National Autonomous University of Mexico

Faculty of Chemistry

Master's and Doctoral Program in Chemical Sciences

A chemical library of natural products from Latin America (LANaPDB)

Tutor
Dr. José Luis Medina Franco

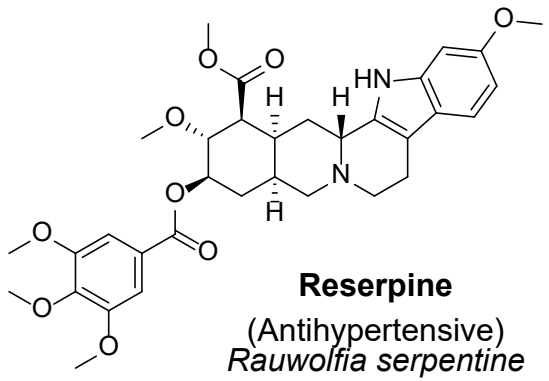
Student
Alejandro Gómez García

INDEX

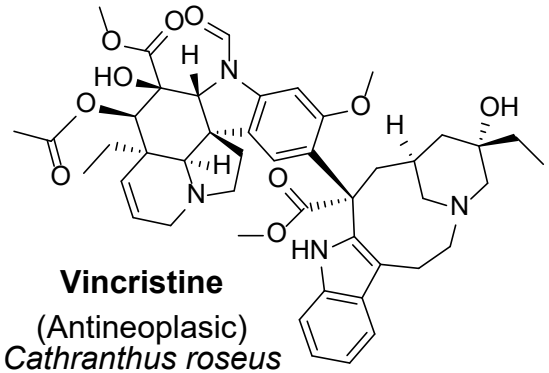
NATURAL PRODUCTS

Source

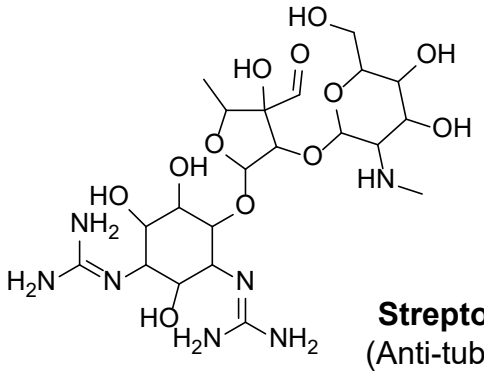
Substances produced by plants, microorganisms, and animals.



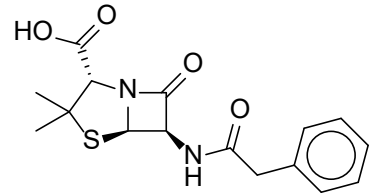
Reserpine
(Antihypertensive)
Rauwolfia serpentine
Plant



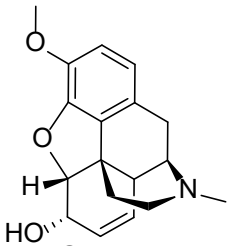
Vincristine
(Antineoplasic)
Cathranthus roseus
Plant



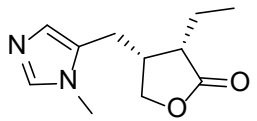
Streptomycin
(Anti-tubercular)
Streptomyces griseus
Bacteria



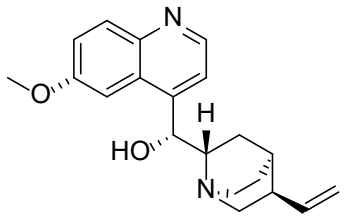
Penicillin G
(Bactericide)
Penicillium chrysogenum
Fungus



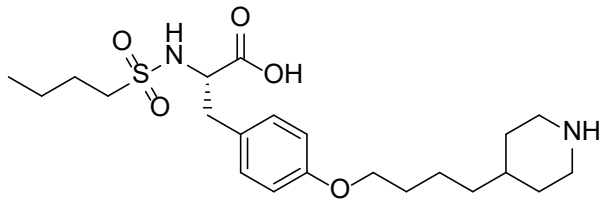
Codein
(Analgesic/antitussive)
Papaver somniferum
Plant



Pilocarpine
(Intraocular pressure lowering)
Pilocarpus jaborand
Plant



Quinine
(Antimalarial)
Chinchina trees
Plant



Tirofiban
(Platelet aggregation inhibitor)
Echis carinatus saw-scaled viper
Terrestrial animal

NATURAL PRODUCTS IN THE PHARMACEUTICAL INDUSTRY

Approved antineoplasic drugs

1946-1980

- Total: 70
- NPs: **40** (53.3%):
  - ➢Unaltered NPs.
  - ➢NPs that have undergone a semi-synthetic modification

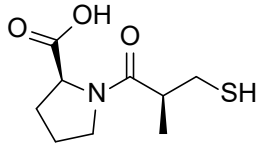
1981-2019

- Total: 185
- NPs: **120** (64.9%):
  - ➢Unaltered NPs.
  - ➢Synthetic compounds, whose core is inspired by NPs.

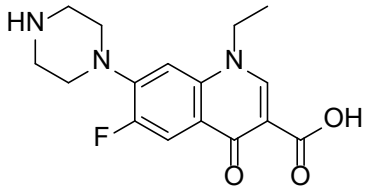
Newman-D, J. *et al*. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of natural products.* **2020**. 83, 770–803.

COMPUTER-AIDED DRUG DESIGN

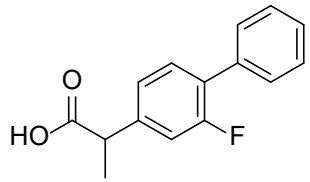
Between 1981 and 2019, more than **70 drugs** were approved, whose development process involved some **computational method**.



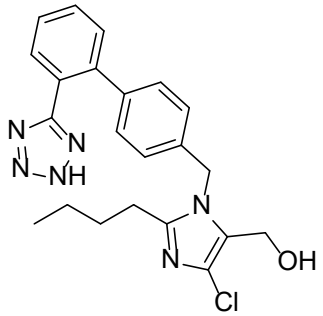
Captopril
(Antihypertensive)
Bothrops jararaca snake
Terrestrial animal
Receptor structure-based



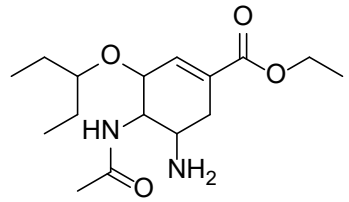
Norfloxacin
(Antibiotic)
Quantitative structure-activity relationship



Flurbiprofen
(Anti-inflammatory)
Molecular docking



Losartan
(Antihypertensive)
Receptor structure-based



Oseltamivir
(Antiviral)
Receptor structure-based

Sabe-T, V. *et al*. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European journal of medicinal chemistry.* **2021**. 224, 113705.

DATABASES

Definition

Organized collection of information in any field of knowledge.

Utility

Organize information

Retrive required information

Masic, I. *et al*. Review of most important biomedical databases for searching of biomedical scientific literature. *Ultrasound in obstetrics and gynecology.* **2012**. 6, 343–361

DATABASES IN THE CHEMISTRY AREA

Database category	Content	Database examples
Chemical information	Chemical and crystal structures spectra Reactions and syntheses Thermophysical data	ChemSpider ChEBI Chemical Universe Database GDB
Bioactivity	Inhibitor constant (K_i) Dissociation constant (K_d) Half maximal inhibitory concentration (IC_{50}) Half maximal effective concentration (EC_{50})	PubChem ChEMBL BindingDB ChemBank PDBbind
Drug	Detailed drug data Comprehensive drug target information	DrugBank
Natural product	Structures	Universal Natural Product Database MeFSAT Natural Product Atlas
Chemical availability	Available compounds offered by chemical vendors	ZINC NCI
Fragments	Structures Physicochemical information Binding site preferences	FDB-17 Fragment Store PADFrag

ChemSpider The free chemical database

PubChem

ChEMBL

DRUGBANK

ZINC DATABASE

Yang, J. *et al*. Freely accessible chemical database resources of compounds for in silico drug discovery. *Current medicinal chemistry.* **2019**. 26, 7581-7597.

REPRESENTATIVE NATURAL PRODUCT DATABASES

Database name	Number of compounds	Accessibility
Collection of Open Natural Products (COCONUT)	411,621	Open access
Universal Natural Product Database	~229,000	Open access
SuperNatural 3.0	449,048	Open access
ZINC	~80,000	Open access
Dictionary of Natural Products	~230,000	Commercial
Scifinder	~300,000	Commercial
Reaxys	~200,000	Commercial
TCM@Taiwan (China)	~58,000	Open access
IMPPAT (India)	~10,000	Open access
AfroDB (Africa)	~1000	Open access
Phyto4Health (Russia)	3128	Open access

Gómez-G, A. and Medina-F, J.L. Progress and impact of Latin American natural product databases. *Biomolecules.* **2022**. 12, 1202.

COMMERCIAL NATURAL PRODUCTS DATABASES

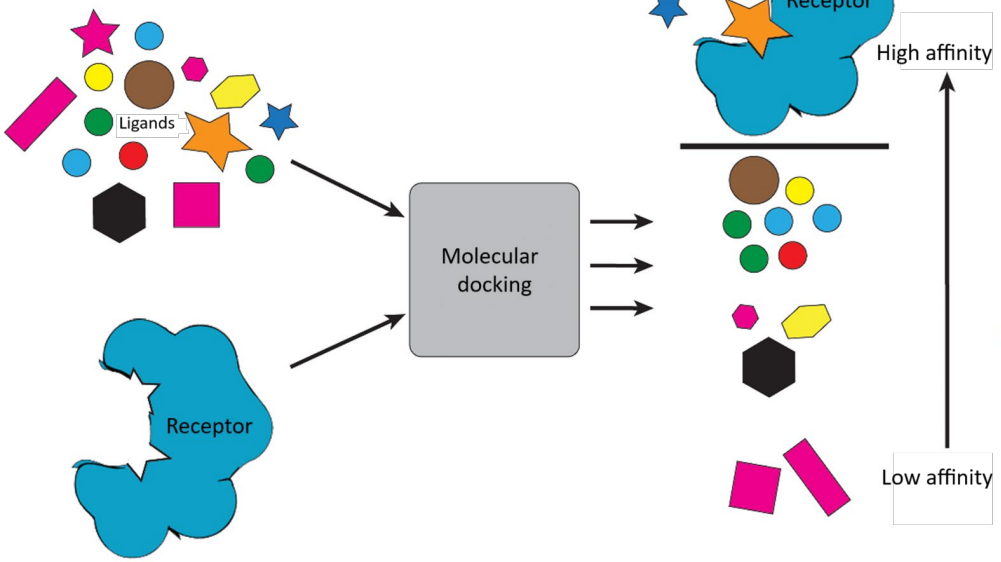
Commercial databases	
Database/ Name of the company	**Source of natural products**
AnalytiCon Discovery	Bacteria, terrestrial animals
Axxam/IMAX Discovery	Bacteria, plants and fungi
BioAustralis	Bacteria and fungi
Biosortia Microbiomics	Bacteria and fungi
Caithness Biotechnologies	Plants
ChromaDex	Fungi
Curia	Bacteria, fungi, terrestrial and marine animals
Cyano Biotech	Cyanobacteria
Greenpharma Natural Compound Library	Bacteria, plants, terrestrial animals
INDOFINE Chemical Company	Plants

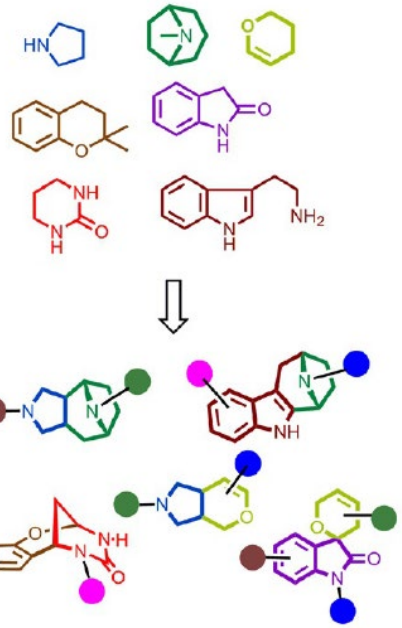
The original list is made up of 25 commercial databases

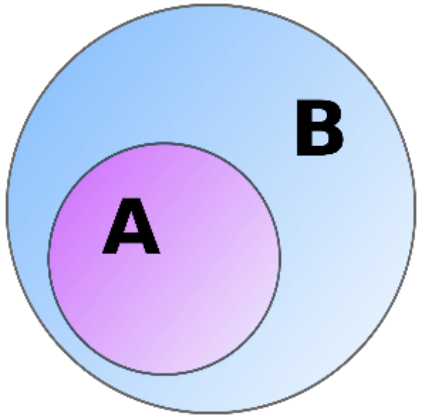
National Center for Complementary and Integrative Health
https://www.nccih.nih.gov/grants/natural-product-libraries

PRACTICAL APPLICATIONS OF NATURAL PRODUCTS DATABASES

- Applications
  - Prediction of bioactive molecules
  - Pseudo natural products
  - Subset selection
  - Ecology

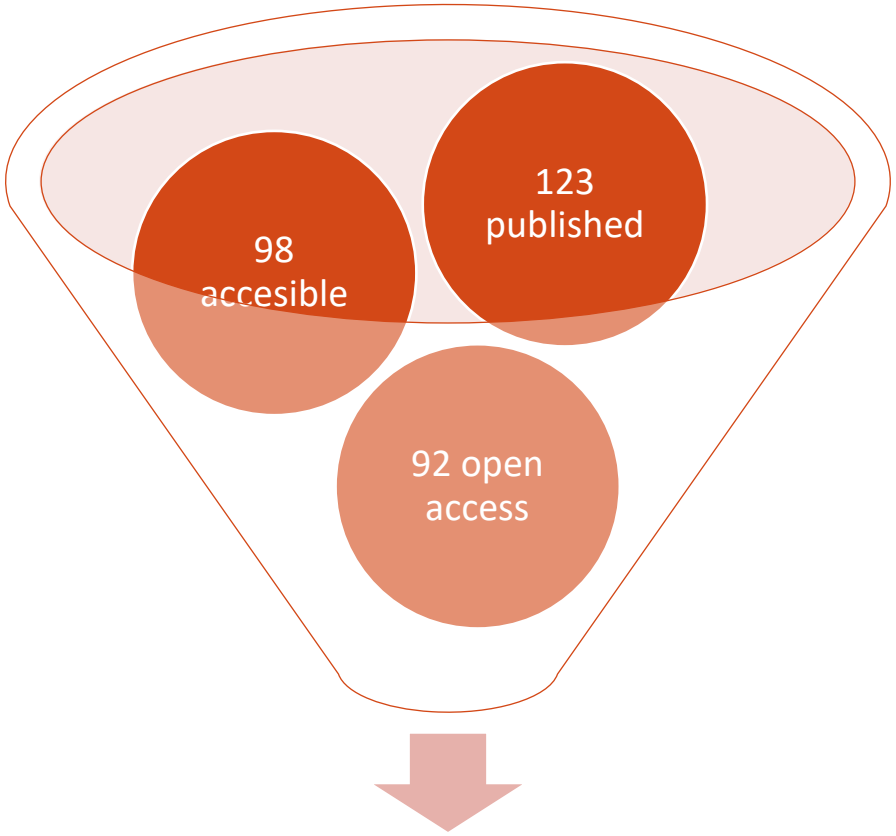






PUBLISHED NATURAL PRODUCTS DATABASES

2000-2019



Only **50** contain structures that can be recovered for chemoinformatics analysis.

Sorokina, M. *et al.* Review on natural products databases: Where to find data in 2020. *Journal of chemoinformatics.* **2020**. 12, 20.

NATURAL PRODUCTS DATABASES

Latin America

It is home to at least a third of global biodiversity.

Biodiversity
Variety of living beings in their environment.

Latin America is a source of bioactive molecules.

Raven, P.H. *et al*. The distribution of biodiversity richness in the tropics. *Science advances.* **2020**. 6, 228.

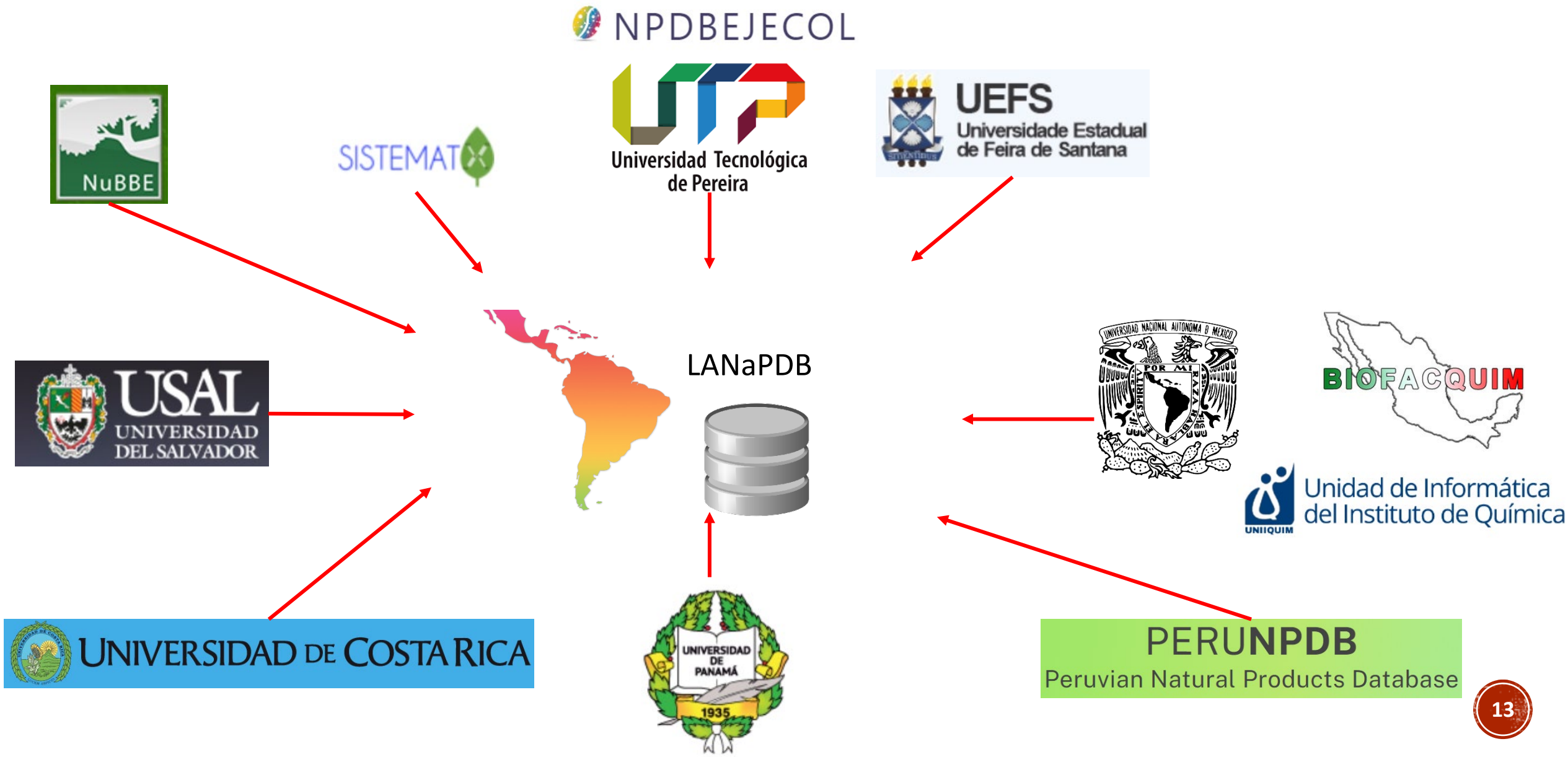
PRACTICAL APPLICATIONS OF THE LATIN AMERICAN NP DATABASES

Database	Country	Disease or symptom	Causal agent	Number of compounds identified	Year of publication
$NuBBE_{DB}$	Brazil	Chagas disease Tuberculosis	*Trypanosoma cruzi* *Mycobacterium tuberculosis*	10 13	2019 2021
SistematX	Brazil	Chagas disease Leishmaniasis Schistosomiasis Coronavirus disease 2019 Alzheimer's disease	*Trypanosoma cruzi* *Leishmania donovani* *Schistosoma mansoni* SARS-CoV-2 -	13 13 5 19 2	2018 2021 2022 2020 2021
UNIIQUIM	Mexico	Pain		6	2017
BIOFACQUIM	Mexico	Obesity Diabetes Hyperlipoproteinemia Cancer Human immunodeficiency virus infection Hepatitis B and C. Age-related diseases		8 3	2020 2021

Gómez-G, A. and Medina-F, J.L. Progress and impact of Latin American natural product databases. *Biomolecules* **2022** 12:1202

PROBLEM

There is not a database that unifies the Latin American NP databases



OBJECTIVE

Build a database of compounds found in natural products from Latin America.

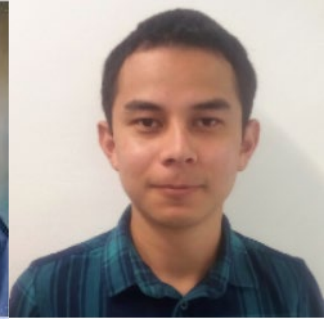
STRATEGY

The database will be built from the **integration** and **unification** of **databases** already established or under construction.

LATIN AMERICAN NATURAL PRODUCTS DATABASES IN LANaPDB

Database	Number of compounds	Country	Accessibility	Source	Year of publication
$NuBBE_{DB}$	2223	Brazil	Open-access	Plants Microorganisms Terrestrial animals Marine animals	2013, 2017
SistematX	9514	Brazil	Open-access	Plants	2018, 2021
UEFS	503	Brazil	Open-access	Plants	There is no associated scientific publication
NPDB EjeCol	236	Colombia	Open-access	Plants	2024
NAPRORE-CR	359	Costa Rica	Access under request	Plants Microorganisms	Not published yet
LAIPNUDELSAV	214	El Salvador	Access under request		There is no associated scientific publication
CIFPMA	454	Panama	Access under request	Plants	2017
PeruNPDB	280	Peru	Open-access	Plants Animales	2023
UNIIQUIM	~1112	Mexico	Open-access	Plants	There is no associated scientific publication
BIOFACQUIM	553	Mexico	Open-access	Plants Fungi Propolis Marine animals	2019, 2020

COUNTRIES COLLABORATING IN THE PROJECT

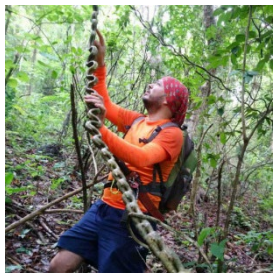


Dr. José Luis Medina Franco

MSc. Alejandro Gómez García

México

BIOFACQUIM



Dr. Marvin José Núñez

El Salvador

LAIPNUDESAL

Dr. Dionisio Antonio Olmedo

Panamá

CIFPMA

Dr. Héctor Fabio Cortés Hernández

MSc. Johny Roberto Rodríguez Pérez

Colombia

Dr. Oscar M. Mosquera Martinez

Dr. Hoover A. Valencia Sánchez

Dr. Miguel Ángel Chávez Fumagalli

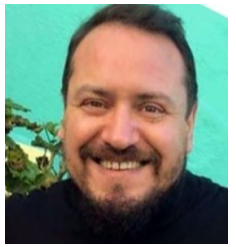
Dra. Haruna L. Barazorda-Ccahuana

Perú

PeruNPDB

Dr. José Ramírez Domenech

Puerto Rico



Dr. Pablo Dands

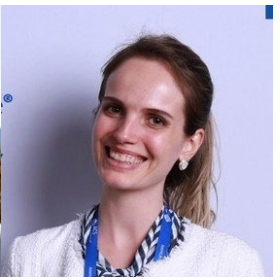
Uruguay

Dr. Willam Zamora

Daniel A Acuña Jiménez

Costa Rica

NAPRORE-CR



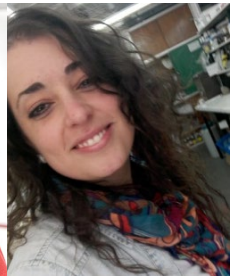
Dra. Vanderlan da Silva Bolzani

Dr. Adriano D. Andricopulo

Dra. Marilia Valli

Brasil

NuBBE$_{DB}$



Dra. Valeria Patricia Sülsen

Dra. Soledad Ravetti

Dra. Manuela Emilia García

Argentina

LANAPDB 2° VERSION

- Total number of compounds: **13,578**
- Total number of Latin American databases in LANaPDB: 10
- Total number of Latin American countries: 7

Content of the database
Chemical structures in SMILES format
References
Commercial availability
Biological activity
Synthetic feasibility
Molecular complexity
Structural classification
Physicochemical properties
Original and new IDs
Cross-references to ChEMBL and PubChem

DATABASE CURATION

Compound standardization
Removal of explicit hydrogen atoms.
Disconnect covalent bonds between organic molecules and metals.
Removal of salts, keeping the largest fragment, which is neutralized.
Reionization (Ensures that the strongest acid is protonated first into partially ionized molecules.)
Preservation of the original stereochemistry.
Elimination of repeating molecules.

CHEMOINFORMATICS ANALYSIS

- Structural classification
- Chemical space visualization

LANaPDB

Cross-reference to:

- ChEMBL
- PubChem

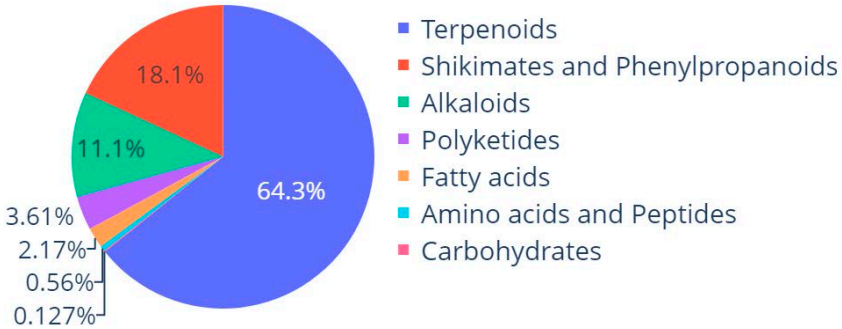
Determination of:

- Physicochemical properties
- Commercial availability
- Biological activity
- Molecular complexity
- Synthetic feasibility

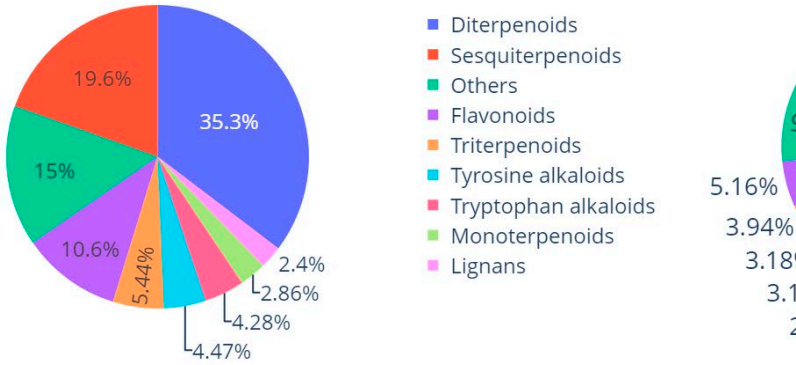
STRUCTURAL CLASSIFICATION

Natural product classifier

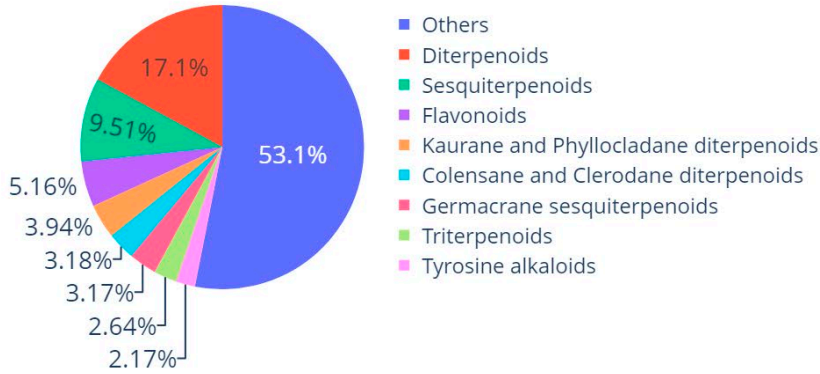
Pathway



SuperClass



Class

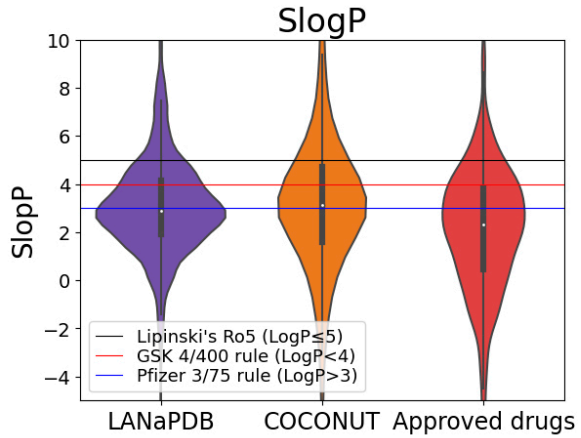


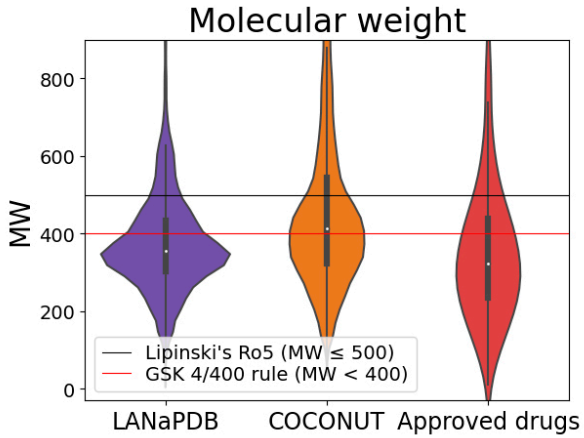
Pathway	Nature of the biosynthetic pathway.
Superclass	Chemical properties or chemotaxonomic information.
Class	Structural details

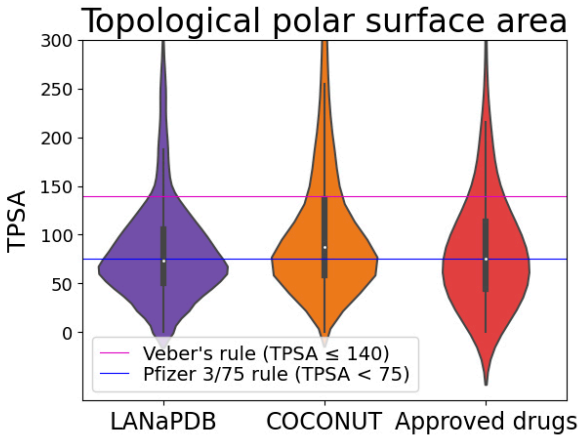
- In all 3 cases, the terpenoid compounds had the highest abundance.

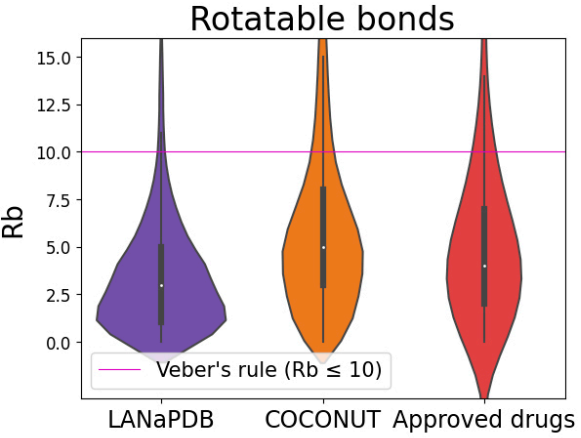
Kim-H,W. *et al*. NPClassifier: A deep neural network-based structural classification tool for natural products. *J. Nat. Prod*. **2021**, 84, 2795–2807.

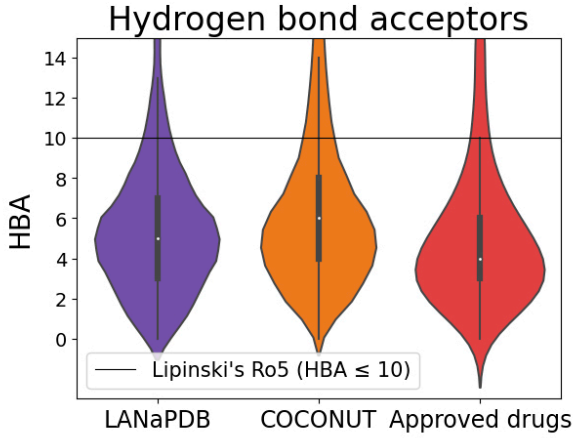
PHYSICOCHEMICAL PROPERTIES

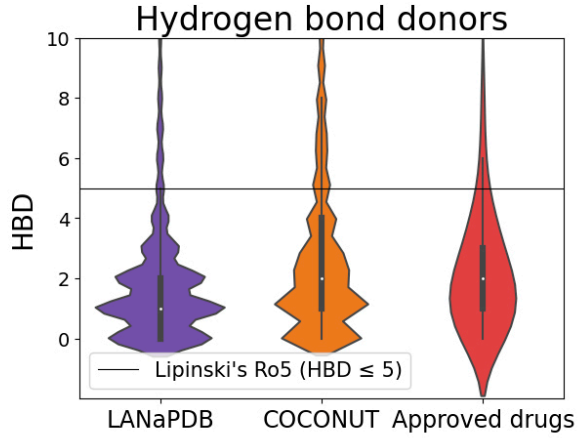


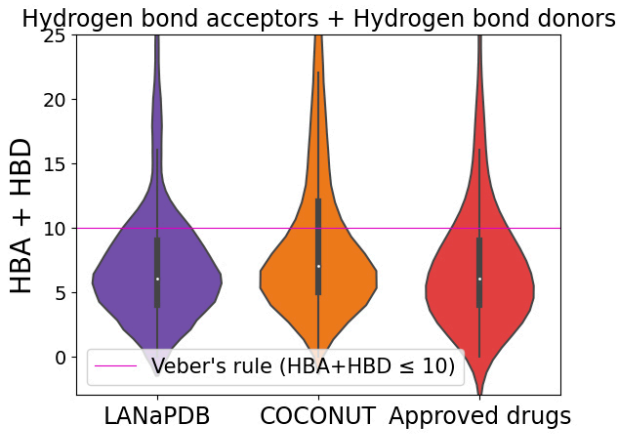












COCONUT

One of the largest open-access natural product databases with more than 411,000 compounds.

FDA-approved small-molecule drugs

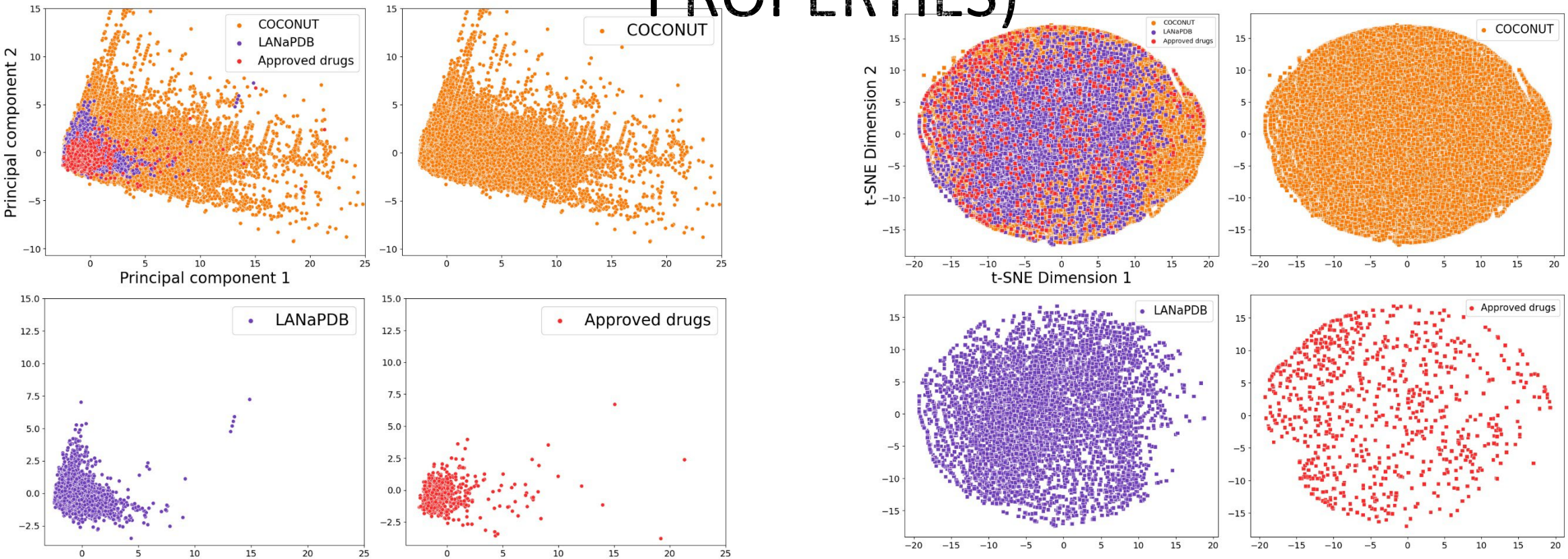
Version 5.1.10 (released by DrugBank in January 2023)

- LANaPDB and COCONUT show a similar distribution in their physicochemical properties.
- In general, overlap the areas where the greatest number of compounds in LANaPDB and approved drugs are concentrated.

VISUALIZATION (PHYSICOCHEMICAL PROPERTIES)

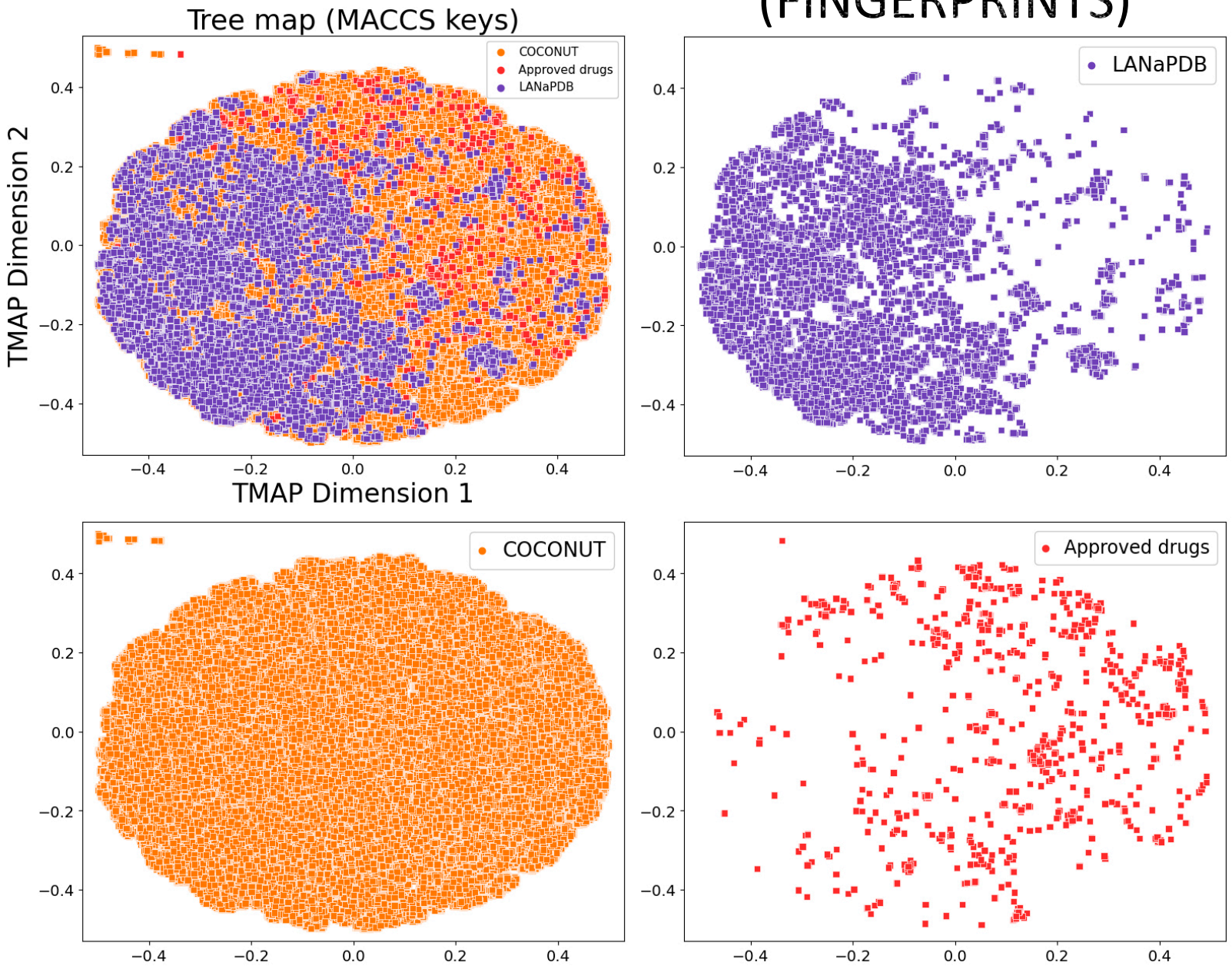
Principal component analysis

t-Distributed Stochastic Neighbor Embedding



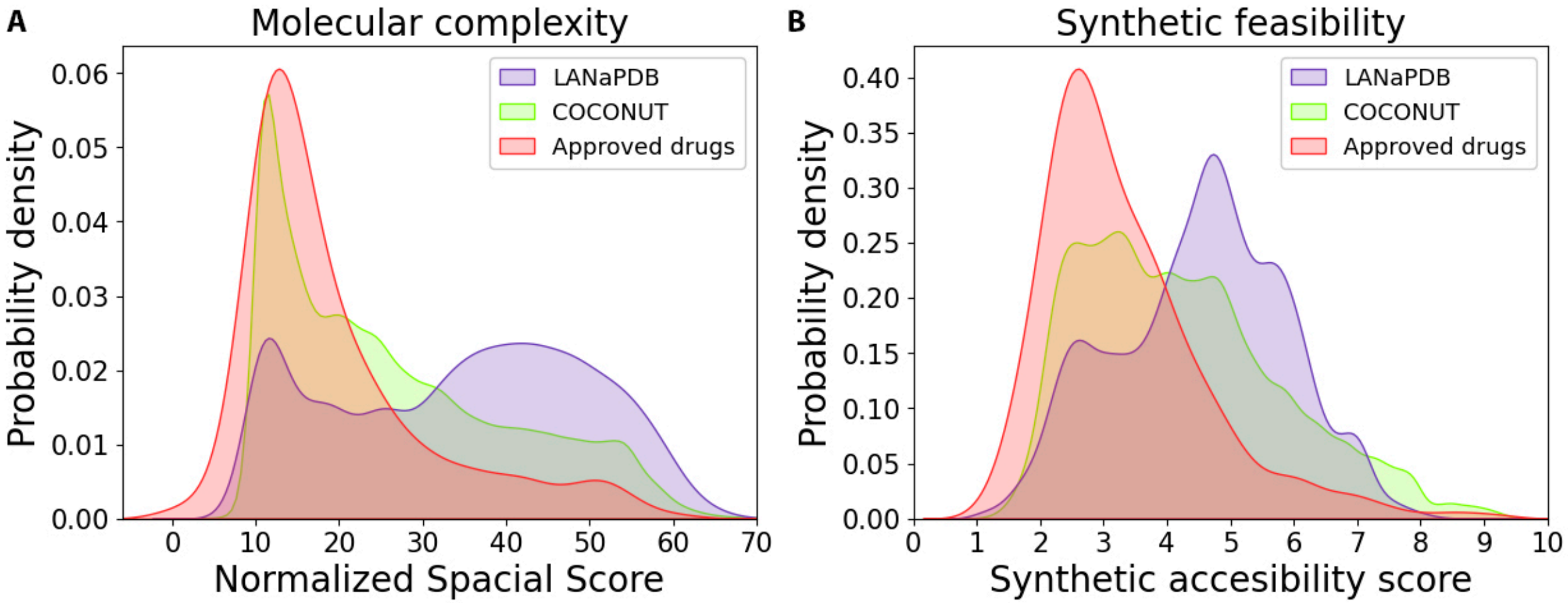
COCONUT (PCA and t-SNE)	Totally overlaps with approved drugs and LANaPDB.
Approved drugs (PCA)	Almost totally overlap with LANaPDB.
Approved drugs (t-SNE)	Are more dispersed, but with a high degree of overlapping to LANaPDB.

CHEMICAL SPACE VISUALIZATION (FINGERPRINTS)



- LANaPDB totally overlaps with COCONUT.
- The approved drugs are more dispersed with a lower degree of overlapping with LANaPDB.

MOLECULAR COMPLEXITY AND SYNTHETIC FEASIBILITY



Krzyzanowski, A. *et al*. Spacial Score–A Comprehensive Topological Indicator for Small-Molecule Complexity. *J. Med. Chem*. **2023**, 66 (18), 12739–12750.

Ertl, P. *et al*. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform*. **2009**, 1 (1), 8.

CROSS-REFERENCE TO ChEMBL AND PubChem

- **ChEMBL** and **PubChem** are two of the biggest publicly available chemical compound databases.

- Was made the addition of the ChEMBL and PubChem IDs to the LANaPDB compounds.

- 71.71% of the LANaPDB compounds have a PubChem ID.

- 23.69% of the LANaPDB compounds have a ChEMBL ID.

COMMERCIAL AVAILABILITY

- Obtained from PubChem.

- 70.5% of the LANaPDB compounds are comercial available.

- The information regard the available vendors for the individual compounds can be consulted in the PubChem website with the PubChem IDs already available for the LANaPDB compound.

BIOACTIVITY

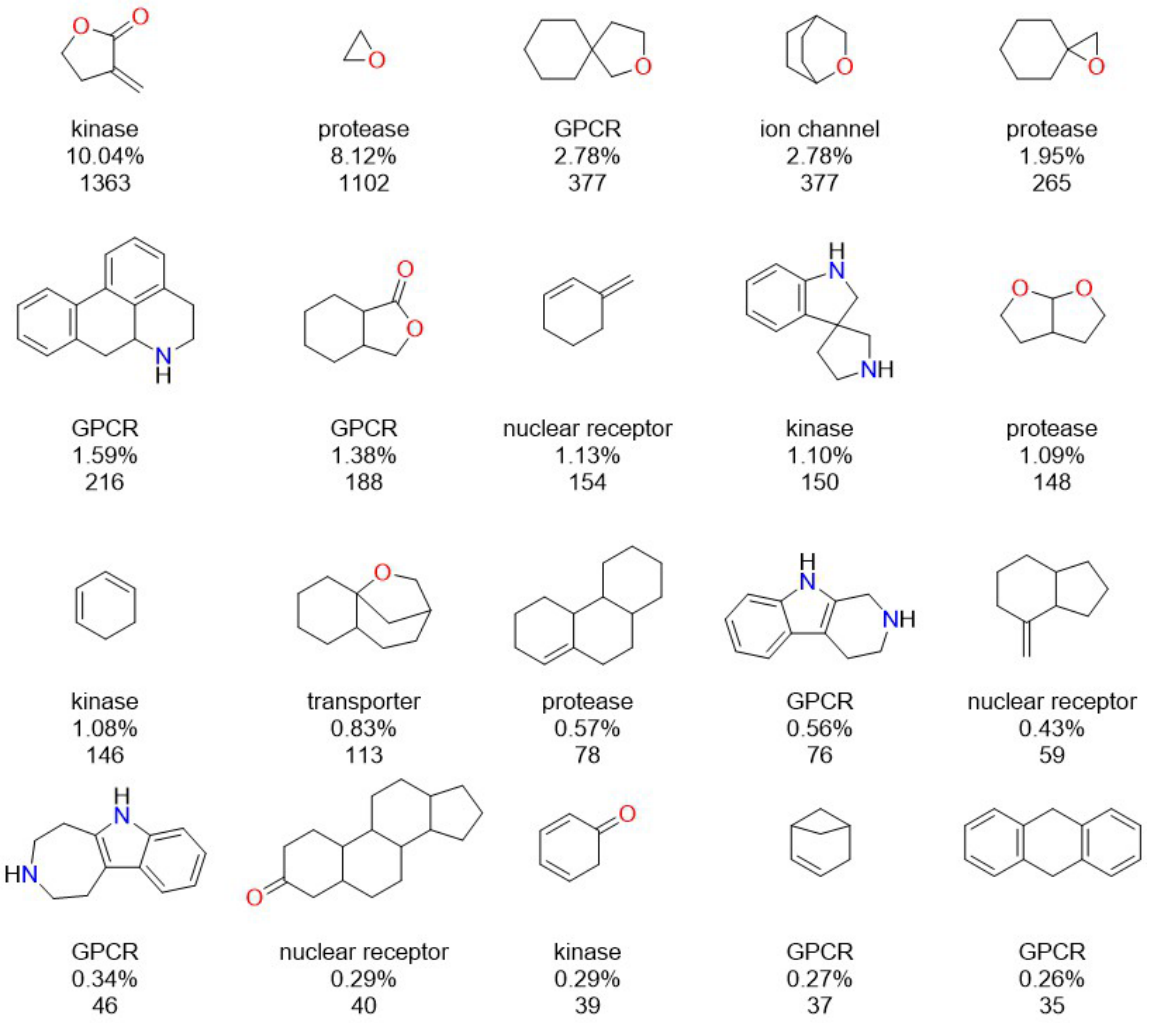
Whole molecules

- Bioactivity data obtained with the ChEMBL API from the InChIKey strings.
- 0.29% of the LANaPDB compounds have reported biological activity in the ChEMBL database.

Scaffolds

- Bioactivity data consulted in a paper which gathers these information from ChEMBL.
- 31.51% of the LANaPDB compounds have bioactive scaffolds.

Most common bioactive scaffolds in LANaPDB



Ertl, P. Magic Rings: Navigation in the Ring Chemical Space Guided by the Bioactive Rings. *J. Chem. Inf. Model*. **2022**, 62 (9), 2164–2170.

pharmaceuticals

MDPI

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

check for updates

Citation: Gómez-García, A.; Jiménez, D.A.A.; Zamora, W.J.; Barazorda-Ccahuana, H.L.; Chávez-Fumagalli, M.Á.; Valli, M.; Andricopulo, A.D.; Bolzani, V.d.S.; Olmedo, D.A.; Solis, P.N.; et al. Navigating the Chemical Space and Chemical Multiverse of a Unified Latin American Natural Product Database: LANaPDB. *Pharmaceuticals* **2023**, *16*, 1388. https://doi.org/10.3390/ph16101388

Academic Editors: Fabrizio Manetti and Paolo Governa

Received: 31 August 2023
Revised: 22 September 2023
Accepted: 26 September 2023
Published: 30 September 2023

Latin American Natural Product Database (LANaPDB): an update

27 August 2024, Version 1

Working Paper

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

Show author details

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

Received: 8 February 2024 | Revised: 3 April 2024 | Accepted: 13 April 2024
DOI: 10.1002/minf.202400052

RESEARCH ARTICLE

molecular informatics

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

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

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

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

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

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

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

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

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

Abstract

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

KEYWORDS
chemical space, chemoinformatics, databases, LANaPDB, natural products

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

biomolecules

MDPI

Review

Progress and Impact of Latin American Natural Product Databases

Alejandro Gómez-García and José L. Medina-Franco *

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico
* Correspondence: medinajl@unam.mx; Tel.: +52-55-5622-3899

Abstract: Natural products (NPs) are a rich source of structurally novel molecules, and the chemical space they encompass is far from being fully explored. Over history, NPs have represented a significant source of bioactive molecules and have served as a source of inspiration for developing many drugs on the market. On the other hand, computer-aided drug design (CADD) has contributed to drug discovery research, mitigating costs and time. In this sense, compound databases represent a fundamental element of CADD. This work reviews the progress toward developing compound databases of natural origin, and it surveys computational methods, emphasizing chemoinformatic approaches to profile natural product databases. Furthermore, it reviews the present state of the art in developing Latin American NP databases and their practical applications to the drug discovery area.

Keywords: chemoinformatics; compound databases; chemical space; diversity; drug discovery; open science; pseudo-natural product

check for updates

Citation: Gómez-García, A.; Medina-Franco, J.L. Progress and Impact of Latin American Natural Product Databases. *Biomolecules* **2022**, *12*, 1202. https://doi.org/10.3390/biom12091202

Academic Editor: Lukasz Kurgan

Received: 11 August 2022
Accepted: 29 August 2022
Published: 30 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Natural products (NPs) are a major source of bioactive molecules, and their importance is invaluable [1]. Between 1981 and 2014, over 50% of newly developed drugs were developed from NPs [2]. Over nearly four decades, they have been a significant resource of bioactive compounds for medicinal chemistry [3]. There are several sources for bioactive molecules, which include marine [4,5], fungal [6,7], bacteria [8], and plants [9]. Endogenous substances produced by humans and animals are another vital source of bioactive compounds [10]. Venoms and poisons produced by different animals are other rich sources [11].

Currently, there is an effort to find bioactive compounds from NPs as starting points for the further development of drug candidates for infectious diseases: antibacterial [12], antiprotozoal [13], antifungal [14], and antiviral [15]. Additionally, NPs are currently employed in medicinal chemistry to develop new chemotherapies, for example, neurodegenerative [16], cancer [17], immune-related [18], liver [19], and kidney [20] diseases, to mention a few examples. 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) [21,22].

Figure 1 shows the chemical structures of representative NPs approved for clinical use. The figure shows the pharmacological effect and the source of the compound. With the exception of captopril, all compounds were approved for clinical use without modifying the original chemical structure of the compound found in the extraction source. Captopril was developed based on the bradykinin potentiating factor in *Bothrops jararaca* snake venom. In 1981, it was the first animal toxin-based drug approved for human use. [23,24]. Digoxin is obtained from the plants of the genus *Digitalis* [25].

COCONUT

Search | Collections | Docs | About us | Download | Log in →

#COLLECTION

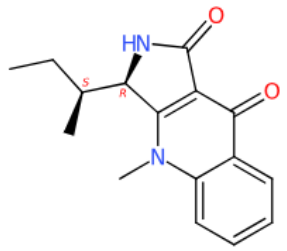
Latin America dataset

Molecules | Organism | Citations

Latin America dataset × Search

Showing 1 to 20 of 11289 results

< 1 2 3 4 5 6 7 8 9 10 ... 564 565 >

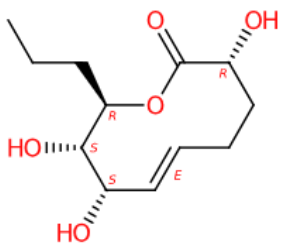


★★★★★

CNP0186650.1

Quinolactacin A1

1 5 2 1

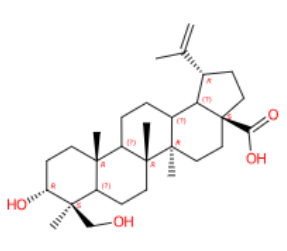


★★★★★

CNP0374713.1

KQJGPGHQDDZVHJ-ZFOCBLLI

1 4 2 1

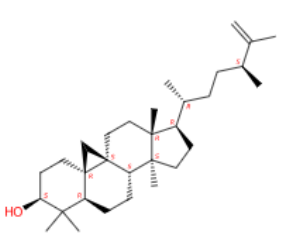


★★★★★

CNP0139464.1

(1R,3aS,5aR,5bR,8S,9R,11aR)-9-I

1 2 2 1



★★★★★

CNP0333329.1

Cyclolaudenol

5 10 3 1

https://coconut.naturalproducts.net/search?type=tags&q=Latin+America+dataset&tagType=dataSource

ACKNOWLEDGEM

Tutoring Committee

- Dr. Rogelio Gregorio Pereda Miranda
- Dr. Ramón Garduño Juárez

- Tutor: Dr. José Luis Medina Franco

➢ All the members of the Latin American research investigation groups that collaborates in this proyect.

DIFACQUIM research group.

CONACYT for the granted scholarship

SUPPORTING SLIDES

SYNTHETIC ACCESSIBILITY SCORE

Training dataset
1 million PubChem molecules

$$\text{SAscore} = \text{fragmentScore} - \text{complexityPenalty}$$

FragmentScore
- The contribution of each fragment.
- The most common fragments in PubChem have a higher contribution, which is positive.
- Less common fragments in PubChem have a negative contribution.

ComplexityPenalty
- It characterizes the presence of complex structural features.
- Take into account:
  1. Complexity of the ring system.
  2. Number of stereogenic centers.
  3. Presence of macrocycles (ring of 12 or more atoms) and size.

NORMALIZED SPACIAL SCORE

$$\mathrm{nSPS} = \frac{1}{a}\sum_{i} h_i s_i r_i n_i^2$$

a=Number of heavy atoms.

h=Atom hybridization

- *h*=3 sp^3
- *h*=2 sp^2
- *h*=1 sp
- *h*=4 other hybridization

s=Stereogenicity

- *s*=2 Atoms involved in *E* or *Z* isomers
- *s*=1 Any other case

r=Ring term

- *r*=2 Atoms of a non-aromatic ring.
- *r*=1 Atoms of an aromatic ring or linear structure.

n=Number of heavy atom neighbors

The branching of the molecular skeleton is accounted for by squaring the *n* term.