Chapter

Chemoinformatic Approach: The Case of Natural Products of Panama

Dionisio A. Olmedo and José L. Medina-Franco

Abstract

Chemoinformatic analysis was used to characterize a compound database of natural products from Panama and other reference collections. Data mining allowed to compare drug-likeness properties with public and commercial software and to achieve a statistical analysis of the physicochemical properties. Visualization of the chemical space in 3D indicates a high structural similarity. Molecular flexibility and complexity were evaluated using 2D descriptors, whereas the molecular scaffold was obtained using the Murcko method, and these showed few differences between the explored data set. In this chapter, we also present and discuss an example of the application of the chemoinformatic approach using the concept of modeling the activity landscape to study the structure-activity relationships (SARs) of compounds with activity against *Plasmodium falciparum*.

Keywords: chemoinformatic, complexity, data mining, physicochemical properties, scaffold

1. Introduction

Natural products (NPs) and their derivatives constitute a significant fraction of approved drugs [1–3], bioactive compounds [4–8], and lead compounds for drug discovery [9]. NP fragment has been used to guide the synthesis of bioactive compounds and generate BIOS combinatorial libraries [10–15]. NPs have structures with different substituent patterns, giving rise to different biological activities for compounds with very similar structures [16–19]. These bioactive metabolites have greater affinity for biological targets and, overall, may have better bioavailability than synthetic compounds, and the presence of pan-assay interference compounds (PAIN) is less frequent in this type of product [20]. The chemoinformatic analysis of several databases of NPs developed by academic institutions and private companies [21] has been carried out in different countries. Thus, the following databases were obtained: BIOFACQUIM [22], CIFPMA [23], NuBBE [24, 25], NANPDB [26], TCM [27], HIT [28], and NPACT [29]. The application of chemoinformatic tools involves the generation, manipulation, and analysis of data set of chemical substances. This allows us through mathematical calculations to order, develop, and evaluate structural information that can be visualized in 2D and 3D [30]. The determination of the physicochemical properties carried out on different databases of NPs and principal component analysis (PCA) was used as an approximation to display the chemical spaces [22–24, 31–37].

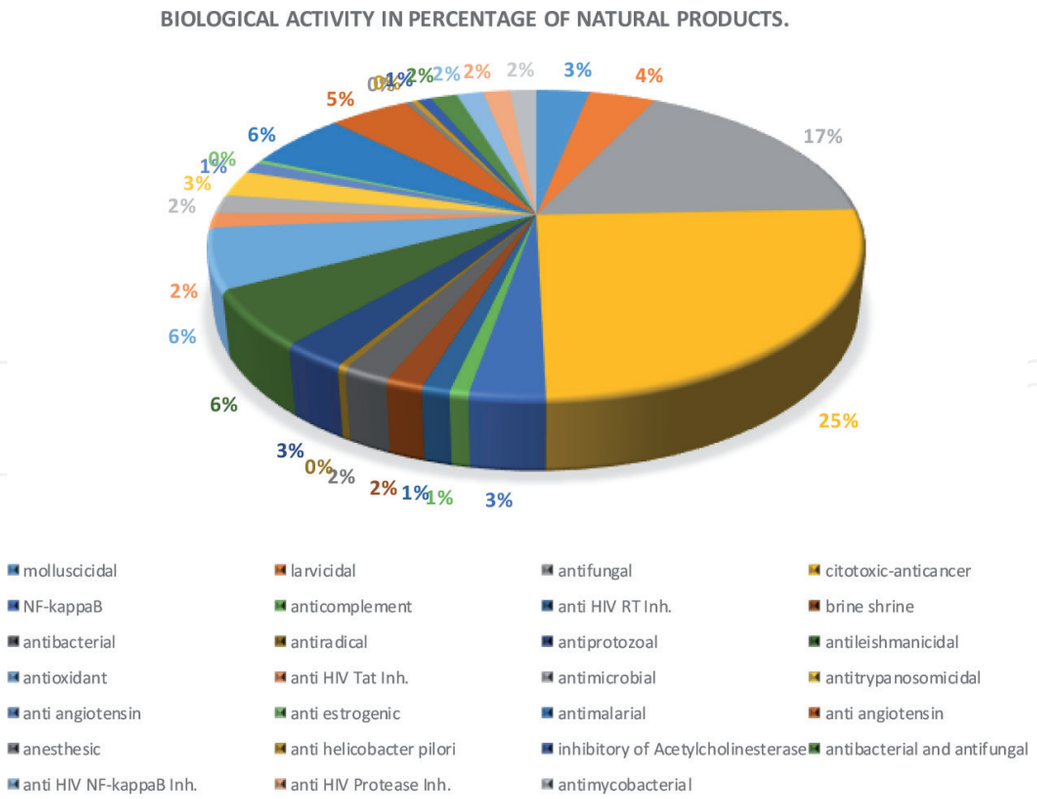


Figure 1.
Biological endpoints and targets in which natural products from Panama present bioactivity.

Computational exploration of NPs has increased in recent years, giving greater relevance to studies that include structural diversity metrics calculated with parameters based on distances such as Euclidean distance, Manhattan distances, and Cosine distance. Other criteria are based on circular fingerprint (ECFP-4, ECFP-6) [22–24, 38–45] and fingerprint based on substructure (MACCS, PubChem) [22–24, 39–45]. Another metric used in NPs is the comparison by similarity that uses the Tanimoto index/Tanimoto coefficient [22–24, 45–49].

In this study, the molecular scaffolds of natural products have been obtained using the Murcko method [22–24, 50–57]. Meanwhile, the molecular complexity is frequently evaluated by descriptors in 2D such as fraction of sp^3 hybridized carbons (Fsp^3) [23], fraction of chiral centers (FCC) [23], and globularity [22–24, 58–63].

An update of the Natural Products Database from the University of Panama (UPMA) containing 454 compounds (Unpublished data) has been evaluated against different therapeutic targets such as cytotoxicity bioassay in cell lines, antifungal assay in vitro, parasites of tropical diseases (*Leishmania* sp., *Plasmodium falciparum*, and *Trypanosoma cruzi*), and the bioassay against HIV-1 virus, demonstrating an inhibitor effect on protease, reverse transcriptase, nuclear factor NFkappaB, and Tat protein affecting the viral replication. These are the most significant biological targets in which the natural products from Panama present bioactivity. The values of their biological activities are represented as percentages in **Figure 1**.

2. Application of chemoinformatic antimalarial databases: case of natural products from Panama

2.1 Preparation curated and processing of data set

In this chapter, we present a chemoinformatic analysis of natural products with antimalarial activities (in vitro), expressed as pIC_{50} against sensitive and resistant

strains. Databases of natural products with antimalarial activity (NPAs) were constructed in-house by reviewing published articles including those compounds that were isolated and characterized by spectroscopic techniques of nuclear magnetic resonance. Around 1312 compounds were compared to 8 reference data sets: an open database, DrugBank (antimalarial drug), European Bioinformatics Institute. (CHEMBL drug indications) (antimalarial activities), Open Source Drug Discovery (OSDD) Malaria, Malaria Box (Medicines for Malaria Venture (MMV)), St. Jude Children's Research Hospital (St. Jude), Novartis (GNF Malaria Box), and GlaxoSmithKline (GSK) Tres Cantos antimalarial set. All data sets were curated using the "Wash" function implemented in the Molecular Operating Environment (MOE2018.0101) software [64]. The structure of the studied compounds was represented by simplified molecular input line entry system (SMILES) notation, thus obtaining 20,364 unique molecules that are summarized in **Table 1**. The difference between initial compounds and unique compounds is due to the fact that during the data preparation (curation process), the duplicate compounds are eliminated, those that have positive or negative partial loads have neutralized their protonation states, the metals are disconnected, and the energy is minimized using the molecular mechanistic force field (MMFF94). The result of the data curation is the reduction of the initial number of molecules present in the databases evaluated in this work.

2.2 Molecular descriptors

The descriptors of physicochemical properties, hydrogen bond acceptors (HBAs), hydrogen bond donors (HBDs), number of rotatable bonds (NRBs), the octanol/water partition coefficient (logP), topological polar surface area (TPSA),

Databases	Initial compounds	Unique compounds	Source
Natural Products Antimalarial (NPAs)	1353	1312	Databases of NP in house
DrugBank Version 5.0. (Drug Antimalarial)	26	4	https://www.drugbank.ca
European Bioinformatics Institute. (CHEMBL Drugs Indications) (Antimalarial activities	27	24	[https://www.ebi.ac.uk/chembl]
Open Source Drug Discovery (OSDD) Malaria	93	88	http://opensourcemalaria.org/
Malaria Box-Medicine of Malaria Venture (MMV)	124	124	https://www.ebi.ac.uk/chembl/malaria/source
St. Jude Children's Research Hospital's	1.478	1.478	https://www.ebi.ac.uk/chemblntd
Novartis-GNF Malaria Box	4.878	4.868	Available in: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3941073/ Available in: https://www.ebi.ac.uk/chemblntd
GlaxoSmithKline Tres Cantos Antimalarial	12.470	12.466	Open Source Malaria (GSK-TCMDC). Available in: https://www.ebi.ac.uk/chemblntd

Table 1.
Databases analyzed with chemoinformatic tools.

and molecular weight (MW), or others such as molar refractivity, are important physicochemical parameters for quantitative structure-activity relationship (QSAR) analysis. These molecular descriptors are based on Lipinski's rule and Verger's rule regarding the prediction of the pharmacological similarity of orally active pharmacological potential [65–67]. The statistical analysis of the physicochemical properties was realized with RStudio Software 1.0.136 AGPL [68].

2.3 3D visualization of chemical space of compounds with antimalarial activity

PCAs were done with MOE software [64], and the dominant characteristics are expressed as covariance and visualized with the corresponding 2D or 3D graphic score plot with DataWarrior program v. 5.0 [69]. **Figures 2–8** showed the distribution of different compounds with antimalarial activities in the chemical spaces.

In **Figures 2–8** we observed that NPs, drugs, and synthetic compounds occupy, in general, similar chemical space and are overlapping in most of the evaluated databases.

2.4 Molecular diversity based on fingerprints

Three binary molecular fingerprints were calculated with RStudio package rcdk: Extended connectivity fingerprints with diameter 4 (ECFP-4) for similarity searching, molecular access system (MACCS) keys of 166 bits for determining similarity and molecular diversity, and PubChem keys of 881 bits for encoding molecular fragment information [42–44]. The similarity of fingerprints by structural pairs of compounds was calculated with the Tanimoto coefficient and analyzed with the cumulative distribution function (CDF). This approach has been used to calculate, measure, and represent the molecular variety of compound data sets [23].

Figures 9–11 show the CDFs of the pairwise similarity of the different data sets evaluated with Tanimoto coefficient and ECPF-4, MACCS keys, and PubChem fingerprints, respectively.

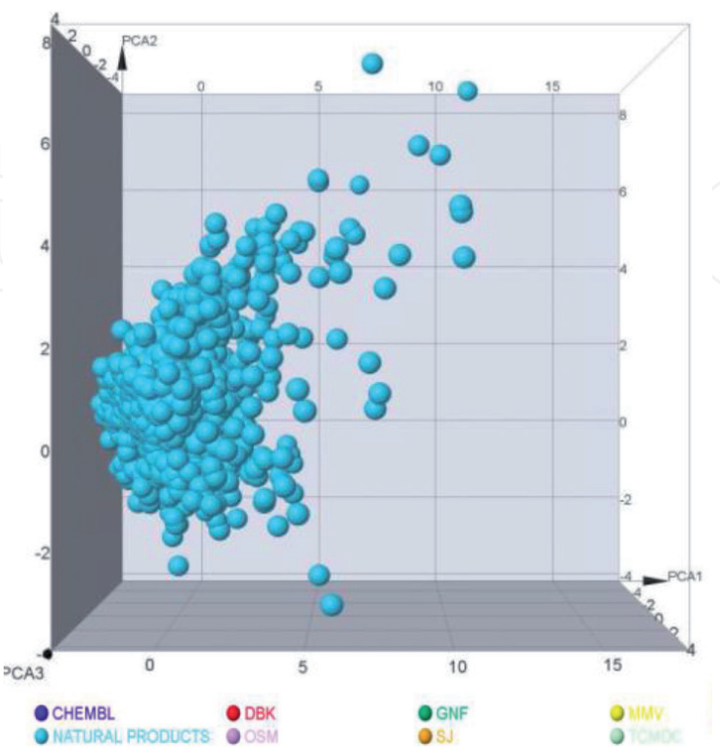


Figure 2.
3D visualization of the chemical space of natural product databases.

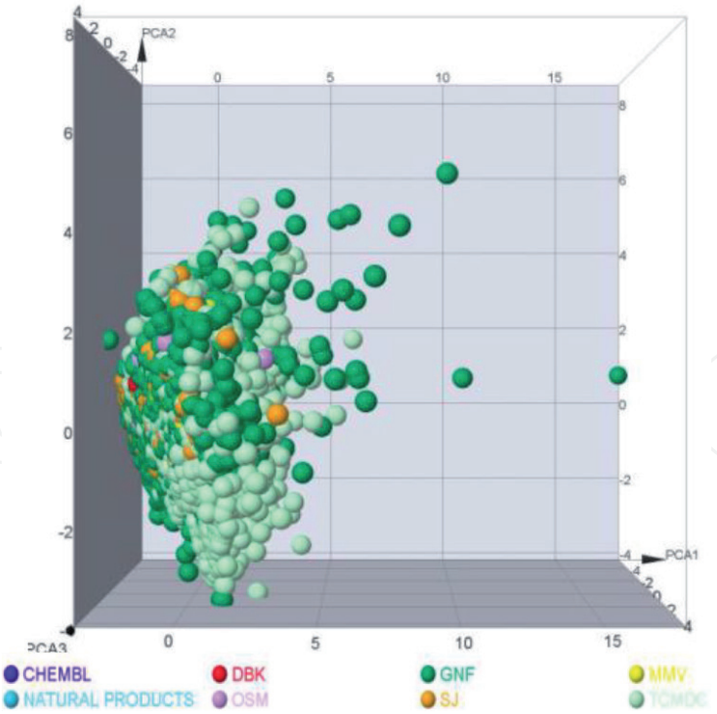


Figure 3.
3D visualization of the chemical space of synthetic compounds.

Figures 9–11 provide information on the structural diversity of the six databases. Similar approach has been previously published [23]; the curves obtained with ECFP-4 did not prove to be a suitable fingerprint representation for these data sets. In the three similarity graphs based on fingerprints, it is shown that the database of natural products with antimalarial activity, OMS, and MMV has the lowest molecular diversity, while GSK DB was the most diverse.

In **Tables 2–4**, the statistical values of the pairwise Tanimoto similarity with the data sets analyzed are shown. In these tables, CHEMBL and DrugBank databases are excluded from our analysis, due to the small amount of data.

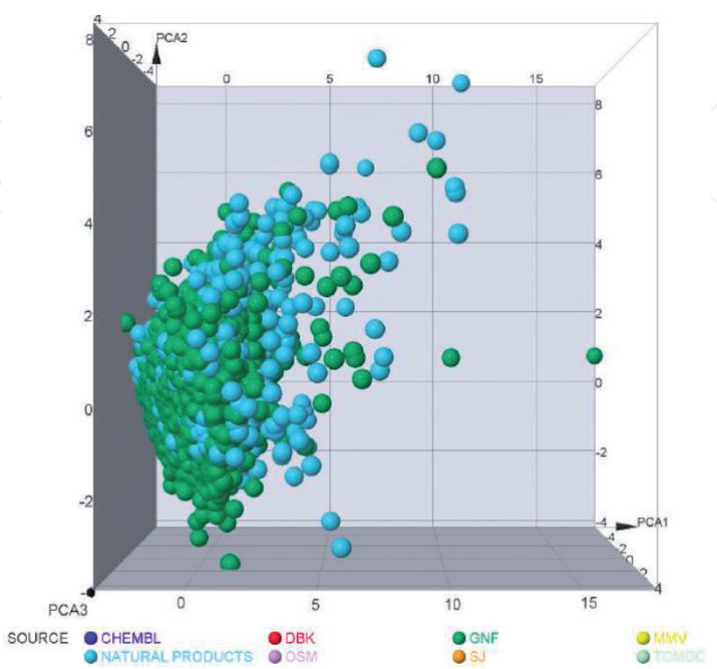


Figure 4.
3D visualization of the chemical spaces of natural products and GNF DBs.

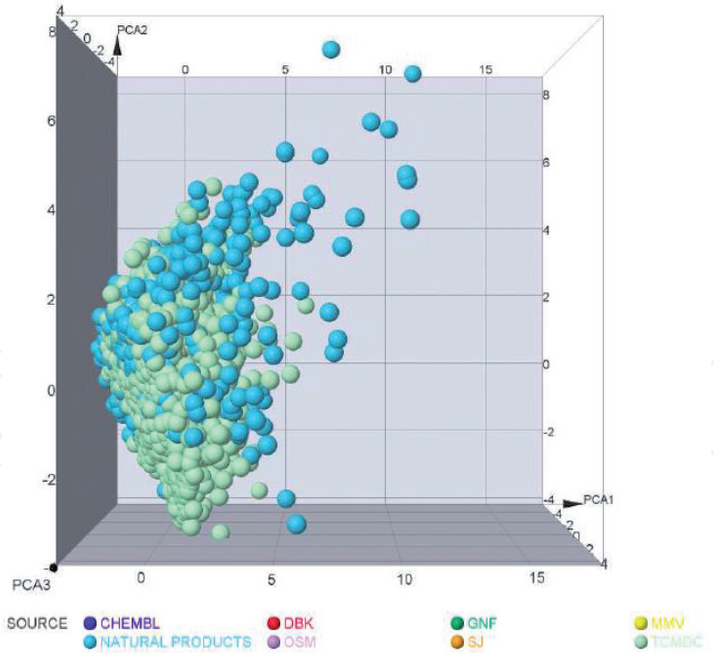


Figure 5.
3D visualization of the chemical spaces of natural products and TCMDC DBs.

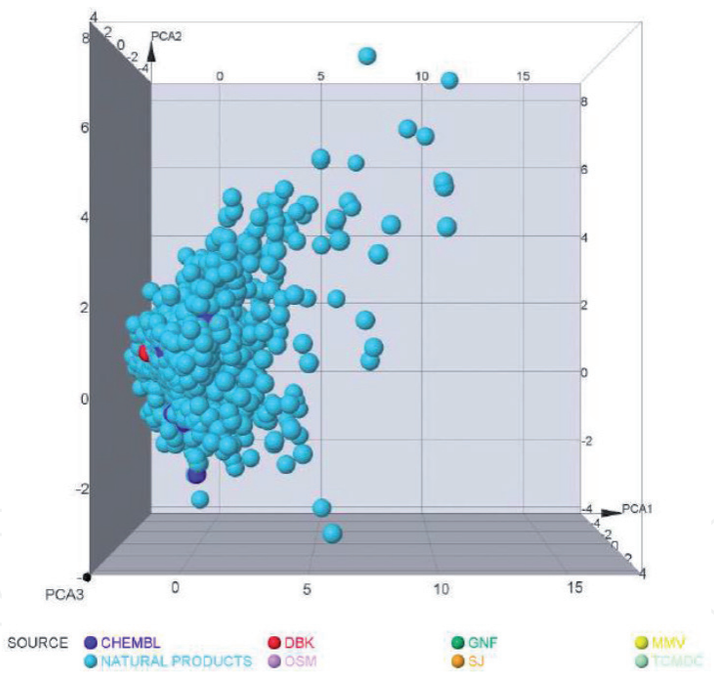


Figure 6.
3D visualization of the chemical spaces of natural products and DBK DBs.

2.5 Molecular scaffolds: content and diversity

2.5.1 Scaffold content

Murcko scaffolds were calculated with the program Molecular Equivalent Indices (MEQI) [50, 51] and DataWarrior program [69]. MEQI has been used to obtain the codes corresponding to the chemotypes most frequently analyzed in the databases. [23, 45, 52–55]. The distribution and diversity of the molecular scaffolds present in the data sets were calculated and analyzed using the cyclic system retrieval (CSR) curves [42]. These curves were obtained by plotting the fraction of

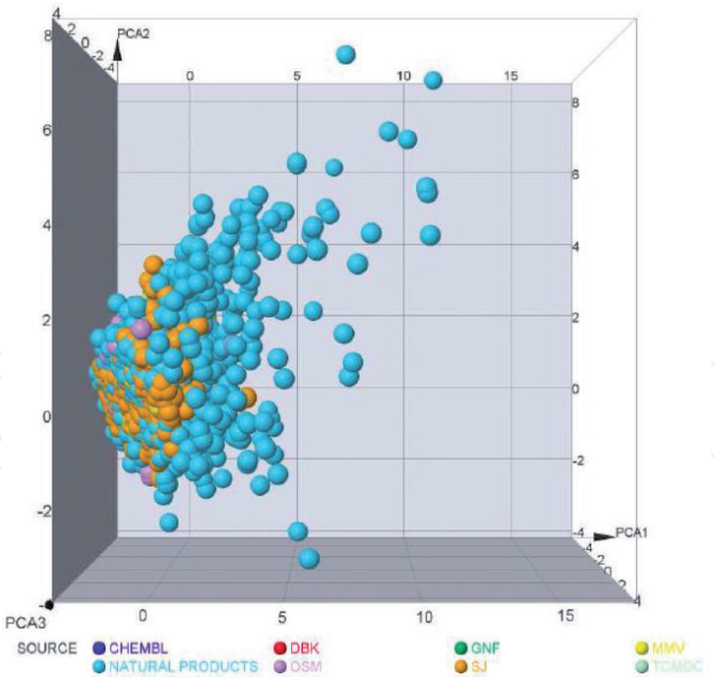


Figure 7.
3D visualization of the chemical spaces of natural products, OSM and St. Jude.

scaffold and the fraction of compounds that contain cyclic systems [43, 44].

Table 5 indicates that the MMV DB (0.491) was the most diverse in scaffold content taken as reference the F_{50} values compared to the data set from GSK (0.183), NPs (0.168), and GNF (0.161), respectively. CSR curves on **Figure 12** further confirm the relative scaffold variety of the eight databases. The analysis of area under curve (AUC) metrics associated with the CSR curves is reported in **Table 5**. The CSR curves showed that MMV has more variety in scaffold content with AUC value of 0.507. In contrast OSM, NPs, GNF, GSK, St. Jude, and CHEMBL were the least diverse (e.g., AUC scores of 0.745, 0.712, 0.705, 0.698, 0.655 and 0.607,

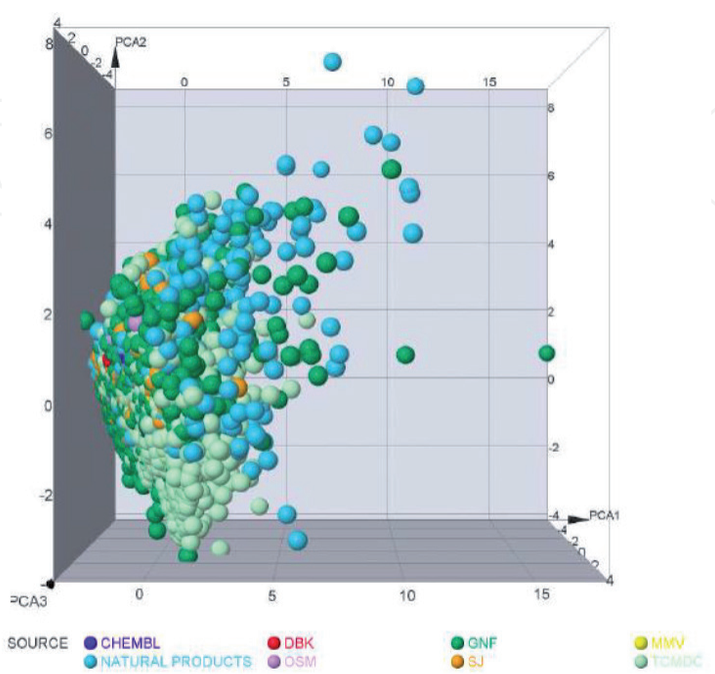


Figure 8.
3D visualization of the chemical spaces of all databases.

respectively). The CSR curves provide information on the diversity of the most frequent scaffolds in all databases.

2.5.2 Shannon entropy (SE) and scaled Shannon entropy (SSE)

The Shannon entropy has been adapted to measure the scaffold diversity based on the (**N**) number of most recurrent scaffolds [70]. The scaled Shannon entropy is a normalized value that measures the most common chemotypes present in a

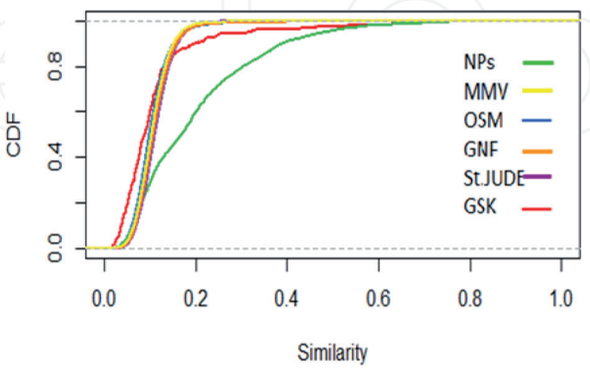


Figure 9.
Curve for cumulative frequency distribution (CFD) based on ECFP-4.

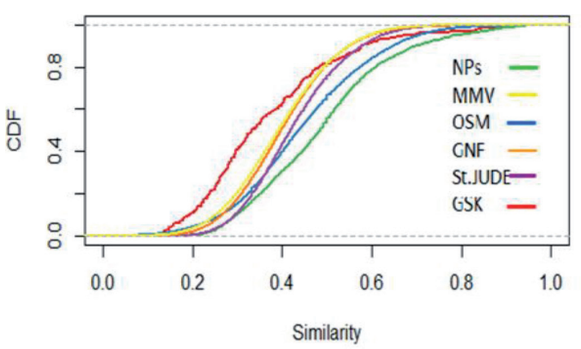


Figure 10.
Curve for cumulative frequency distribution based on MACCS keys.

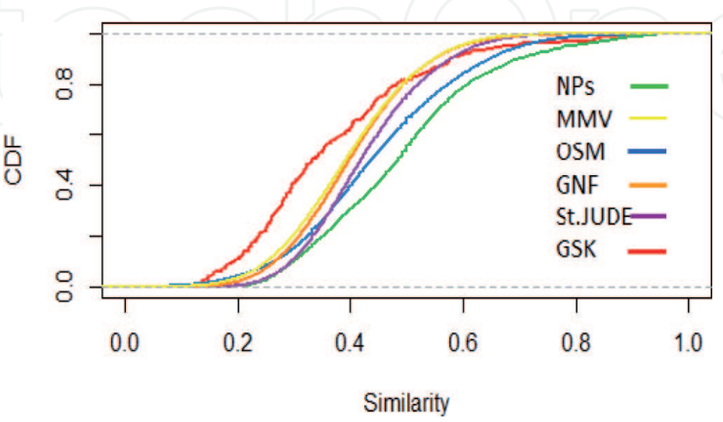


Figure 11.
Curve for cumulative frequency distribution based on PubChem.

database. Thus, SSE closer to 1 indicates higher scaffold diversity, while SSE closer to zero (0) indicates lower diversity. In this study, we calculated the SSE for values ranging from **N** = 10 to **N** = 40.

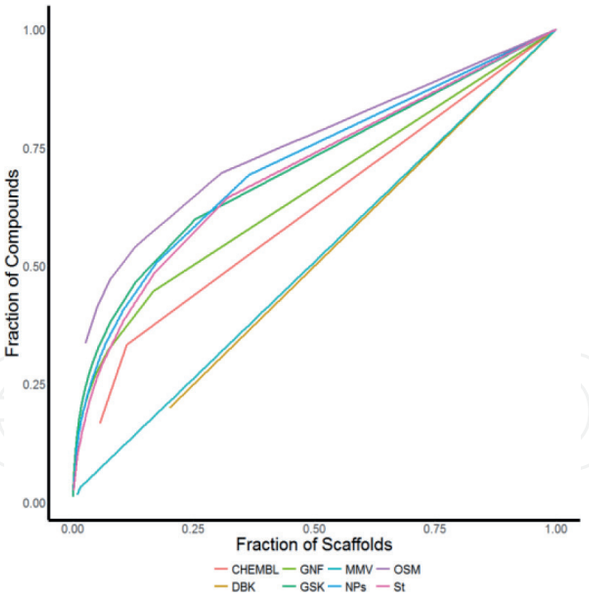


Figure 12.
Cyclic system retrieval curves for all databases evaluated in this study.

Similarity ECFP-4/Tanimoto coefficient						
DBs	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
GSK	0.01724	0.05789	0.08844	0.11490	0.12245	0.82353
NPs	0.00000	0.07826	0.09910	0.10565	0.12389	1.00000
OSM	0.00000	0.07826	0.09917	0.10607	0.12397	1.00000
MMV	0.00000	0.07826	0.09924	0.10615	0.12403	1.00000
ST JUDE	0.00000	0.08197	0.10345	0.10980	0.12857	1.00000
GNF	0.00000	0.08209	0.10345	0.10772	0.12739	1.00000

Table 2.
The statistical values of the similarity of the Tanimoto coefficient with ECFP-4.

Similarity MACCS keys/Tanimoto coefficient						
DBs	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
GSK	0.07813	0.25682	0.33333	0.37009	0.45581	0.92683
NPs	0.00000	0.34426	0.43636	0.44673	0.54545	1.00000
OSM	0.00000	0.34483	0.43636	0.44693	0.54545	1.00000
MMV	0.00000	0.34483	0.43636	0.44677	0.54412	1.00000
ST JUDE	0.00000	0.33333	0.41250	0.42313	0.50000	1.00000
GNF	0.00000	0.31746	0.39437	0.39999	0.47619	1.00000

Table 3.
The statistical values of the similarity of the Tanimoto coefficient with MACCS keys.

Figure 13 shows a histogram with the distribution of the 40 most populated scaffolds in NPAs. The histogram includes the corresponding chemotype code. The comparison of the scaffolds of the NPAs allowed the identification of the 68MBD chemotype as one of the most active compounds in this database.

Similarity PubChem/Tanimoto coefficient						
DBs	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
GSK	0.08125	0.24500	0.37555	0.40263	0.54002	1.00000
NPs	0.03684	0.32298	0.43802	0.46184	0.58621	1.00000
OSM	0.03684	0.32340	0.43902	0.46253	0.58730	1.00000
MMV	0.03684	0.32444	0.44033	0.46321	0.58791	1.00000
ST JUDE	0.03684	0.38224	0.47143	0.47624	0.56195	1.00000
GNF	0.00000	0.40598	0.48117	0.47800	0.55446	1.00000

Table 4.
The statistical values of the similarity of the Tanimoto coefficient with PubChem.

DBs	Number of Compounds (M)	Unique chemotypes (N)	FN/M	NSING	FNSING/M	FNSING/NS	AUC	F_{50}
NPs	1298	629	0.4846	400	0.3082	0.6359	0.7125	0.1685
DBK	5	5	1.0000	5	1.0000	1.0000	0.4800	0.4000
CHEMBL	24	18	0.7500	16	0.6667	0.8889	0.6072	0.3333
OSM	89	39	0.4382	27	0.3034	0.6923	0.7453	0.1025
MMV	124	122	0.9839	120	0.9677	0.9836	0.5079	0.4918
St. JUDE	915	479	0.5235	325	0.3552	0.6785	0.6551	0.2474
GNF	4860	3229	0.6644	2690	0.5535	0.8331	0.7054	0.1615
GSK	12,463	6703	0.5378	5009	0.4019	0.7473	0.6982	0.1837

M = number of molecules in the BD, N = number of chemotypes or substructures, FN/M = chemotype diversity fraction, NSING = singleton number, FNSING/M = singleton fraction between total molecules, FNSING/N = fraction of singleton among total chemotypes, AUC = area under the curve, F50 = fraction of chemotype required to recover 50% of the molecules.

Table 5.
Summary of the scaffold diversity of the eight databases analyzed in this work.

2.5.3 Molecular complexity and flexibility

The structural descriptors used to quantify fraction of sp^3 hybridized carbons (Fsp^3) [23, 58, 63, 70], fraction of chiral centers (CCF) [23, 59, 63, 70], fraction of aromatic atoms (Faro-atm), globularity [60], principal moments of inertia (PMI), normalized principal moments of inertia ratio (NRP) [61, 62], molecular complexity, shape index of Kier, and molecular flexibility were calculated with DataWarrior program [69] and MOE 2018.0101 [64]. **Figures 14–19** showed the descriptors utilized to evaluate the complexity and the molecular flexibility.

Tables 6–8 summarize the statistics of the distribution of Fsp^3, FCC, and Faro-atm of NPs and reference data sets. These results indicate that the NP data set has the largest complexity molecular in Fsp^3 (0.63) and CCF (0.16) and a low distribution of Faro-atm (0.67–0.78). In contrast, GNF, MMV, St. Jude, and GSK DBs are very similar in these three metrics with values between 0.25 and 0.37, 0.27 and 0.37, and 0.014 and 0.025, respectively. In contrast, the structural flexibility was evaluated with the index of form presenting all databases in the range of 0.41–0.58 indicating that many of the compounds present sphericity and intermediate molecular flexibility (data not presented).

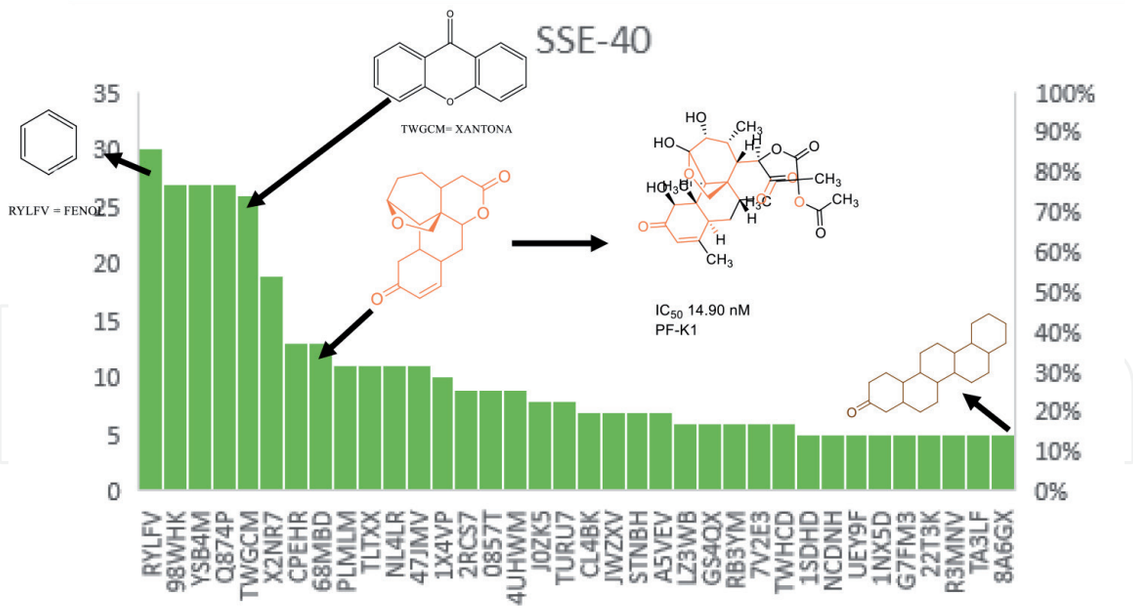


Figure 13.
Scaled Shannon entropy of the most frequent scaffolds with values ranging from 10 to 40 in natural products.

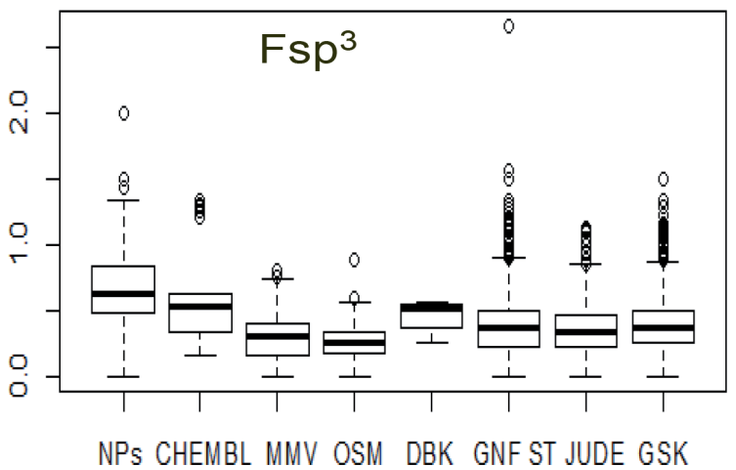


Figure 14.
Distribution of the fraction of sp^3 hybridized carbons in different databases.

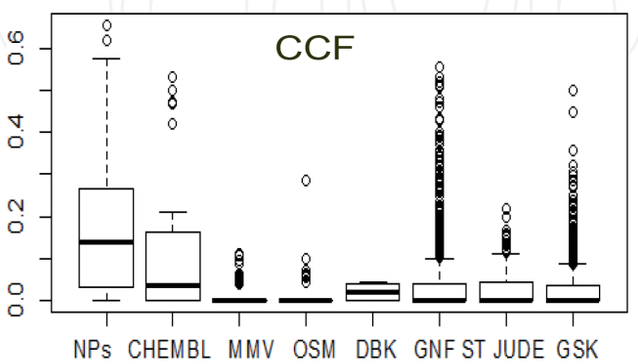


Figure 15.
Distribution of the fraction of chiral centers in different databases.

The descriptors globularity, PMI, and NRP did not prove to be suitable metrics to measure and differentiate the molecular complexity in the data sets evaluated. This is because the corresponding values computed for all data sets were very low

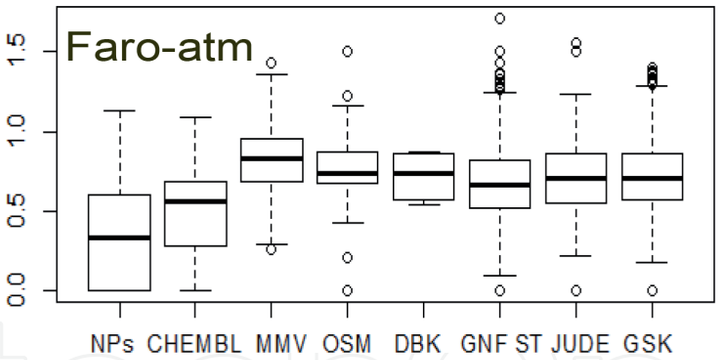


Figure 16.
Distribution of the fraction of aromatic atoms (Faro-atm) in different databases.

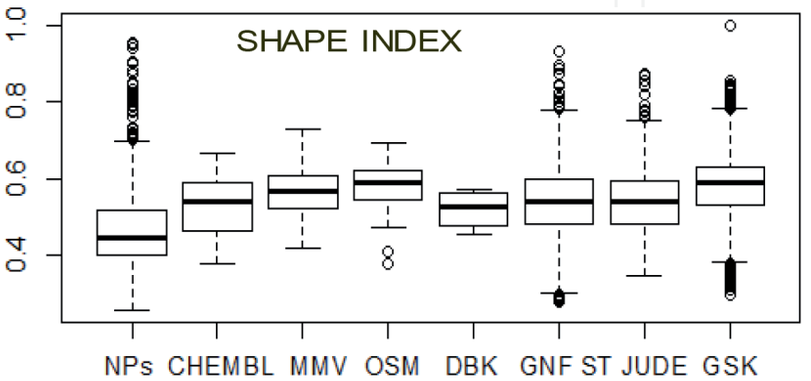


Figure 17.
Shape index distribution of different databases.

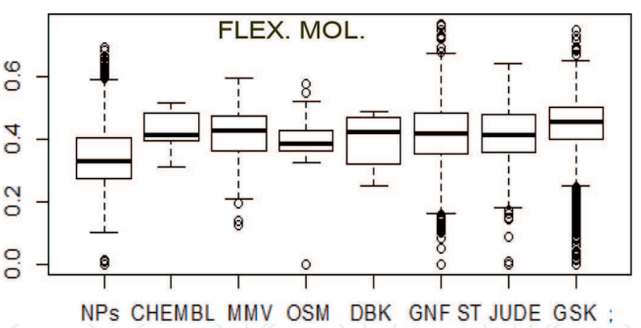


Figure 18.
Distribution of the molecular flexibility in different databases.

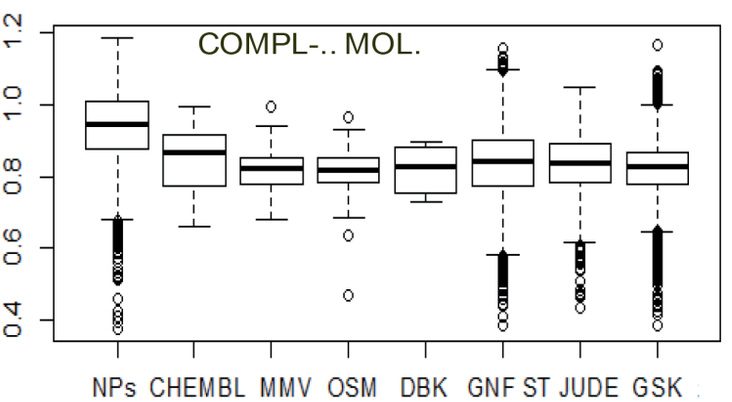


Figure 19.
Distribution of the molecular complexity in different databases.

Fraction of sp^3 hybridized atoms (Fsp^3)							
DBs	**Min**	**1qst**	**median**	**mean**	**3qrt**	**max**	**dev.st**
NPs	0.000	0.481	0.636	0.656	0.833	2.000	0.254
CHEMBL	0.167	0.342	0.536	0.621	0.627	1.333	0.374
MMV	0.000	0.167	0.300	0.316	0.402	0.800	0.190
OSM	0.000	0.174	0.255	0.277	0.338	0.893	0.145
DBK	0.250	0.438	0.519	0.463	0.545	0.565	0.175
GNF	0.000	0.227	0.364	0.377	0.500	2.667	0.207
STJUDE	0.000	0.222	0.333	0.353	0.471	1.136	0.178
GSK	0.000	0.250	0.375	0.372	0.500	1.500	0.180

Table 6.
Distribution of Fsp^3 in different databases.

Fraction of chiral centers (CCF)							
DBs	**min**	**1qst**	**median**	**mean**	**3qrt**	**max**	**dev.st**
NPs	0.000	0.033	0.139	0.161	0.267	0.656	0.145
CHEMBL	0.000	0.000	0.036	0.128	0.141	0.533	0.192
MMV	0.000	0.000	0.000	0.014	0.000	0.111	0.028
OSM	0.000	0.000	0.000	0.008	0.000	0.286	0.035
DBK	0.000	0.000	0.019	0.020	0.040	0.043	0.024
GNF	0.000	0.000	0.000	0.025	0.040	0.556	0.053
STJUDE	0.000	0.000	0.000	0.024	0.045	0.217	0.037
GSK	0.000	0.000	0.000	0.017	0.034	0.500	0.033

Table 7.
Distribution of FCC in different databases.

Fraction of aromatic atoms (Faro-atm)							
DBs	**min**	**1qst**	**median**	**mean**	**3qrt**	**max**	**dev.st**
NPs	0.000	0.000	0.324	0.341	0.600	1.133	0.294
CHEMBL	0.000	0.299	0.556	0.509	0.690	1.091	0.321
MMV	0.261	0.682	0.826	0.817	0.956	1.429	0.230
OSM	0.000	0.677	0.733	0.786	0.860	1.500	0.232
DBK	0.538	0.591	0.733	0.720	0.862	0.875	0.171
GNF	0.000	0.522	0.667	0.670	0.818	1.714	0.235
STJUDE	0.000	0.553	0.712	0.708	0.857	1.556	0.216
GSK	0.000	0.571	0.706	0.713	0.857	1.400	0.208

Table 8.
Distribution of fraction of aromatic atoms.

(close to zero) and did not differentiate the data sets (data not shown). The large molecular complexity of NPs measured is in agreement with previous studies using similar metrics [23, 63, 71].

3. Activity landscape modeling

The methods of modeling the landscape based on properties of the compounds (property landscape modeling (PLM)) is at the interface between experimental sciences and computational chemistry, being a frequent strategy to systematically describe the structure-property relationships (SPR) of the compound data set [72]. PLM have been used in medicinal chemistry in the stages of drug discovery with a quantitative, descriptive, and statistical approach to activity cliffs [72–74]. Structure-activity relationships (SARs), using the concept of modeling the activity landscape (activity landscape modeling ALM), are an increasing common practice in the drug discovery process to identify the activity cliffs, guide the optimization of compound hits, and to avoid the deleterious effects of the activity cliffs in the studies of the classic models of QSAR and in the search of structural similarity. In this

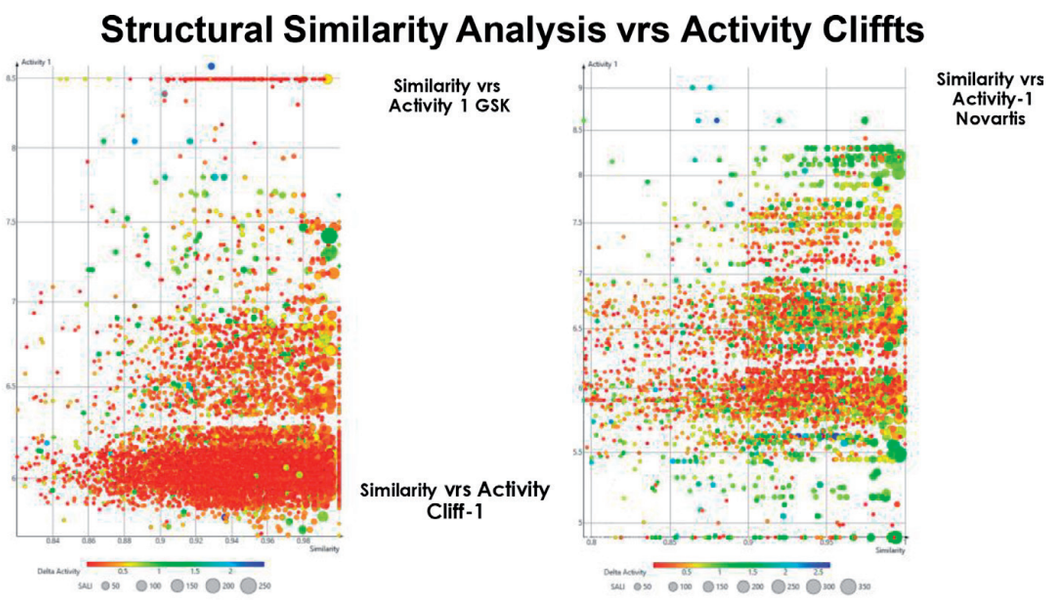


Figure 20.
Structural similarity compared with activity cliffs in NPAs.

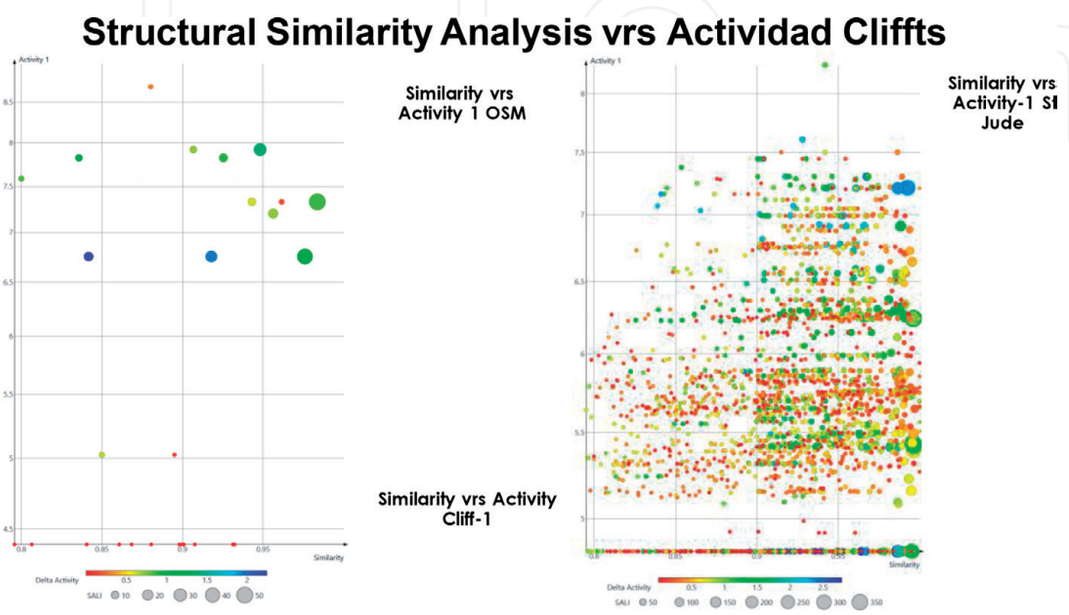
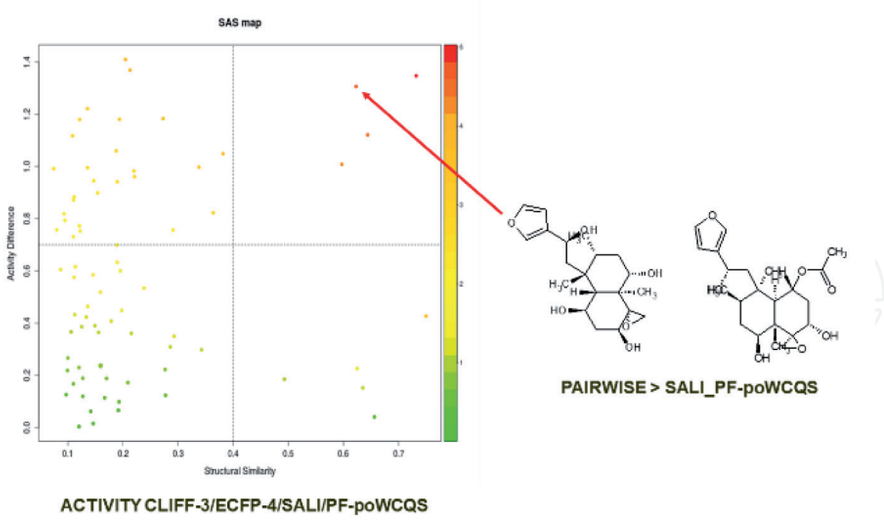
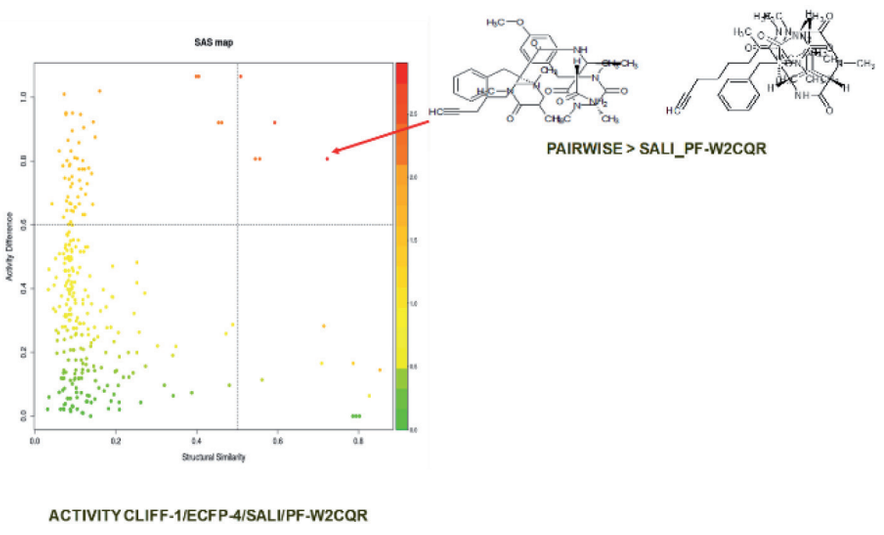


Figure 21.
Structural similarity compared with activity cliffs in GSK and Novartis (GNF).

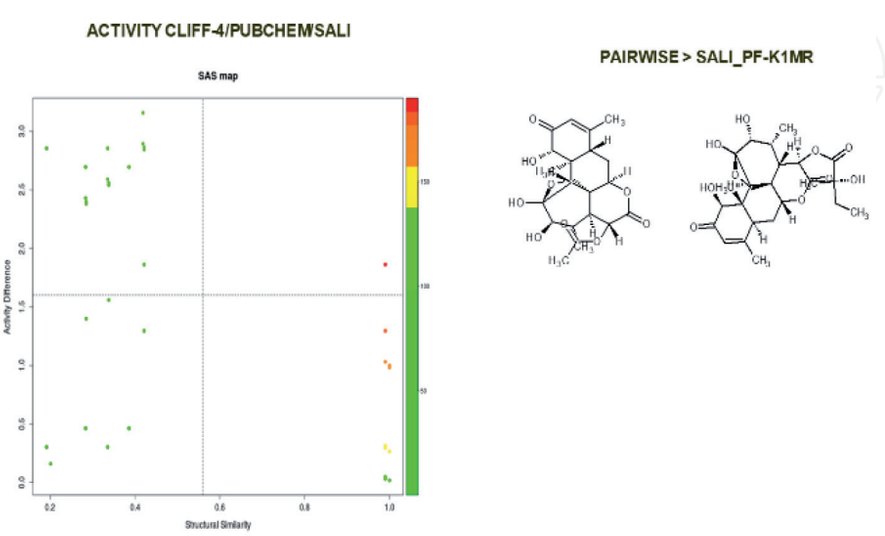
research we analyze, through the web tool Activity Landscape Plotter (ALP) [72], a set of data from NPs from Panama with antimalarial activity against four strains of



(a)



(b)



(c)

Figure 22.
SAS maps of compounds with antimalarial activity ((a), (b), and (c)) through the web tool activity landscape plotter.

Plasmodium falciparum in the erythrocyte gametocyte stage (**Figures 20** and **24**).

The generation and comparison of structure-activity pairs, by structure-activity similarity maps (SAS map). The SAS map has been used to link up structure and biological activity, based on a systematic pairwise comparison of all the compounds in a data set analyzed. We compare the values of structure-activity similarity, the activity difference, and structure-activity landscape index (SALI) to find the pairs of compounds with high molecular similarity and the activity difference that are located in the upper right quadrant of the SAS map (activity cliffs) [72–76]. **Figures 17–19** show SAS map in NP of Panama, NP published, GSK, and GNF. In SAS maps, data points are colored by density (**Figure 22**).

The SAS maps using the molecular fingerprints EFCP-4, MACCS keys, and PubChem led to the identification of a total of 26 pairs of compounds with structure-activity similarity ratios >0.50 and structure-activity landscape index values varying between 0.3 and 5.0. The web application Activity Landscape Plotter [72] is a tool that allows us to perform QSAR. The SAS generated represent 55 natural products isolated in Panama with antimalarial activity which were analyzed and compared the biological activities against strains of *Plasmodium falciparum* sensitive, resistant and multiresistant. The analysis with the parameters the (SAS / Tanimoto index / ECFP-4), a total of twenty-six pairs of compounds showed similarity values greater than 70%, sixteen pairs greater than 80% and only two pairs of compounds gave a similarity greater than 85%. While with activity cliffs, only three pairs of compounds show structural similarity correlated with the values of pIC50 activity [72, 77].

SAS maps are color-coded according to their intensity and we observe that most pairs of compounds with antimalarial activity show an intense red color. A nalyzed are located in the region of little structural similarity, indicating that the natural products have high structural diversity and low difference in activity, attributed to having similar functional groups in their molecules.

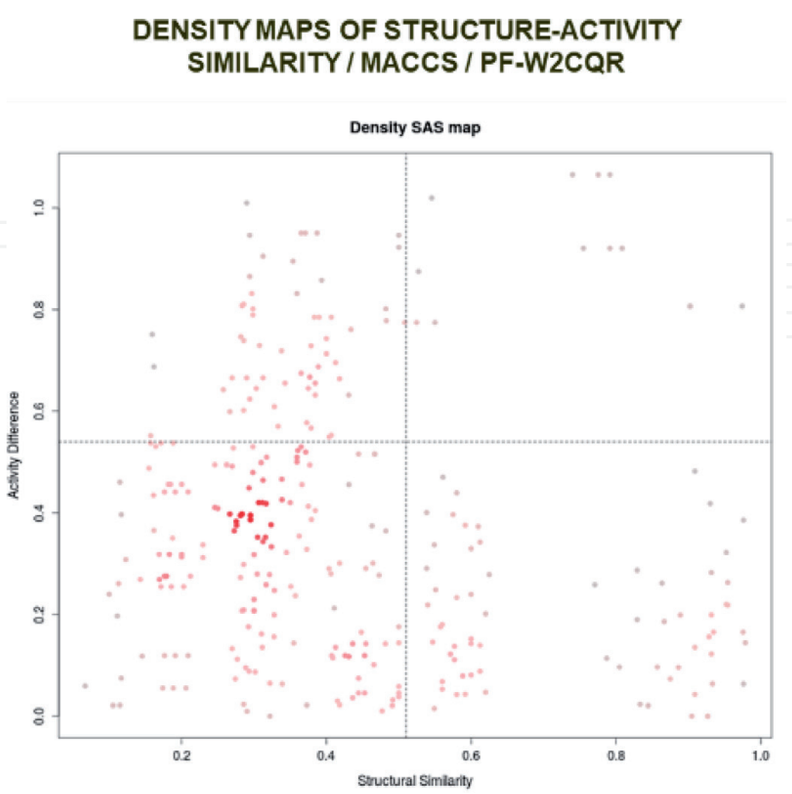


Figure 23.
DAS map with MACCS key fingerprint.

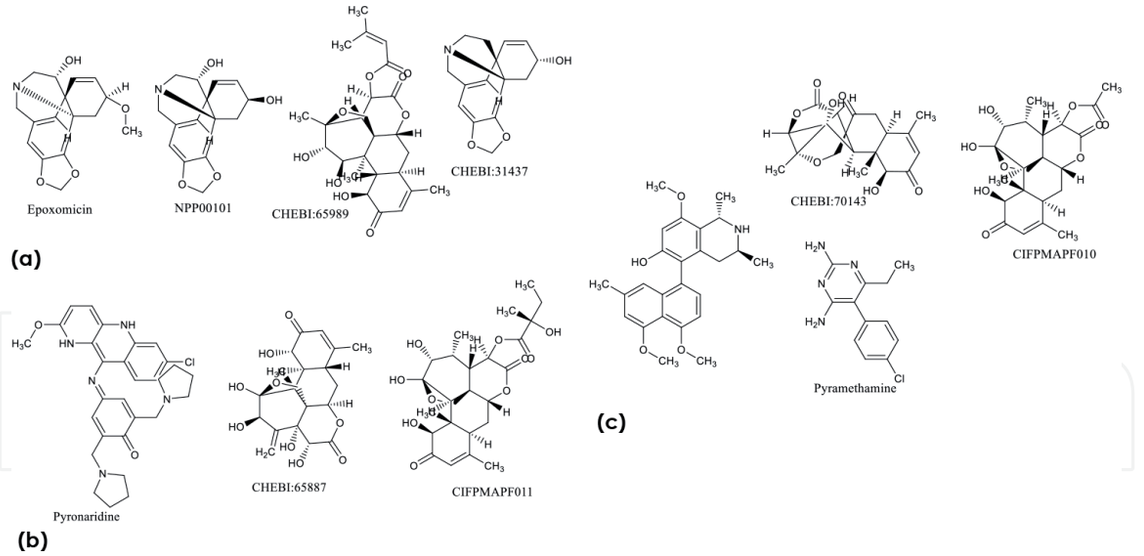


Figure 24.
Antimalarial compounds in NPs from Panama.

DAS maps represent the pairwise activity differences for each possible pair of compounds in an evaluated data set, against two biological targets. These maps permitted to differentiate if a structural modification can increase or decrease the activity under one target or other (**Figure 23**).

With this web application, we have carried out a QSAR study in a fast, simple, and easily interpretable way, obtaining three natural products as leading computational compounds for their optimization as ***Plasmodium falciparum*** blockers, which exhibit a gametocidal activity [78] (**Figure 24**).

4. Conclusion

The chemoinformatic analysis of the 20,364 compounds (1312 NPs and 19,052 synthetic (MMV, OSM, GNF, St. Jude, GSK, CHEMBL, and DrugBank)) indicates that so many natural products and synthetic products (S) share the same chemical space showing molecules that have similar structural properties. NPs present a greater diversity based on fingerprint than the synthetic compounds. Also, NPs have a higher proportion of chiral carbons and atoms with sp^3 hybridization and greater complexity, while synthetic products contain a greater proportion of aromatic atoms. Finally, concerning the properties related to cyclicity, relative shape, and flexibility, all have very similar values, which could explain the antimalarial activity of computationally determined compound hits in this work against *Plasmodium falciparum*-sensitive (3D7, D6, poW, D10) and chloroquine-resistant strains (W2, Dd).

Acknowledgements

The DAO acknowledges the SNI 2018 awards from SENACYT of Panama.

Conflict of interest

The authors declare that there are no financial or commercial conflicts of interest.

Author details

Dionisio A. Olmedo[1*] and José L. Medina-Franco[2]

1 CIFLORPAN Center for Pharmacognostic Research on Panamanian Flora, College of Pharmacy, University of Panama, Panama City, Panama

2 DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico (UNAM), Mexico City, Mexico

*Address all correspondence to: ciflorp4@up.ac.pa

IntechOpen

References

[1] Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. Journal of Natural Products. 2016;**9**:629-661. DOI: 10.1021/acs.jnatprod.5b01055

[2] Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. Journal of Natural Products. 2016;**75**:311-335. DOI: 10.1021/np200906s

[3] Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. Journal of Natural Products. 2007;**70**:461-477. DOI: 10.1021/np068054v

[4] Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A, et al. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. International Journal of Molecular Sciences. 2018;**19**:1578. DOI: 10.3390/ijms19061578

[5] Gurnani N, Mehta D, Gupta M, Mehta BK. Natural products: Source of potential drugs. African Journal of Basic & Applied Sciences. 2014;**6**:171-186. DOI: 10.5829/idosi.ajbas.2014.6.6.21983

[6] Hong J. Role of natural product diversity in chemical biology. Current Opinion in Chemical Biology. 2011;**15**:350-354. DOI: 10.1016/j.cbpa.2011.03.004

[7] Schreiber SL. Organic chemistry: Molecular diversity by design. Nature. 2009;**457**:153-154. DOI: 10.1038/457153a

[8] Schneider G, Grabowski K. Properties and architecture of drugs and natural products revisited. Current Chemical Biology. 2007;**1**:115-127. DOI: 10.2174/2212796810701010115

[9] Cragg GM, Newman DJ. Natural products: A continuing source of novel drug leads. Biochimica et Biophysica Acta. 2013;**1830**:3670-3695. DOI: 10.1016/j.bbagen.2013.02.008

[10] Sen S, Prabhu G, Bathula C, Hati S. Diversity-oriented asymmetric synthesis. Synthesis. 2014;**46**:2099-2121. DOI: 10.1055/s-0033-1341247

[11] van Hattum H, Waldmann H. Biology-oriented synthesis: Harnessing the power of evolution. Journal of the American Chemical Society. 2014;**136**:11853-11859. DOI: 10.1021/ja505861d

[12] Welsch ME, Snyder SA, Stockwell BR. Privileged scaffolds for library design and drug discovery. Current Opinion in Chemical Biology. 2010;**14**:347-361. DOI: 10.1016/j.cbpa

[13] Wetzel S, Bon RS, Kumar K, Waldmann H. Biology-oriented synthesis. Angewandte Chemie (International Ed. in English). 2011;**50**:10800-10826. DOI: 10.1002/anie.201007004

[14] Ertl P, Roggo R, Schuffenhauer A, Natural Product-likeness A. Score and its application for prioritization of compound libraries. Journal of Chemical Information and Modeling. 2008;**48**:68-74. DOI: 10.1021/ci700286x

[15] Mang C, Jakupovic S, Schunk S, Ambrosi H-D, Schwarz O, Jakupovic J. Natural products in combinatorial chemistry: An andrographolide-based library. Journal of Combinatorial Chemistry. 2006;**8**(2):268-274. DOI: 10.1021/cc050143n

[16] Wach JY, Gademann K. Reduce to the maximum: Truncated natural products as powerful modulators of biological processes. Synlett. 2012;**23**:163-170. DOI: 10.1055/s-0031-1290125

[17] Feher M, Schmidt JM. Property distribution: Differences between

drugs, natural products, and molecule from combinatorial chemistry. Journal of Chemical Information and Modeling. 2003;**43**:218-227. DOI: 10.1021/ci0200467

[18] Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK. Quantifying biogenic bias in screening libraries. Nature Chemical Biology. 2009;**5**:479-483. DOI: 10.1038/nchembio.180

[19] Schenone M, Dancik V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nature Chemical Biology. 2013;**9**:232-240. DOI: 10.1038/nchembio.1199

[20] Baell JB. Feeling nature's PAINS: Natural products, natural product drugs, and pan assay interference compounds (PAINS). Journal of Natural Products. 2016;**79**(3):616-628. DOI: 10.1021/acs.jnatprod.5b00947

[21] Egieyeh S, Syce J, Christoffels A, Malan SF. Exploration of scaffolds from natural products with antiplasmodial activities, currently registered antimalarial drugs and public malarial screen data. Molecules. 2016;**21**:104. DOI: 10.3390/molecules21010104

[22] Pilón-Jiménez B, Saldivar-Gonzalez F, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A mexican compound database of natural products. Biomolecules. 2019;**9**:31. DOI: 10.3390/biom9010031

[23] Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL. Cheminformatic characterization of natural products from Panama. Molecular Diversity. 2017;**21**(4):779-789. DOI: 10.1007/s11030-017-9781-4

[24] Pilon AC, Valli M, Dametto AC, MEF P, Freire RT, Castro-Gamboa I, et al. NuBBEDB: An updated database to uncover chemical and biological information from Brazilian biodiversity. Scientific Reports. 2017;7:7215-1-7215-12. DOI: 10.1038/s41598-017-07451-x

[25] Valli M, dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, et al. Development of a natural products database from the biodiversity of Brazil. Journal of Natural Products. 2013;**76**(3):439-444. DOI: 10.1021/np3006875

[26] Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, Moumbock AF, Malange YI, et al. NANPDB: A resource for natural products from northern african sources. Journal of Natural Products. 2017;**80**(7):2067-2076. DOI: 10.1021/acs.jnatprod.7b00283

[27] Chen CY-C. TCM database@ Taiwan: The world's largest traditional chinese medicine database for drug screening in silico. PLoS ONE. 2011;**6**(1):e15939. DOI: 10.1371/journal.pone.0015939

[28] Ye H, Ye L, Kang H, Zhang D, Tao L, Tang K, et al. HIT: Linking herbal active ingredients to targets. Nucleic Acids Research. 2011;**39**:D1055-D1059. DOI: 10.1093/nar/gkq1165

[29] Mangal M, Sagar P, Singh H, Raghava GPS, Agarwal SM. NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database. Nucleic Acids Research. 2013;**41**:D1124-D1129. DOI: 10.1093/nar/gks1047

[30] Bhalerao SA, Verna DR, D'Souza LR, Teli NC, Didwana VS. Chemoinformatics: The application of informatic methods to solve chemical problem. Research Journal of Pharmaceutical, Biological and Chemical Sciences. 2007;**4**(3):475-499

[31] Wenderski TA, Stratton CF, Bauer RA, Kopp F, Tan DS. Principal component analysis as a tool for library design: A case study investigating natural products, brand-name drugs, natural product-like libraries,

and drug-like libraries. Methods in Molecular Biology. 2015;**1263**:225-242. DOI: 10.1007/978-1-4939-2269-7_18

[32] Medina-Franco JL, Mayorga-Martínez K, Giulianotti MA, Houghten RA, Pinilla C. Visualization of the chemical space in drug discovery. Current Computer-Aided Drug Design. 2008;**4**:322-333. DOI: 10.2174/157340908786786010

[33] Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, Zefirov NS. Progress in visual representations of chemical space. Expert Opinion on Drug Discovery. 2015;**10**(9):959-973. DOI: 10.1517/17460441

[34] Ringner M. What is principal component analysis? Nature Biotechnology. 2018;**26**:303-304. DOI: 10.1038/nbt0308-303

[35] Jolliffe I. Principal component analysis. In: Everitt BS, Howell DC, editors. Encyclopedia of Statistics in Behavioral Science. Vol. 3. Aberdeen, Chichester, UK: University of Aberdeen, John Wiley and Sons, Ltd; 2005. pp. 1580-1584. DOI: 10.1002/0470013192.bsa501

[36] Clemons PA, Wilson JA, Dančík V, Muller S, Carrinski HA, Wagner BK, et al. Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. Proceedings of the National Academy of Sciences of the United States of America. 2011;**108**:6817-6822. DOI: 10.1073/pnas.1015024108

[37] Djuric SW, Akritopoulou-Zanze I, Cox PB, Galasinski S. Compound collection enhancement and paradigms for high-throughput screening-an update. Annual Reports in Medicinal Chemistry. 2010;**45**:409-428

[38] Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling. 2010;**50**:742-754. DOI: 10.1021/ci100050t

[39] Medina-Franco JL, Martínez-Mayorga K, Peppard TL, Del Rio A. Chemoinformatic analysis of GRAS (generally recognized as safe) flavor chemicals and natural products. PLoS One. 2012;**7**(11):e50798. DOI: 10.1371/journal.pone.0050798

[40] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences. 2002;**42**(6):1273-1280

[41] Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated platform of small molecules and biological activities. Annual Reports in Computational Chemistry. 2008;**31**(4):217-241

[42] Rogers D, Brown R, Hahn M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening. Journal of Biomolecular Screening. 2005;**10**:682-686. DOI: 10.1177/1087057105281365

[43] Hert J, Willet P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, et al. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. Organic and Biomolecular Chemistry. 2004;**2**:3256-3266. DOI: 10.1039/B409865J

[44] Barnard JM, Downs GM. Chemical fragment generation and clustering software. Journal of Chemical Information and Computer Sciences. 1997;**37**:141-142. DOI: 10.1021/ci960090k

[45] González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL. Consensus diversity plots: A global diversity analysis of chemical libraries. Journal of Cheminformatics. 2016;**8**:63. DOI: 10.1186/s13321-016-0176-9

[46] Bajusz D, Rácz A, Károly Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal

of Cheminformatics. 2015;**7**:20. DOI: 10.1186/s13321-015-0069-3

[47] Owen JR, Nabney IT, Medina-Franco JL, López-Vallejo F. Visualization of molecular fingerprints. Journal of Chemical Information and Modeling. 2011;**51**:1552-1563. DOI: d10.1021/ci1004042

[48] Skinnider MA, Dejong CA, Franczak BC, McNicholas PD, Nathan A, Magarvey NA. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. Journal of Cheminformatics. 2017;**9**:46. DOI: 10.1186/s13321-017-0234-y

[49] Medina-Franco JL. Discovery and development of lead compounds from natural sources using computational approaches. In: Mukherjee P, editor. Evidence-Based Validation of Herbal Medicine. Amsterdam, The Netherlands: Elsevier; 2015. pp. 455-475

[50] Xu Y-J, Johnson M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. Journal of Chemical Information and Computer Sciences. 2002;**42**:912-926. DOI: 10.1021/ci025535l

[51] Xu Y-J, Johnson M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. Journal of Chemical Information and Computer Sciences. 2001;**41**:181-185. DOI: 10.1021/ci0003911

[52] Lopez-Vallejo F, Castillo R, Yepez-Mulia L, Medina-Franco JL. Benzotriazoles and indazoles are scaffolds with biological activity against Entamoeba histolytica. Journal of Biomolecular Screening. 2011;**16**:862-868. DOI: 10.1177/1087057111414902

[53] Hu Y, Bajorath J. Quantifying the tendency of therapeutic target proteins to bind promiscuous or selective compounds. PLoS ONE. 2015;**10**:e0126838. DOI: 10.1371/journal.pone.0126838

[54] Lipkus AH, Yuan Q, Lucas KA, Funk SA, Bartelt WF 3rd, Schenck RJ, et al. Structural diversity of organic chemistry. A scaffold analysis of the CAS registry. The Journal of Organic Chemistry. 2008;**73**:4443-4451. DOI: 10.1021/jo8001276

[55] Krier M, Bret G, Rognan D. Assessing the scaffold diversity of screening libraries. Journal of Chemical Information and Modeling. 2006;**46**:512-524. DOI: 10.1021/ci050352v

[56] Medina-Franco JL, Martínez-Mayorga K, Bender A, Scior T. Scaffold diversity analysis of compound data sets using an entropy-based measure. QSAR and Combinatorial Science. 2009;**28**:1551-1560. DOI: 10.1002/qsar.200960069

[57] Yongye AB, Waddell J, Medina-Franco JL. Molecular scaffold analysis of natural products databases in the public domain. Chemical Biology & Drug Design. 2012;**80**:717-724. DOI: 10.1111/cbdd.12011

[58] Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Germany: Wiley-VCH; 2009

[59] Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, et al. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. Proceedings of the National Academy of Sciences of the United States of America. 2010;**107**:18787-18792. DOI: 10.1073/pnas.1012741107

[60] Meyer AY. Molecular mechanics and molecular shape. III. Surface area and cross-sectional areas of organic molecules. Journal of Computational Chemistry. 1986;**7**:144-152. DOI: 10.1002/jcc.540070207

[61] Sauer WHB, Schwarz MK. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. Journal of Chemical Information and Computer Sciences. 2003;**43**:987-1003. DOI: 10.1021/ci025599w

[62] Rolfe A, Lushington GH, Hanson PR. Reagent based DOS: A 'click, click, cyclize strategy to probe chemical space. Organic & Biomolecular Chemistry. 2010;**8**:2198-2203. DOI: 10.1039/b927161a

[63] Méndez-Lucio O, Medina-Franco JL. The many roles of molecular complexity in drug discovery. Drug Discovery Today. 2017;**22**:120-126. DOI: 10.1016/j.drudis.2016.08.009

[64] Molecular Operating Environment (MOE). 2018.0101. 910-1010 Sherbrooke, St. W. Montreal, QC H3A 2R7; Canada: Chemical Computing Group, Corporate Headquarters Montreal

[65] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews. 2001;**46**(1-3):3-26. DOI: 10.1016/S0169-409X(00)00129-0

[66] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. Journal of Medicinal Chemistry. 2002;**45**(12):2615-2623. DOI: 10.1021/jm020017n

[67] Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nature Reviews. Drug Discovery. 2007;**6**(11):881-890. DOI: 10.1038/nrd2445

[68] RStudio Team. RStudio: Integrated Development for R. Boston: RStudio, Inc.; 2015. Available form: http://www.rstudio.com/

[69] Sander T, Freyss J, Von Korff M, Rufener C. Datawarrior: An open-source program for chemistry aware data visualization and analysis. Journal of Chemical Information and Modeling. 2015;**55**:460-473. DOI: 10.1021/ci500588j

[70] Godden JW, Bajorath J. Analysis of chemical information content using Shannon entropy. In: Lipkowitz KB, Cundari TR., editors. Reviews in Computational Chemistry. Hoboken: John Wiley & Sons, Inc.; 2007. pp. 263-289. DOI: 10.1002/9780470116449.ch5

[71] Lovering F, Bikker J, Humblet C. Escape from flatland: Increasing saturation as an approach to improving clinical success. Journal of Medicinal Chemistry. 2009;**52**:6752-6756. DOI: 10.1021/jm901241

[72] González-Medina M, Prieto-Martínez FD, Naveja J, Méndez-Lucio O, El-Elimat T, Pearce CJ, et al. Chemoinformatic expedition of the chemical space of fungal products. Future Medicinal Chemistry. 2016;**8**:1399-1412. DOI: 10.4155/fmc-2016-0079

[73] González-Medina M, Méndez-Lucio O, Medina-Franco JL. Activity landscape plotter: A web-based application for the analysis of structure–activity relationships. Journal of Chemical Information and Modeling. 2017;**57**:397-402

[74] Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discovery Today. 2014;**19**(8):1069-1080

[75] Bajorath J. Modeling of activity landscapes for drug discovery. Expert Opinion on Drug Discovery. 2012;**7**(6):463-473

[76] Garcia-Sanchez MO, Cruz-Monteagudo M, Medina-Franco JL. Challenges and advances in computational chemistry on physics quantitative structure-epigenetic activity relationship. In: Lezcznski J, Roy K, editors. Advances in QSAR Modeling: Application in Pharmaceutical, Chemical, Foods, Agricultural and Environmental Science, 24. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer Nature. Springer International Publishing AG; 2017. pp. 303-338

[77] Medina-Franco JL. Scanning structure–activity relationships with structure–activity similarity and related maps: From consensus activity cliffs to selectivity switches. Journal of Chemical Information and Modeling. 2012;**52**(10):2485-2493. DOI: 10.1021/ci300362x

[78] Kiszewski AE. Blocking plasmodium falciparum malaria transmission with drugs: The gametocytocidal and sporontocidal properties of current and prospective antimalarials. Pharmaceuticals. 2011;**4**(1):44-68. DOI: 10.3390/ph4010044