




Article

# Sampling and Mapping Chemical Space with Extended Similarity Indices

Kenneth López-Pérez <sup>1</sup>, Edgar López-López <sup>2,3</sup> , José L. Medina-Franco <sup>2,\*</sup>   
and Ramón Alain Miranda-Quintana <sup>1,\*</sup> 

<sup>1</sup> Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA; klopezperez@chem.ufl.edu

<sup>2</sup> DIFACQUIM Research Group, Department of Pharmacy, National Autonomous University of Mexico, Mexico City 04510, Mexico; elopez.lopez@cinvestav.mx

<sup>3</sup> Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico City 07000, Mexico

\* Correspondence: medinajl@unam.mx (J.L.M.-F.); quintana@chem.ufl.edu (R.A.M.-Q.)

**Abstract:** Visualization of the chemical space is useful in many aspects of chemistry, including compound library design, diversity analysis, and exploring structure–property relationships, to name a few. Examples of notable research areas where the visualization of chemical space has strong applications are drug discovery and natural product research. However, the sheer volume of even comparatively small sub-sections of chemical space implies that we need to use approximations at the time of navigating through chemical space. ChemMaps is a visualization methodology that approximates the distribution of compounds in large datasets based on the selection of satellite compounds that yield a similar mapping of the whole dataset when principal component analysis on a similarity matrix is performed. Here, we show how the recently proposed extended similarity indices can help find regions that are relevant to sample satellites and reduce the amount of high-dimensional data needed to describe a library’s chemical space.

**Keywords:** ChemMaps; chemical space; data visualization; extended similarity; similarity; sampling



**Citation:** López-Pérez, K.; López-López, E.; Medina-Franco, J.L.; Miranda-Quintana, R.A. Sampling and Mapping Chemical Space with Extended Similarity Indices. *Molecules* **2023**, *28*, 6333. <https://doi.org/10.3390/molecules28176333>

Academic Editors: Cheng Fang and Zhiyan Xiao

Received: 25 July 2023

Revised: 24 August 2023

Accepted: 26 August 2023

Published: 30 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Chemical space is an intuitive concept that has become a cornerstone in many areas of chemistry, traditionally in drug discovery but with increasing applications in other areas, such as natural product and food chemical research, organic synthesis planning, and library enumeration and design [1]. In recent years, the increasing number of compounds in large and ultra-large chemical libraries (most of them virtual) demands the development of fast and reliable visualization methods [2].

Although there are several intuitive concepts in chemistry (e.g., molecular similarity or “chemical beauty” [3]), there is no unique and “best” definition of chemical space. Different definitions have been reviewed recently that could be roughly divided into descriptor-independent and descriptor-dependent [4]. In the former, chemical space has been associated with the number of chemical structures that could possibly exist (in many instances, the definitions are focused on small organic compounds but, in principle, the chemicals could be of any type) [5]. Other definitions of chemical space are focused on the descriptor space in which the compounds are represented [6]. In this second case, the number of possible descriptors to define the chemical space of the same set of compounds could be very large, and that has led to the so-called *chemical multiverse* [4]. The choice of descriptors usually is set to the goals of this study, the type or nature of the compounds (e.g., small organic molecules, peptides, organometallic molecules, etc.), and the amounts of chemicals to describe (where large and ultra-large libraries require fast and as accurate as possible descriptors). One of the simplest chemical spaces for a given set of descriptors

would be defined by one, two, or three variables (in which spaces could be easily visualized in one-, two-, or three-dimensions using scatter plots). However, in most instances, the descriptor space is large, so dimension-reduction methods or networks are applied.

Visualization techniques that graphically represent the chemical space have been reviewed [7,8]. Examples of very common visualization methods that represent chemical spaces, in addition to principal component analysis [9–12], are self-organizing maps [13–15], *t*-distributed stochastic neighbor embedding [10,12,16], and generative topographic mapping [17–19]. Open-source tools to both compute molecular descriptors and generate graphical representations of chemical space have been reviewed [20].

One of the most recent developments in chemical space visualization is the Chemical Library Networks, which are very valuable resources for visualizing the chemical space of very large libraries [21]. Other visualization methods such as ChemGPS [22,23], ChemMaps [24], and Similarity Mapplet [25], rely on reference or “chemical satellite” compounds. Recent frameworks have also been proposed to dissect the properties of DNA-encoded libraries [26,27]. In principle, “satellites” are molecules whose similarities to the rest of the compounds in the library give enough information to generate a visualization of their chemical space [24]. In ChemGPS, chemical satellites have extreme properties or descriptor values that place them as outliers or reference compounds, with the purpose of reaching as much of the chemical space as possible. An obvious challenge is defining a set of generic or “universal” reference compounds because there is a large variety of compounds that a user might explore, e.g., organic drug-like compounds, natural products, peptides (that can vary significantly in size, etc.), inorganic molecules to name a few examples. In our previous attempts to address this issue, a subset of the database to be represented is used as adaptive satellite compounds in ChemMaps [24]. In that proof-of-concept study conducted with small datasets, it was concluded that ChemMaps is a feasible approach to produce reliable visualizations of the chemical space based on principal component analysis (PCA) of similarity matrices. The methodology worked better for relatively less diverse datasets but remained robust when used with diverse datasets. For compound datasets with small diversity, fewer satellites were enough to generate a reliable visualization of the chemical space. However, the applicability of ChemMaps to larger datasets was not as clearly established, with the adaptive satellite sampling remaining a difficult problem to tackle. Of note, Borrel et al. have developed an interactive webserver called “ChemMaps.com” to navigate visually the chemical space of large chemical databases, although it is not based on the concept of chemical satellites [28,29].

The goal of this work is to propose ways to dissect molecular libraries using sampling methods based on extended similarities (*vide infra*). With this, we aim to find the relevant regions of a library’s chemical space that are key to sample as “chemical satellites”, to generate a PCA visualization reminiscent of the entire library, and to provide the opportunity to study large chemical libraries in a more computationally efficient way. The extended similarity indices prove very versatile and efficient for this task, quickly identifying critical regions in the chemical space.

## 2. Theory

### 2.1. Extended Similarity

A key tool in our formalism is the notion of *extended similarity* [30,31]. Originally proposed as a way to speed up the comparison of drug-like molecules represented by binary fingerprints [32], these indices have since been generalized to deal with arbitrary categorical variables [33] and real-value inputs (like Cartesian coordinates [34,35] and molecular properties [36]). Despite this versatility, the central idea behind these variants is the same: comparing multiple objects at the same time, instead of performing pairwise comparisons. This results in a key advantage, since now comparing  $N$  objects only demands an  $O(N)$  scaling, as opposed to the traditional  $O(N^2)$ .

Due to its relatively recent introduction, we provide a brief description of how to calculate the extended similarity indices [30,31] used in this work. For a set of molecules rep-

represented by (binary) fingerprints, we first need to calculate a vector  $\Sigma = [\sigma_1, \sigma_2, \dots, \sigma_M]$ , representing the sum of every fingerprint bit position in the set. In order to see how each of the  $\sigma_k$  contributes to the similarity or dissimilarity of the  $N$  molecules set, we use the quantity  $\Delta_{\sigma_k} = |2\sigma_k - N|$ . This is combined with the coincidence threshold,  $\gamma$ , following a simple set of rules: (i) if  $2\sigma_k - N > \gamma$ , we have a 1-similarity, (ii) if  $N - 2\sigma_k > \gamma$ , we have a 0-similarity, and (iii) otherwise, we have a dissimilarity. In this way,  $\gamma$  effectively acts as an indicator determining at what point we can consider that the elements in a bit position are distributed uniformly. The final step is then to properly weight the cases in which, despite having assigned a similarity or dissimilarity, we do not have a perfect coincidence of “on” or “off” bits. We performed this with functions  $f_s$  and  $f_d$ , which could be conveniently defined as shown in Equation (1) below:

$$f_s(\Delta_{\sigma_k}) = \frac{\Delta_{\sigma_k}}{N}; f_d(\Delta_{\sigma_k}) = 1 - \frac{\Delta_{\sigma_k} - N \bmod 2}{N} \quad (1)$$

These steps lead to the natural generalization of many pairwise similarity indices. In particular, the (extended) Jaccard-Tanimoto index is given by:

$$s_{eJT} = \frac{\sum_{1-s} f_s(\Delta_{\sigma})}{\sum_{1-s} 1 + \sum_d 1} \quad (2)$$

Note how  $s$ ,  $1-s$ ,  $0-s$ , and  $d$  represent summations over the similar, 1-similar, 0-similar, and dissimilar columns, respectively. The sums in the denominator of Equation (2) indicate adding over all of the 1-similarity and dissimilarity columns, respectively.

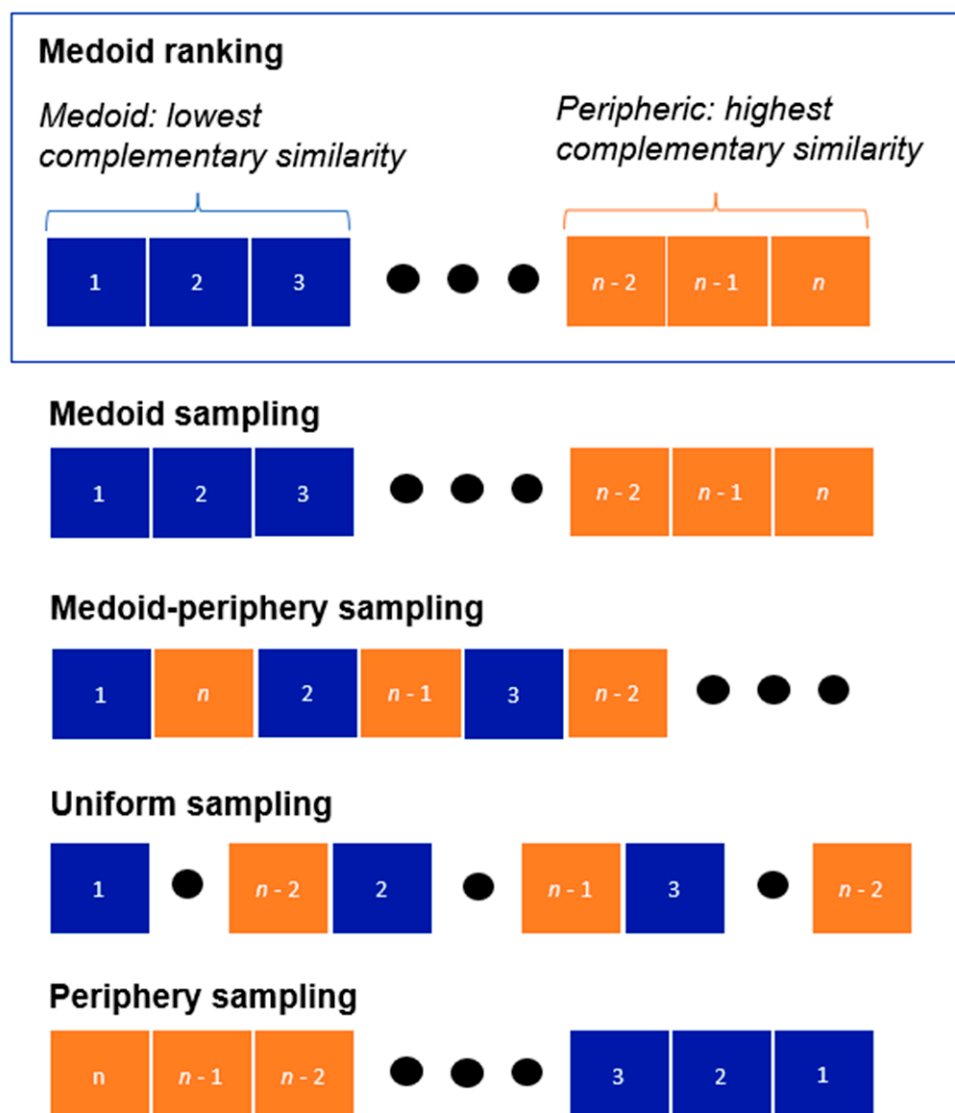
## 2.2. Sampling Techniques

The extended similarity measures provide a very convenient way to explore different regions in chemical space. If we calculate the effect of removing a single molecule from a library, this will indicate if said compound was part of a region of high density or low density of molecules. This can be performed by calculating the extended similarity of the set after removing the given molecule, which we have termed: the complementary similarity of a molecule. This is a very simple task since we only need to calculate the vector of column sums  $\Sigma = [\sigma_1, \sigma_2, \dots, \sigma_M]$ , subtract from it the fingerprint of the  $i$ th molecule,  $m_i = [s_{1i}, s_{2i}, \dots, s_{Mi}]$ , and then repeat the next steps described in the previous section over the vector  $\Sigma - m_i = [\sigma_1 - s_{1i}, \sigma_2 - s_{2i}, \dots, \sigma_M - s_{Mi}]$ , but taking into account that the new set now has  $N - 1$  molecules. It is important to highlight that the two most time-consuming steps in this algorithm: calculating  $\Sigma$  and calculating all of the  $\Sigma - m_i$  terms; both scale linearly, so this is a very efficient procedure. Then, after this process, we can identify molecules with low complementary similarity as belonging to the high-density (or “central”) region of the library, while molecules with bigger complementary similarity can be identified as outliers of the set (essentially, as points in the periphery or low-density region).

Here, we will use the ranking provided by the complementary similarity to explore four different ways to sample chemical satellites in the chemical space of a given compound dataset. The four approaches are schematically shown in Figure 1 and are detailed hereunder:

1. Medoid sampling: selecting molecules in increasing order of their complementary similarity values (sampling chemical space from the center-to-the-outside).
2. Medoid-periphery sampling: selecting molecules in an alternating pattern, with odd selections (1, 3, 5, ...) coming from the medoid region, and even selections (2, 4, 6, ...) coming from the outlier region.
3. Uniform sampling: the data are separated into five batches, and then we take one molecule from each of them in increasing order of complementary similarity within each batch.

4. Periphery sampling: selecting molecules in decreasing order of their complementary similarity values (sampling chemical space from the outside-to-the-center).



**Figure 1.** Schematic representation of the medoid, medoid–periphery, uniform, and periphery chemical satellite sampling.

### 3. Results

#### 3.1. Overall Diversity

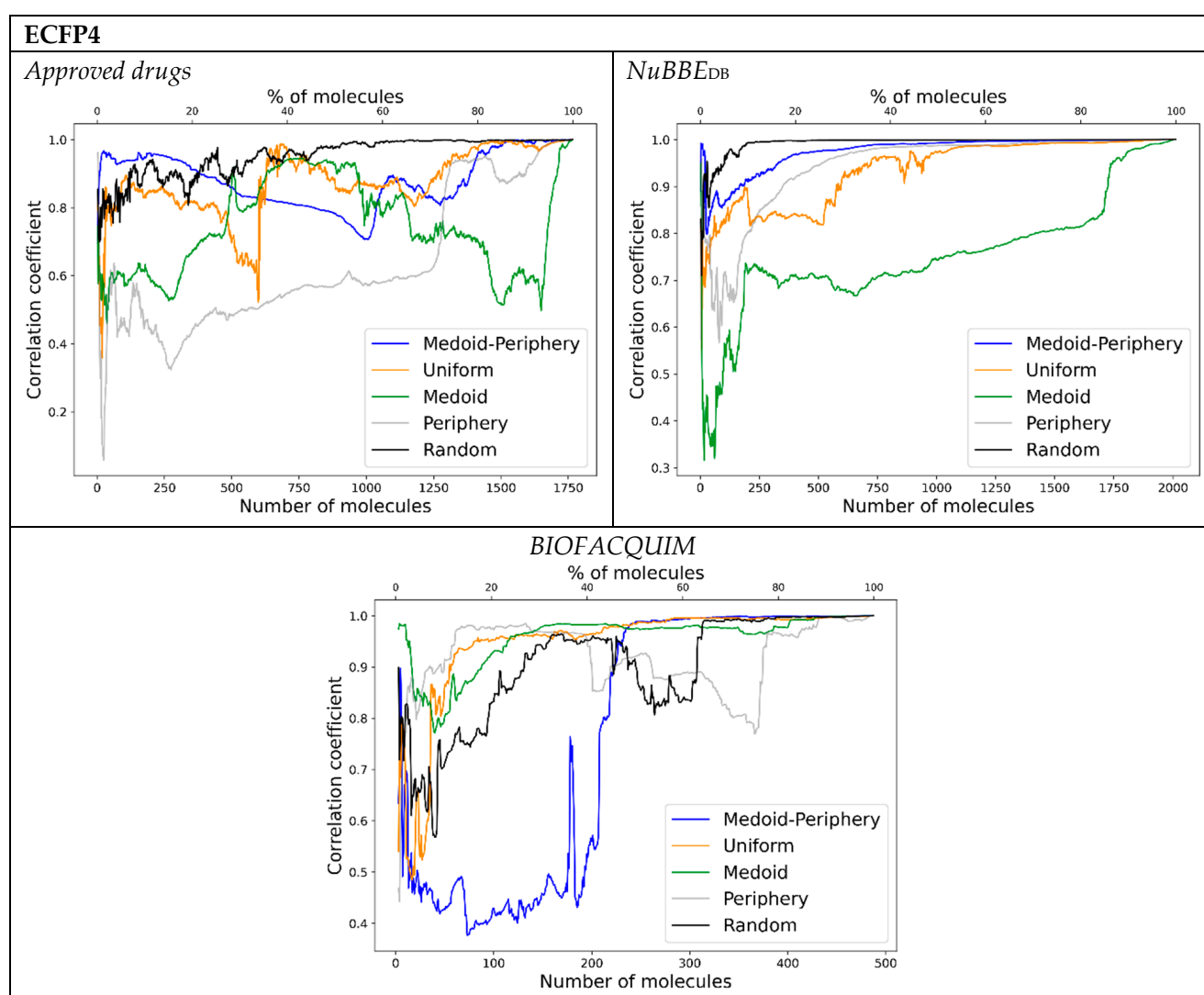
Our first task was to characterize the overall diversity of the analyzed datasets. In Table 1 we show the combined results for all the fingerprints considered. Note that since we had to calculate the pairwise similarity matrix between all the molecules in each set in order to perform the backward approach, we used these values to evaluate the chemical diversity of each library. It is reassuring that this measure provides consistent results, with *approved drugs* being identified as the most diverse library, while BIOFACQUIM is the least diverse. As reported, the current version of BIOFACQUIM is a rather small set of natural products from one country developed over the last few years [37]. NuBBE<sub>DB</sub> has been developed for a decade and contains four times more compounds than BIOFACQUIM [38]. Not surprisingly, approved drugs cover a broad range of diverse chemical structures.

**Table 1.** Average pairwise similarity and PCA distance for the studied libraries.

Dataset	N	Average Pairwise Similarity			Average PCA Distance		
		ECFP4	RDKit	MACCS Keys	ECFP4	RDKit	MACCS Keys
Approved drugs	1768	0.09	0.21	0.32	1.39	4.48	4.67
NuBBE <sub>DB</sub>	2013	0.12	0.24	0.42	2.34	5.93	7.42
BIOFACQUIM	488	0.12	0.25	0.46	1.21	2.88	3.10

### 3.2. ChemMaps with Backward Approach

Figure 2 shows the correlation coefficient between the distances of the satellites in the ChemMaps and the distances in the whole similarity matrix PCA versus the percentage of the library used as satellites, using the backward approach and the five satellite sampling methods.

**Figure 2.** Cont.

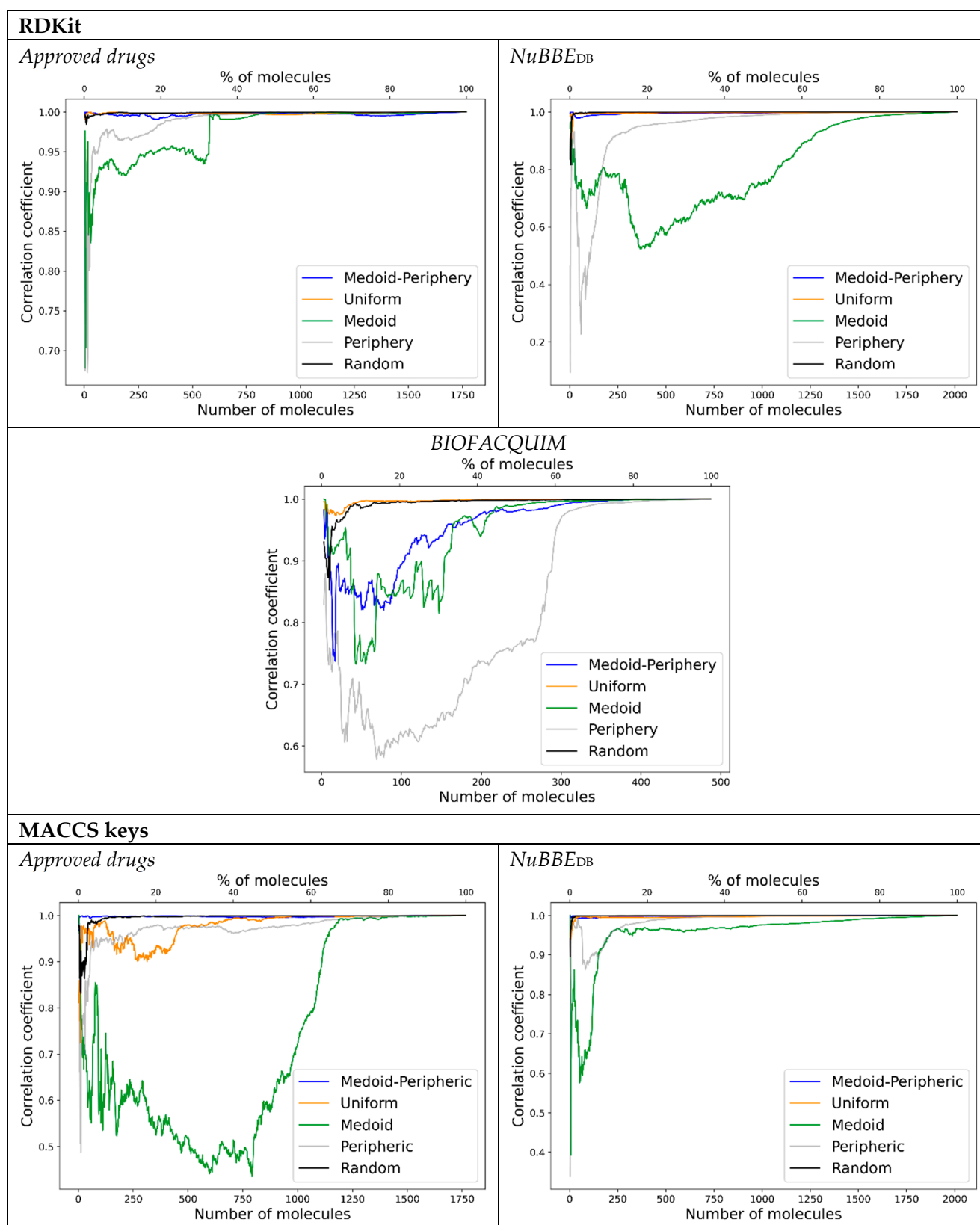
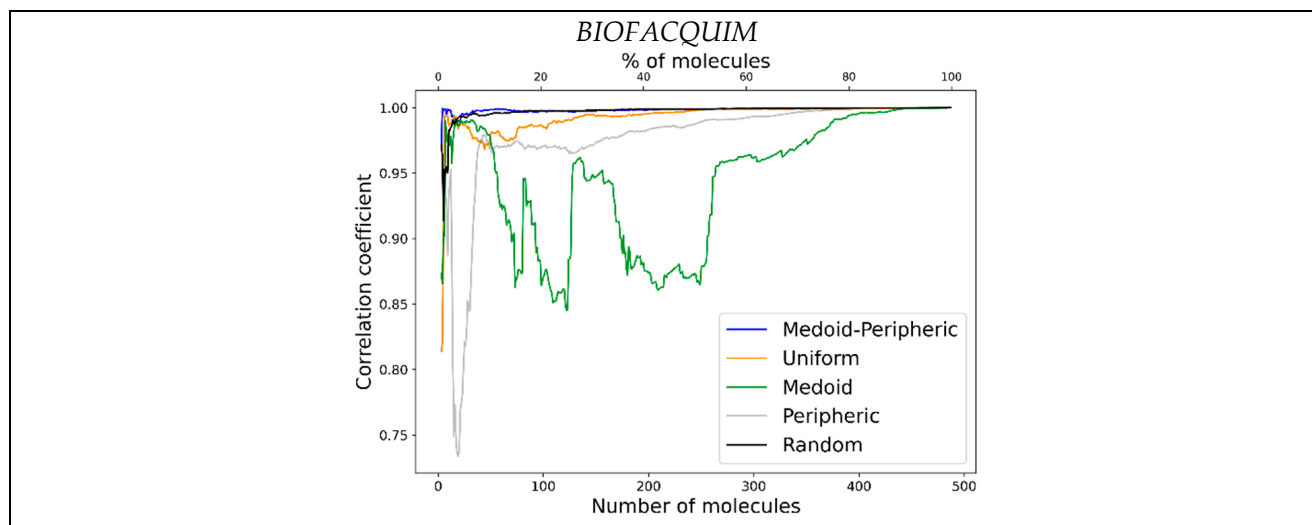


Figure 2. Cont.



**Figure 2.** Backward correlation plots for all the libraries and fingerprints considered.

The backward results (Figure 2) show marked differences in behavior depending on the type of dataset and fingerprint used. In general, for the three types of fingerprints, the medoid–periphery sampling provides the best results for small numbers of molecules over the approved drugs and NuBBEDB libraries. When using RDKit, except for medoid and periphery, all the other sampling methods provide essentially equivalent results over the two mentioned libraries. In the case of BIOFACQUIM, ECFP4 shows a preference for periphery sampling for small numbers of molecules, while the medoid–periphery performs rather badly. This agrees with the original spirit of using satellite molecules that are essentially “outliers” in the data as good reference points unto which project the relations of the larger molecular set. In most cases, it is notable how the medoid and periphery samplings tend to show poor correlations for almost all numbers of molecules selected. This emphasizes that while sampling chemical space we should not focus on a single region (either the “central” or “outlier” parts of a library), and that a balanced exploration (even if performed randomly) is preferred and provides a better description of the underlying correlations between the species.

Figure 3 shows the ChemMaps (blue) and the whole similarity matrix PCA (orange) graphs for the three libraries. The best and most consistent sampling method at a lower percentage of satellites was used in each case. It can be noted that the shapes resemble each other with only using 25% of the library as satellites; however, they are not aligned and are not oriented in the same direction. The reason for this is the use of  $R^2$  as metric, it only depends on the distances between points and not the orientation. This supports the hypothesis that a lower number of compounds can be used to resemble the visualization of the whole library’s chemical space. From the ChemMaps of BIOFACQUIM, it can be noted that the shapes generated by the whole matrix PCA scoring plot are not “filled” with points, which explains why periphery sampling has high correlations in the case of the ECFP4 fingerprint.



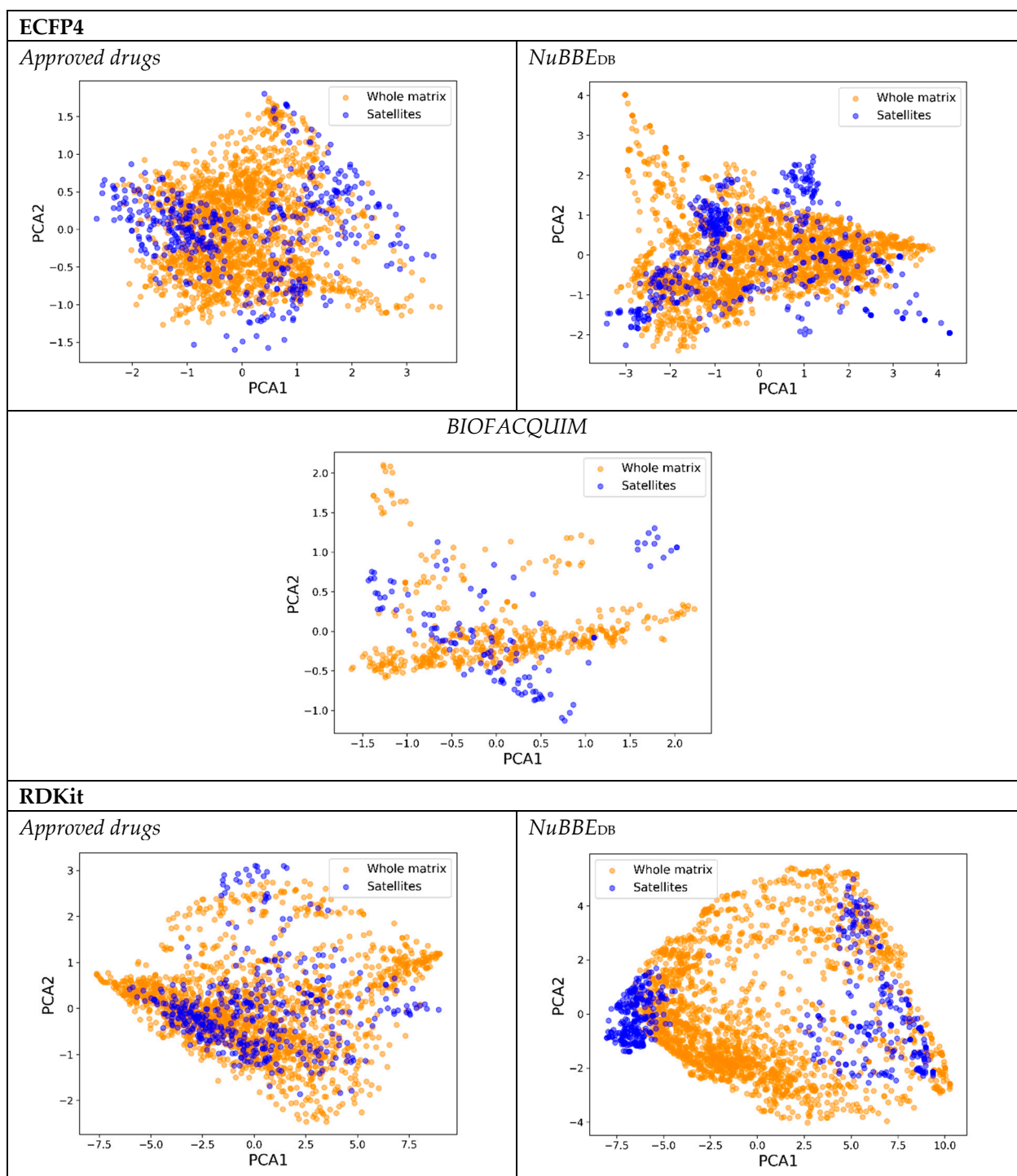
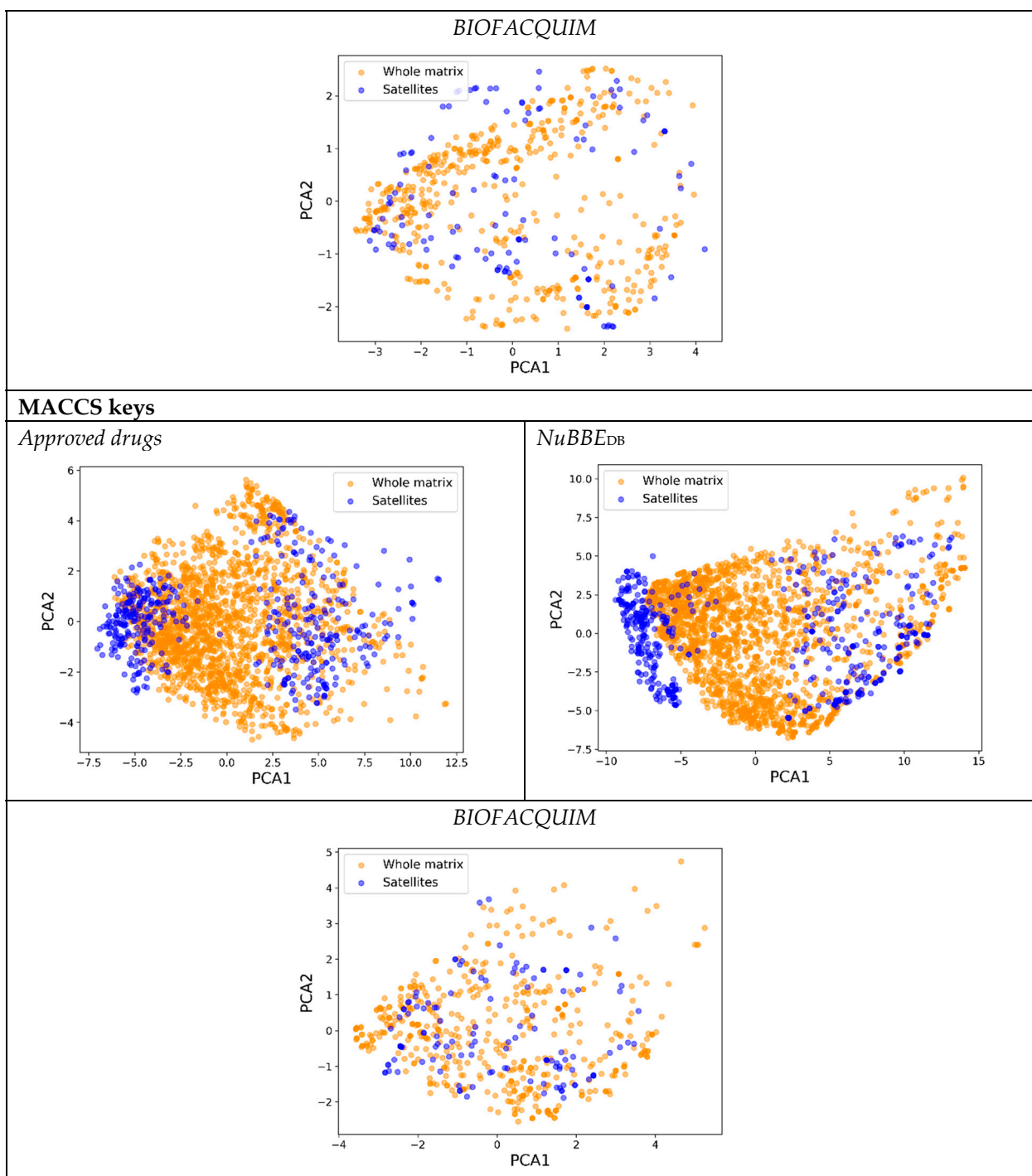


Figure 3. Cont.





**Figure 3.** PCA scoring plots for the whole database's binary similarity matrix with different fingerprints and ChemMaps using 25% of the database as satellites sampled with medoid–periphery for approved drugs, medoid–periphery for NuBBE, and uniform for BIOFACQUIM.

### 3.3. ChemMaps with Forward Approach

The forward results (Figure 4) showed similar trends as the backward approach, with the very attractive outcome that, in most cases, a small number of molecules are enough to obtain a high correlation coefficient, i.e., a small number of compound satellites are enough to obtain a reliable representation of the chemical space. ECFP4 and MACCS keys, once again, tend to favor the medoid–periphery sampling for a small number of molecules, especially for the approved drugs and NuBBEDB libraries. In this case, RDKit

shows virtually the same preference for the uniform sampling and medoid–periphery, and also for the approved drugs and NuBBE<sub>DB</sub>, and mostly for a small number of molecules. It is surprising how medoid sampling is consistently the worst in almost all cases considered. Even the periphery sampling outperforms the medoid-only selections, indicating that a diverse selection (even if only from the outlier region) is preferred over a selection strongly biased towards the central part of the library. This justifies the traditional strategy of selecting very diverse satellites to represent sectors of chemical space [22,23].

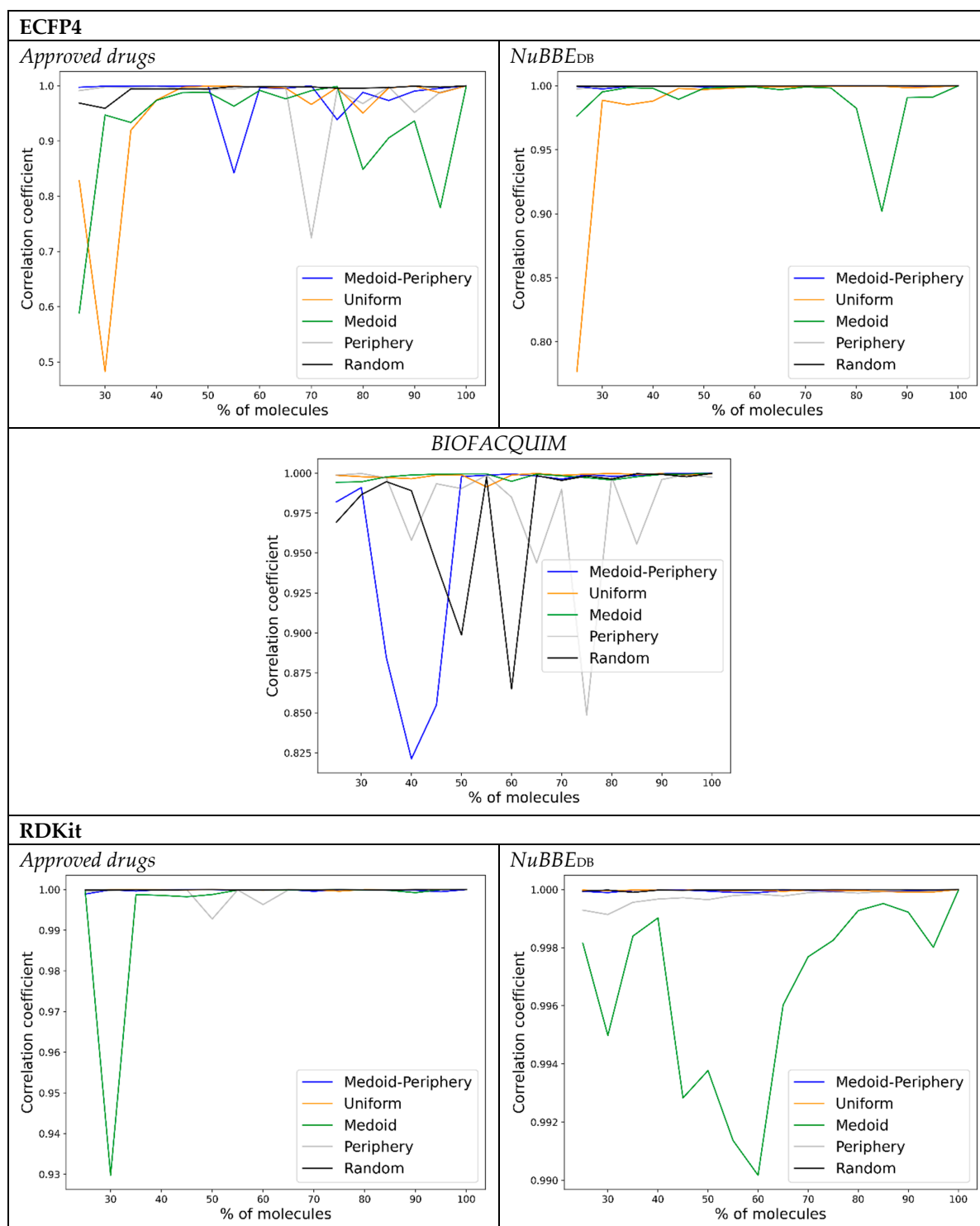
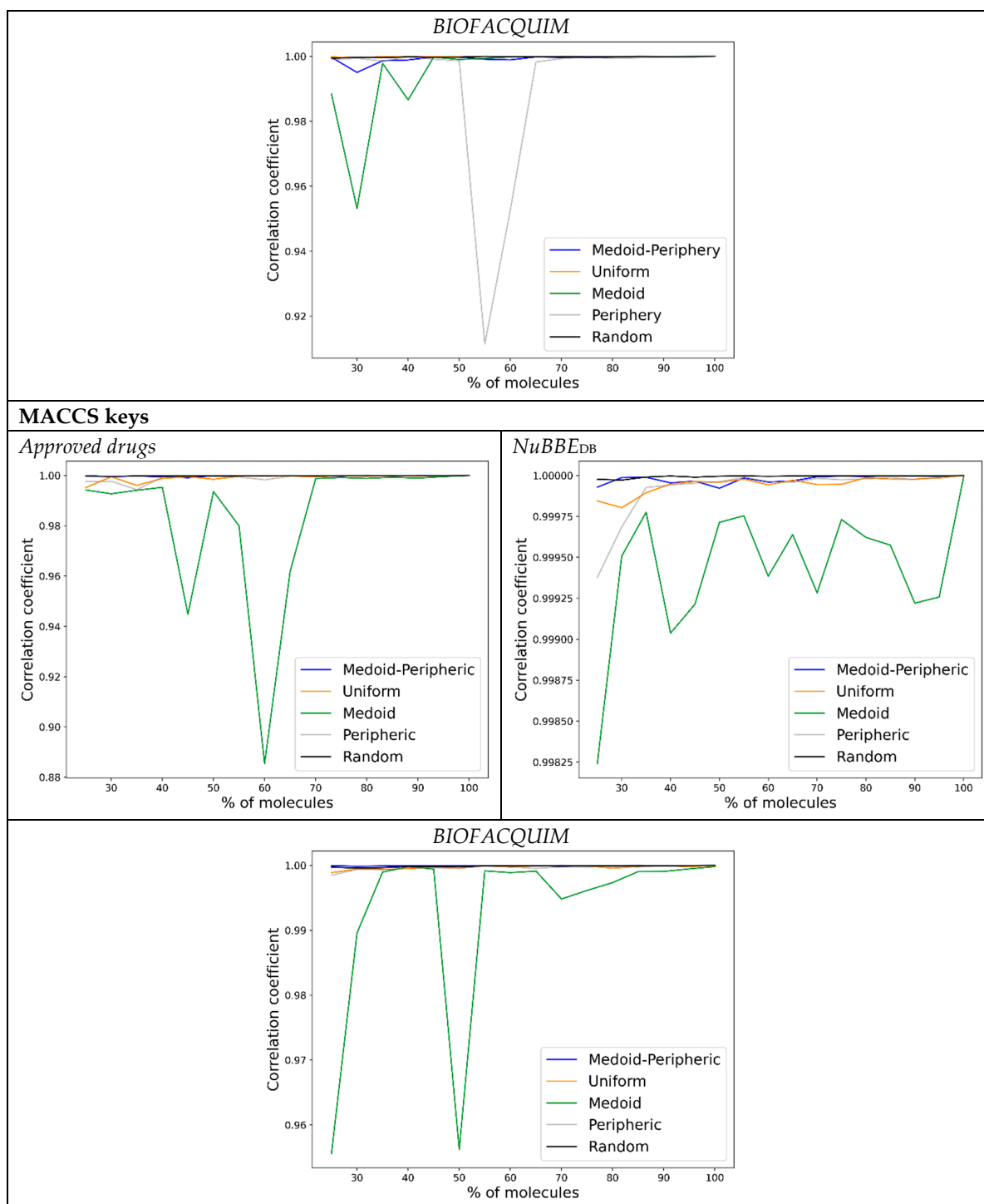


Figure 4. Cont.



**Figure 4.** Forward correlation plots for all the libraries and fingerprints considered.

One of the main practical outcomes of this work is to have a simple, yet robust and systematic, methodology for identifying a small set of compounds within a database to generate a visual representation of the chemical space based on PCA and structural fingerprints. This essentially provides an “embedding” of the data within the satellite space, thus showing that a reduced number of “degrees of freedom” is enough to capture a large fraction of the correlation in the original set. This provides the enticing possibility of describing bigger datasets while reducing the computational cost. Based on these findings, it is proposed to select satellite molecules using the medoid–periphery approach.

Fingerprints of different designs (e.g., ECFP4, RDKit, MACCS keys) can be employed to generate visual representations of the chemical space for compound datasets. As recently discussed, it is not necessary to identify the “best” fingerprint for visual analysis of the chemical space but to analyze a group of alternative chemical spaces of compound datasets or “chemical multiverse”.

#### 4. Methods

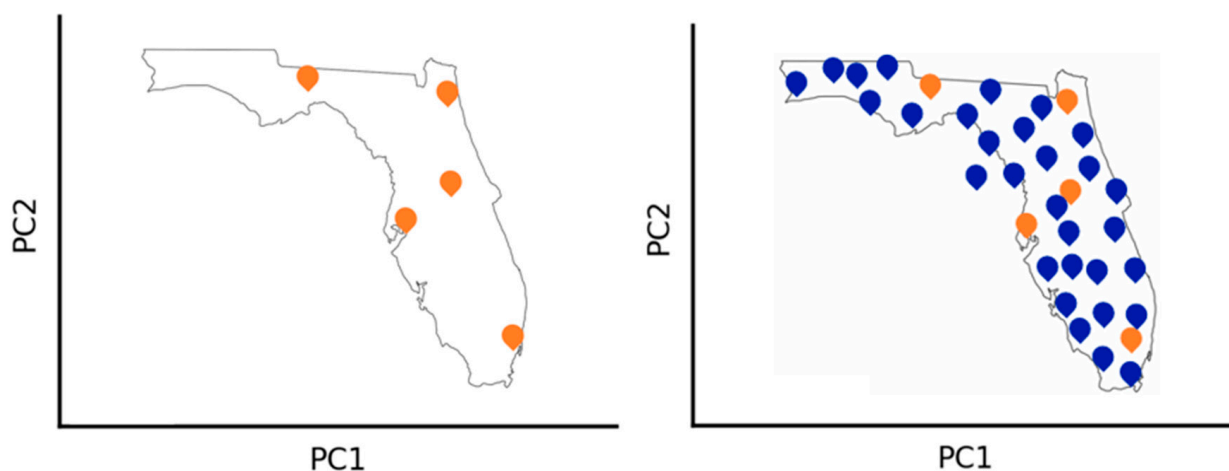
##### 4.1. Molecular Libraries and Computational Conditions

The vastness of chemical space contains a large structural diversity of chemical compounds that explore different regions of it. In this work as a case study, we explored the “druggable” chemical space using the approved drugs available in the DrugBank database V. 5.1.10 (1768 compounds) [39]; for the rest of the work, we will name *approved\_drugs* this library. Also, we studied the chemical space of two public natural products databases: NUBBE<sub>DB</sub> (2013 compounds) [40] and BIOFACQUIM (488 compounds) [37] (freely available databases from Brazil and Mexico, respectively). The SMILES code [41] for each compound was computed, and for all databases, their duplicated SMILES codes were removed. All the information discussed here pertains to these datasets by 10 July 2023.

Each SMILES code has been used to represent the chemical structure of each compound for each dataset using different fingerprints: MACCs keys (166 bits), ECFP4 (1024 bits), and RDKit (2048 bits). The fingerprints were computed using the RDKit module implemented by the python programming language [42]. Finally, from each fingerprint of each molecule, the extended similarity values were calculated using the Jaccard-Tanimoto similarity index [43] with the code freely available from <https://github.com/ramirandaq/MultipleComparisons> (18 March 2022). The curated libraries used in this work can be found at <https://doi.org/10.6084/m9.figshare.23654316.v1> (accessed on 18 March 2022) for *approved\_drugs* and NuBBE and <https://doi.org/10.6084/m9.figshare.11312702.v1> (accessed on 18 March 2022) for BIOFACQUIM.

##### 4.2. ChemMaps

The main goal of ChemMaps is to resemble the chemical space of a database using only a portion of it as satellites. ChemMaps uses PCA on a pairwise similarity matrix of only the satellites and calculates the distances based on the PC scores. The correlation between the ChemMaps distances and the ones derived from the whole matrix is used as the metric to prove if the proposed map resembles the whole chem space picture [24]. A cartoon on the ChemMaps principle is shown in Figure 5.



**Figure 5.** Schematic representation of ChemMaps idea, left side. Orange points (left) represent the “satellites” or references that can be used to represent the chemical space of the whole dataset (right).

Two main approaches were conducted. The steps used ChemMaps [24] as a guide for incorporating satellite sampling techniques. The backward approach, as proof of principle, compares the ChemMaps satellite distances with the ones derived from the whole matrix. As reported in the former paper, the use of two principal components (PCs) gives a good correlation of the ChemMap distances with the whole library mapping [24].

#### Backward approach

1. Generate the  $N \times N$  similarity matrix using the Jaccard-Tanimoto index.
2. Perform PCA on the given matrix with two PCs.
3. Compute all pairwise Euclidean distances based on PC scorings. These distances will be used as reference values.
4. Choose the first three satellites ( $S$ ) according to the sampling method chosen. (i.e., for medoid sampling the three compounds with the lowest complementary similarity).
5. Perform PCA with the  $S \times N$  similarity matrix and obtain the pairwise Euclidean distances based on those PC scores.
6. Calculate the correlation between distances with the whole matrix (step 3) and the satellite's matrix (step 5).
7. Iterate over steps 4 to 6 adding one satellite at the time, based on the chosen sampling method.
8. Establish the proportion of satellites required to preserve a high correlation (of at least 0.90).

The forward approach uses an initial portion of the set and adds a smaller portion, and compares the distances generated in consecutive steps to avoid computing of the PCA on the complete matrix.

#### Forward approach

1. Start taking 25% of the database as satellites ( $S$ ) by the sampling method of choice, having then a  $S \times N$  similarity matrix. This percentage is used as demonstrated in previous work to be the lowest percentage needed to render high correlation coefficients [24].
2. Perform PCA with 2 PCs and use the scorings to calculate the Euclidean distances.
3. Add the next 5% to the satellites according to the sampling method and perform step 2 with the updated satellite matrix.
4. Calculate the correlation between the updated satellite Euclidean distances (of the elements in common) and the distances from the former satellite matrix (i.e., 30–25%).
5. Repeat steps 3 and 4 until a high correlation (greater than 0.90) or to 100%.

All the combinations of backward/forward approaches, datasets, fingerprints and sampling methods were computed with the goal of evaluating what regions of chemical space are important to sample as satellites so we can obtain a meaningful ChemMap. Overall, we showed that the extended similarity-based sampling methods offer a variety of options for sampling different regions of chemical space.

**Author Contributions:** Conceptualization, J.L.M.-F. and R.A.M.-Q.; Methodology, K.L.-P.; Software, K.L.-P. and R.A.M.-Q.; Validation, K.L.-P. and E.L.-L.; Formal analysis, K.L.-P., E.L.-L., J.L.M.-F. and R.A.M.-Q.; Investigation, K.L.-P., E.L.-L. and J.L.M.-F.; Data curation, K.L.-P. and J.L.M.-F.; Writing—original draft, K.L.-P., J.L.M.-F. and R.A.M.-Q.; Writing—review & editing, K.L.-P., E.L.-L., J.L.M.-F. and R.A.M.-Q.; Supervision, J.L.M.-F.; Project administration, J.L.M.-F. and R.A.M.-Q.; Funding acquisition, J.L.M.-F. and R.A.M.-Q. All authors have read and agreed to the published version of the manuscript.

**Funding:** UFII Seed Award (R.A.M.-Q., K.L.-P.); DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grants No. IN201321 (J.L.M.-F.); Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCyT), Mexico, scholarship No. CVU: 894234 (E.L.-L.)

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The code used to calculate the extended similarity indices is freely available at <https://github.com/ramirandaq/MultipleComparisons> (accessed on 24 July 2023).

**Acknowledgments:** R.A.M.-Q. and K.L.-P. thank the University of Florida for a UFII Seed Award. J.L.M.-F. thanks DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grants No. IN201321. E.L.-L. thanks the Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCyT), Mexico, for the scholarship No. CVU: 894234.

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

**Sample Availability:** Not applicable.

## References

1. Rarey, M.; Nicklaus, M.C.; Warr, W. Special Issue on Reaction Informatics and Chemical Space. *J. Chem. Inf. Model.* **2022**, *62*, 2009–2010. [[CrossRef](#)]
2. Warr, W.A.; Nicklaus, M.C.; Nicolaou, C.A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62*, 2021–2034. [[CrossRef](#)] [[PubMed](#)]
3. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98. [[CrossRef](#)] [[PubMed](#)]
4. Medina-Franco, J.L.; Chávez-Hernández, A.L.; López-López, E.; Saldívar-González, F.I. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inform.* **2022**, *41*, 2200116. [[CrossRef](#)]
5. Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432*, 855–861. [[CrossRef](#)]
6. Virshup, A.M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D.N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303. [[CrossRef](#)]
7. Osolodkin, D.I.; Radchenko, E.V.; Orlov, A.A.; Voronkov, A.E.; Palyulin, V.A.; Zefirov, N.S. Progress in Visual Representations of Chemical Space. *Expert. Opin. Drug Discov.* **2015**, *10*, 959–973. [[CrossRef](#)]
8. Medina-Franco, J.L.; Naveja, J.J.; López-López, E. Reaching for the Bright StARs in Chemical Space. *Drug Discov. Today* **2019**, *24*, 2162–2169. [[CrossRef](#)] [[PubMed](#)]
9. Bro, R.; Smilde, A.K. Principal Component Analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [[CrossRef](#)]
10. Gaytán-Hernández, D.; Chávez-Hernández, A.L.; López-López, E.; Miranda-Salas, J.; Saldívar-González, F.I.; Medina-Franco, J.L. Art Driven by Visual Representations of Chemical Space. *ChemRxiv Chem. Educ.* **2023**. [[CrossRef](#)]
11. Viarengo-Baker, L.A.; Brown, L.E.; Rzepiela, A.A.; Whitty, A. Defining and Navigating Macrocyclic Chemical Space. *Chem. Sci.* **2021**, *12*, 4309–4328. [[CrossRef](#)]
12. Cihan Sorkun, M.; Mullaj, D.; Koelman, J.M.V.A.; Er, S. ChemPlot, a Python Library for Chemical Space Visualization\*\*. *Chem. Methods* **2022**, *2*, e202200005. [[CrossRef](#)]
13. Bonachera, F.; Marcou, G.; Kireeva, N.; Varnek, A.; Horvath, D. Using Self-Organizing Maps to Accelerate Similarity Search. *Bioorg. Med. Chem.* **2012**, *20*, 5396–5409. [[CrossRef](#)] [[PubMed](#)]
14. Takács, G.; Sándor, M.; Szalai, Z.; Kiss, R.; Balogh, G.T. Analysis of the Uncharted, Druglike Property Space by Self-Organizing Maps. *Mol. Divers.* **2022**, *26*, 2427–2441. [[CrossRef](#)]
15. Achenbach, J.; Klingler, F.-M.; Blöcher, R.; Moser, D.; Häfner, A.-K.; Rödl, C.B.; Kretschmer, S.; Krüger, B.; Löhr, F.; Stark, H.; et al. Exploring the Chemical Space of Multitarget Ligands Using Aligned Self-Organizing Maps. *ACS Med. Chem. Lett.* **2013**, *4*, 1169–1172. [[CrossRef](#)] [[PubMed](#)]
16. Andronov, M.; Fedorov, M.V.; Sosnin, S. Exploring Chemical Reaction Space with Reaction Difference Fingerprints and Parametric T-SNE. *ACS Omega* **2021**, *6*, 30743–30751. [[CrossRef](#)]
17. Gaspar, H.A.; Baskin, I.I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2015**, *55*, 84–94. [[CrossRef](#)] [[PubMed](#)]
18. Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping in Drug Design. *Drug Discov. Today Technol.* **2019**, *32–33*, 99–107. [[CrossRef](#)]
19. Kireeva, N.; Baskin, I.I.; Gaspar, H.A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012**, *31*, 301–312. [[CrossRef](#)]
20. Medina-Franco, J.L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B.I. Progress on Open Chemoinformatic Tools for Expanding and Exploring the Chemical Space. *J. Comput. Aided Mol. Des.* **2022**, *36*, 341–354. [[CrossRef](#)]
21. Dunn, T.B.; Seabra, G.M.; Kim, T.D.; Juárez-Mercado, K.E.; Li, C.; Medina-Franco, J.L.; Miranda-Quintana, R.A. Diversity and Chemical Library Networks of Large Data Sets. *J. Chem. Inf. Model.* **2022**, *62*, 2186–2201. [[CrossRef](#)]
22. Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space. *J. Nat. Prod.* **2007**, *70*, 789–794. [[CrossRef](#)] [[PubMed](#)]
23. Oprea, T.I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157–166. [[CrossRef](#)] [[PubMed](#)]
24. Naveja, J.J.; Medina-Franco, J.L. ChemMaps: Towards an Approach for Visualizing the Chemical Space Based on Adaptive Satellite Compounds. *F1000Research* **2017**, *6*, 1134. [[CrossRef](#)]
25. Awale, M.; Reymond, J.-L. Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2015**, *55*, 1509–1516. [[CrossRef](#)]



26. Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. The Chemical Library Space and Its Application to DNA-Encoded Libraries. *ChemRxiv Theor. Comput. Chem.* **2023**. [CrossRef]
27. Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case. *J. Chem. Inf. Model.* **2023**, *63*, 4042–4055. [CrossRef]
28. Borrel, A.; Conway, M.; Nolte, S.Z.; Unnikrishnan, A.; Schmitt, C.P.; Kleinstreuer, N.C. ChemMaps.Com v2.0: Exploring the Environmental Chemical Universe. *Nucleic Acids Res.* **2023**, *51*, W78–W82. [CrossRef]
29. Borrel, A.; Kleinstreuer, N.C.; Fourches, D. Exploring Drug Space with ChemMaps.Com. *Bioinformatics* **2018**, *34*, 3773–3775. [CrossRef] [PubMed]
30. Miranda-Quintana, R.A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 2: Speed, Consistency, Diversity Selection. *J. Cheminform.* **2021**, *13*, 33. [CrossRef]
31. Miranda-Quintana, R.A.; Bajusz, D.; Rácz, A.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 1: Theory and Characteristics<sup>†</sup>. *J. Cheminform.* **2021**, *13*, 32. [CrossRef] [PubMed]
32. Flores-Padilla, E.A.; Juárez-Mercado, K.E.; Naveja, J.J.; Kim, T.D.; Alain Miranda-Quintana, R.; Medina-Franco, J.L. Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. *Mol. Inform.* **2022**, *41*, 2100285. [CrossRef]
33. Bajusz, D.; Miranda-Quintana, R.A.; Rácz, A.; Héberger, K. Extended Many-Item Similarity Indices for Sets of Nucleotide and Protein Sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3628–3639. [CrossRef]
34. Rácz, A.; Mihalovits, L.M.; Bajusz, D.; Héberger, K.; Miranda-Quintana, R.A. Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices. *J. Chem. Inf. Model.* **2022**, *62*, 3415–3425. [CrossRef]
35. Chang, L.; Perez, A.; Miranda-Quintana, R.A. Improving the Analysis of Biological Ensembles through Extended Similarity Measures. *Phys. Chem. Chem. Phys.* **2022**, *24*, 444–451. [CrossRef]
36. Rácz, A.; Dunn, T.B.; Bajusz, D.; Kim, T.D.; Miranda-Quintana, R.A.; Héberger, K. Extended Continuous Similarity Indices: Theory and Application for QSAR Descriptor Selection. *J. Comput. Aided Mol. Des.* **2022**, *36*, 157–173. [CrossRef]
37. Pilón-Jiménez, B.; Saldívar-González, F.; Díaz-Eufracio, B.; Medina-Franco, J. BIOFACQUIM: A Mexican Compound Database of Natural Products. *Biomolecules* **2019**, *9*, 31. [CrossRef] [PubMed]
38. Valli, M.; dos Santos, R.N.; Figueira, L.D.; Nakajima, C.H.; Castro-Gamboa, I.; Andricopulo, A.D.; Bolzani, V.S. Development of a Natural Products Database from the Biodiversity of Brazil. *J. Nat. Prod.* **2013**, *76*, 439–444. [CrossRef] [PubMed]
39. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef]
40. Pilon, A.C.; Valli, M.; Dametto, A.C.; Pinto, M.E.F.; Freire, R.T.; Castro-Gamboa, I.; Andricopulo, A.D.; Bolzani, V.S. NuBBEDB: An Updated Database to Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep.* **2017**, *7*, 7215. [CrossRef]
41. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101. [CrossRef]
42. Landrum, G.; Penzotti, J. RDKit. 2018. Available online: <http://www.rdkit.org/> (accessed on 17 January 2022).
43. Dunn, T.B.; López-López, E.; Kim, T.D.; Medina-Franco, J.L.; Miranda-Quintana, R.A. Exploring Activity Landscapes with Extended Similarity: Is Tanimoto Enough? *Mol. Inform.* **2023**, *42*, 2300056. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.