "Drug-likeness" properties of natural compounds. *Phys. Sci. Rev.* **2019**, *4*, 20180169. [CrossRef]

110. Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminf.* **2020**, *12*, 12. [CrossRef] [PubMed]

111. Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *J. Cheminform.* **2020**, *12*, 43. [CrossRef] [PubMed]

112. 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. [CrossRef] [PubMed]

113. Waskom, M. Seaborn: Statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]

114. Open-Source Chemoinformatics and Machine Learning. RDKit: Open-Source Cheminformatics Software. Available online: https://www.rdkit.org (accessed on 8 February 2023).

115. MolVS: Molecule Validation and Standardization. Available online: https://molvs.readthedocs.io/en/latest/index.html (accessed on 9 February 2023).

116. 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. [CrossRef]

117. Plotly Technologies Inc. *Collaborative Data Science Publisher: Plotly Technologies Inc.*; Plotly Technologies Inc.: Montréal, QC, USA, Canada, 2015.

118. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz information miner. *SIGKDD Explor. Newsl.* **2009**, *11*, 26. [CrossRef]

119. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. *arXiv* **2012**. [CrossRef]

120. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef]

121. Bajusz, D.; Rácz, A.; Héberger, K. Chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching. In *Comprehensive Medicinal Chemistry III.*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 329–378, ISBN 9780128032015.

122. Willighagen, E.L.; Mayfield, J.W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **2017**, *9*, 33. [CrossRef] [PubMed]

123. Miranda-Salas, J.; Peña-Varas, C.; Valenzuela Martínez, I.; Olmedo, D.A.; Zamora, W.J.; Chávez-Fumagalli, M.A.; Azevedo, D.Q.; Castilho, R.O.; Maltarollo, V.G.; Ramírez, D.; et al. Trends and challenges in chemoinformatics research in Latin America. *Artif. Intell. Life Sci.* **2023**, *3*, 100077. [CrossRef]