Artificial Intelligence in the Life Sciences 3 (2023) 100066

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Short Communications

Natural products subsets: Generation and characterization

Ana L. Chávez-Hernández, José L. Medina-Franco*

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, México City 04510, Mexico

ARTICLE INFO

Keywords:
Artificial intelligence
Chemical space
Chemical multiverse
Chirality
De novo design
Deep learning
Molecular fingerprint
Molecular representation
Natural products

ABSTRACT

Natural products are attractive for drug discovery applications because of their distinctive chemical structures, such as an overall large fraction of sp^3 carbon atoms, chiral centers (both features associated with structural complexity), large chemical scaffolds, and diversity of functional groups. Furthermore, natural products are used in *de novo* design and have inspired the development of pseudo-natural products using generative models. Public databases such as the Collection of Open NatUral ProdUcTs and the Universal Natural Product database (UNPD) are rich sources of structures to be used in generative models and other applications. In this work, we report the selection and characterization of the most diverse compounds of natural products from the UNPD using the MaxMin algorithm. The subsets generated with 14,994, 7,497, and 4,998 compounds are publicly available at https://github.com/DIFACQUIM/Natural-products-subsets-generation. We anticipate that the subsets will be particularly useful in building generative models based on natural products by research groups, particularly those with limited access to extensive supercomputer resources.

1. Introduction

Natural products and fragments derived from natural products have been attractive in drug design and development because of their distinctive chemical structures. For example, natural products have, in general, an overall larger diversity of functional groups and larger structural complexity than synthetic molecules [1–4]. However, a drawback for some natural products, particularly those with sizeable structural complexity, is that they can be challenging to synthesize. A workaround for this issue is the so-called pseudo-natural products which are synthetically feasible compounds generated through a *de novo* combination of natural product fragments [3]. Pseudo-natural products allow the exploration of uncharted areas of biologically relevant chemical space that are different from the chemical space covered by the compounds from which they are generated.

The Collection of Open NatUral ProdUcTs (COCONUT) and the Universal Natural Product Database (UNPD) are two large compound databases. COCONUT [5] is arguably the most extensive public natural product database, with 389,184 unique structures. UNPD [6], with 153,375 natural products, is the second-largest public natural product database. A distinctive feature of UNPD compared to COCONUT is that compounds in UNPD contain chirality information. In Latin America, several public natural products databases compile the compounds isolated and characterized from the country of origin. Examples are $NuBBE_{DB}$ [7,8], SistematX [9,10] from Brazil; CIFPMA [11,12] from Panama; PeruNPDB [13] from Peru; and BIOFACQUIM [14,15] from Mexico. The latter database contains the structures of 531 natural products.

De novo design generates virtual molecules from scratch. It filters structures generated using several scoring functions and assesses synthetic chemical feasibility to remove reactive and unrealistic compounds [16]. *De novo* design based on generative algorithms such as deep learning involves using neural networks [17,18] and databases with many compounds [19]. In the source compound databases, many compounds with three-dimensional information (*e.g.*, stereochemistry) are relevant to build robust generative models [20]. However, for several research groups, it is difficult to access supercomputer resources to handle many compounds to obtain appropriate subsets that impact the model prediction [21,22].

This Communication reports the selection and characterization of the most diverse compounds from UNPD. The subsets were selected using a dissimilarity-based compound selection (DBCS) method, the MaxMin algorithm [23]. The compound subsets were characterized by the Natural Product Likeness (NPL) score [24], structural diversity, and distribution in chemical space. The structural diversity was assessed with the Tanimoto coefficient and molecular fingerprints of different designs. Chemical space analysis was performed through the visual representation of the chemical multiverse [25] using T-distributed Stochastic Neighbor

Abbreviations: COCONUT, Collection of Open NatUral ProdUcTs; DNMT, DNA methyltransferase; ECFP, extended connectivity fingerprint; TMAP, Tree MAP; NPL, natural product-likeness; UNPD, Universal Natural Product database.

* Correspondence author.
E-mail address: medinajl@unam.mx (J.L. Medina-Franco).

https://doi.org/10.1016/j.ailsci.2023.100066
Received 20 December 2022; Received in revised form 5 February 2023; Accepted 21 February 2023
Available online 26 February 2023
2667-3185/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

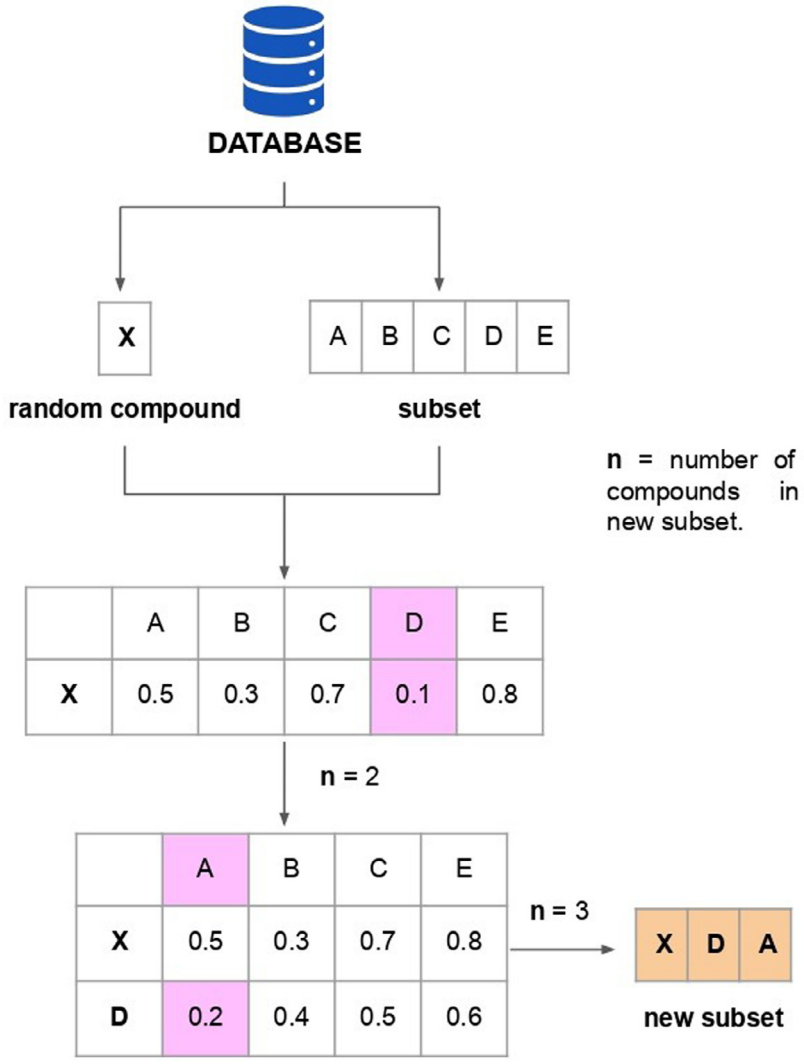

Fig. 1. MaxMin algorithm implemented in this work. A database (UNPD) was split into a given number of subsets. From each subset, a random compound (X) was selected. The molecular similarity was calculated between X and each compound (A, B, C, D, and E) from the subset using the Tanimoto coefficient and the ECFP4 fingerprint. In this figure, only a subset (that contains the compounds A, B, C, D, and E) is shown for illustrative purposes. The compound with the smallest molecular similarity was chosen and deleted from the original subset. The process was repeated to generate a new subset with the number of compounds desired.

Embedding t-SNE [26] and the recently proposed Tree MAP (TMAP) [27]. We anticipate that the natural product subsets generated and thoroughly characterized in this work will be helpful to the scientific community in a variety of tasks including building generative models, in particular to those research groups with limited access to large supercomputer resources.

2. Materials and methods

2.1. Data sets

For this study, different types of databases with compounds from various origins were used including three data sets of natural products, and a data set focused on a specific molecular target of pharmaceutical relevance. Natural product databases from different sources were employed to compare the NPL score of the UNPD subsets (described in the *Natural product-likeness* section). The data set focused on a specific molecular target and the natural products from different sources were used to evaluate and compare the molecular diversity of UNPD subsets. Using a focused database gives a value of low structural diversity, which serves as a reference to compare the structural diversity of the UNPD subsets. Likewise, the chemical structures from other natural product databases served as reference compound collections to evaluate the diversity of the data sets newly generated. Each type of data set and its contents are described below.

Three data sets of natural products were used, namely, COCONUT, UNPD, and BIOFACQUIM. COCONUT [5] had 389,184 unique structures, not including chirality information. UNPD [6] with 153,372 natural products, encoding their chirality. BIOFACQUIM [14,15] is a database with 531 natural products isolated and characterized in Mexico, and the chirality of the compounds is included. A data set with 715 compounds focused on DNA methyltransferase 1 (DNMT1) inhibitors. DNMT1 is an epigenetic target relevant to drug discovery [28,29]. The data set of DNMT1 inhibitors was obtained from the ChEMBL database, release 31 [30,31].

2.2. Data set standardization

The preparation of compounds, encoded in SMILES strings [32] was performed using the open-source cheminformatics toolkit RDKit [33] and MolVS [34]. Compounds with valence errors or any chemical element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I, were removed. Stereochemistry information, when available, was retained. Compounds with multiple components were split, and the largest component was retained. The remaining compounds were neutralized and reionized to generate the corresponding canonical tautomer.

2.3. Natural product-likeness

The NPL score, introduced by Ertl et al. [24], was computed for all compounds in COCONUT, UNPD, and BIOFACQUIM. The NPL score

Table 1
Descriptive statistics of NPL scores computed for COCONUT, UNPD, and BIOFACQUIM.

Dataset	COCONUT	UNPD	BIOFACQUIM
count	389,184	153,372	531
mean	0.89	1.81	1.73
[a]std	1.38	0.96	0.90
[b]min	−3.53	−2.51	−0.36
[c]Q1	−0.32	1.14	1.03
median	0.97	1.87	1.65
[d]Q3	2.01	2.56	2.45
[e]max	4.08	4.08	3.87

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

ranges between −5 (if the compound is more similar to a synthetic compound) and 5 (if the compound is more similar to a natural product).

2.4. Selection of sets of diverse compounds

Dissimilarity-based compound selection (DBCS) methods allow the identification of diverse compound data sets by directly calculating distances or dissimilarities between the chemical structures [23]. Among the DBCS methods, the MaxMin algorithm reduces the number of compounds choosing the molecules with the largest diversity from the original databases. Three subsets of UNPD were generated using the MaxMin algorithm as follows (Fig. 1): The initial 153,372 compounds from UNPD were split into three ways: (A) thirty random subsets of 5,000 compounds each; (B) fifteen random subsets of 10,000 compounds each; and (C) ten random subsets of 15,000 compounds each. A new diverse set with 500 compounds was selected from each one using MaxMin [23]. First, a random compound was picked from each subset. The binary similarity between the compound selected (query set) and the remaining compounds (target set) was calculated. A new compound was selected from the target set if this had the lowest similarity value and then removed from the target set. The iteration process continued until the number of compounds desired set to 500 was reached. In total, three diverse subsets from UNPD were generated: UNPD-A (15,000 compounds), UNPD-B (7,500), and UNPD-C (5,000). For the diversity selection, we used the Tanimoto coefficient and the extended connectivity fingerprint (ECFP) [35] of 1024-bits and diameter 4 (ECFP4). For the diversity set calculations, we used an E5–2670v1 processor, 16 cores, and 64 Gbytes of RAM). The maximum calculation time for each initial subset of the natural product subsets were: UNPD-A (15,000 compounds), 19,989.21 s; UNPD-B (7,500 compounds), 102,569.61 s, and UNPD-C (5,000 compounds), 209,241.25 s.

2.5. Structural diversity

The structural diversity of three UNPD subsets, UNPD, BIOFACQUIM, and DNMT1, was compared through the distribution of the pairwise similarity values generated with the Tanimoto coefficient using three molecular fingerprints Molecular ACCes System (MACCS) Keys (166-bits) [36], extended connectivity fingerprint (ECFP) [35] of 1024-bits with diameter 6 (ECFP6) and diameter 4 (ECFP4).

2.6. Chemical multiverse visualization

The chemical multiverse [25] of the three UNPD subsets was compared to the chemical multiverse of the entire UNPD collection. A chemical multiverse is a group of numerical vectors that differently describe a set of molecules depending on the molecular representation [25]. So, each chemical space is an M-dimensional cartesian space in which compounds are located by a set of M descriptors. Each type of molecular

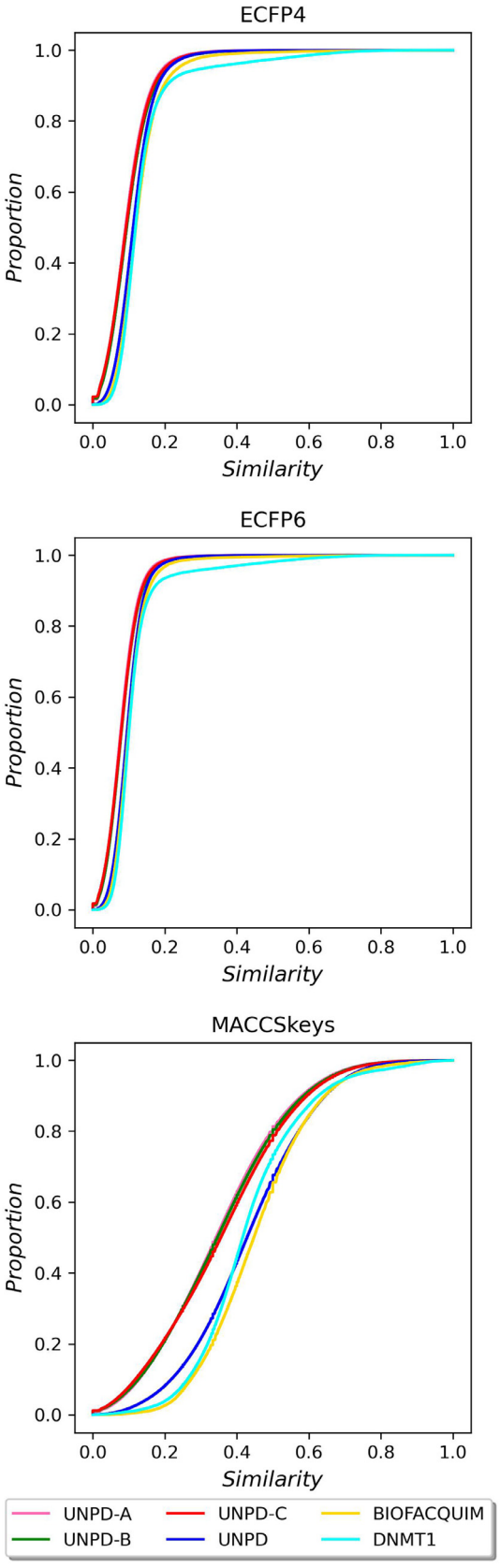

Fig. 2. Cumulative distribution functions of the pairwise Tanimoto similarity using MACCS keys (166-bits), ECFP4, and ECFP6, as molecular representations. The datasets are UNPD-A (pink line), UNPD-B (green), UNPD-C (red), and UNPD (blue).

Table 2
Summary of the structural diversity of the three UNPD subsets, BIOFACQUIM, and DNMT1 datasets. The number of initial and unique compounds from each database is indicated.

Dataset	Compounds	Unique compounds	MACCS keys (166-bits)[a]	ECFP4[a]	ECFP6[a]
UNPD-A	15,000	14,994	0.341	0.091	0.077
UNPD-B	7,500	7,497	0.346	0.094	0.08
UNPD-C	5,000	4,998	0.356	0.092	0.079
UNPD	153,372	153,372	0.43	0.111	0.094
BIOFACQUIM	531	531	0.447	0.119	0.099
DNMT1	714	714	0.417	0.119	0.1

[a] The structural diversity is reported as the median value of the distribution of the pairwise comparison using the Tanimoto coefficient and molecular fingerprints (MACCS keys, ECFP4 and ECFP6). A full diversity assessment is presented at the cumulative distribution functions of the pairwise similarity values in Fig. 2.

Table 3
Descriptive statistics of the number of hydrogen bond donors and acceptors (HBD, HBA) of the UNPD subsets and the entire UNPD.

Property	HBD				HBA			
Dataset	UNPD-A	UNPD-B	UNPD-C	UNPD	UNPD-A	UNPD-B	UNPD-C	UNPD
count	14,994	7,497	4,998	153,372	14,994	7,497	4,998	153,372
mean	2.51	2.6	2.69	3.15	5.58	5.65	5.94	7.05
[a]std	3.17	3.10	3.40	3.55	4.95	4.80	5.46	5.68
[b]min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[c]Q1	0.00	1.00	0.00	1.00	2.00	3.00	2.00	3.00
median	2.00	2.00	2.00	2.00	4.00	5.00	5.00	5.00
[d]Q3	3.00	3.00	4.00	4.00	7.00	7.00	8.00	9.00
[e]max	36.00	33.00	33.00	36.00	53.00	53.00	52.00	54.00

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

Table 4
Descriptive statistics of the number of rotatable bonds (RB) and LogP values of the UNPD subsets and the entire UNPD.

Property	RB				LogP			
Dataset	UNPD-A	UNPD-B	UNPD-C	UNPD	UNPD-A	UNPD-B	UNPD-C	UNPD
count	14,994	7,497	4,998	153,372	14,994	7,497	4,998	15,372
mean	4.74	5.34	4.81	5.97	2.94	2.98	2.94	2.94
[a]std	6.02	6.66	5.69	6.08	3.02	3.00	3.13	3.12
[b]min	0.00	0.00	0.00	0.00	−18.53	−14.10	−18.16	−20.82
[c]Q1	1.00	1.00	1.00	2.00	1.46	1.40	1.43	1.33
median	3.00	3.00	3.00	4.00	2.87	2.79	2.90	2.96
[d]Q3	6.00	7.00	6.00	8.00	4.32	4.24	4.45	4.58
[e]max	59.00	63.00	59.00	63.00	24.43	23.11	21.63	25.12

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

Table 5
Descriptive statistics of the topological surface area (TPSA) and molecular weight (MW) of the UNPD subsets and the entire UNPD.

Property	TPSA				MW			
Dataset	UNPD-A	UNPD-B	UNPD-C	UNPD	UNPD-A	UNPD-B	UNPD-C	UNPD
count	14,994	7,497	4,998	153,372	14,994	7,497	4,998	153,372
mean	90.78	93.58	94.78	114.63	371.94	369.57	388.43	450.55
[a]std	82.74	80.35	90.04	93.94	196.43	194.20	210.40	228.63
[b]min	0.00	0.00	0.00	0.00	16.04	16.04	17.03	16.04
[c]Q1	40.46	46.53	39.10	54.37	246.31	248.32	252.28	302.28
median	69.67	74.60	69.92	85.97	330.29	328.45	342.47	396.20
[d]Q3	112.05	116.20	119.61	145.91	445.60	444.48	466.74	534.66
[e]max	877.36	927.43	900.36	927.43	1,887.28	1,889.30	1,875.31	1,907.36

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

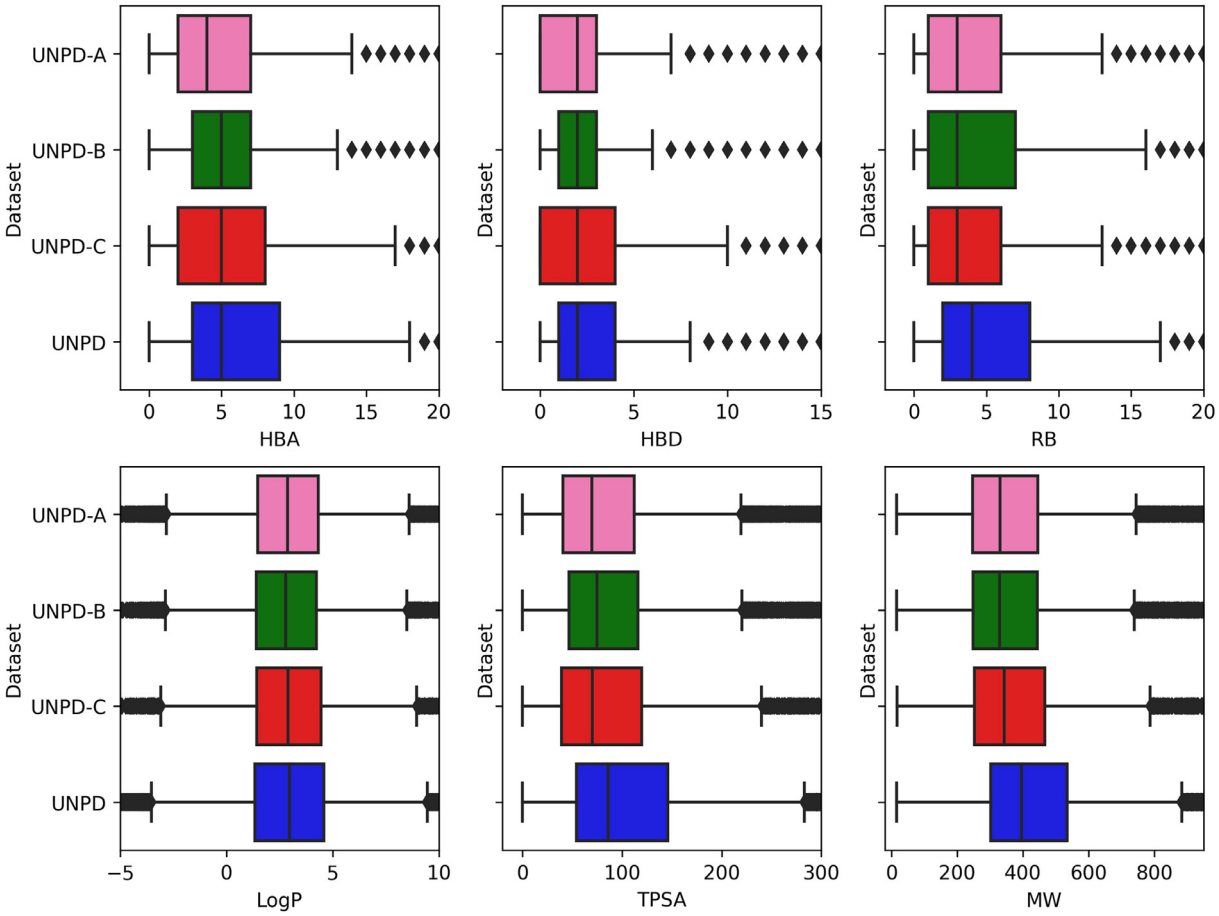

Fig. 3. Box-whisker plots of six properties of pharmaceutical relevance: hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of rotatable bonds (RB), molecular weight (MW), and partition coefficient octanol/water (LogP). The datasets are represented in different colors, UNPD-A (pink), UNPD-B (green), UNPD-C (red), and UNPD (blue). Black diamonds represent outliers.

fingerprint generates a distinct chemical space, and the sets of chemical spaces comprise the chemical multiverse. For this work, the visual representation of the chemical multiverse was done using t-SNE [26] and TMAPs [27] as visualization methods, and MACCS keys (166 bits), ECFP4, and ECFP6 as molecular representations.

The t-SNE generates plots that organize compounds. Similar compounds form clusters and dissimilar compounds are distant from each other. TMAP allows visualizing of many chemical compounds through the distance between clusters. Local Sensitive Hashing (LSH) allows each compound to be grouped hierarchically according to common substructures using molecular fingerprints. The number of nearest neighbors, $k = 50$, and the factor used by the augmented query algorithm, kc = 10, were used to develop the TMAP graphs. In previous studies, we used TMAP to describe the molecular diversity of natural products such as COCONUT [2], BIOFACQUIM [15], and a focused library of HIV-1 protease inhibitors using natural fragments from COCONUT [37].

3. Results and discussion

3.1. Natural-product-likeness

Table 1 summarizes the descriptive statistics for NPL scores calculated for BIOFACQUIM, UNPD, and COCONUT. The COCONUT, UNPD, and BIOFACQUIM had NPL score values in the range of [−3.53,4.08], [−2.51,4.08], and [−0.36,3.87], respectively. As anticipated, natural product compounds from COCONUT, UNPD, and BIOFACQUIM had NPL scores *ca.* 4. Only 25% of COCONUT´s compounds (min=−3.53, Q1=−0.32) had NPL scores *ca.* −3.53, and 25% of UNPD´s compounds (min=−2.51, Q1=1.14) had NPL close to −2.51 meaning that COCONUT had compounds with more chemical structures similar to synthetic compounds regarding the UNPD because of NPL scores *ca.* −5 is associated with compounds derived from the synthetic origin [24]. In general, natural products (COCONUT, UNPD, BIOFACQUIM) had NPL values closer to 5.

3.2. Molecular diversity

Three subsets of UNPD (A-C) were generated using the MaxMin algorithm, as described in Section 2.4. UNPD-A, UNPD-B, and UNPD-C had 15,000, 7,500, and 5,000 compounds, respectively. Then, the new subsets were curated, as described in Section 2.2. Table 2 shows that six, three, two, and one duplicated compounds were found in UNPD-A, UNPD-B, and UNPD-C, respectively (between 0.04% and 0.02%). Fig. 2 shows the cumulative distribution functions (CDF) of the pairwise Tanimoto similarity using MACCS keys (166-bits), ECFP4, and ECFP6, as molecular representations. The UNPD subsets (A-C) are represented in continous line as UNPD-A (pink), UNPD-B (green), and UNPD-C (red); UNPD (blue); BIOFACQUIM (yellow); and DNMT1 (cyan). The CDF with ECFP4 and ECFP6 shows that UNPD subsets (A-C) were the most diverse data sets (median=0.09); followed by the UNPD (median=0.1), BIOFACQUIM (median=0.12), and DNMT1 (median=0.12). The CDF with MACCS keys shows that UNPD subsets (A-C) were the most diverse (median=0.3) and UNPD (median=0.4), followed by DNMT1 (median=0.4) and BIOFACQUIM (median=0.43). Likewise, Table 2 shows that UNPD subsets were more diverse than UNPD, DIFACQUIM, and DNMT1. The UNPD-A was the most di-

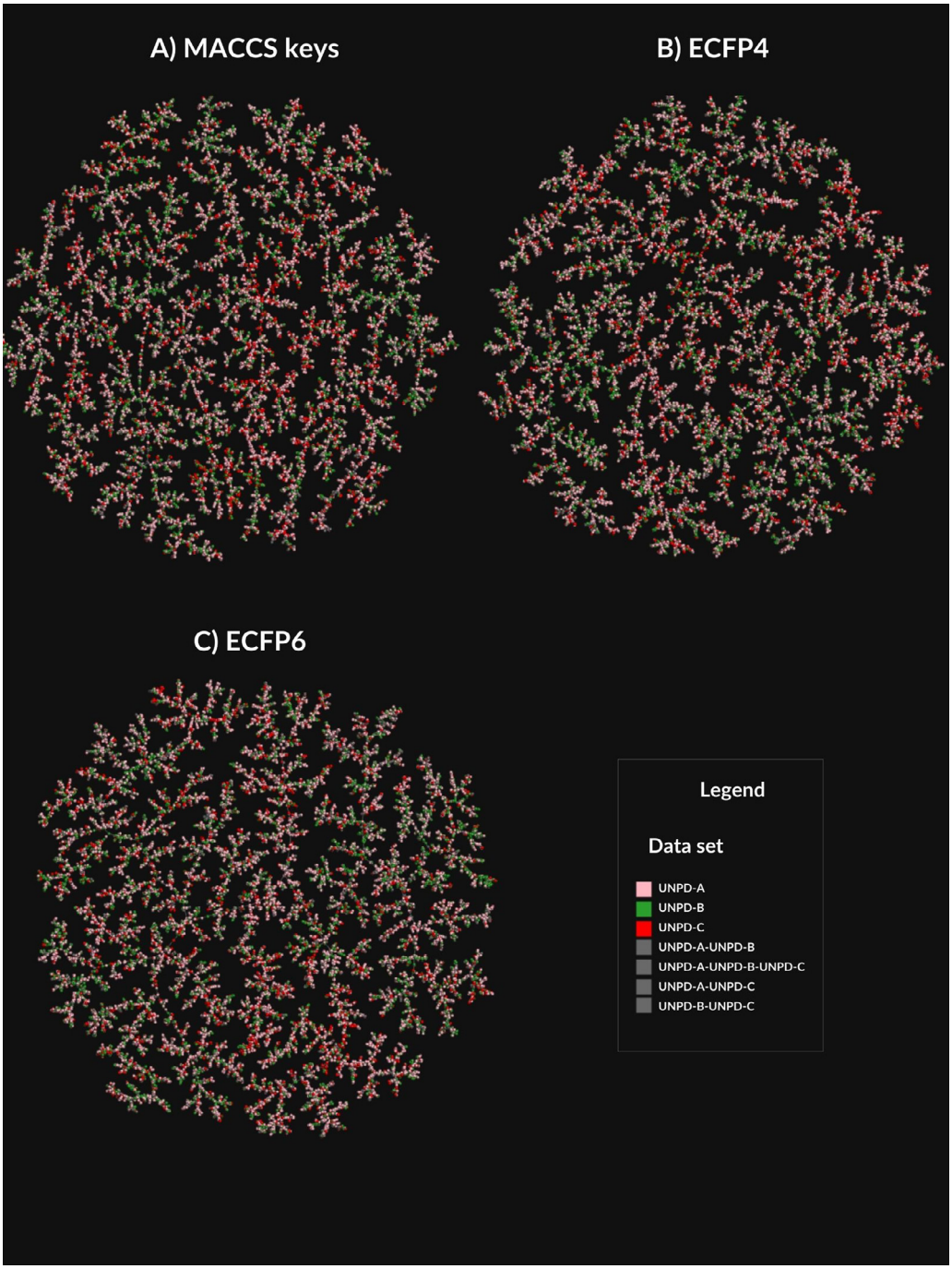

Fig. 4. Chemical multiverse visualization of UNDP subsets using TMAPs and MACCS keys (166-bits), ECFP4 and ECFP6, as molecular representations. The data sets are represented as UNPD-A (pink), UNPD-B (green), and UNPD-C (red). The overlap between natural product subsets is represented in gray.

verse dataset and had the lowest median similarity value (0.091, 0.077) calculated with ECFP4 and ECFP6, respectively, followed by the UNPD-B (0.094, 0.08), and UNPD-C (0.092, 0.079); and the reference databases: UNPD (0.111, 0.094), BIOFACQUIM (0.119, 0.099) and DNMT1 (0.119, 0.1). The structural diversity calculated with MACCS keys (Table 2 and Fig. 2) indicated that UNPD subsets had the same trend, being the UNPD-A the most diverse (median=0.341) followed by UNPD-B (median=0.346), and UNPD-C (median=0.356); followed by UNPD (median=0.43), BIOFACQUIM (median=0.447), and DNMT1 (median=0.417). For UNPD, the data set generated from a larger number of subsets (thirty *versus* fifteen or ten subsets) was the most diverse, *i.e.*, UNPD-A according to the ECFP4, ECFP6, and MACCS keys fingerprints.

3.3. Properties of pharmaceutical relevance

The natural products subsets were further characterized by means of the distribution of six properties of pharmaceutical relevance: hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of rotatable bonds (RB), molecular weight (MW), and partition coefficient octanol/water (LogP). Fig. 3 shows box-whisker plots of the distribution of the property values calculated for the three UNPD subsets and the entire UNPD (UNPD-A (pink), UNPD-B (green), UNPD-C (red), and UNPD (blue)). Tables 3-5 summarize the descriptive statistics. Fig. 3 and Tables 3-5 show that 75% of compounds of UNPD subsets and UNPD had the same range of LogP values, namely UNPD-A: LogP<=4.32, UNPD-

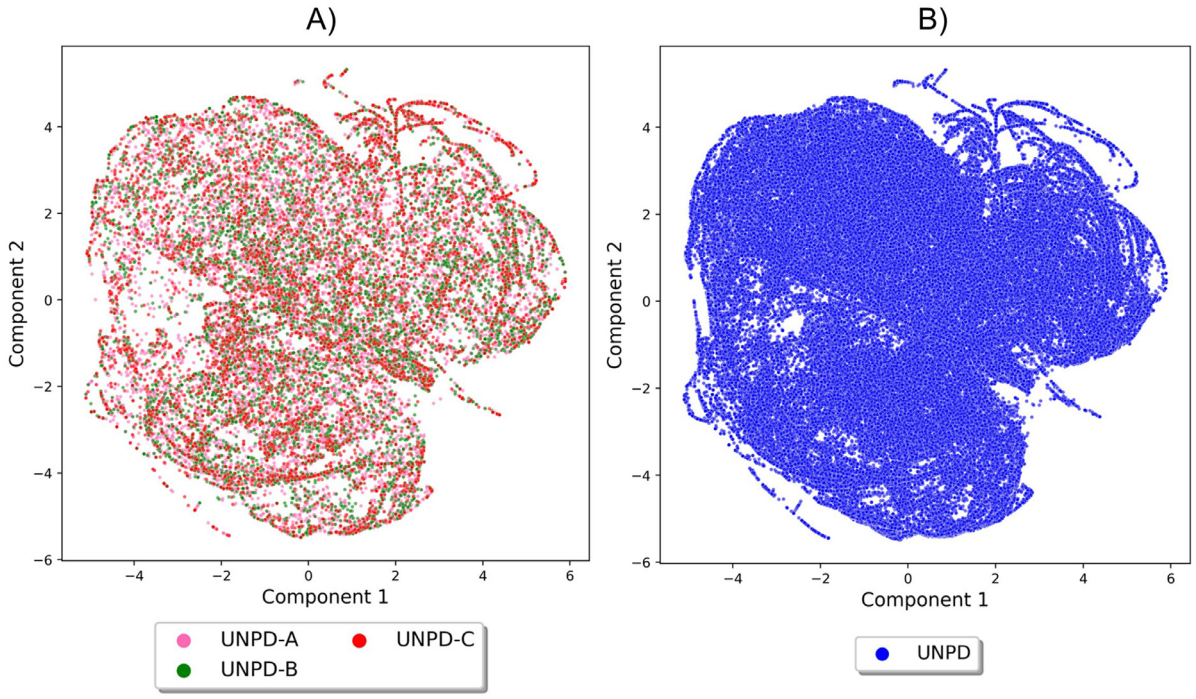

Fig. 5. Chemical space visualization of UNPD and UNPD subsets using t-SNE based on the properties: hydrogen bond donors, hydrogen bond acceptors, topological polar surface area, number of rotatable bonds, molecular weight, and partition coefficient octanol/water. The datasets are represented in scatter points as UNPD-A (pink), UNPD-B (green), UNPD-C (red), and UNPD (blue). Panel A shows the UNPD subsets (A-C), and panel B shows the entire UNPD.

B: LogP<=4.24, UNPD-C: LogP<=4.45, and UNPD: LogP<=4.58. Regarding MW and TPSA, 75% of UNPD's compounds (TPSA<=145.91 and MW<=534.66) were more diverse than 75% of UNPD subsets (A-C): UNPD-A's compounds (TPSA<=112.05, MW<=445.60); UNPD-B's compounds (TPSA<=116.20, MW<=444.48), and UNPD-C's compounds (TPSA<=116.61, MW<=466.74). Regarding RB, UNPD's compounds (RB<=8.0) and UNPD-B's compounds (RB<=7.0) were more diverse than UNPD-A's compounds (RB<=6.0) and UNPD-C's compounds (RB<=6.0). Regarding HBA and HBD, UNPD's compounds (HBD<=4.0, HBA<=9.0) and UNPD-C's compounds (HBD<=4.0, HBA<=8.0) were more diverse than UNPD-A (HBD<=3.0, HBA<=7.0) and UNPD-B (HBD<=4.0, HBA<=7.0). Overall, the UNPD-C dataset was the most diverse in terms of the properties of pharmaceutical relevance after the entire UNPD.

3.4. Visualization of the chemical space and chemical multiverse

As described in Section 2.6, a chemical multiverse is conceptualized as a group of molecular representations that each describe a compound dataset (in contrast to a chemical space that is defined by only one set of descriptors). Fig. 4 shows a visual representation of the chemical multiverse of the UNPD datasets using TMAPs and MACCS keys (166-bits), ECFP4, and ECFP6 as molecular representations. In this study, the chemical multiverse is comprised of three molecular fingerprints: one based on structural keys, MACCS keys (Fig. 4A); and the hashed molecular fingerprint, ECFP4 and ECFP6 (Figs. 4B and C). The data sets are represented in scatter points with different colors as UNPD-A (pink), UNPD-B (green), and UNPD-C (red). The overlap between different data sets is depicted in gray. The TMAP generated with MACCS keys shows fewer clusters with more compounds than ECFP4 and ECFP6. For natural products, the TMAP generated with ECFP6 shows that the compounds of the UNPD subsets are more evenly distributed with respect to the TMAPs constructed with ECFP4 and MACCS keys. The chemical multiverse represented by ECFP6 is more accurate in describing the structural diversity than ECFP4 because ECFP6 encodes molecular fragments in more detail than ECFP4 [35] and thus has an impact on quantifying structural diversity.

We also visualize the chemical space of the datasets using t-SNE based on the six physicochemical properties of pharmaceutical interest discussed in Section 3.3. Fig. 5A shows the chemical space of UNPD subsets (A-C). Each data point represents a compound: UNPD-A (pink), UNPD-B (green), and UNPD-C (red). As reference, Fig. 5B depicts a visualization of the chemical space of the entire UNPD (blue data points). t-SNE in Fig. 5A shows that the three UNPD subsets overlap with diverse regions of the chemical space of UNPD (Fig. 5B). UNPD-C was the most diverse regarding the six physicochemical properties of pharmaceutical interest, as described before in Section 3.3.

3.5. Applications of selection of subsets in drug discovery

The herein diverse subsets derived from natural products annotated with chirality information can be used in several different ways. For example, they can be used as reference sets for diversity analysis and coverage of chemical space of other compound libraries. For instance, Vivek-Ananth et al. recently reported a comparative analysis of more than 1800 secondary metabolites of medicinal fungi [38]. In that work, the authors included nine reference databases including natural products datasets. The same research group has reported and characterized an extensive database of Indian Medicinal Plants, Phytochemistry, and Therapeutics [39]. Having diverse subsets from UNDP included herein would further expand the outcome of such diversity studies. Other applications of the diverse subsets include its use in the development of generative models, such as *de novo* design. This approach generates new chemical entities with the properties desired, and recently uses algorithms of deep learning [16]. Deep learning algorithms require a large number of compounds, but it means more computer resources. The rational design of drugs involves quality data [40] to develop good models with the best predictions [21], and computer scientists advise the use of algorithms that can detect meaningful patterns in small data sets characteristics of the early stage of drug discovery can generate prospective studies [41]. For instance, a first approach to *de novo* design is to start

from small data sets of compounds with diverse structures and diverse properties of pharmaceutical relevance herein generated; and add to the distinctive structural complexity and diversity of the natural products as a larger fraction of sp^3 carbon atoms and chiral centers [3,4].

4. Conclusions

We report the selection and characterization of the three subsets with the most diverse compounds from UNPD using the MaxMin algorithm. Three subsets with 14,994, 7497, and 4998 compounds selected from the UNPD contain the most structurally diverse natural products. Unlike compounds in the COCONUT database, molecules in UNPD are annotated with chirality. The structural diversity of compounds is not affected by the number of subsets derived from the original database from which a new database is generated, and an analysis of the chemical multiverse supports that UNPD subsets contain the most diverse molecules. During the study, we also concluded that the visualization of the chemical space described with ECFP6 is more accurate to describe the structural diversity of compounds compared with ECFP4 and MACCS keys (166 bits). The natural product subsets had a large diversity of chemical compounds with different structural features and properties of pharmaceutical relevance. The NPL score supports that the chemical structures of natural products are very different and diverse as defined by the threshold of the NPL score, and as expected, natural products (COCONUT, UNPD, and BIOFACQUIM) had NPL scores values close to 5.

A significant perspective of this work is that the natural product subsets derived from the UNPD can be used to develop generative models that use deep learning algorithms and require the most diverse compounds, such as *de novo* design. The natural products subsets can also be used to develop predictive models; for virtual screening; and reference databases for evaluating the structural diversity or similarity to a specific subset, among other applications. The public availability of the natural product subsets can save costly computational resources for research groups with limited accessibility to supercomputer means.

Supplementary material

The MaxMin algorithm and structural diversity implemented in Python language, the interactive TMAPs, and the three subsets generated from UNPD with 14,994, 7,497, and 4,998 compounds with stereochemical information, are publicly available at GitHub: https://github.com/DIFACQUIM/Natural-products-subsets-generation.

Funding

Authors thank the Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), UNAM, for the computational resources to use Miztli supercomputer at UNAM under project LANCAD-UNAM-DGTIC-335; and the innovation space UNAM-HUAWEI the computational resources to use their supercomputer under project No. 7 "*Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer*". We also thank the funding of DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant No. IN201321.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is public at GitHub.

Acknowledgments

A.L.C.-H. is thankful to CONACyT, Mexico, for the Ph.D. scholarship number 847870.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100066.

References

[1] Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. A fragment library of natural products and its comparative chemoinformatic characterization. Mol Inform 2020;39:e2000050.
[2] Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. Fragment library of natural products and compound databases for drug discovery. Biomolecules 2020;10:1518.
[3] Grigalunas M, Brakmann S, Waldmann H. Chemical evolution of natural product structure. J Am Chem Soc 2022;144:3314–29.
[4] Atanasov AG, Zotchev SB, Dirsch VM. International Natural product sciences taskforce, C.T. Supuran, natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 2021;20:200–16.
[5] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: collection of Open Natural Products database. J Cheminform 2021;13:2.
[6] Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. PLoS ONE 2013;8:e62839.
[7] Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. Sci Rep 2017;7:7215.
[8] Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL. Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. J Chem Inf Model 2019;59:74–85.
[9] Costa RPO, Lucena LF, Silva LMA, Zocolo GJ, Herrera-Acevedo C, Scotti L, Da–Costa FB, Ionov N, Poroikov V, Muratov EN, Scotti MT. The SistematX web portal of natural products: an update. J Chem Inf Model 2021;61:2516–22.
[10] Scotti MT, Herrera-Acevedo C, Oliveira TB, Costa RPO, de O Santos SYK, Rodrigues RP, Scotti L, Da-Costa FB. SistematX, an online web-based cheminformatics tool for data management of secondary metabolites. Molecules 2018;23:103.
[11] Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL. Cheminformatic characterization of natural products from Panama. Mol Divers 2017;21:779–89.
[12] Olmedo DA, Medina-Franco JL. Chemoinformatic approach: the case of natural products of panama. Cheminformatics and its applications. IntechOpen; 2020.
[13] H.L. Barazorda-Ccahuana, L.G. Ranilla, M.A. Candia-Puma, E.G. Cárcamo-Rodriguez, A.E. Centeno-Lopez, G.D. Del-Carpio, J.L. Medina-Franco, M.A. Chávez-Fumagalli, PeruNPDB: the Peruvian Natural Products Database for in silico drug screening, bioRxiv. (2023) 2023.01.15.524152. 10.1101/2023.01.15.524152.
[14] Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: a Mexican compound database of natural products. Biomolecules 2019;9:31.
[15] N. Sánchez-Cruz, B.A. Pilón-Jiménez, J.L. Medina-Franco, Functional group and diversity analysis of BIOFACQUIM: a Mexican natural product database, F1000Res. 8 (2019) (Chem Inf Sci) 2071.
[16] Palazzesi F, Pozzan A. Deep learning applied to ligand-based de novo drug design. Methods Mol Biol 2022;2390:273–99.
[17] Hessler G, Baringhaus K-H. Artificial intelligence in drug design. Molecules 2018;23:2520.
[18] Sousa T, Correia J, Pereira V, Rocha M. Generative deep learning for targeted compound design. J Chem Inf Model 2021;61:5343–61.
[19] Jing Y, Bian Y, Hu Z, Wang L, Xie X-Q. Deep Learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J 2018;20:58.
[20] Miljković F, Rodríguez-Pérez R, Bajorath J. Impact of artificial intelligence on compound discovery, design, and synthesis. ACS Omega 2021;6:33293–9.
[21] Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow Jr RA, Fisher J, Jansen JM, Duca JS, Rush TS, Zentgraf M, Hill JE, Krutoholow E, Kohler M, Blaney J, Funatsu K, Luebkemann C, Schneider G. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 2020;19:353–64.
[22] Bajorath J, Chávez-Hernández AL, Duran-Frigola M, Fernández-de Gortari E, Gasteiger J, López-López E, Maggiora GM, Medina-Franco JL, Méndez-Lucio O, Mestres J, Miranda-Quintana RA, Oprea TI, Plisson F, Prieto-Martínez FD, Rodríguez-Pérez R, Rondón-Villarreal P, Saldívar-Gonzalez FI, Sánchez-Cruz N, Valli M. Chemoinformatics and artificial intelligence colloquium: progress and challenges in developing bioactive compounds. J Cheminform 2022;14:82.
[23] . Selecting diverse sets of compounds. In: Leach AR, Gillet VJ, editors. An introduction to chemoinformatics. Netherlands, Dordrecht: Springer; 2007. p. 119–39.
[24] Ertl P, Roggo S, Schuffenhauer A. Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model 2008;48:68–74.
[25] Medina-Franco JL, Chávez-Hernández AL, López-López E, Saldívar-González FI. Chemical multiverse: an expanded view of chemical space. Mol Inform 2022;41:e2200116.
[26] G. Hinton, Visualizing Data using t-SNE, (2008). https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl (accessed February 4, 2023).

[27] Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminform 2020;12:12.
[28] Prado-Romero DL, Medina-Franco JL. Advances in the exploration of the epigenetic relevant chemical space. ACS Omega 2021;6:22478–86.
[29] Conery AR, Rocnik JL, Trojer P. Small molecule targeting of chromatin writers in cancer. Nat Chem Biol 2022;18:124–33.
[30] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47:D930–40.
[31] Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 2015;43:W612–20.
[32] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28:31–6.
[33] RDKit, (n.d.). https://www.rdkit.org (accessed 08 January 08 2022).
[34] MolVS, (n.d.). https://molvs.readthedocs.io/en/latest/(accessed 08 accessed January 2022).
[35] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54.
[36] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 2002;42:1273–80.
[37] Chávez-Hernández AL, Juárez-Mercado KE, Saldívar-González FI, Medina-Franco JL. Towards the de novo design of HIV-1 protease inhibitors based on natural products. Biomolecules 2021;11:1805.
[38] Vivek-Ananth RP, Sahoo AK, Baskaran SP, Samal A. Scaffold and structural diversity of the secondary metabolite space of medicinal fungi. ACS Omega 2023;8:3102–13.
[39] Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, Samal A. IMPPAT: a curated database of Indian medicinal plants, phytochemistry and therapeutics. Sci Rep 2018;8:4329.
[40] Perron Q, da Silva VBR, Atwood B, Gaston-Mathé Y. Key points to succeed in Artificial Intelligence drug discovery projects. Chem Int 2022;44:19–21.
[41] Schneider G, Clark DE. Automated de novo drug design: are we nearly there yet? Angew Chem Int Ed Engl 2019;58:10792–803.