Natural Products Repository of Costa Rica (NAPRORE-CR): An open-access database

Daniel A. Acuña-Jiménez,[1] Jose Rodríguez-Zúñiga,[2] Daniela Gutiérrez-Ramírez,[2] Ricardo Quesada-Grosso,[2] Valery Conejo-López,[2] Kelvin Arce-Villalobos,[2] William J. Zamora,[2-3,*] 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, México.*

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

[3] *Laboratory of Computational Toxicology and Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Costa Rica*

* Correspondence authors: william.zamoraramirez@ucr.ac.cr (W.J.Z); medinajl@unam.mx (J.L.M-F.).

Abstract

Natural products are outstanding resources of bioactive compounds with potential applications not only in drug discovery but also in the cosmetic industry and natural pesticides. Costa Rica is among the most biologically diverse countries in terms of the number of known species per unit of area, even above conventionally considered megadiverse countries. In this work, we introduce the Natural Products Repository of Costa Rica (NAPRORE-CR): the first dedicated database that compiles structural representations and predicted properties of natural products found and/or characterized in Costa Rica. The first version of this collection comprises 1161 compounds, annotated with structural classifications, calculated structural and physicochemical properties (MW, HBD, HBA, RB, AlogP, and TPSA), and complex descriptors (SA score, QED, nSPS). The diversity and chemical space coverage of compounds in NAPRORE-CR were compared to drugs, pesticides, and cosmetics. Through the analysis and visualization of chemical space coverage and diversity, it was found that NAPRORE-CR has a property profile compatible with applications in all three fields, and that its compounds are structurally similar to those of approved drugs and natural pesticides. Cross-referencing NAPRORE-CR with PubChem and ChEMBL, combined with activity predictions, facilitated the identification of both known applications of the included compounds and potential new areas of study. In favour of open science and FAIR principles for data sharing, NAPRORE-CR is freely available at https://doi.org/10.5281/zenodo.7858061.

Keywords: chemoinformatics; cosmetics; drug discovery; natural products; pesticides.

Abbreviations: CADD, computer-aided drug design; EPA, United States Environmental Protection Agency; FAIR, Findable, Accessible, Interoperable, and Reusable; FDA, United States Food and Drug Administration; NAPRORE-CR, Natural Products Repository of Costa Rica; NPs, natural products; PCA, principal component analysis; t-SNE, t-distributed Stochastic Neighbor Embedding; QSAR, quantitative structure-activity relationship models; TMPAs, Tree MAPs; VS, virtual screening.

1. Introduction

The ever-increasing need to discover novel and effective compounds with potential applications has led to the development of computational methods to aid in the discovery of novel bioactive entities. The primary field benefiting from this technology is drug discovery, where chemoinformatic tools such as virtual screening (VS), quantitative structure-activity relationship models (QSAR), and *in silico* property predictions are utilized to identify lead compounds.[1] Using these and other techniques encompassed in the field of computer-aided drug design (CADD), more than 70 new drugs have been developed and approved for clinical use.[2] Although heavily applied in drug discovery, *in silico* methods have also permeated other industries, such as cosmetics, with QSAR and physiology-based kinetic models for risk assessment of exposure to cosmetic ingredients.[3] Another noteworthy example is environmental chemistry, where structure-based molecular design techniques, such as VS and molecular docking, have been applied to find lead compounds for novel pesticide design.[4]

Specialized compound databases are a key component of computer-aided compound research and are a cornerstone in chemoinformatics and artificial intelligence, which depend on high-quality data.[5] Compound databases provide information on a vast variety of molecules and their properties, from which potential lead compounds with desirable traits and activities can be derived.[6] Public databases such as PubChem,[7] BindingDB,[8] ChEMBL,[9], and ZINC20,[10] to name a few, have facilitated access to relevant data for medicinal chemistry on hundreds of thousands of compounds and their sources.[11] Other examples in the fields mentioned above include platforms like CCIBP and COSMOS,[12,13] for information on cosmetic ingredient regulations, physicochemical properties, metabolic routes and *in vitro/in vivo* toxicity; open databases for pesticide discovery, such as the Pesticide Properties Database, Bio-Pesticides Database, and Pesticide-Target Interaction Database;[14] or even resources relevant to several fields like the Antimicrobial TTC Project, that collects antimicrobial compounds useful as food and cosmetic preservatives, pesticides and others.[15]

Among the compound categorizations found in databases, natural products (NPs) stand out as an abundant source of bioactive molecules of interest. In drug design, during the 1981 to 2019 period, an average of 32% of the new drugs approved each year by the United States Food and Drug Administration (FDA) and other similar governmental organizations worldwide were unaltered NPs, botanical drugs (natural extracts with defined composition,) or compounds derived from NPs. Additionally, 28% of the small-molecule therapeutic agents approved at the end of the third quarter of 2019 also fall into these categories.[16] In cosmetics, market tendencies show growing interest in products containing NPs with antioxidant, anti-inflammatory, and regenerative properties, among other benefits.[17] Plants are the primary source of these compounds, with over 5,000 different species being used in cosmetic and cosmeceutical formulations globally.[18] However, marine organisms have picked up relevance as an alternative source of active ingredients, due to the unique conditions of their ecosystems.[19,20] In pesticides, NPs represent the source of 21.1% of the new compounds approved for crop protection during the period from 1997 to 2010 by the United States Environmental Protection

Agency (EPA). This figure increases to 35,7% when including biopesticides (living organisms or substances directly extracted from them with pest control capabilities).[21] Bacterial compounds have recently surged in interest in this field, representing the source of 42% of first-in-class pesticides in the past 30 years.[22]

Due to the potential applications of NPs, numerous databases have been developed focused on specific regions, industries, source organisms, and compound families. A total of 123 of these resources have been mentioned in literature between 2000 and 2020, of which 92 are open-access and 50 include readily available molecular structures.[23] Notable mentions include COCONUT, an open NPs repository with 695,133 compounds unified from 63 open-access databases;[24] SuperNatural 3.0, a freely available database of NPs and derivatives including more than 449,000 entries;[25] and NP Atlas, a database centered on microbially-derived NPs with over 32,000 unique structures.[26] Many of these databases are centered around region-specific species and knowledge of their use. Examples are TCM@Taiwan, the largest library of NPs obtained in China;[27] ANPDB, a merged database of natural compounds found in Northern and Eastern Africa;[28] and IMPPAT, a manually curated database of phytochemicals extracted from Indian medicinal plants.[29]

In Latin America, an enormous, largely unexplored potential lies in the vast natural wealth of the region, estimated to be a third of the global biodiversity.[30] The biodiversity hotspot status of several areas of the continent (Mesoamerica, Cerrado, Tropical Andes, among others), which means these zones are home to an outstanding number of species while at the same time facing severe loss of habitat, generates urgency to study and conserve these ecosystems as sources for bioactive compounds.[31,32] In this regard, multiple efforts have been made in Latin American countries to establish repositories of compounds found in their territories. From north to south, examples are BIOFACQUIM (Mexico),[33] CIFPMA (Panama),[34] NPDBEjeCol (Colombia),[35] $NuBBE_{DB}$ (Brazil),[36] PeruNPDB (Peru),[37], and NaturAr (Argentina).[38] It is worth noting the ongoing international effort to create a unified resource encompassing all Latin American NP databases, known as LANaPDB.[39]

Costa Rica is worldwide known for its vast natural wealth, housing close to 100 000 different identified species of plants, fungi, animals, microorganisms, and other taxonomic groups, which represent 4,9% of the total known species in the planet, going up to over a million species and 6% respectively when considering all expected species.[40] Factoring in its small share of land and marine extension of the globe (0,03% and 0,16% respectively), Costa Rica sits among the most biologically diverse countries in the world in terms of number of known species per unit of area, even above conventionally considered megadiverse countries such as Brazil and China.[41] Numerous research projects aimed at the discovery, characterization, identification, and development of potential applications for the species inhabiting their territory and their derived products can be found in the literature. These projects range from novel antibiotic discovery and tumor cell inhibiting compounds, to nanoparticle generation for improved bioavailability and biomass management and utilization. These and several other related lines of work are compiled in Table 1. Chemical structures of representative

NPs from Costa Rica are shown in Figure 1. Nonetheless, despite its brimming biodiversity and the evident interest in exploiting its resources, Costa Rica does not currently have a dedicated repository that encompasses the properties and molecular structures of the NPs obtained from species found in the country.

Table 1. Examples of research lines related to the utilization of natural products and resources found in Costa Rica.

Topic	References
Tumor growth inhibitors	Cruz et al.,[42] Cao et al.,[43] Vilariño et al.[44]
Biofilm inhibitors	Park et al.[45]
Novel antibacterial/antibiotic/antiviral agents	Mike et al.,[46] Tripathi et al.,[47] Cheng et al.,[48] Ymele-Leki et al.[49]
Antioxidant/cytotoxic compound extraction	Navarro et al.,[50] Villalobos-Vega et al.[51]
Natural product biosynthesis	Montero-Zamora et al.[52]
Nanoparticles for property enhancement	Quirós-Fallas et al.,[53] Castillo-Henríquez et al.,[54] Araya-Sibaja et al.[55]
Biomass management and utilization	Rosales-López et al.,[56] Syedd-León et al.[57]
Natural herbicides	Portuguez-García et al.[58]
Food contamination by mycotoxins	Granados-Chinchilla et al.[59]
Characterization of traditional medicinal remedies	Navarro et al.,[60] Doyle et al.[61]

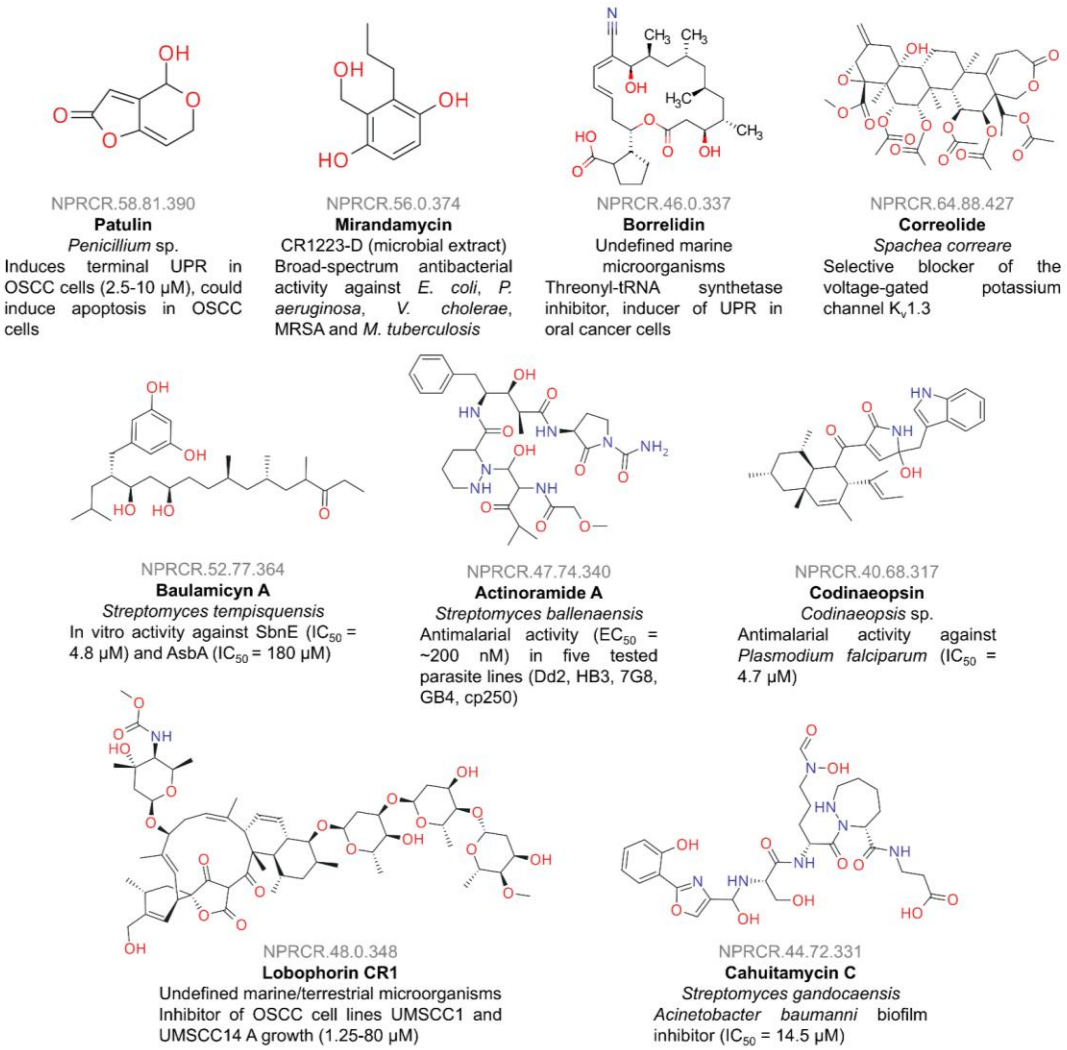


Figure 1. Chemical structures of exemplary natural products from Costa Rica and their reported activities in the scientific literature.

The main objective of this study was to introduce the Natural Products Repository of Costa Rica (NAPRORE-CR) as the first dedicated database that compiles structural representations and predicted properties of NPs found and/or characterized in Costa Rica. The data collection, curation, and standardization processes are described, as well as the compound structural classification, calculation of predicted physicochemical properties, and the characterization and visualization of its chemical space. We compared NAPRORE-CR against representative reference compound sets of approved drugs, cosmetic ingredients, and natural pesticides. The selected reference data sets enable comparisons between the NPs space and the compounds used in these industries, highlighting their prevalence as sources of valuable bioactive compounds.

2. Methods

2.1. Construction and curation of the NAPRORE-CR data set

The NAPRORE-CR was assembled through a literature search of scientific manuscripts related to the discovery, characterization, and application of NPs obtained from native species of the country. This bibliographic research was conducted in three phases: phase one was focused on lines of research developed by researchers of state universities of Costa Rica, mainly from the BIODESS Laboratory at School of Chemistry, Research Centre in Natural Products (CIPRONA), and the Research Centre in Grains and Seeds (CIGRAS) from the University of Costa Rica (UCR), and the Phytochemistry Laboratory (LAFIT) and the Natural Products and Biological Assays Laboratory (LAPRONEB) from the National University of Costa Rica (UNA). The papers included were extracted from the researchers' profiles found on the ResearchGate and Google Scholar platforms, after verifying that they were published in indexed journals, either in Spanish or English, and that the methods section indicated that the NPs were obtained from organisms found in Costa Rica or endemic to the country.

The second phase consisted of searches in PubMed, Scopus, and Google Scholar, using the keywords "Costa Rica," "natural products," and "University of Costa Rica." In addition, papers cited in the body of these sources that provided information on the compounds present in the studied species were included. The same review criteria were considered to select the papers to be included. The first two phases amounted to a total of 99 scientific publications.

In the third and last phase, several master's thesis manuscripts from the Chemistry Postgraduate Programme at the University of Costa Rica were added to the database. These theses were provided by professors from the School of Chemistry of UCR who participated in the writing of these works, with topics related to the extraction, isolation, and structural characterization of NPs. A total of six sources were included, resulting in 105 documents in the current version of NAPRORE-CR. It is worth mentioning that this is just the first published version of the database, as this repository will be expanded with scientific articles not included or published after the release of this work. One of the future goals of the NAPRORE-CR project is to serve as an updatable and collaborative resource where researchers can deposit new structures as new research is published in the field of NPs from Costa Rica.

The following information was manually mined from the selected manuscripts (when available): compound name, scientific name of the source organism, structural representation, year of publication, and bibliographic information (DOI or permanent link). The structural representations were stored using SMILES strings,[62] and obtained by searching the PubChem database,[7] using the compound name as a query for automated requests to its PUG-REST service,[63] prioritising the isomeric representation when available. Whenever these searches showed no results, the freely available online service DECIMER was employed to generate the corresponding SMILES.[64] This deep-learning-based tool is capable of transforming an image of a compound's skeletal formula into several text-based notation formats. All the SMILES generated by DECIMER were manually checked to confirm their fidelity to the sources.

The data set curation was performed using the Python programming language (version 3.10.12), employing the RDKit (version 2024.03.5),[65] MolVS (version 0.1.1),[66] and pandas (version 2.2.2) libraries,[67] for the structure manipulation, standardization, and data management, respectively. The default standardization routine of the MolVS module was applied to all the SMILES strings, which includes several corrections: normalization of functional groups, recombination of separated charges, breaking of bonds to metal atoms (and elimination of disconnected metallic centers), competitive reionization to ensure the correct ionization order in partially ionized molecules, standardization of stereochemistry information, and validation to identify molecules with unusual and/or problematic characteristics. Several other functions were applied subsequently to remove salts and keep the largest fragment, neutralize charges by addition/removal of hydrogens where possible, ionize partially ionized fragments, and determine a canonical tautomer. Duplicate entries were removed by generating a numeric identifier for each SMILES, scientific name, and reference, and combining them to create a unique entry ID. The SMILES identifier was also used to combine entries referring to the same compound found in multiple references or species.

2.2. Structural classification (NPClassifier)

The structural classification of the NAPRORE-CR data set was performed using the NPClassifier online service, a freely available deep-learning tool designed to automatically generate a classification for NPs based on an extensive ontology manually constructed from traditional labels found in the literature.[68] A script was constructed in the Python programming language (version 3.10.12), utilizing the pandas (version 2.2.2) and Requests (version 2.32.3) libraries to make API calls to the NPClassifier server.[69] The server response includes the assigned categories for every compound at three hierarchical levels: pathway (biosynthetic routes), superclass (metabolite classes, molecular shapes, and/or biosynthetic information), and class (compound families, functional groups, shared scaffolds). The category distribution for every hierarchical level was visually represented in pie charts using the Matplotlib library (version 3.8.0).[70]

2.3. Reference data sets

2.3.1. Approved drugs

The FDA-approved drugs data set was obtained from DrugBank, due to the significant presence of NPs in the set and the resource's relevance as a standard list for drug research. The data set contains 2769 compounds in the version used for this work (5.1.12),[71] and was last accessed in October 2024 through the DrugBank website (https://go.drugbank.com) with a free Academic License.

2.3.2. Pesticides

A natural pesticides data set was compiled from the “New active ingredient EPA registrations for conventional pesticides from 1997 to 2010” list, as compiled by Cantrell et al.[21], and the list of

Biopesticide Active Ingredients reported on the official EPA website (https://www.epa.gov),[72] last accessed in October 2024. The SMILES strings were obtained from PubChem using the compound names as search queries, as described in section 2.1. In the case of mixtures of compounds, such as essential oils, only the main components or those with reported pesticide activity were included. A total of 322 structures were obtained to represent the natural pesticides space.

2.3.3. Cosmetics

The COSMOS TTC data set was selected to represent cosmetic ingredients. This resource comprises cosmetic-related compounds and was created to compile the lowest observed adverse effect levels, based on the concept of the Threshold of Toxicological Concern (TTC).[73] The total data set includes 966 structures and was last accessed in October 2024 through the official COSMOS website (https://cosmosdb.eu/cosmosdb.v2).

2.4. Chemical property spaces

2.4.1. Physicochemical and structural properties

The property space of the NAPRORE-CR data set was constructed with the following descriptors, commonly used to profile compounds with therapeutic interest:[74] molecular weight (MW), hydrogen bond donors (HBD) and acceptors (HBA), rotatable bonds (RB), calculated octanol/water partition coefficient (AlogP) and topological polar surface area (TPSA). All these descriptors were generated using the Descriptors module from the RDKit library (version 2024.03.5) for the Costa Rican NPs database and the three reference data sets.

To include a descriptor representative of the ionization profile of the molecules, an acidic (AG) and basic (BG) group counter was constructed using Python (version 3.10.12) with the pandas (version 2.2.2), NumPy (version 1.26.4),[75] and RDKit libraries. A SMARTS list of acidic and basic functional groups was obtained from AstraZeneca's Peptide Tools GitHub repository,[76] which was used to build a script to search for these substructures in all the compounds in the NAPRORE-CR and reference data sets and count the ionizable centers found in each structure. The ionization profile significantly influences properties such as lipophilicity, polarity, and permeability; however, due to the difficulty in accurately determining it, it is often underrepresented in chemical property spaces.[77] In contrast, it is estimated that up to 77.5% of drugs show ionizable centers, which exacerbates the relevance of the inclusion of this property.[78]

To visualize the value distributions for all the calculated properties, a Python script (version 3.10.12) was constructed using the pandas (version 2.2.2), NumPy (version 1.26.4), Matplotlib (version 3.8.0), and Seaborn (version 0.13.2) libraries. Violin plots were generated for the eight chosen descriptors, with separate graphs for each dataset to visualize their value distributions better.

2.4.2. Synthetic accessibility, drug-like character, and molecular complexity

The Synthetic Accessibility Score (SAScore) was included as a preliminary assessment of the synthetic feasibility of the studied compounds. Compounds with an SAScore ≤ 6 are considered synthetically accessible.[79] The quantitative estimate of drug-likeness (QED) serves as a measurement of drug-likeness based on the calculated physicochemical and structural descriptors of the dataset. A QED > 0.67 represents "attractive", drug-like structures.[80] Molecular complexity is represented by the size-normalized spacial score (nSPS), which complements the analysis of the library's potential as a source of synthetically achievable, drug-like compounds.[81] An nSPS in the range of 10-20 coincides with most FDA-approved drugs, as recommended by the authors of this scoring system.[82] All three descriptors were calculated using Python (version 3.10.12) with the RDKit library (version 2024.03.5), alongside pandas (version 2.2.2) for data manipulation. Histograms were generated for each of these descriptors to visualize their value distributions, using Matplotlib (version 3.8.0) and Seaborn (version 0.13.2).[83]

2.5. Commercial availability

Commercial availability of the NAPRORE-CR compounds was determined using the PUG-REST service to retrieve chemical vendor information. A Python (version 3.10.12) script was constructed using the urllib module and the Requests (version 2.32.3) and pandas (version 2.2.2) libraries to cross-reference NAPRORE-CR compounds with PubChem to obtain their CIDs, which were then used as a search term to find all available vendors included in PubChem for each compound. The number of unique vendors was counted and stored when data was available; otherwise, a value of zero was reported when no results were found or no CID could be obtained.

2.6. Bioactivity profile

2.6.1. Experimental bioactivity profile

The ChEMBL API was consulted via a Python script (version 3.10.12) to cross-reference with the NAPRORE-CR database and retrieve the ChEMBL ID of the available compounds, utilizing the chembl_webresource_client library (version 0.10.9),[84] alongside Requests (version 2.32.3). The obtained IDs were then used to retrieve all the available activity data for the Costa Rican NPs using the *activity* module. A promiscuity ratio was calculated by dividing the number of activity assays marked as Active under the "*activity_comment*" section by the total number of activity assays found for every compound. This metric was then used to represent the bioactivity profile of the NAPRORE-CR molecular scaffolds in a constellation plot.[85]

2.6.2. Predicted bioactivity profile (target fishing)

To explore the potential of studying specific biological targets, the NetInfer web server was utilized for an inverse virtual screening (target fishing) analysis. NetInfer works via network-based inference

methods to make affinity predictions based on bioactivity data from ChEMBL and BindingDB chemical substructures, by associating them with substructures extracted from a training set and comparing them to the input compounds in the SMILES format. The website (https://lmmd.ecust.edu.cn/netinfer) allows for batch processing and generates a report with information related to the top-ranked targets based on a scoring function correlated with binding affinities. In this study, the weighted substructure-drug target network-based inference method was used with the default parameters and configuration (Morgan fingerprints, $\alpha = 0.4$, $\beta = 0.2$, $\gamma = -0.5$, $\delta = 20$, $\varepsilon = 4$, $k = 2$, 50 predictions per compound) to get the top 50 protein targets with the best predicted affinities for every compound. The list was further shortened to the top 5 results per compound, and their gene symbols and protein families were used to construct a frequency bar plot using Python (version 3.10.12) with the pandas (version 2.2.2) and Matplotlib (version 3.10.3) packages. The *target* module from the chembl_webresource_client library (version 0.10.9) was used to retrieve the UniProt identifier for the targets found in the NAPRORE-CR experimental activity results and match them with the NetInfer results to identify coincidences between predicted and experimental activity profiles.

2.7. Chemical space visualisation

To generate visualizations of the chemical space of NAPRORE-CR and the reference datasets, principal component analysis (PCA) was chosen to represent the physicochemical and structural descriptor space, and Tree MAPs (TMAPs) for the structural similarity space. PCA uses the original high-dimensional space variables to generate new, uncorrelated components by linearly combining them, thereby maximizing the data's variability explained by each component. [86] TMAPs generates minimum spanning trees to better represent the structural relationships between compounds in large datasets, preserving both global and local features for interpretation.[87]

2.7.1. PCA

The same libraries and versions mentioned in Section 2.4.1 were used, along with the SciKit-learn library (version 1.5.2), for PCA.[88] A cumulative explained variance graph was generated to choose the number of principal components necessary to represent at least 70% of the data sets' variance; this threshold was chosen because it is a common cut-off value in this type of analysis.[89] Finally, the PCA graphs were generated for NAPRORE-CR and all sets combined, with their respective descriptor vectors plotted on top of a scatter plot. The dots in each graph are colored to represent the respective pathway (NAPRORE-CR) and data set (combined sets) of the corresponding compound.

2.7.2. TMAP

A Python (version 3.9.20) virtual environment was set up to generate the tree MAPs, utilizing the TMAP (version 1.0.6), pandas (version 2.2.3), Numpy (version 1.26.4), Matplotlib (version 3.9.2), Seaborn (version 0.13.2), and RDKit (version 2024.3.5) libraries. MACCS keys for all the compounds

were generated with RDKit's MACCSkeys module. A 166-bit MinHash function was applied to the fingerprint vectors, and an LSHForest with 128 iterations was generated from them. A k-nearest neighbors search was conducted with k = 10 for the NAPRORE-CR set and k = 20 for the combined sets. These values were chosen following the examples found in the TMAP library documentation and considering the number of compounds in each space. A list of weighted vectors obtained from the neighbors search was generated to create a graph, which was then used to find the minimum spanning tree and set the layout for the final TMAP. Colors were assigned to each data point to represent pathways (NAPRORE-CR) or data sets (combined sets).

2.7.3. Constellation Plot

The Constellation Plot was created as previously proposed, utilizing the "*get-cores.py*" program to wash the NAPRORE-CR structures, extract the cores of every molecule, and generate the analog series, employing custom scripts provided by the authors in the original publication.[85] After data processing, the "*constellation-plots.ipynb*" notebook was used to generate the t-SNE low-dimensional chemical space and produce the final plot, using the following modules and libraries: Math, Itertools, Scripts, RDKit (version 2024.03.5), NumPy (version 2.2.6), pandas (version 2.2.3), SciKit-learn (version 1.5.2), and Matplotlib (version 3.10.3). The following default parameters were used in the t-SNE function: perplexity = 40, iterations = 3000, minimum molecules per core = 2. The promiscuity ratio generated with the activity assay information extracted from ChEMBL (section 2.4) was used to map the average promiscuity of every core cluster onto the constellation plot.

2.8. Global diversity analysis

To assess and compare the diversity of the compound data sets using multiple representations, a Consensus Diversity Plot (CDP) was generated as described by González-Medina et al.[90] A Python script (version 3.10.12) was built with the same libraries and versions mentioned in section 2.4.2, plus RDKit (version 2024.03.5) and SciPy (1.13.1).[91] Tanimoto coefficients were calculated for each possible pair of compounds in each data set using their MACCS keys, and a mean value was calculated for each data set. Murcko scaffolds were generated using RDKit and used to group the compounds in each dataset by their common scaffolds. The groups were ranked in descending order by number of compounds, and the cumulative fraction of scaffolds was calculated. Cyclic system recovery curves were built for each data set, and their corresponding area under the curve (AUC) was obtained as a measure of scaffold diversity. The physicochemical and structural descriptors generated in section 2.2 were normalized to calculate the mean intra-set Euclidean distance of the property space for each dataset. The CDP was constructed by plotting the mean Tanimoto coefficients on the X-axis, the AUC values on the Y-axis, and the mean intra-set Euclidean distances of the descriptors as a continuous color scale. Each data set was represented as a dot, the size of which was proportional to the number of compounds.

3. Results and Discussion

3.1. Database construction and curation

From the 105 manuscripts included in the current version of NAPRORE-CR, which span the 1988-2024 period, a total of 2121 compound entries were extracted prior to data curation and merging of duplicate entries. After applying the curation protocol described in Section 2.1, the compound list was reduced to 1161 entries, including stereoisomers identified in the consulted literature. The number of compounds retained and discarded in each data set after applying the molecular weight filter is shown in Table 2, corresponding to the sets used for the subsequent chemical space analysis. The complete NAPRORE-CR data set can be freely accessed through its Zenodo repository at https://doi.org/10.5281/zenodo.7858061.[92]

Table 2. Number of compounds per data set after filtering.

Data set	Total compounds	MW < 1500 Da	Filtered out
NAPRORE-CR	1161	1161	0
DrugBank	2769	2521	248
EPA	322	320	2
COSMOS	966	961	5

3.2. Structural classification (NPClassifier)

The compounds included in NAPRORE-CR show representation of the seven biosynthetic pathways of the NPClassifier ontology, as well as forty-four of its superclasses and one hundred seventy of its classes; distributions by level are shown in Figures 2 and 3. The most frequently found pathways, arranged in decreasing order of frequency, were terpenoids (48.7%), followed by shikimate-phenylpropanoids (22.7%) and fatty acids (12.0%). The most represented superclasses were sesquiterpenoids (28.9%), monoterpenoids (10.6%), and flavonoids (9.0%). Lastly, the most common classes were germacrane sesquiterpenoids (7.3%), cadinane sesquiterpenoids (4.3%), and wax monoesters (4.3%). These observations are consistent with the sources used to build the database, as these were primarily studies on the phytochemistry and biochemistry of marine microorganisms. The high presence of terpenoids in the database was expected, as the cyclization reactions of polyolefinic carbocations found in plant metabolism give rise to over 20000 polycyclic base structures.[93] This great structural diversity has been valuable in the discovery of antiparasitic compounds throughout history, such as the case of artemisinin and malaria.[94] Shikimate-phenylpropanoids are another common compound type in NP databases with a high presence of plant-derived compounds, like NAPRORE-CR, as they are vital for pest resistance (tannins), reproduction (pigments such as flavonoids and anthocyanins), and structural stability (biopolymers such as lignin and suberin) of plants.[95] Finally, the

notable presence of fatty acids in the database can be associated with the multiple studies on essential oils of several plant genera (e.g., *Croton, Lippia, Ocotea*) and compound extraction from fruits such as berries (e.g., *Piper nigrum, Psidium friedrichsthalianum).* A small portion of compounds could not be classified by NPClassifier (2,2%), with most structures corresponding to benzofurans and phenolic compounds.

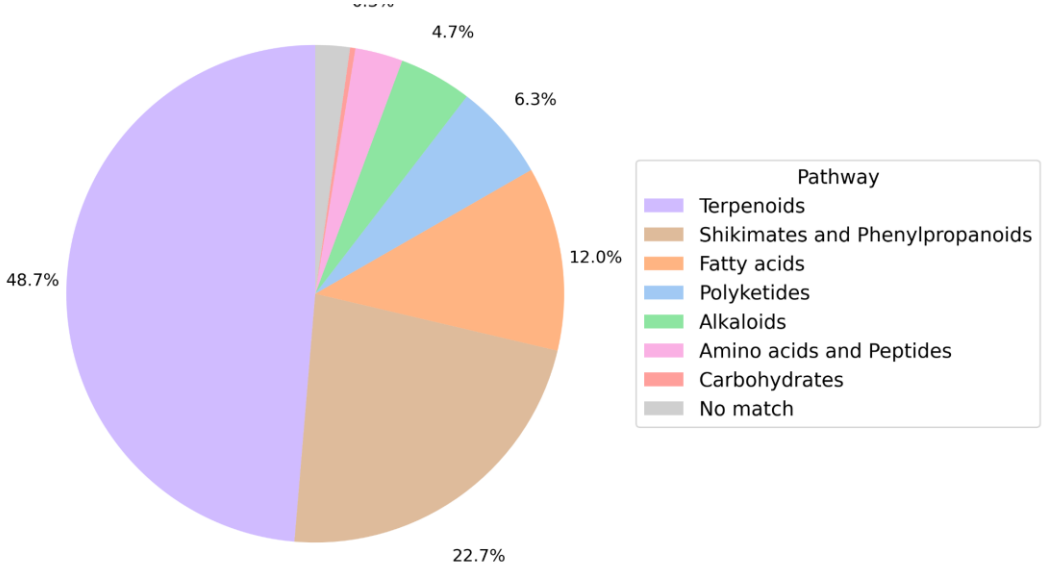


Figure 2. Structural classification of NAPRORE-CR compounds by biosynthetic pathway, according to the NPClassifier ontology.

The predominant superclasses follow the same pattern observed in the biosynthetic pathways, since they correspond to subcategories of terpenoids (sesquiterpenoids, monoterpenoids, diterpenoids), shikimate-phenylpropanoids (flavonoids, phenolic acids, phenylpropanoids), and fatty acids (fatty acyls, fatty esters). The most frequently found classes (sesquiterpenes and wax monoesters) denote the presence of a high number of compounds derived from essential oils and volatile fractions of Costa Rican flora studied by Cicció,[96–98] which comprise a notable portion of the manuscripts included in the current version of NAPRORE-CR. Previous analysis of the LANaPDB database, which included a preliminary set of NAPRORE-CR compounds, revealed the predominance of certain fragment combinations in the molecules of the Costa Rican NP space, suggesting an uneven representation of plants versus other types of organisms (e.g., bacteria, fungi, animals). Future updates should aim to diversify the chemical space by incorporating compounds derived from species found in other taxonomic kingdoms, thereby improving its representativeness.

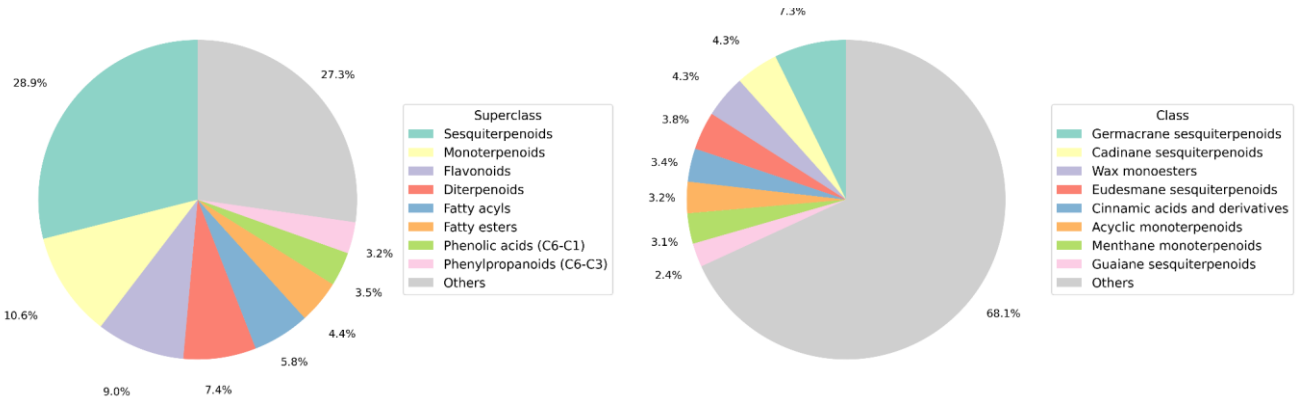


Figure 3. Structural classification of NAPRORE-CR compounds according to the NPClassifier ontology: a) Superclasses; b) Classes.

3.3. Chemical property spaces

3.3.1. Physicochemical and structural properties

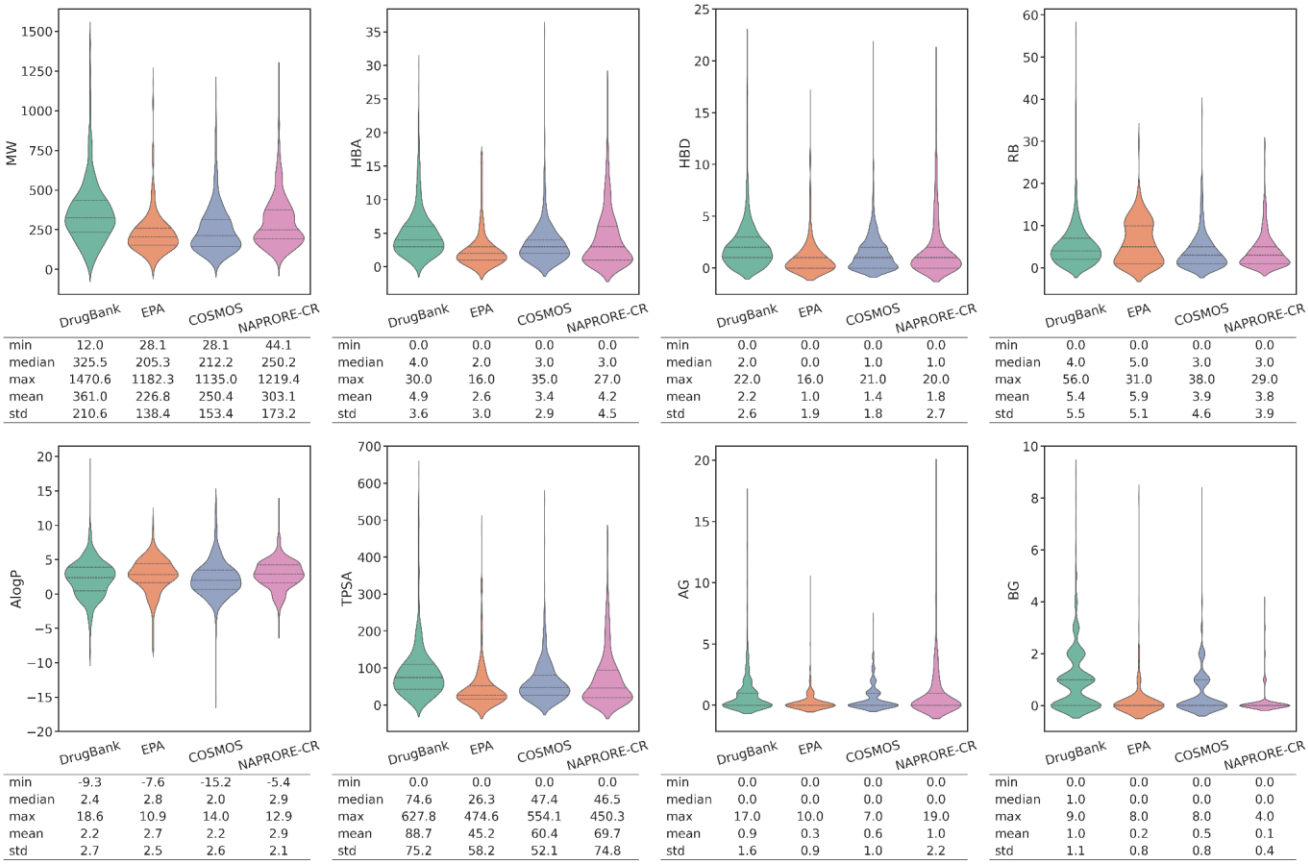


Figure 4. Violin plots for the selected structural and physicochemical descriptors of NAPRORE-CR and the reference data sets (n = 2521 DrugBank; 320 EPA; 961 COSMOS, 1161 NAPRORE-CR). Dotted lines indicate the first and third quartiles, dashed lines indicate the median values. A summary of statistical parameters is included below each plot. MW: molecular weight; HBA: hydrogen bond acceptors; HBD: hydrogen bond donors; RB: rotatable bonds; AlogP: atomic octanol/water partition coefficient (calculated); TPSA: total polar surface area; AG: acid groups; BG: basic groups.

Figure 4 shows the value distributions of the selected properties to represent the size (MW), flexibility (RB), polarity (AlogP, TPSA, HBA, HBD), and ionizability (AG, BG) of the compounds in the NAPRORE-CR and reference spaces. The MW plots showed that the NAPRORE-CR compounds have a value distribution more similar to the pesticides and cosmetics reference sets, in contrast with the approved drugs set. The median and mean values (250.2 g/mol and 303.1 g/mol, respectively) are the second highest of the analyzed sets, second only to DrugBank (325.5 g/mol and 361.0 g/mol, respectively). The abundance of molecules with a higher MW in the approved drugs data set might be derived from the presence of synthetic compounds with complex and diverse structures. Overall, NAPRORE-CR compounds are more similar to the cosmetic and natural pesticide spaces in terms of molecular size.

In terms of flexibility, the RB violin plots showed that NAPRORE-CR has a similar distribution and central tendency measures (median 3.0; mean 3.8) to the COSMOS database compounds (median 3.0; mean 3.9). Both data sets showed their maximum frequency near the first quartile and an unimodal distribution. Although it was expected to find a similar distribution in the EPA data set, as it is composed of NPs used as pesticides, it showed a bimodal distribution with a more uniform value distribution up to 15 RB. This result could be traced to the low sample size of this dataset, as it was the smallest space analyzed, which might have resulted in an underrepresented chemical space.

Referring to the polarity measurements, the AlogP showed a higher similarity between the NAPRORE-CR space and the natural pesticides data set, as their central tendency measures are the closest among the compared sets (NAPRORE-CR: median 2.9; mean 2.9; EPA: median 2.8; mean 2.7). As previously commented, this was expected, as these two sets are entirely composed of NPs. The TPSA descriptor showed similar value profiles for NAPRORE-CR and COSMOS, characterized by a low mode, a high frequency band exceeding 100 Å², and close median values (NAPRORE-CR: 46.5 Å²; COSMOS: 47.4 Å²). For the third and last polarity measure, the HBA and HBD plots for NAPRORE-CR were most similar to the DrugBank ones, with higher frequencies in high HBA values (>10 acceptors) that gave them higher mean values (DrugBank: 4.9; EPA: 2.6; COSMOS: 3.4; NAPRORE-CR: 4.2) compared to the other sets, and similar distributions and mean values for HBD as well (DrugBank: 2.2; NAPRORE-CR: 1.8). Depending on the selected individual descriptor, all reference data sets have similar polarity profiles when compared to NAPRORE-CR.

Ionizability showed certain similarities between NAPRORE-CR and DrugBank in terms of their acid group counts, while the EPA set is closer in terms of basic group counts. While the similarity between NAPRORE-CR and EPA can be attributed once again to the origin of their compounds, the differences from DrugBank can be explained by the structural modifications commonly performed on drug candidates to improve properties such as ligand affinity. By adding basic functional groups, these can interact with acidic amino acid residues in the protein's active site.[99] Furthermore, the biosynthetic routes found in the NAPRORE-CR compounds related to basic moieties were mainly alkaloids, polyketides, and amino acids & peptides, which together represent only a minor fraction of the data set.

Notably, the standard deviation values for these properties were significantly larger than their corresponding mean values, making it inadequate to draw conclusive observations, however, it is worth noting that the low abundance of ionizable groups in NAPRORE-CR directly correlates with classical phytochemical extraction techniques, where the use of non-polar organic solvents preferentially isolates neutral compounds.

In summary, when comparing the selected descriptors individually, NAPRORE-CR shows similarity with COSMOS' cosmetic ingredients in terms of size, flexibility, and polarity, and with EPA's natural pesticides in the molecular weight and AlogP properties. The most dissimilar dataset was the approved drugs set, especially in its MW and TPSA distributions. The presence of an essential portion of synthetic and modified molecules could explain the peculiarities of this dataset. Nonetheless, the central tendency values for all properties fell under the parameters defined by Lipinski's rule of five (MW < 500 Da, HBD < 5, HBA < 10, AlogP < 5), which suggests good potential for finding compounds with sufficient bioavailability. The natural pesticides data set's size resulted in a limitation for the analysis of its properties.

3.3.2. Synthetic accessibility, drug-like character, and molecular complexity

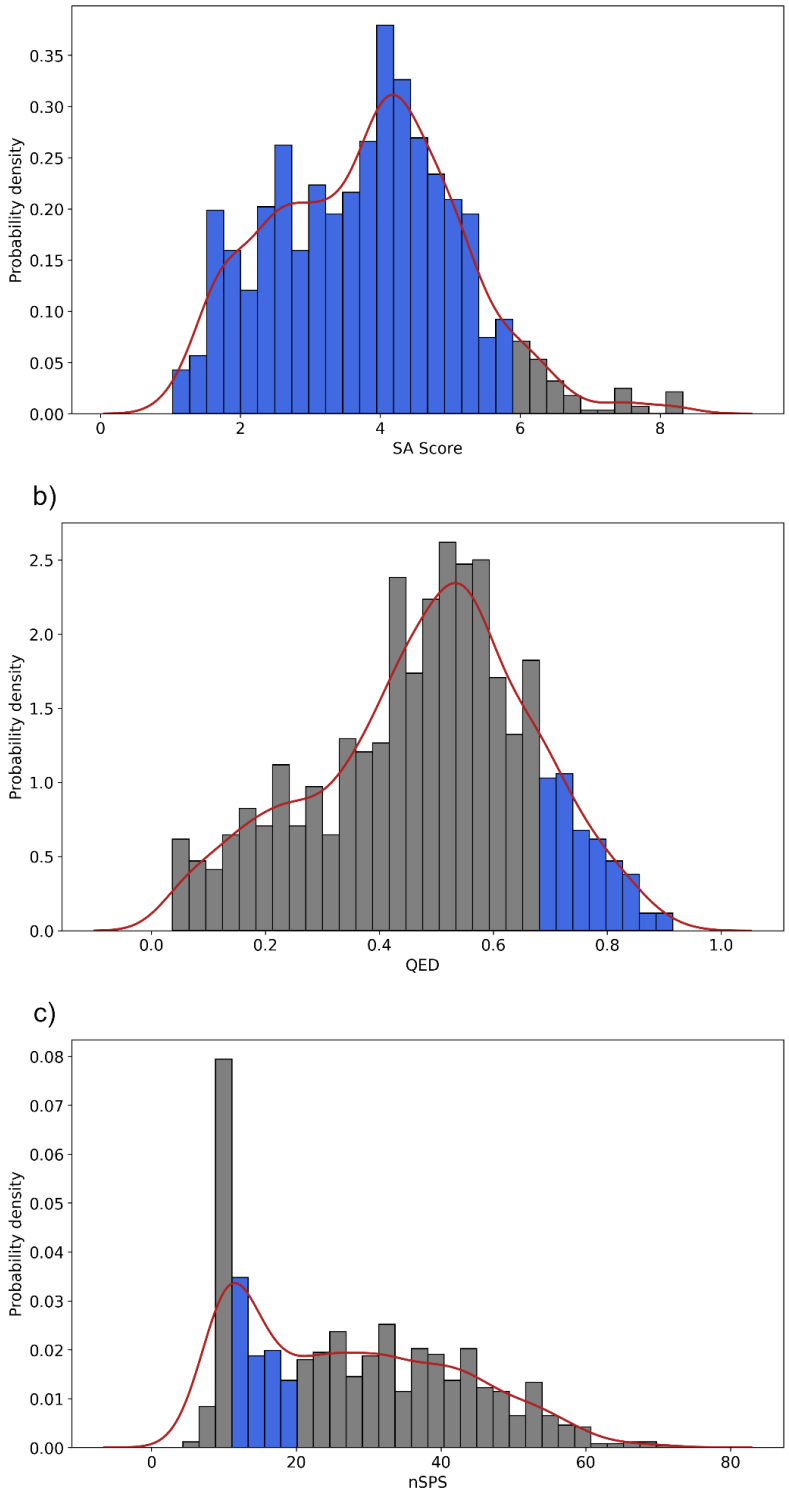


Figure 5. Normalized value distributions for the selected complex descriptors calculated for NAPRORE-CR (n = 1161 compounds): a) SA Score; b) QED; c) nSPS. All plots include a histogram (bins = 30) and a kernel density estimate (red line); the highlighted bins correspond to the value ranges associated with favorable synthetic or drug-like characteristics.

Figure 5 shows the value distributions for the three chosen descriptors to evaluate the synthetic feasibility and drug-likeness of the NAPRORE-CR chemical space. Most of the compounds in the database have SA score values equal to or under 6 (94.92%), indicating that almost the entire set of compounds should be synthetically feasible. The QED score suggests that only a small fraction of compounds (14.21%) exhibit the typical characteristics of a drug-like molecule. This is expected of a set of NPs, as their structural complexity often translates to violations of the common parameters found in already approved drugs, i.e., rule of five. Low QED values do not necessarily mean that this chemical space lacks potential for the exploration of new bioactive compounds, as many NPs are used as drugs despite having one or more violations of these empirical rules; rather, they emphasize the importance of further optimization to improve their drug-like character.[100] The nSPS distribution further emphasizes this observation, as a higher portion of the analyzed compounds (30.15%) showed values in the 10-20 range, associated with most approved drugs, compared to the QED values. When expanding the range to 20-40, as recommended by the authors of this method for exploring promising bioactive compounds with appropriate potency and selectivity, the percentage of included compounds increases to 37.73%, representing a significant proportion of the space.[81] By this analysis, it can be concluded that the NAPRORE-CR space shows promising characteristics for the potential synthesis and study of its compounds for medicinal chemistry applications.

3.4. Commercial availability

A total of 751 out of the 1161 compounds in the NAPRORE-CR database had an associated PubChem CID, which is 64.68% of the database. Out of these, 599 had at least one reported chemical vendor, which represents 79.76% of the cross-referenced compounds and 51.59% of the entire data set. The number of unique vendors for every compound can be consulted in the NAPRORE-CR Zenodo repository.[92]

3.5. Bioactivity profile

3.5.1. Experimental bioactivity profile

A total of 612 out of 1161 (52.71%) compounds of NAPRORE-CR were cross-referenced with ChEMBL and their IDs retrieved to find activity data. 577 of these molecules had associated activity information (91.01% of ChEMBL-available compounds, 49.70% of NAPRORE-CR), while only 284 had at least one assay reported as active (46.40% of ChEMBL-available compounds, 24.46% of NAPRORE-CR). The entire list of assays retrieved from ChEMBL for NAPRORE-CR compounds can be freely accessed from its Zenodo repository.[92]

3.5.2. Predicted bioactivity profile (NetInfer-based target fishing)

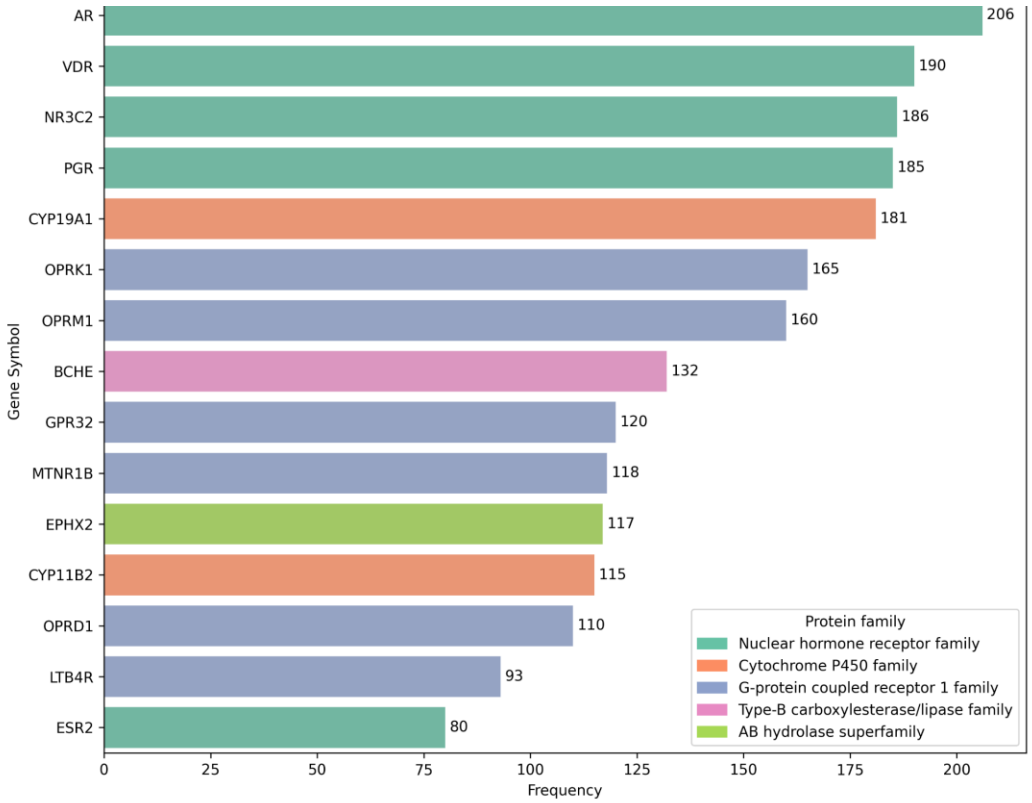
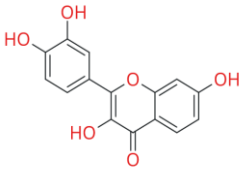
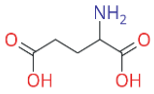
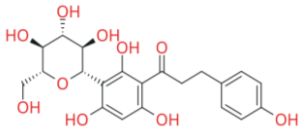


Figure 6. Frequency of occurrence of protein targets in binding affinity predictions of NAPRORE-CR compounds by the NetInfer web service. Colors correspond to protein family.

The top 15 protein targets that appeared more frequently in the top five results for every compound were tallied and are presented in Figure 6. The top four targets all correspond to the nuclear hormone receptor family, which suggests great potential for further evaluating the utility of NAPRORE-CR compounds in the study of inhibitors for treating related cancers, such as prostate and breast cancer.[101] An example of this potential is the top one result, the androgen receptor (AR), which has reports of the usage of NPs present in NAPRORE-CR, such as curcumin, apigenin, quercetin, fisetin, luteolin, kaempferol, berberine, eugenol, gingerol and ellagic acid, for treatment of castration resistance in patients with advanced stage prostate cancer.[102] As this is just a proof-of-concept to explore the possible applications of the database, the complete results sheet for the target fishing is available on the NAPRORE-CR Zenodo repository for further scrutiny.

Table 3. NetInfer predicted targets matched with ChEMBL reported activity for NAPRORE-CR compounds.

NPRCR ID	Name	Structure	Predicted target (UniProt ID)	Reported activity
NPRCR.41.69.319	Fisetin		Aurora kinase B (Q96GD4)	Inhibition of recombinant full-length human aurora B kinase expressed in baculovirus system.
NPRCR.97.146.962	Glutamic acid		Glutamate receptor ionotropic, kainate 2 (Q13002) Glutamate receptor 2 (P42262)	Agonist activity at MAG conjugated light-activated iGluR6 L439C mutant expressed in human HEK293 cells. Binding affinity to wild-type Glutamate receptor 2 S1S2 ligand binding domain.
NPRCR.99.146.1008	Nothofagin		Sodium/glucose cotransporter 2	Inhibition of human SGLT2 expressed in HEK293 cells.

Only three NAPRORE-CR compounds with reported activity in ChEMBL yielded matches with their respective top five predicted matches from the NetInfer service. Fisetin has been reported as an inhibitor of the Aurora B kinase, capable of overriding mitotic arrest and perturbing spindle checkpoint signaling, which causes mitotic inhibition and premature initiation of chromosome segregation in human cancer cells.[103] Glutamic acid is a non-essential amino acid that acts as an excitatory neurotransmitter, interacting with ligand-gated ion channels (glutamate receptors) in its anionic form (glutamate). When binding to ionotropic receptors, such as the kainate 2 subunit, membrane depolarization is induced in postsynaptic cells, causing signal transmission.[104,105] Nothofagin has been studied as a selective inhibitor of SGLT2 over SGLT1, to inhibit glucose reabsorption in the kidneys for the treatment of type 2 diabetes, while minimizing adverse effects such as hypoglycemic episodes and urinary tract infections.[106] Although this initial exploration yielded few matches between experimental and predicted activities, future studies can explore other resources for assay information and target prediction to

identify more known applications of Costa Rica's NPs and potential subspaces that may lead to compounds of medicinal interest.

3.6. Chemical space visualization

3.6.1. Physicochemical and structural properties similarly (PCA)

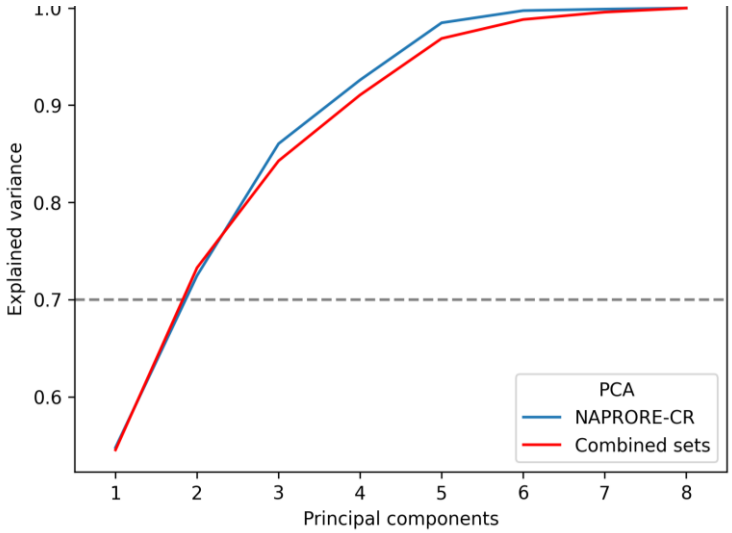


Figure 7. Cumulative explained variance in terms of the number of principal components generated for NAPRORE-CR and the combined data sets. The combined sets include NAPRORE-CR, approved drugs (DrugBank), natural pesticides (EPA), and cosmetic ingredients (COSMOS), which were obtained as described in Section 2.3.

The cumulative explained variance plot used to choose the number of dimensions for the PCA of NAPRORE-CR and the combined reference data sets is shown in Figure 7. n = 2 was selected as the optimal number of principal components in both cases, with a cumulative explained variance of 72.47% for NAPRORE-CR and 73.27% for the combined sets. Although adding a third component increased these values to over 80% in both cases, a bidimensional representation was preferred to facilitate visualisation and interpretation.

Figure 8 shows the plotted result of the NAPRORE-CR PCA, with the respective biosynthetic routes for each compound denoted by the color of its marker. No clear separation between routes is observed, indicating that the properties of compound groups are not dissimilar. However, certain routes show visible tendencies; for instance, shikimate-phenylpropanoids extend horizontally, which suggests properties such as hydrogen bond donors/acceptors and TPSA describe their variance more efficiently. In contrast, fatty acids extend vertically, in the general direction of the AlogP and rotatable bonds property vectors, two important descriptors for this class of compounds (lipophilicity, unsaturations). Other routes, such as polyketides, alkaloids, and amino acids & peptides, exhibit higher dispersion; however, as these groups have a low compound population, their tendencies are harder to determine.

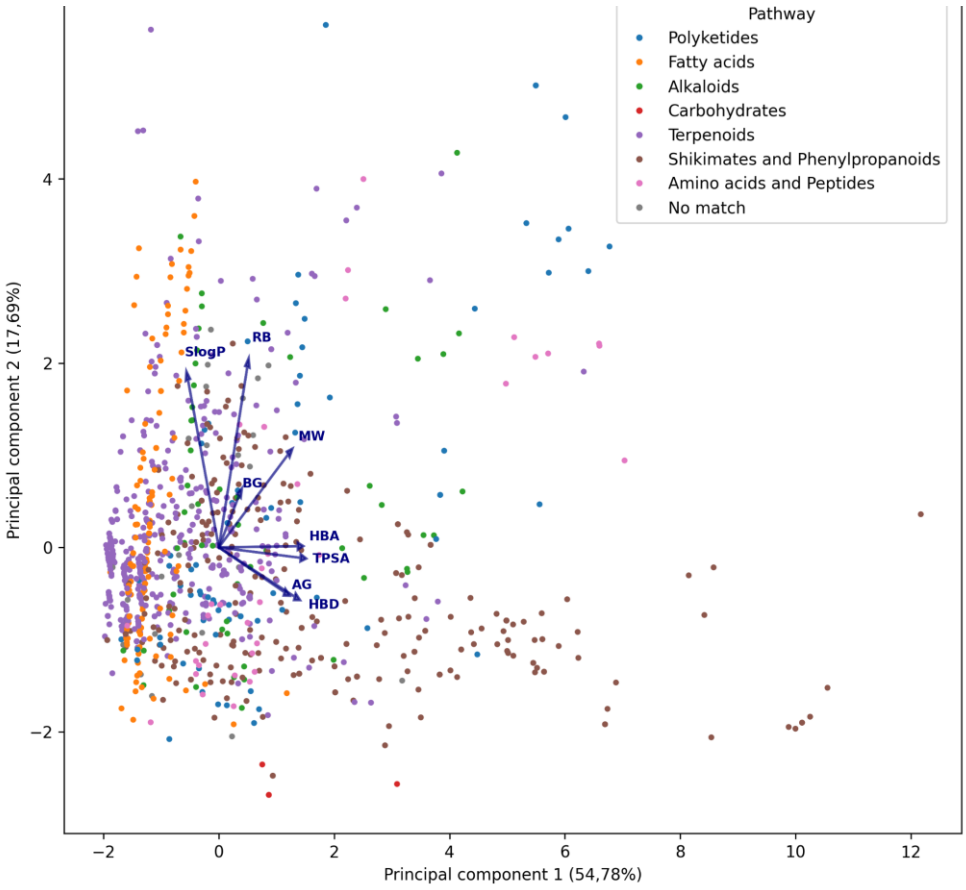


Figure 8. Chemical space visualisation of the physicochemical and structural properties of NAPRORE-CR, generated by PCA (*n* = 1161 compounds). Vectors are scaled to represent the individual contributions of each property to the principal components.

The plotted PCA obtained from the combined datasets is shown in Figure 9, with individual subplots for each dataset for readability purposes. The combined plot shows that the compounds of all four data sets occupy the same general property space with no clear separation between them. This suggests that the NAPRORE-CR compounds possess a physicochemical and structural property profile compatible for exploration in the drug, pesticide, and/or cosmetic ingredient spaces. Once again, the presence of NAPRORE-CR compounds (shikimate-phenylpropanoids) in a wide PC1 range is highlighted, which is a characteristic shared only with DrugBank. This is especially relevant when the size of these data sets is factored in: although COSMOS has a similar compound population to NAPRORE-CR, it does not show this same behaviour in the PCA, which denotes a particularly high potential for exploring this compound pathway in the Costa Rica NPs space in the search for novel drugs, as they exhibit a similar property profile to compounds related to this field.

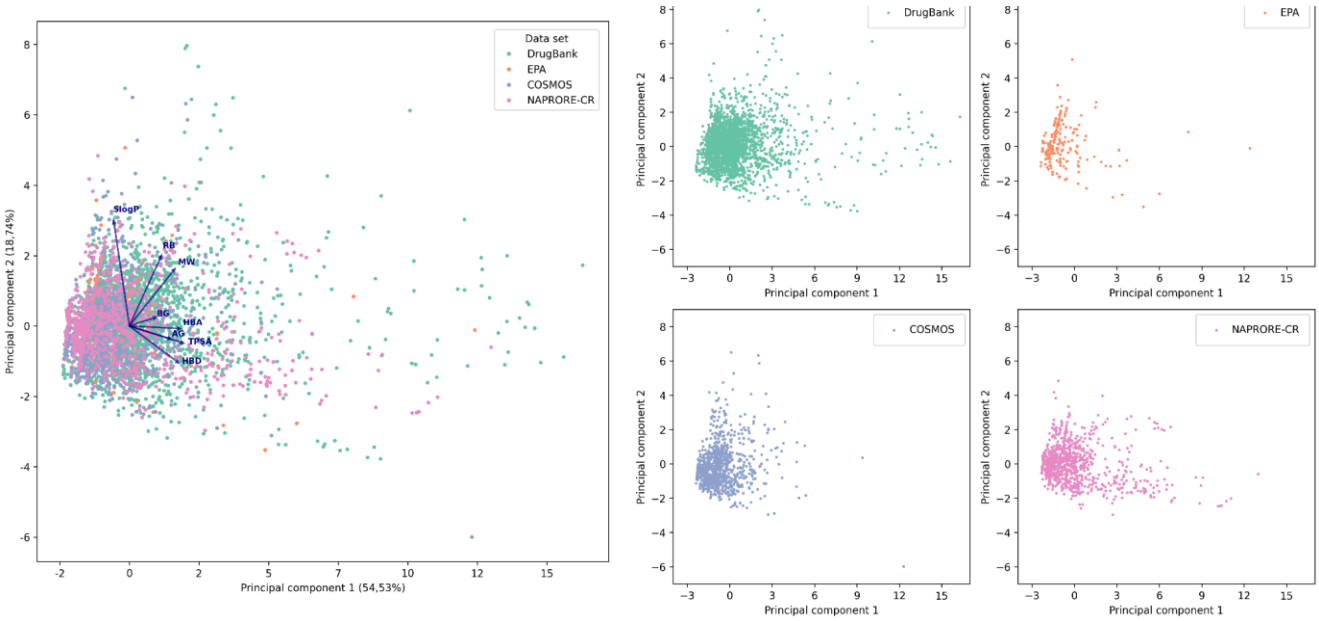


Figure 9. Chemical space visualisation of the physicochemical and structural properties of DrugBank, EPA, COSMOS, and NAPRORE-CR, generated by PCA (*n* = 4963 compounds). Vectors are scaled to represent the individual contributions of each property to the principal components. The main plot encompasses all analyzed databases; additionally, a 2x2 grid displays individual plots for each dataset.

3.6.2. Structural similarity space (TMAP)

Figure 10 contains the plotted result of the generated TMAP for NAPRORE-CR, with the corresponding biosynthetic pathways of each compound represented by the color of its marker. Contrary to the PCA results, a clear separation by pathway can be observed, associated with the capability of molecular fingerprints to represent the structural characteristics of each compound family, which originates from their biotransformation processes and the metabolites that make up their structures.

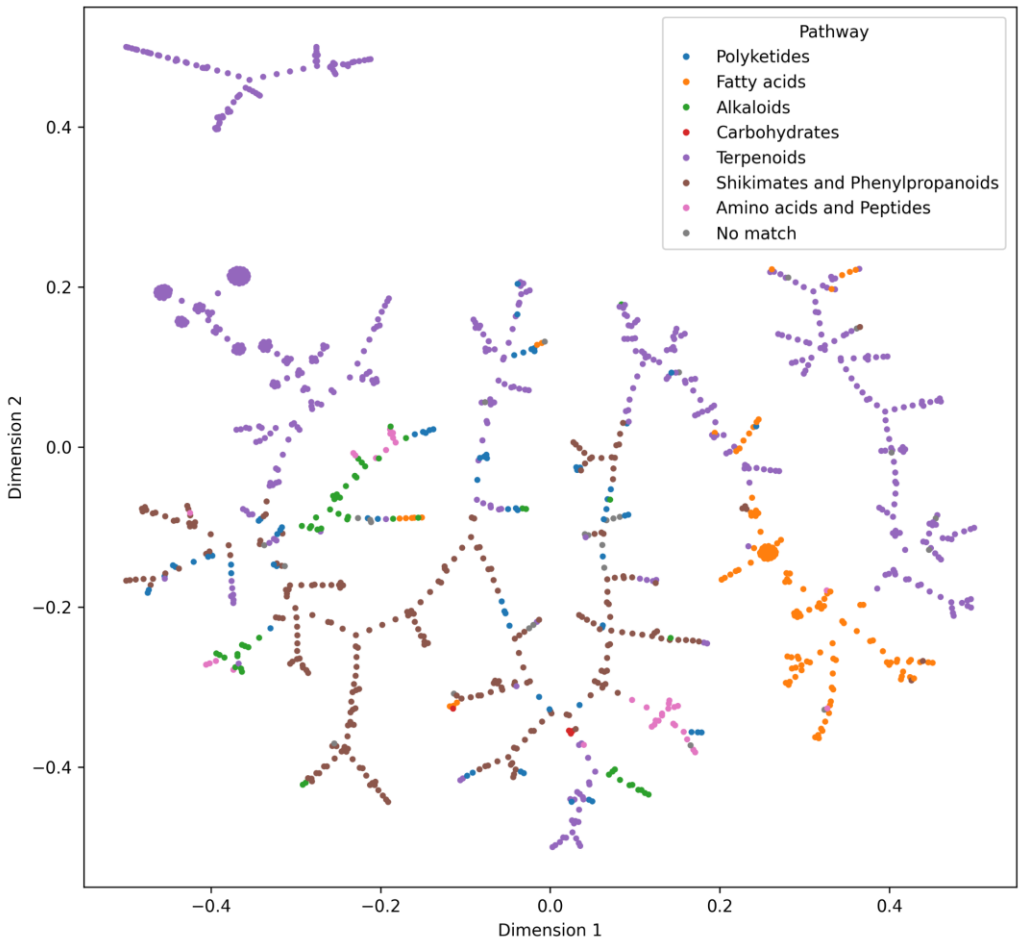


Figure 10. Chemical space visualisation of the structural similarity based on MACCS Keys molecular fingerprints of NAPRORE-CR, generated by TMAP (*n* = 1161 compounds).

The pathways with the higher number of compounds in the database show more clearly defined branches that are independent of other groups, especially in the case of shikimate-phenylpropanoids and fatty acids, which have most of their compounds grouped in a single cluster. Since terpenoids have the highest share of representation by a large margin and exhibit great scaffold diversity, as mentioned previously, they are distributed across multiple branches, divided into other clusters, including a disconnected branch. This isolated cluster is composed mostly of terpenoid classes with low representation in the data set, such as aromadendrene, tujane, trichothecene, and cubebene sesquiterpenoids. Due to the low number of similar structures, the TMAP algorithm groups them in a cluster separate from the tree, which exemplifies its capability of isolating compounds with low structural similarity instead of forcing relationships with them. Polyketides are found dispersed throughout the entire tree, typically near shikimate-phenylpropanoid branches, with which they share scaffolds featuring multiple aromatic rings. Alkaloids share space with amino acids & peptides, due to the presence of nitrogen atoms in their structures; a small fraction of glycoalkaloids grouped with terpenic glycosides was also identified. Unclassified compounds were primarily found in phenylpropanoid

branches, which coincides with the aforementioned observation about the nature of these structures in section 3.1. Overall, the TMAP visualisation associated the less represented compound groups with other biosynthetic routes by their shared functional groups and similar scaffolds, which in turn might reveal a weakness of the MACCS keys system for this implementation, as it is based on predefined substructure dictionaries.

The TMAP for the combined sets is presented in Figure 11 and helps evaluate the structural similarity of the NAPRORE-CR compounds with the reference data sets. Once again, branches with a clear majority of compounds of a single set can be identified, especially for DrugBank and NAPRORE-CR, as these are the two biggest databases analyzed. NAPRORE-CR compounds are primarily located in the top right section of the plot, where they share space with most natural pesticides and a fraction of approved drugs; they also coincide with an EPA branch in the lower right quadrant (~0.2 in dimension 1). The apparent dispersion of the COSMOS and DrugBank sets may be attributed to a higher structural diversity resulting from the presence of modified or synthetic molecules, compared to the entirely natural sets from NAPRORE-CR and EPA, which are comparatively more clustered in a single region of the space. Regardless, it is apparent that the NPs from Costa Rica share an important area of this chemical space with compounds used as natural pesticides or drugs, which suggests they possess a significant structural similarity with these data sets.

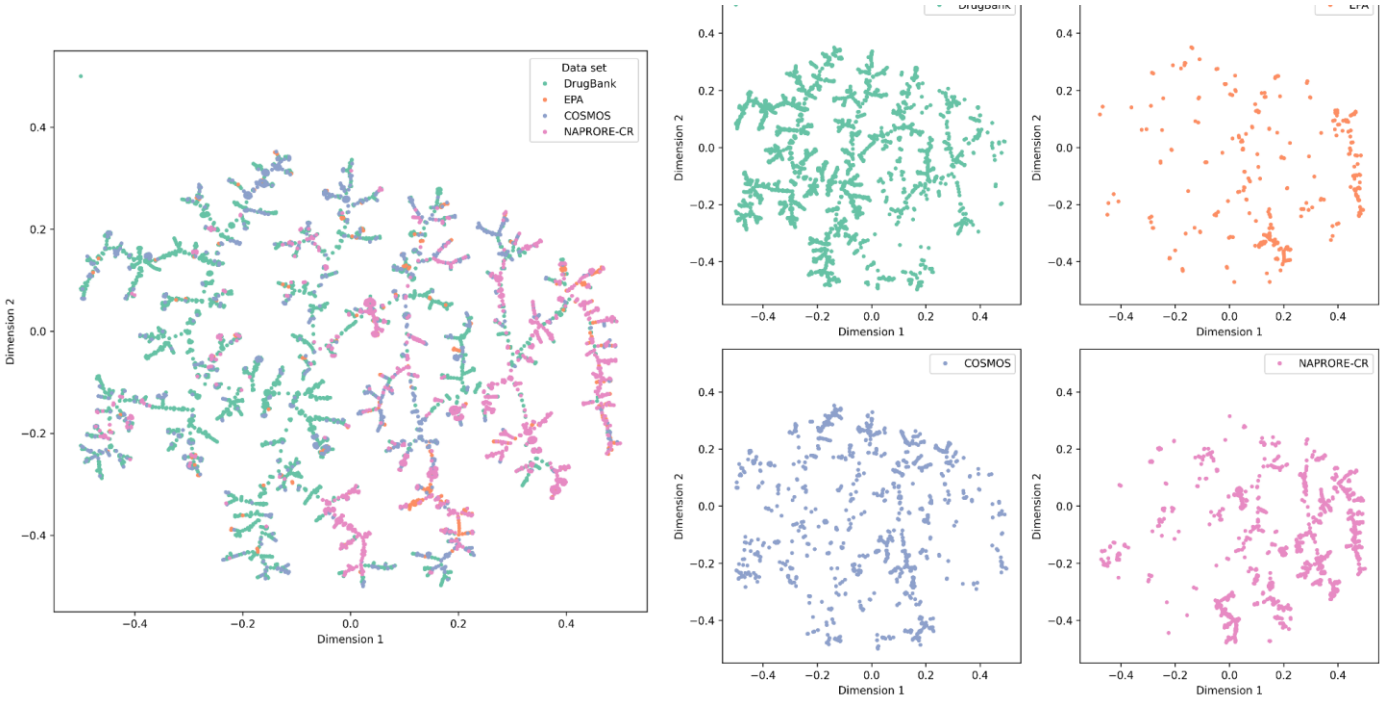


Figure 11. Chemical space visualization of the structural similarity based on MACCS Keys molecular fingerprints of DrugBank, EPA, COSMOS, and NAPRORE-CR, generated by TMAP (*n* = 4963 compounds). The main plot encompasses all analyzed databases; additionally, a 2x2 grid displays individual plots for each dataset.

3.6.3. Constellation Plot

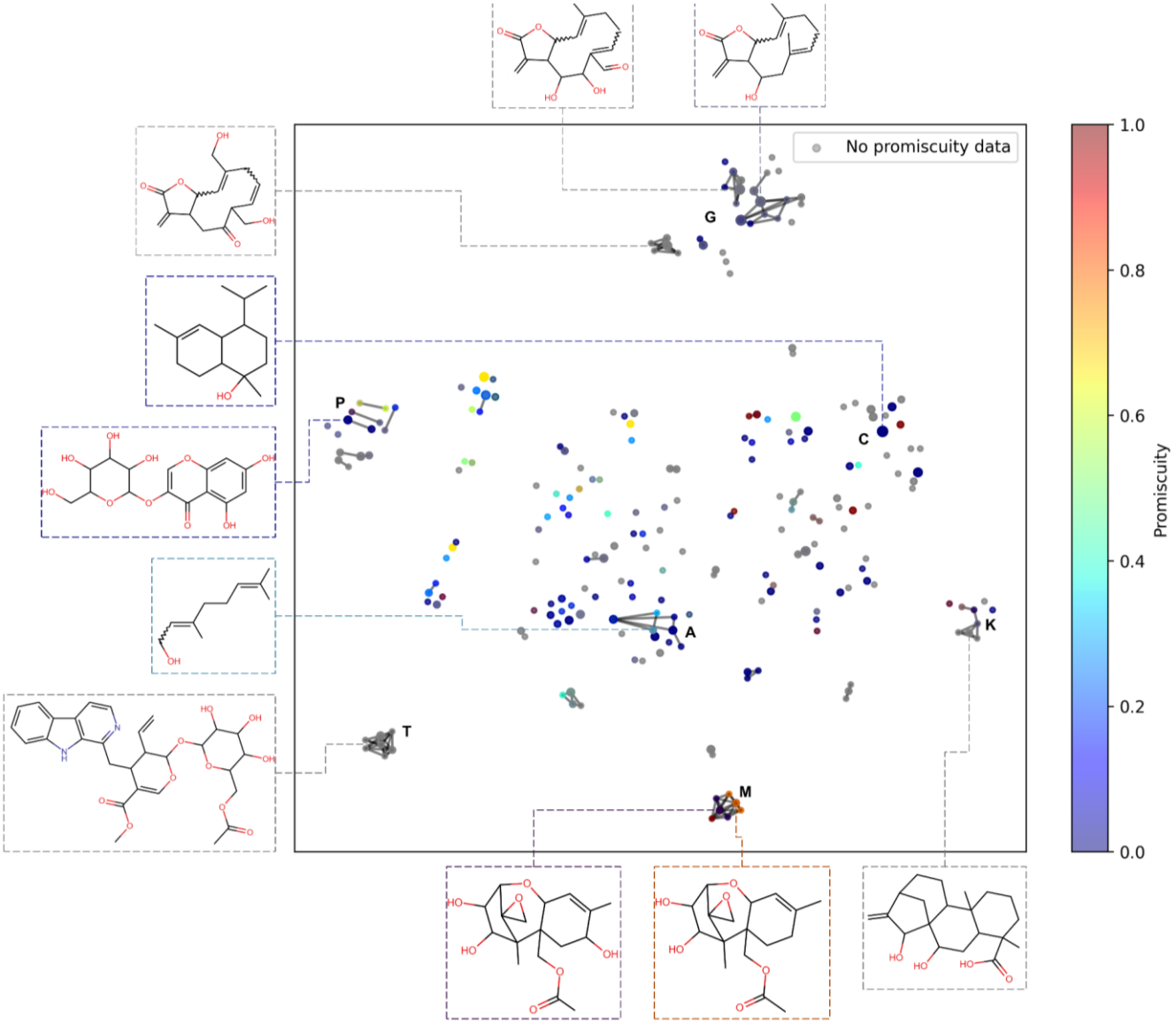


Figure 12. Constellation plot for the NAPRORE-CR data set. Marker size corresponds to the number of related compounds. Color scale indicates the promiscuity ratio calculated from activity assay information retrieved from ChEMBL; gray points indicate no activity data.

Figure 12 shows the Constellation Plot generated for the NAPRORE-CR data set. Data points represent molecular scaffolds with at least two related compounds, while marker size indicates the number of related compounds. Data points are color-coded in a continuous scale from blue (low target promiscuity) to red (high target promiscuity); gray points indicate related compounds had no activity data. Selected chemical structures are highlighted to better visualize the relationships between the analog series. Out of 1100 identified cores, 873 were removed because they had only one related compound, leaving 227 cores related to 445 out of the 1161 compounds (38.33%). Since more than half of the compounds in the database lacked associated activity information in the ChEMBL database, a significant portion of this chemical space lacks promiscuity data. Three examples of this are the series

on the bottom-left side (T), related to glycosylated tryptophan alkaloids, the series on the bottom-right side (K), composed of kaurane diterpenoids, and the series on the top-center side (G), formed by germacrane sesquiterpenoids, the single most represented class of compound in the database. The absence of activity data could represent an opportunity to discover applications for unexplored compound families.

A noteworthy series is the one found at the bottom-center of the plot (M), represented by a trichothecene sesquiterpenoid scaffold and corresponding to a series of mycotoxins, which exhibits significant promiscuity differences caused by the addition or removal of hydroxyl groups and/or acetyl/hydroxyl group replacements. This apparent promiscuity cliff exemplifies the utility of Constellation Plots for exploring the relationships between structural analog series and activity,[85], and it represents a possible research line worth considering in future studies. Another notable series is located at the mid-center of the plot (A). It is one of the most numerous, as it is related to acyclic monoterpenoids, one of the most abundant classes of compounds in the database. Similarly, marker C corresponds to cadinane sesquiterpenoids. The clustered pairs on the mid-left area (P) correspond to glycosylated aromatic polycycles, such as polyketides and flavonoids, with low to middling promiscuities. As these compound groups have shown great importance in the discovery of antitumoral, antimicrobial, antiparasitic, and other classes of medicinal agents,[107,108] expanding on these pathways could prove promising for the discovery of selective drug-like entities.

3.7. Global diversity analysis

The cyclic system recovery (CSR) curves, generated to quantify scaffold diversity in the four studied data sets, are shown in Figure 13. These plots suggest that the sets with the highest scaffold diversity are DrugBank and NAPRORE-CR, as these two have the lowest areas under their respective curves compared to EPA and COSMOS. Using this measure, a dataset with a unique scaffold for every molecule (maximum diversity) would result in a straight line in the CSR plot and an AUC of 0.5, while less diverse sets would show higher values.[90]

The CDP obtained from the intra-set Tanimoto coefficient and AUC calculations for every set is shown in Figure 14. As the Tanimoto coefficient measures structural similarity, lower values are expected for more diverse compound sets. The CDP shows that NAPRORE-CR and DrugBank possess a higher scaffold diversity and a lower molecular fingerprint diversity (NAPRORE-CR: mean Tanimoto coefficient = 0.32; AUC = 0.74. DrugBank: mean Tanimoto coefficient = 0.28; AUC = 0.72) compared with COSMOS and EPA (COSMOS: mean Tanimoto coefficient = 0.18; AUC = 0,83. EPA: mean Tanimoto coefficient = 0.24; AUC = 0.85). This suggests that the Costa Rican natural products space exhibits structural diversity comparable to that of the approved drugs space, despite being considerably smaller in terms of the number of compounds. The primary difference between these two sets lies in the higher molecular fingerprint diversity in DrugBank, which can be attributed to the aforementioned

structural modifications commonly applied to drug candidates to enhance their properties, bioavailability, and biological activity.

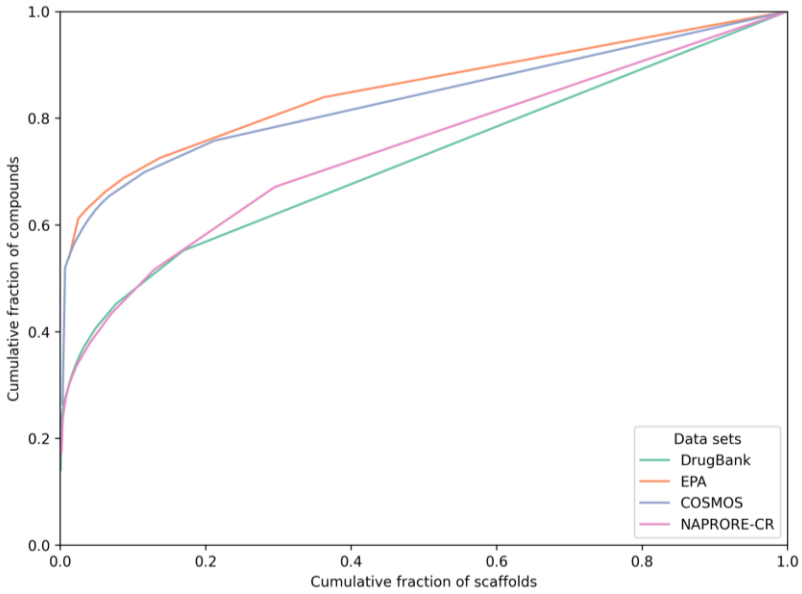


Figure 13. Cyclic recovery system curves for the calculation of a scaffold diversity metric for the analyzed compound databases

The COSMOS data set, although of similar size to NAPRORE-CR, exhibits higher molecular fingerprint diversity, attributed to the variety of types and sources of compounds included in the database (botanical, animal-derived, synthetic, among others). Due to this, COSMOS also includes halogenated species, siloxanes, and other substances with heteroatoms not present in NAPRORE-CR. The lower scaffold diversity of COSMOS can be associated with a higher proportion of molecules that have an aliphatic chain or benzene as their scaffold. These two structures represent more than half of the molecules in COSMOS (51.9%), versus 24.0% for DrugBank and 23.6% for NAPRORE-CR. The EPA data set shares these characteristics with COSMOS: higher source diversity compared to NAPRORE-CR (vegetal, animal, microorganisms) and lower scaffold diversity dominated by the same two scaffolds (61.6% of its compounds, with 54.4% corresponding to aliphatic chains), although this case is harder to assess due to the low sample size.

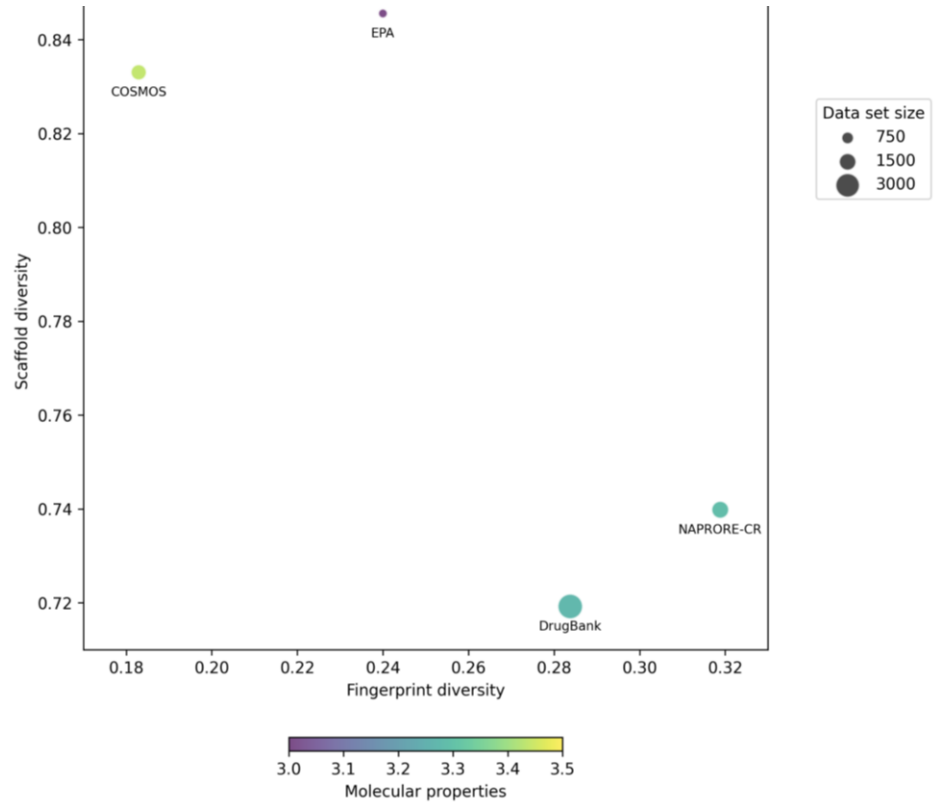


Figure 14. Consensus diversity plot of DrugBank, EPA, COSMOS, and NAPRORE-CR, based on MACCS Keys molecular fingerprints and Murcko molecular scaffolds. X-axis values represent molecular fingerprint diversity (mean Tanimoto coefficient, with lower values indicating higher diversity), and Y-axis values correspond to molecular scaffold diversity (AUC from the cyclic recovery system curves, with lower values indicating higher diversity). Marker size represents the number of compounds in each dataset, while the continuous color scale relates to the physicochemical and structural properties calculated in Section 3.3.1 (mean Euclidean intra-set distance).

4. Conclusions

The first version of the Natural Products repository of Costa Rica was introduced, featuring 1161 compounds obtained from 105 manuscripts, including scientific papers and theses from the country's leading public universities and research centers dedicated to the study of NPs. Thus, NAPRORE-CR aligns with open science initiatives and adheres to the FAIR principles for data sharing, making the rich molecular diversity of Costa Rica's biodiversity accessible to the global scientific community.

According to the NPClassifier service, the most frequent compound families in the database are terpenoids, shikimates and phenylpropanoids, and fatty acids; their corresponding superclasses and classes demonstrate the predominance of plants as the primary source of compounds included in this resource. The physicochemical and structural profiles of the data set showed similarities with the COSMOS cosmetic ingredients database in terms of molecular size, flexibility and polarity, and with EPA natural pesticides in the MW and AlogP properties. Overall, most of the NAPRORE-CR data set complies with empirical rules for drug oral bioavailability (Lipinski's Ro5). Calculated complexity

descriptors indicate that although the structural complexity associated with NPs might prove challenging, the space contains a significant amount of synthetically feasible, potentially bioactive compounds for medicinal chemistry applications. Cross-referencing NAPRORE-CR with the PubChem and ChEMBL databases revealed that 51.59% of its compounds are commercially available, and that 49.70% have associated bioactivity assay data; however, only 24.26% of these compounds showed positive activity results. Predicted target affinity suggests that the nuclear hormone receptor family is a potential area of study of the Costa Rican NPs space, with reported experimental evidence of compounds present in NAPRORE-CR that exhibit activity as treatments for prostate cancer.

The visual representation of the chemical space with PCA showed that NAPRORE-CR shares a similar property space with all the analyzed reference datasets (DrugBank, COSMOS, EPA), with a notable tendency for the shikimate-phenylpropanoids pathway to extend into a region of space only shared with approved drugs. The TMAP visualization revealed structural similarities with compounds used as natural pesticides or drugs, suggesting the possibility of deriving analog structures with applications in these fields. The Constellation Plot highlighted analog series with potential for further study, including an apparent promiscuity cliff in a series of mycotoxins, as well as several polyketides and flavonoids with low target promiscuity. The consensus diversity plot revealed that the NAPRORE-CR database is comparable to the DrugBank database in terms of scaffold, fingerprint, and property diversity, despite being significantly smaller in terms of the number of compounds. We expect NAPRORE-CR to aid current work and inspire future research lines for the rational utilization of the country's natural wealth.

Data availability statement

All the dataset files used to generate the plots and chemical space visualizations for NAPRORE-CR and the compiled EPA pesticides list are freely available at the corresponding Zenodo repository: https://doi.org/10.5281/zenodo.7858061. The COSMOS dataset can be freely obtained through its website after registration: https://cosmosdb.eu/cosmosdb.v2. The DrugBank dataset can be accessed for academic purposes, by applying for a license through the official website: https://go.drugbank.com. The Python scrips built to generate the plots and visualizations can be freely obtained from the CBio^3 Group's GitHub repository: https://github.com/cbio3lab/.

Acknowledgements

D. A. A-J. thanks the Secretaría de la Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico, for the scholarship number 4047090. W. Z. R. thanks to the students of the chemoinformatics course for their collaboration in data collection and the Vice Chancellor for Research of the University of Costa Rica for grant via the research project 115-C2-126. W. Z. R. also extends sincere thanks to all the Costa Rican researchers who worked on the separation and structural elucidation of the compounds present in the database, especially to the memory of his former mentor M.Sc. Victor Castro, as without his valuable work, this repository would not be possible.

References

(1) Martinez-Mayorga, K.; Madariaga-Mazon, A.; Medina-Franco, J. L.; Maggiora, G. The Impact of Chemoinformatics on Drug Discovery in the Pharmaceutical Industry. *Expert Opin. Drug Discov.* **2020**, *15* (3), 293–306. https://doi.org/10.1080/17460441.2020.1696307.

(2) 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. https://doi.org/10.1016/j.ejmech.2021.113705.

(3) Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Rathman, J. F.; Richarz, A.-N.; Yang, C. A Review of in Silico Toxicology Approaches to Support the Safety Assessment of Cosmetics-Related Materials. *Comput. Toxicol.* **2022**, *21*, 100213. https://doi.org/10.1016/j.comtox.2022.100213.

(4) Zhao, W.; Huang, Y.; Hao, G.-F. Pesticide Informatics Expands the Opportunity for Structure-Based Molecular Design and Optimization. *Adv. Agrochem* **2022**, *1* (2), 139–147. https://doi.org/10.1016/j.aac.2022.11.006.

(5) Saifi, I.; Bhat ,Basharat Ahmad; Hamdani, S. S.; Bhat, U. Y.; Lobato-Tapia, C. A.; Mir, M. A.; Dar, T. U. H.; Ganie, S. A. Artificial Intelligence and Cheminformatics Tools: A Contribution to the Drug Development and Chemical Science. *J. Biomol. Struct. Dyn.* **2024**, *42* (12), 6523–6541. https://doi.org/10.1080/07391102.2023.2234039.

(6) Miller, M. A. Chemical Database Techniques in Drug Discovery. *Nat. Rev. Drug Discov.* **2002**, *1* (3), 220. https://doi.org/10.1038/nrd745.

(7) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2025 Update. *Nucleic Acids Res.* **2025**, *53* (D1), D1516–D1525. https://doi.org/10.1093/nar/gkae1059.

(8) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44* (D1), D1045–D1053. https://doi.org/10.1093/nar/gkv1072.

(9) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(10) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.

(11) Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (16), 6987–7002. https://doi.org/10.1021/jm300501t.

(12) Gong, L.; Zhang, R.; Han, M.; Hu, Q.-N. CCIBP: A Comprehensive Cosmetic Ingredients

Bioinformatics Platform. *Bioinformatics* **2023**, *39* (7). https://doi.org/10.1093/bioinformatics/btad416.

(13) Yang, C.; Cronin, M. T. D.; Arvidson, K. B.; Bienfait, B.; Enoch, S. J.; Heldreth, B.; Hobocienski, B.; Muldoon-Jacobs, K.; Lan, Y.; Madden, J. C.; Magdziarz, T.; Marusczyk, J.; Mostrag, A.; Nelms, M.; Neagu, D.; Przybylak, K.; Rathman, J. F.; Park, J.; Richarz, A.-N.; Richard, A. M.; Ribeiro, J. V.; Sacher, O.; Schwab, C.; Vitcheva, V.; Volarath, P.; Worth, A. P. COSMOS next Generation – A Public Knowledge Base Leveraging Chemical and Biological Data to Support the Regulatory Assessment of Chemicals. *Comput. Toxicol.* **2021**, *19*, 100175. https://doi.org/10.1016/j.comtox.2021.100175.

(14) Wang, D.; Deng, H.; Zhang, T.; Tian, F.; Wei, D. Open Access Databases Available for the Pesticide Lead Discovery. *Pestic. Biochem. Physiol.* **2022**, *188*, 105267. https://doi.org/10.1016/j.pestbp.2022.105267.

(15) Yang, C.; Cheeseman, M.; Rathman, J.; Mostrag, A.; Skoulis, N.; Vitcheva, V.; Goldberg, S. A New Paradigm in Threshold of Toxicological Concern Based on Chemoinformatics Analysis of a Highly Curated Database Enriched with Antimicrobials. *Food Chem. Toxicol.* **2020**, *143*, 111561. https://doi.org/10.1016/j.fct.2020.111561.

(16) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

(17) Liu, J.-K. Natural Products in Cosmetics. *Nat. Prod. Bioprospecting* **2022**, *12* (1), 40. https://doi.org/10.1007/s13659-022-00363-y.

(18) Desam, N. R.; Al-Rajab, A. J. The Importance of Natural Products in Cosmetics. In *Bioactive Natural Products for Pharmaceutical Applications*; Pal, D., Nayak, A. K., Eds.; Advanced Structured Materials; Springer International Publishing: Cham, 2021; pp 643–685. https://doi.org/10.1007/978-3-030-54027-2_19.

(19) Espinosa-Leal, C. A.; Garcia-Lara, S. Current Methods for the Discovery of New Active Ingredients from Natural Products for Cosmeceutical Applications. *Planta Med.* **2019**, *85* (7), 535–551. https://doi.org/10.1055/a-0857-6633.

(20) Martins, A.; Vieira, H.; Gaspar, H.; Santos, S. Marketed Marine Natural Products in the Pharmaceutical and Cosmeceutical Industries: Tips for Success. *Mar. Drugs* **2014**, *12* (2), 1066–1101. https://doi.org/10.3390/md12021066.

(21) Cantrell, C. L.; Dayan, F. E.; Duke, S. O. Natural Products As Sources for New Pesticides. *J. Nat. Prod.* **2012**, *75* (6), 1231–1242. https://doi.org/10.1021/np300024u.

(22) Sparks, T. C.; Sparks, J. M.; Duke, S. O. Natural Product-Based Crop Protection Compounds—Origins and Future Prospects. *J. Agric. Food Chem.* **2023**, *71* (5), 2259–2269. https://doi.org/10.1021/acs.jafc.2c06938.

(23) Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020.

J. Cheminformatics **2020**, *12* (1), 20. https://doi.org/10.1186/s13321-020-00424-9.

(24) Chandrasekhar, V.; Rajan, K.; Kanakam, S. R. S.; Sharma, N.; Weißenborn, V.; Schaub, J.; Steinbeck, C. COCONUT 2.0: A Comprehensive Overhaul and Curation of the Collection of Open Natural Products Database. *Nucleic Acids Res.* **2025**, *53* (D1), D634–D643. https://doi.org/10.1093/nar/gkae1063.

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

(26) van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; McCadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M.-L.; Duncan, K. R.; Linington, R. G. The Natural Products Atlas 2.0: A Database of Microbially-Derived Natural Products. *Nucleic Acids Res.* **2022**, *50* (D1), D1317–D1323. https://doi.org/10.1093/nar/gkab941.

(27) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. *PLOS ONE* **2011**, *6* (1), e15939. https://doi.org/10.1371/journal.pone.0015939.

(28) Simoben, C. V.; Qaseem, A.; Moumbock, A. F. A.; Telukunta, K. K.; Günther, S.; Sippl, W.; Ntie-Kang, F. Pharmacoinformatic Investigation of Medicinal Plants from East Africa. *Mol. Inform.* **2020**, *39* (11), 2000163. https://doi.org/10.1002/minf.202000163.

(29) Vivek-Ananth, R. P.; Mohanraj, K.; Sahoo, A. K.; Samal, A. IMPPAT 2.0: An Enhanced and Expanded Phytochemical Atlas of Indian Medicinal Plants. *ACS Omega* **2023**, *8* (9), 8827–8845. https://doi.org/10.1021/acsomega.3c00156.

(30) Raven, P. H.; Gereau, R. E.; Phillipson, P. B.; Chatelain, C.; Jenkins, C. N.; Ulloa Ulloa, C. The Distribution of Biodiversity Richness in the Tropics. *Sci. Adv.* **2020**, *6* (37), eabc6228. https://doi.org/10.1126/sciadv.abc6228.

(31) Mittermeier, R. A.; Turner, W. R.; Larsen, F. W.; Brooks, T. M.; Gascon, C. Global Biodiversity Conservation: The Critical Role of Hotspots. In *Biodiversity Hotspots*; Zachos, F. E., Habel, J. C., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 3–22. https://doi.org/10.1007/978-3-642-20992-5_1.

(32) Kingston, D. G. I. Modern Natural Products Drug Discovery and Its Relevance to Biodiversity Conservation. *J. Nat. Prod.* **2011**, *74* (3), 496–511. https://doi.org/10.1021/np100550t.

(33) Pilón-Jiménez, B. A.; Saldívar-González, F. I.; Díaz-Eufracio, B. I.; Medina-Franco, J. L. BIOFACQUIM: A Mexican Compound Database of Natural Products. *Biomolecules* **2019**, *9* (1), 31. https://doi.org/10.3390/biom9010031.

(34) Olmedo, D.; Medina-Franco, J. Chemoinformatic Approach: The Case of Natural Products of Panama; 2019. https://doi.org/10.5772/intechopen.87779.

(35) Rodríguez-Pérez, J. R.; Valencia-Sánchez, H. A.; Mosquera-Martínez, O. M.; Gómez-Garcia, A.;

Medina-Franco, J. L.; Cortes-Hernández, H. F. NPDBEjeCol: A Natural Products Database from Colombia. *ACS Omega* **2025**, *10* (9), 9778–9792. https://doi.org/10.1021/acsomega.5c00936.

(36) Pilon, A. C.; Valli, M.; Dametto, A. C.; Pinto, M. E. F.; Freire, R. T.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. NuBBEDB: An Updated Database to Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep.* **2017**, *7* (1), 7215. https://doi.org/10.1038/s41598-017-07451-x.

(37) Barazorda-Ccahuana, H. L.; Ranilla, L. G.; Candia-Puma, M. A.; Cárcamo-Rodriguez, E. G.; Centeno-Lopez, A. E.; Davila-Del-Carpio, G.; Medina-Franco, J. L.; Chávez-Fumagalli, M. A. PeruNPDB: The Peruvian Natural Products Database for in Silico Drug Screening. *Sci. Rep.* **2023**, *13* (1), 7577. https://doi.org/10.1038/s41598-023-34729-0.

(38) Martínez Heredia, L.; Quispe, P. A.; Fernández, J. F.; Lavecchia, M. J. NaturAr: A Collaborative, Open-Source Database of Natural Products from Argentinian Biodiversity for Drug Discovery and Bioprospecting. *J. Chem. Inf. Model.* **2025**. https://doi.org/10.1021/acs.jcim.4c01507.

(39) Gómez-García, A.; Acuña Jiménez, D. A.; Zamora, W. J.; Barazorda-Ccahuana, H. L.; Chávez-Fumagalli, M. Á.; Valli, M.; Andricopulo, A. D.; Bolzani, V. da S.; Olmedo, D. A.; Solís, P. N.; Núñez, M. J.; Rodríguez Pérez, J. R.; Valencia Sánchez, H. A.; Cortés Hernández, H. F.; Mosquera Martinez, O. M.; Medina-Franco, J. L. Latin American Natural Product Database (LANaPDB): An Update. *J. Chem. Inf. Model.* **2024**. https://doi.org/10.1021/acs.jcim.4c01560.

(40) Rojas, T. B.; Acuña, V. O. Biodiversidad en cifras: avances en el conocimiento de especies en Costa Rica. *Biocenosis* **2021**, *32* (2). https://doi.org/10.22458/rb.v32i2.3899.

(41) Sistema Nacional de Áreas de Conservación SINAC. *IV Informe de País al Convenio Sobre La Diversidad Biológica*; GEF-PNUD: San José, Costa Rica. https://www.cbd.int/doc/world/cr/cr-nr-04-es.pdf (accessed 2025-06-30).

(42) Cruz, P. G.; Fribley, A. M.; Miller, J. R.; Larsen, M. J.; Schultz, P. J.; Jacob, R. T.; Tamayo-Castillo, G.; Kaufman, R. J.; Sherman, D. H. Novel Lobophorins Inhibit Oral Cancer Cell Growth and Induce Atf4- and Chop-Dependent Cell Death in Murine Fibroblasts. *ACS Med. Chem. Lett.* **2015**, *6* (8), 877–881. https://doi.org/10.1021/acsmedchemlett.5b00127.

(43) Cao, S.; McMillin, D. W.; Tamayo, G.; Delmore, J.; Mitsiades, C. S.; Clardy, J. Inhibition of Tumor Cells Interacting with Stromal Cells by Xanthones Isolated from a Costa Rican Penicillium Sp. *J. Nat. Prod.* **2012**, *75* (4), 793–797. https://doi.org/10.1021/np2009863.

(44) Vilariño, M.; García-Sanmartín, J.; Ochoa-Callejero, L.; López-Rodríguez, A.; Blanco-Urgoiti, J.; Martínez, A. Macrocybin, a Natural Mushroom Triglyceride, Reduces Tumor Growth In Vitro and In Vivo through Caveolin-Mediated Interference with the Actin Cytoskeleton. *Molecules* **2020**, *25* (24). https://doi.org/10.3390/MOLECULES25246010.

(45) Park, S. R.; Tripathi, A.; Wu, J.; Schultz, P. J.; Yim, I.; McQuade, T. J.; Yu, F.; Arevang, C.-J.; Mensah, A. Y.; Tamayo-Castillo, G.; Xi, C.; Sherman, D. H. Discovery of Cahuitamycins as Biofilm Inhibitors Derived from a Convergent Biosynthetic Pathway. *Nat. Commun.* **2016**, *7* (1),

10710. https://doi.org/10.1038/ncomms10710.

(46) Mike, L. A.; Tripathi, A.; Blankenship, C. M.; Saluk, A.; Schultz, P. J.; Tamayo-Castillo, G.; Sherman, D. H.; Mobley, H. L. T. Discovery of Nicoyamycin A, an Inhibitor of Uropathogenic Escherichia Coli Growth in Low Iron Environments. *Chem. Commun. Camb. Engl.* **2017**, *53* (95), 12778–12781. https://doi.org/10.1039/c7cc07732g.

(47) Tripathi, A.; Schofield, M. M.; Chlipala, G. E.; Schultz, P. J.; Yim, I.; Newmister, S. A.; Nusca, T. D.; Scaglione, J. B.; Hanna, P. C.; Tamayo-Castillo, G.; Sherman, D. H. Baulamycins A and B, Broad-Spectrum Antibiotics Identified as Inhibitors of Siderophore Biosynthesis in Staphylococcus Aureus and Bacillus Anthracis. *J. Am. Chem. Soc.* **2014**, *136* (4), 1579–1586. https://doi.org/10.1021/ja4115924.

(48) Cheng, K. C.-C.; Cao, S.; Raveh, A.; MacArthur, R.; Dranchak, P.; Chlipala, G.; Okoneski, M. T.; Guha, R.; Eastman, R. T.; Yuan, J.; Schultz, P. J.; Su, X.; Tamayo-Castillo, G.; Matainaho, T.; Clardy, J.; Sherman, D. H.; Inglese, J. Actinoramide A Identified as a Potent Antimalarial from Titration-Based Screening of Marine Natural Product Extracts. *J. Nat. Prod.* **2015**, *78* (10), 2411–2422. https://doi.org/10.1021/acs.jnatprod.5b00489.

(49) Ymele-Leki, P.; Cao, S.; Sharp, J.; Lambert, K. G.; McAdam, A. J.; Husson, R. N.; Tamayo, G.; Clardy, J.; Watnick, P. I. A High-Throughput Screen Identifies a New Natural Product with Broad-Spectrum Antibacterial Activity. *PLOS ONE* **2012**, *7* (2), e31307. https://doi.org/10.1371/journal.pone.0031307.

(50) Navarro, M.; Moreira, I.; Arnaez, E.; Quesada, S.; Azofeifa, G.; Vargas, F.; Alvarado, D.; Chen, P. Flavonoids and Ellagitannins Characterization, Antioxidant and Cytotoxic Activities of Phyllanthus Acuminatus Vahl. *Plants* **2017**, *6* (4), 62. https://doi.org/10.3390/plants6040062.

(51) Villalobos-Vega, M. J.; Rodríguez-Rodríguez, G.; Armijo-Montes, O.; Jiménez-Bonilla, P.; Álvarez-Valverde, V. Optimization of the Extraction of Antioxidant Compounds from Roselle Hibiscus Calyxes (Hibiscus Sabdariffa), as a Source of Nutraceutical Beverages. *Molecules* **2023**, *28* (6), 2628. https://doi.org/10.3390/molecules28062628.

(52) Montero-Zamora, J.; Fernández-Fernández, S.; Redondo-Solano, M.; Mazón-Villegas, B.; Mora-Villalobos, J. A.; Barboza, N. Assessment of Different Lactic Acid Bacteria Isolated from Agro-Industrial Residues: First Report of the Potential Role of Weissella Soli for Lactic Acid Production from Milk Whey. *Appl. Microbiol.* **2022**, *2* (3), 626–635. https://doi.org/10.3390/applmicrobiol2030048.

(53) Quirós-Fallas, M. I.; Wilhelm-Romero, K.; Quesada-Mora, S.; Azofeifa-Cordero, G.; Vargas-Huertas, L. F.; Alvarado-Corella, D.; Mora-Román, J. J.; Vega-Baudrit, J. R.; Navarro-Hoyos, M.; Araya-Sibaja, A. M. Curcumin Hybrid Lipid Polymeric Nanoparticles: Antioxidant Activity, Immune Cellular Response, and Cytotoxicity Evaluation. *Biomedicines* **2022**, *10* (10), 2431. https://doi.org/10.3390/biomedicines10102431.

(54) Castillo-Henríquez, L.; Alfaro-Aguilar, K.; Ugalde-Álvarez, J.; Vega-Fernández, L.; Montes de

Oca-Vásquez, G.; Vega-Baudrit, J. R. Green Synthesis of Gold and Silver Nanoparticles from Plant Extracts and Their Possible Applications as Antimicrobial Agents in the Agricultural Area. *Nanomaterials* **2020**, *10* (9), 1763. https://doi.org/10.3390/nano10091763.

(55) Araya-Sibaja, A. M.; Wilhelm-Romero, K.; Vargas-Huertas, F.; Quirós-Fallas, M. I.; Alvarado-Corella, D.; Mora-Román, J. J.; Vega-Baudrit, J. R.; Sánchez-Kopper, A.; Navarro-Hoyos, M. Hybrid Nanoparticles of Proanthocyanidins from Uncaria Tomentosa Leaves: QTOF-ESI MS Characterization, Antioxidant Activity and Immune Cellular Response. *Plants* **2022**, *11* (13), 1737. https://doi.org/10.3390/plants11131737.

(56) Rosales-López, C.; Vargas-López, A.; Monge-Artavia, M.; Rojas-Chaves, M. Evaluation of Conditions to Improve Biomass Production by Submerged Culture of Ganoderma Sp. *Microorganisms* **2022**, *10* (7), 1404. https://doi.org/10.3390/microorganisms10071404.

(57) Syedd-León, R.; Solano-Campos, F.; Campos-Rodríguez, J.; Pereira-Arce, D.; Villegas-Peñaranda, L. R.; Sandoval-Barrantes, M. Fungal Extracellular Lipases from Coffee Plantation Environments for the Sustainable Management of Agro-Industrial Coffee Biomass. *Biomass* **2022**, *2* (2), 62–79. https://doi.org/10.3390/biomass2020005.

(58) Portuguez-García, M. P.; Agüero-Alvarado, R.; González-Lutz, M. I. Herbicidal activity of three natural products on four weed species. *Agron. Mesoam.* **2021**, *32* (3), 991–999. https://doi.org/10.15517/AM.V32I3.41394.

(59) Granados-Chinchilla, F.; Redondo-Solano, M.; Jaikel-Víquez, D. Mycotoxin Contamination of Beverages Obtained from Tropical Crops. *Beverages* **2018**, *4* (4), 83. https://doi.org/10.3390/beverages4040083.

(60) Navarro, M.; Moreira, I.; Arnaez, E.; Quesada, S.; Azofeifa, G.; Alvarado, D.; Monagas, M. J. Proanthocyanidin Characterization, Antioxidant and Cytotoxic Activities of Three Plants Commonly Used in Traditional Medicine in Costa Rica: Petiveria Alliaceae L., Phyllanthus Niruri L. and Senna Reticulata Willd. *Plants* **2017**, *6* (4), 50. https://doi.org/10.3390/plants6040050.

(61) Doyle, B. J.; Frasor, J.; Bellows, L. E.; Locklear, T. D.; Perez, A.; Gomez-Laurito, J.; Mahady, G. B. Estrogenic Effects of Herbal Medicines from Costa Rica Used for the Management of Menopausal Symptoms. *Menopause* **2009**, *16* (4), 748–755. https://doi.org/10.1097/gme.0b013e3181a4c76a.

(62) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(63) Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Res.* **2018**, *46* (W1), W563–W570. https://doi.org/10.1093/nar/gky294.

(64) Rajan, K.; Brinkhaus, H. O.; Agea, M. I.; Zielesny, A.; Steinbeck, C. DECIMER.Ai: An Open Platform for Automated Optical Chemical Structure Identification, Segmentation and Recognition

in Scientific Publications. *Nat. Commun.* **2023**, *14* (1), 5045. https://doi.org/10.1038/s41467-023-40782-0.

(65) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; gedeck; Jones, G.; NadineSchneider; Kawashima, E.; Nealschneider, D.; Dalke, A.; Swain, M.; Cole, B.; Turk, S.; Savelev, A.; Vaucher, A.; Wójcikowski, M.; Take, I.; Scalfani, V. F.; Probst, D.; Ujihara, K.; Walker, R.; Pahl, A.; godin, guillaume; tadhurst-cdd; Lehtivarjo, J.; Bérenger, F.; Bisson, J. Rdkit/Rdkit: 2024_03_5 (Q1 2024) Release, 2024. https://doi.org/10.5281/zenodo.12782092.

(66) MolVS: Molecule Validation and Standardization. https://github.com/mcs07/MolVS (accessed 2025-02-21).

(67) The pandas development team. Pandas-Dev/Pandas: Pandas, 2024. https://doi.org/10.5281/zenodo.10957263.

(68) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **2021**, *84* (11), 2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399.

(69) Requests: Python HTTP for Humans. https://requests.readthedocs.io (accessed 2025-02-21).

(70) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.

(71) Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E. L.; Strawbridge, S. A.; Garcia-Patino, M.; Kruger, R.; Sivakumaran, A.; Sanford, S.; Doshi, R.; Khetarpal, N.; Fatokun, O.; Doucet, D.; Zubkowski, A.; Rayat, D. Y.; Jackson, H.; Harford, K.; Anjum, A.; Zakir, M.; Wang, F.; Tian, S.; Lee, B.; Liigand, J.; Peters, H.; Wang, R. Q. R.; Nguyen, T.; So, D.; Sharp, M.; da Silva, R.; Gabriel, C.; Scantlebury, J.; Jasinski, M.; Ackerman, D.; Jewison, T.; Sajed, T.; Gautam, V.; Wishart, D. S. DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **2024**, *52* (D1), D1265–D1275. https://doi.org/10.1093/nar/gkad976.

(72) US EPA. *Biopesticide Active Ingredients*. United States Environmental Protection Agency. https://www.epa.gov/ingredients-used-pesticide-products/biopesticide-active-ingredients (accessed 2024-11-02).

(73) Yang, C.; Barlow, S. M.; Muldoon Jacobs, K. L.; Vitcheva, V.; Boobis, A. R.; Felter, S. P.; Arvidson, K. B.; Keller, D.; Cronin, M. T. D.; Enoch, S.; Worth, A.; Hollnagel, H. M. Thresholds of Toxicological Concern for Cosmetics-Related Substances: New Database, Thresholds, and Enrichment of Chemical Space. *Food Chem. Toxicol.* **2017**, *109*, 170–193. https://doi.org/10.1016/j.fct.2017.08.043.

(74) Di, L.; Kerns, E. H. Chapter 4 - Prediction Rules for Rapid Property Profiling from Structure. In *Drug-Like Properties (Second Edition)*; Di, L., Kerns, E. H., Eds.; Academic Press: Boston, 2016; pp 29–38. https://doi.org/10.1016/B978-0-12-801076-1.00004-6.

(75) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.;

Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.

(76) Frolov, A. I.; Chankeshwara, S. V.; Abdulkarim, Z.; Ghiandoni, G. M. pIChemiSt — Free Tool for the Calculation of Isoelectric Points of Modified Peptides. *J. Chem. Inf. Model.* **2022**, *63* (1), 187. https://doi.org/10.1021/acs.jcim.2c01261.

(77) Ermondi, G.; Vallaro, M.; Goetz, G.; Shalaeva, M.; Caron, G. Updating the Portfolio of Physicochemical Descriptors Related to Permeability in the beyond the Rule of 5 Chemical Space. *Eur. J. Pharm. Sci.* **2020**, *146*, 105274. https://doi.org/10.1016/j.ejps.2020.105274.

(78) Manallack, D. T. The pKa Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2007**, *1*, 25–38.

(79) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.

(80) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.

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

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

(83) Waskom, M. L. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.

(84) Chembl-Webresource-Client: Python Client Fot Accessing ChEMBL Webservices. https://www.ebi.ac.uk/chembl/api/data/docs (accessed 2025-06-17).

(85) Naveja, J. J.; Medina-Franco, J. L. Finding Constellations in Chemical Space Through Core Analysis. *Front. Chem.* **2019**, *7*. https://doi.org/10.3389/fchem.2019.00510.

(86) Bro, R.; K. Smilde, A. Principal Component Analysis. *Anal. Methods* **2014**, *6* (9), 2812–2831. https://doi.org/10.1039/C3AY41907J.

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

(88) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.

(89) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Transact. A Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

(90) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus Diversity Plots: A Global Diversity Analysis of Chemical Libraries. *J. Cheminformatics* **2016**, *8* (1), 63. https://doi.org/10.1186/s13321-016-0176-9.

(91) Scipy: Fundamental Algorithms for Scientific Computing in Python. https://scipy.org/ (accessed 2025-06-17).

(92) Zamora, W. J.; Pinheiro, S.; Acuña, D.; Zuñiga, J. NAPRORE-CR (NAtural PROducts REpository - Costa Rica), 2025. https://zenodo.org/records/14910594 (accessed 2025-06-16).

(93) Anarat-Cappillino, G.; Sattely, E. S. The Chemical Logic of Plant Natural Product Biosynthesis. *Curr. Opin. Plant Biol.* **2014**, *19*, 51. https://doi.org/10.1016/j.pbi.2014.03.007.

(94) 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, 2018; Vol. 59, pp 227–250. https://doi.org/10.1016/B978-0-444-64179-3.00007-4.

(95) Vogt, T. Phenylpropanoid Biosynthesis. *Mol. Plant* **2010**, *3* (1), 2–20. https://doi.org/10.1093/mp/ssp106.

(96) Chaverri, C.; Cicció, J.; Diaz, C. Chemical Composition of Aiouea Costaricensis (Lauraceae) Essential Oils from Costa Rica and Their Cytotoxic Activity on Cell Lines. *J. Essent. Oil Res. - J ESSENT OIL RES* **2010**, *22*, 524–529. https://doi.org/10.1080/10412905.2010.9700389.

(97) Montero-Villegas, S.; Crespo, R.; Rodenak-Kladniew, B.; Castro, M. A.; Galle, M.; Cicció, J. F.; García de Bravo, M.; Polo, M. Cytotoxic Effects of Essential Oils from Four Lippia Alba Chemotypes in Human Liver and Lung Cancer Cell Lines. *J. Essent. Oil Res.* **2018**, *30* (3), 167–181. https://doi.org/10.1080/10412905.2018.1431966.

(98) Díaz, C.; Quesada, S.; Brenes, O.; Aguilar, G.; Cicció, J. F. Chemical Composition of Schinus Molle Essential Oil and Its Cytotoxic Activity on Tumour Cell Lines. *Nat. Prod. Res.* **2008**, *22* (17), 1521–1534. https://doi.org/10.1080/14786410701848154.

(99) Manallack, D. T.; Yuriev, E.; Chalmers, D. K. The Influence and Manipulation of Acid/Base Properties in Drug Discovery. *Drug Discov. Today Technol.* **2018**, *27*, 41–47. https://doi.org/10.1016/j.ddtec.2018.04.003.

(100) Simoben, C. V.; Babiaka, S. B.; Moumbock, A. F. A.; Namba-Nzanguim, C. T.; Eni, D. B.; Medina-Franco, J. L.; Günther, S.; Ntie-Kang, F.; Sippl, W. Challenges in Natural Product-Based Drug Discovery Assisted with in Silico-Based Methods. *RSC Adv.* **2023**, *13* (45), 31578–31594. https://doi.org/10.1039/D3RA06831E.

(101) Edwards, D. M.; Speers, C.; Wahl, D. R. Targeting Noncanonical Regulators of the DNA Damage Response to Selectively Overcome Cancer Radiation Resistance. *Semin. Radiat. Oncol.* **2022**,

32 (1), 64–75. https://doi.org/10.1016/j.semradonc.2021.09.006.

(102) Singla, R. K.; Sai, C. S.; Chopra, H.; Behzad, S.; Bansal, H.; Goyal, R.; Gautam, R. K.; Tsagkaris, C.; Joon, S.; Singla, S.; Shen, B. Natural Products for the Management of Castration-Resistant Prostate Cancer: Special Focus on Nanoparticles Based Studies. *Front. Cell Dev. Biol.* **2021**, *9*. https://doi.org/10.3389/fcell.2021.745177.

(103) Salmela, A.-L.; Pouwels, J.; Varis, A.; Kukkonen, A. M.; Toivonen, P.; Halonen, P. K.; Perälä, M.; Kallioniemi, O.; Gorbsky, G. J.; Kallio, M. J. Dietary Flavonoid Fisetin Induces a Forced Exit from Mitosis by Targeting the Mitotic Spindle Checkpoint. *Carcinogenesis* **2009**, *30* (6), 1032–1040. https://doi.org/10.1093/carcin/bgp101.

(104) Hayes, D. M.; Braud, S.; Hurtado, D. E.; McCallum, J.; Standley, S.; Isaac, J. T. R.; Roche, K. W. Trafficking and Surface Expression of the Glutamate Receptor Subunit, KA2. *Biochem. Biophys. Res. Commun.* **2003**, *310* (1), 8–13. https://doi.org/10.1016/j.bbrc.2003.08.115.

(105) Institute of Medicine (US) Forum on Neuroscience and Nervous System Disorders. *Glutamate-Related Biomarkers in Drug Development for Disorders of the Nervous System: Workshop Summary*; The National Academies Collection: Reports funded by National Institutes of Health; National Academies Press (US): Washington (DC), 2011.

(106) Jesus, A. R.; Vila-Viçosa, D.; Machuqueiro, M.; Marques, A. P.; Dore, T. M.; Rauter, A. P. Targeting Type 2 Diabetes with C-Glucosyl Dihydrochalcones as Selective Sodium Glucose Co-Transporter 2 (SGLT2) Inhibitors: Synthesis and Biological Evaluation. *J. Med. Chem.* **2017**, *60* (2), 568–579. https://doi.org/10.1021/acs.jmedchem.6b01134.

(107) Wang, J.; Zhang, R.; Chen, X.; Sun, X.; Yan, Y.; Shen, X.; Yuan, Q. Biosynthesis of Aromatic Polyketides in Microorganisms Using Type II Polyketide Synthases. *Microb. Cell Factories* **2020**, *19* (1), 110. https://doi.org/10.1186/s12934-020-01367-4.

(108) Ullah, A.; Munir, S.; Badshah, S. L.; Khan, N.; Ghani, L.; Poulson, B. G.; Emwas, A.-H.; Jaremko, M. Important Flavonoids and Their Role as a Therapeutic Agent. *Molecules* **2020**, *25* (22), 5243. https://doi.org/10.3390/molecules25225243.

TABLE OF CONTETNS