Exploring activity landscapes with extended similarity: is Tanimoto enough?

Timothy B. Dunn,[1‡] Edgar López- López,[2‡] Taewon David Kim,[1] José L. Medina-Franco,[2*] Ramón Alain Miranda-Quintana[1*]

1. Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States

2. DIFACQUIM Research Group, Department of Pharmacy, National Autonomous University of Mexico, Mexico City 04510, Mexico

‡ These authors contributed equally

Email: quintana@chem.ufl.edu , medinajl@unam.mx

ABSTRACT

Understanding structure-activity landscapes is essential in drug discovery. Similarly, it has been shown that the presence of activity cliffs in compound data sets can have a substantial impact not only on the design progress but also can influence the predictive ability of machine learning models. With the continued expansion of the chemical space and the currently available large and ultra-large libraries, it is imperative to implement efficient tools to analyze the activity landscape of compound data sets rapidly. The goal of this study is to show the applicability of the *n*-ary indices to quantify the structure-activity landscapes of large compound data sets using different types of structural representation rapidly and efficiently. We also discuss how a recently introduced medoid algorithm provides the foundation to finding optimum correlations between similarity measures and structure-activity rankings. The applicability of the *n*-ary indices and the medoid algorithm is shown by analyzing the activity landscape of 10 compound data sets with pharmaceutical relevance using three fingerprints of different designs, 16 extended similarity indices, and 11 coincidence thresholds.

Key words: structure-activity relationships; chemical space; eSALI, similarity; extended similarity; molecular fingerprints

1. INTRODUCTION

Exploring structure-activity and structure-property relationships of compound data sets is one of the first basic steps in drug discovery. To this end, one of the approaches is quantifying the activity (property) landscapes and identifying, if any, activity (or more general property) cliffs and similarity cliffs.[1, 2] Activity cliffs point to key and small structural features associated with large changes in activity (information that can be used in lead optimization). Also, the presence of activity cliffs in data sets hamper the performance of predictive models.[3] In contrast to activity cliffs, similarity cliffs point to different chemical features associated with very similar or identical biological activity and provide information for scaffold or R-group hopping.

Over the past several years, different approaches have emerged and evolved to analyze visually and quantitatively activity (property) landscapes and identify activity cliffs.[4-9] For instance, Aldeghi et al. recently introduced a roughness index (ROGI) to characterize quantitatively the property topology of a given data set.[10] The metric ROGI can take values between zero and one and is intended to capture the total roughness (or flatness) of a normalized data set. Also recently, van Tilborg et al. evaluated the performance of twenty-four machine and deep learning approaches on curated bioactivity data from 30 macromolecular targets in the presence of activity cliffs. The authors concluded that the presence of activity cliffs in compound data sets affect the performance of the predictive models.[11]

One of the earliest and straightforward approaches to capture the activity landscapes of compound data sets and rapidly detect activity cliffs is the Structure-Activity Landscape Index (SALI) proposed by Guha and Van Drie:[9]

$$\mathrm{SALI}(i,j) = \frac{\left|P_i - P_j\right|}{1 - s(i,j)} \tag{1}$$

where P_i and P_j are the property values of molecules i and j, respectively, and $s(i,j)$ is the similarity of i and j. In most published applications of SALI, $s(i,j)$ has been computed with the Tanimoto coefficient using molecular fingerprints as representation, but it can be quantified by any other combination of molecular representation and similarity index. However, for large and ultra large data sets calculating SALI can be inefficient. The key issue is that all these roughness measures are based on binary (pairwise) comparisons,[12-16] so they will require $O(N^2)$ operations to study a library with N molecules, thus quickly becoming unmanageable. That is, while the standard molecule-to-molecule relations are very efficient to calculate (and, in many cases, pivotal to help

us navigate through chemical space), using the SALI (or SALI-inspired) measure to study full libraries is too computationally expensive.

Recently, we proposed a new class of similarity indices: extended similarity indices,[17-19] that allow comparing any number of molecules at the same time, with a much more attractive O(*N*) scaling. These new measures have been successfully applied in clustering tasks,[20] studying phylogenetic trees,[19] epigenetic-focused libraries,[21] in feature selection,[22] sampling of molecular dynamics simulations,[23] and many other studies.[24-27] More notably, it has been shown that these indices can provide a very efficient way to quantify the chemical diversity of large libraries.[28] However, up to this moment these indices have not been connected to molecular properties. The aggregative nature of the extended indices means that they are very attractive to study the global properties of large libraries (from quantifying their diversity, to their representation in chemical space). However, and perhaps a bit surprisingly, these new indices also allow us to zoom-in on these same large collections of molecules and analyze them at the local level. In this study, we will combine these two aspects. First, we will show how the *n*-ary indices can be used to provide a very efficient measure of heterogeneity, applicable at the library level. Then, we will study how our medoid algorithm provides the foundation to finding optimum correlations between similarity measures and structure-activity rankings.

2. EXTENDED SIMILARITY FRAMEWORK

The first step to calculate the *n*-ary indices is to accommodate all the molecular fingerprints in a matrix-like arrangement. (Notice that in this work we will be only considering binary fingerprints, but the extended indices could be calculated for other molecular representations as well, including arbitrary strings of characters, atomic coordinates, or latent space representations from arbitrary descriptors.) We then just have to compute the vector $\Sigma = [\sigma_1, \sigma_2, ..., \sigma_N]$, where σ_k indicates the sum of all the elements in the *k*th column. Next, we calculate the indicators $\Delta_\sigma = |2\sigma - N|$, which we use to determine if each column contributes to the similarity or dissimilarity of the set. This is done with the help of a coincidence threshold, γ, which is essentially a hyperparameter indicating how many "characters of the same type" do we need in a column to consider it as similar. In more mathematical terms, whenever $\Delta_\sigma > \gamma$, we will assign the column as similar (in particular, "1-similar" if $2\sigma - N > \gamma$ and a "0-similar" if $N - 2\sigma > \gamma$),

which means that $\Delta_\sigma \le \gamma$ will correspond to a dissimilar column. If we would like to penalize the partial (e.g., not perfect) coincidence in a column we just need to include weight functions f_s and f_d to deal with the similar and dissimilar cases, respectively. There are many ways to choose these weights, but in all our studies so far we have only considered perhaps their simplest realization:

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

With this simple recipe we can then simply "translate" large classes of binary similarity indices, so they could be used now to quantify the similarity of *N* objects at the same time. For instance, the (extended) Jaccard-Tanimoto (or simply, Tanimoto) and Russell-Rao indices are given by:

$$s_{eJT} = \frac{\sum_{1-s} f_s(\Delta_\sigma)}{\sum_{1-s} C_s + \sum_d C_d} \tag{3}$$

$$s_{eRR} = \frac{\sum_{1-s} f_s(\Delta_\sigma)}{\sum_{s} C_s + \sum_d C_d} \tag{4}$$

s, 1-*s*, 0-*s*, and *d* represent summations over the similar, 1-similar, 0-similar, and dissimilar columns, respectively.

The most time-consuming step involved in the calculation of the *n*-ary indices is the formation of vector Σ. However, this only involves a sum over independent columns, so it scales linearly with the number of rows. In other words, this algorithm scales as O(*N*), in stark contrast with the inherent O(N^2) scaling of methods based on pairwise similarity indices. For instance, if we have a set of molecules $M = \{m_1, m_2, ..., m_N\}$ and we want to estimate the roughness of this chemical (sub-)space with the standard SALI we would have to calculate some version of the following expression:

$$\beta \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{\left|P_i - P_j\right|}{1 - s(m_i, m_j)} \tag{5}$$

which obviously demands $\sim \frac{N(N-1)}{2}$ similarity (and property difference) calculations (with β an arbitrary normalization).

We propose to alleviate this problem by instead using an extended SALI (eSALI) index calculated by:

$$e\mathrm{SALI}(M) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|P_i - \langle P\rangle\right|}{1 - s_e(M)} \tag{6}$$

Here, s_e indicates the extended similarity (notice that the input to this index is the whole library, *M*), $\langle P\rangle = \frac{1}{N}\sum_{i=1}^{N} P_i$ is just the average of the property values over all the set, and the $\frac{1}{N}$ is just a convenient normalization factor. Since $s_e(M), \langle P\rangle, \sum_{i=1}^{N}\left|P_i - \langle P\rangle\right|$ scale at worst as O(*N*), this means that Eq. (6) is dramatically more efficient than Eq. (5). Notice that when we only consider two molecules, $M = \{m_1, m_2\}$, Eqs. (5) and (6) are trivially related, since:

$$\left|P_1 - P_2\right| = \left|P_1 - \langle P\rangle\right| + \left|P_2 - \langle P\rangle\right| \tag{7}$$

and, by construction:

$$\forall m_1, m_2 : s(m_1, m_2) = s_e(\{m_1, m_2\}) \tag{8}$$

Hence:

$$e\mathrm{SALI}(\{m_1, m_2\}) = \frac{1}{2\beta} SALI(m_1, m_2) \tag{9}$$

So, an inconsequential factor aside, the eSALI can be interpreted as a natural extension of the SALI to any number of molecules.

Another attractive feature of the *n*-ary indices is their ability to find the medoid of a set or, more generally, rank every element in a set depending on whether they are more "important" or "central" (e.g., medoid or medoid-like) to those molecules that are "less important" or "outliers". The main ingredient here is the concept of the *complementary similarity* of a molecule in a set. Formally speaking, the complementary similarity of the *i*th molecule, $\overline{s_i}$, will be given by:

$$\overline{s_i} = s_e(M / \{m_i\}) \tag{10}$$

In other words, it is the extended similarity of the set after removing the *i*th molecule.

The algorithm to perform this calculation is very simple. If each molecule is represented by a fingerprint *F*, then we just need to calculate $\overline{s_i} = s_e(\Sigma - F_i)$, with Σ being the column-sum vector introduced above. Then, molecules with the lowest complementary similarity values will be closer

to the medoid regions of the library. This algorithm is essentially composed of two O(*N*) steps (e.g., calculating Σ, and performing the N $\Sigma - F_i$ subtractions) so, just like the standard s_e calculations, it will scale linearly.

We will be using this medoid algorithm to propose a workflow aimed to dissect structure-activity relations. The central idea is to correlate different rankings induced in the dataset both by means of the property values, and by the structural features of the molecules. The property ranking will be given by how much the molecules differ from the average value of a given property. That is, we will order the molecules from central to outlier as an increasing function of $\left|P_i - \bar{P}\right|$. On the other hand, the structural ranking will be given by the previously explained complementary similarity calculations. It is to be expected that a similarity index will be all the more suitable for a particular problem if the structural ranking it induces has a good correlation with the associated property ranking. We will analyze this correlation using two different tools. In both cases, we begin by generating the two rankings, and then we proceed to select a subset from each of them containing a fraction of the data (e.g., 10%, 20%, etc. of the total number of molecules). In the first approach, we investigate how many molecules appear at the same time in the property subset and the structural subset. For this, we use the set theory version of the Tanimoto index (which, to avoid confusions with the extended Tanimoto index, we will either refer to as "set Tanimoto" or Jaccard index). That is, given two sets A and B, the set Tanimoto (or Jaccard) index will be given by:

$$set\_Tanimoto = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \tag{11}$$

The second analysis aims to see not the identical coincidence of some molecules between the rankings, but to quantify, overall, how similar are the molecules in each subset. That is, we average the pairwise similarities between molecules in the property subset and the structural subset (e.g., $\langle s(i,j) \rangle$, with *i* in the property and *j* in the structural subset, respectively). Since the extended indices are very often very well correlated with their binary counterparts (they are *externally consistent*), we must be careful of not biasing our analysis with an incorrect selection of the pairwise index. Here we bypassed this issue by working with the pairwise cosine similarity, which was not included in the extended indices considered.

3. MOLECULAR LIBRARIES AND COMPUTATIONAL CONDITIONS

From ChEMBL V.31[29-32] compounds tested against ten representative anti-diabetic and anti-cancer targets (Table 1) were retrieved.[33-37] Only compounds associated with quantitative inhibitory values (expressed in pIC_{50}) were retrieved on the final data set (used in this manuscript). Duplicate compounds for the same target were removed and the most potent compounds per case (lower pIC_{50} values) were conserved. Finally, 28,289 compounds were filtered and considered in the final data set.

Each molecule has been associated with their respective SMILES[38] that has been used to represent their chemical structures using different fingerprints: MACCs keys (166 bits), ECFP4 (1024 bits), and RDKit (2048 bits). The fingerprints were computed using the RDKit module implemented by python programming language.[39]

Finally, from each fingerprint of each molecule, the extended similarity values were calculated using different similarity indices (as been previously described the section 2). The extended similarity indices (AC: Austin-Colwell, BUB: Baroni-Urbani-Buser, CT1: Consoni-Todeschini 1, CT2: Consoni-Todeschini 2, CT3: Consoni-Todeschini 3, CT4: Consoni-Todeschini 4, Fai: Faith, Gle: Gleason, JT: Jaccard-Tanimoto, Ja0: Jaccard, RR: Russel-Rao, RT: Rogers-Tanimoto, SM: Sokal-Michener, SS1: Sokal-Sneath 1, SS2: Sokal-Sneath 2) were computed using the code freely available at https://github.com/ramirandaq/MultipleComparisons.

Table 1. Overview of the ten compound datasets used in this work.

Identifier	Database	Property	# of Molecules after curation
1	Aldose reductase inhibitors	pIC_{50}	914
2	β-secretase 1 inhibitors		7364
3	Epidermal growth factor receptor erbB1 (EGFR1) inhibitors		9853
4	Free fatty acid receptor 1 (FFA1) inhibitors		66
5	Histone-lysine N-methyltransferase, H3 lysine-9 (HDAC9) inhibitors		250
6	Peroxisome proliferator-activated receptor α (PPARα) inhibitors		1096
7	Peroxisome proliferator-activated receptor γ (PPAR γ) inhibitors		1686

8	Protein-tyrosine phosphatase 1B (PTB1B) inhibitors	3177
9	Serine/threonine-protein kinase AKT inhibitors	3033
10	Tubulin polymerization inhibitors	850

In summary, we studied 3 fingerprint types (MACCS keys (166-bits), RDKit, and ECFP4), 16 extended similarity indices (AC, BUB, CT1, CT2, CT3, CT4, Fai, Gle, JT, Ja0, RR, RT, SM, SS1, SS2), 10 libraries, and 11 coincidence thresholds (from *N*mod2, to 10%, 20%, …, up to 90% of the elements in each library).

4. RESULTS

4.1 Extended SALI

The study of the dependency of the eSALI values with the coincidence thresholds went exactly as expected (Fig. 1). Since increasing γ causes the extended similarity to decrease, the denominator in Eq. (6) will increase, leading to smaller values of eSALI. Notice that, for a given combination of similarity index, fingerprint type, coincidence threshold, and property, bigger eSALI values indicate a prevalence of similar molecules with different property values, while a smaller eSALI corresponds to a more homogeneous activity landscape.

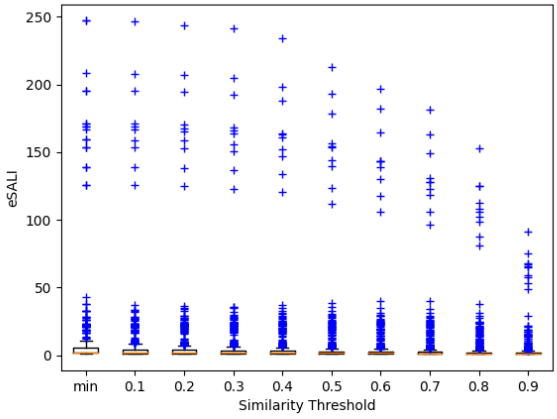

Figure 1: Average eSALI values over the libraries, similarity indices, and fingerprint types vs. coincidence thresholds.

More interestingly, the analysis of the dependency of the eSALI with either fingerprint type or the similarity index provides important insights regarding the optimum way to study the roughness of

property landscapes. This opens the possibility to discuss the impact of fingerprints, similarity index, and threshold of the dataset to quantify the eSALI values. For instance, the individual analysis of the similarity indices highlights some very interesting trends. If we look at the variation of the eSALI values resolved for the different γ values (Fig. 2B) we can see that, overall, most of the indices that include the 0-similarity in their numerators (e.g., AC, BUB, CT1, CT2, Ja0, RT, SM, SS2) tend to result in higher eSALI values and more variability on the other conditions. This can be readily understood in the following way: since these indices reward the absence of common features (e.g., an increase in 0-similar columns will increase the similarity), they could end up giving artificially inflated similarity values, just by virtue of the coincidence of a large number of "off" bits in the molecular representation. Hence, it is easier to have molecules that appear to be structurally similar, even though they do not have similar values in their properties. In other words, property values depend more heavily on the characteristics that the molecules has, than on those that it has not. This heavily implies that similarity indices that only (or heavily) favor the presence of common "on" bits (1-similarity), and do not rely too much on the coincidence of "off" bits, provide more reliable measures of activity cliffs. These indices (e.g., CT3, CT4, Gle, Fai, JT, Ja, RR, SS1), present the same trends when we also average the results over all γ values (Fig. 2A), where we can once again see a tendency towards less variability and lower eSALI values, compared to the 0-similarity cases. Moreover, even within the 1-similarity indices we see some slight variations in the dependency with the computational conditions. Notice that the indices that do not have the 0-similarity in their denominators (e.g., CT4, Gle, JT, Ja, SS1) tend to have more outliers than those that penalize the inclusion of "off" bits (e.g., RR and Fai). The only exception if CT3, but the variability in the CTn indices has been shown to be in part dependent on the length of the molecular representation.

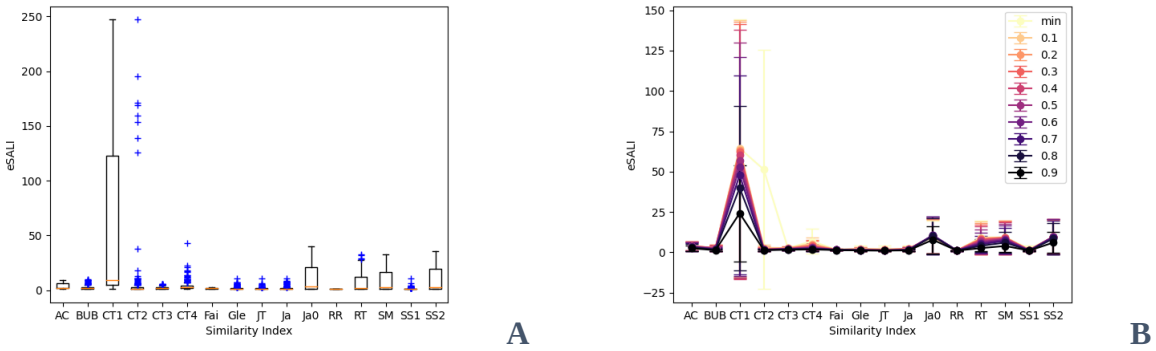

Figure 2: Average eSALI values vs. extended similarity indices. A) Average over the libraries, coincidence thresholds, and fingerprint types; B) γ -resolved averages.

Another factor that is commonly glossed over when studying activity cliffs is the effect of the molecular representation. Fig. 3 includes the analysis of the behavior of 3 popular fingerprint types: MACCS, RDKit, and ECFP4. Here the same general guiding principle holds: we should expect to see lower eSALI values when the given representation is more adept at capturing the structural nuances that dictate the properties of the molecule. In this regard, we would have expected that MACCS would have corresponded to the largest eSALI values. This is so because the reduced number of bits (essentially encoding only 166 features) should make it easier to result in high similarity, given the (on average) higher chance of having bits with the same values across otherwise different molecules (e.g., very much like favoring 0-similarity in the previous example produced bigger similarity scores). However, ECFP4 was actually the representation that showed the highest variability, despite having ~10X more features than MACCS. The comparatively poor performance of ECFP4 mirrors that found by us in chemical diversity studies. The difference ECFP4 and MACCS is yet another reminder that a better description of a system does not depend as much on having more features, but on having the correct features for the problem at hand. On the other hand (and as it was also the case in our previous studies), RDKit fingerprints seem to provide the best way to encode structural information, providing more robust results.

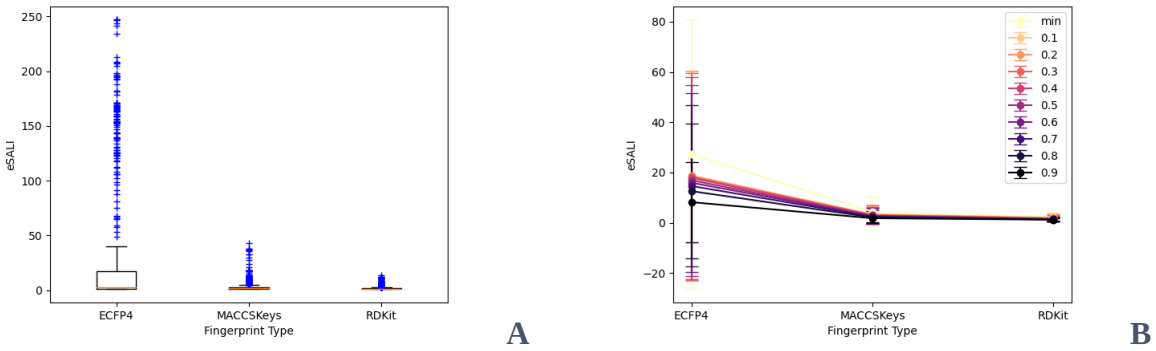

Figure 3: Average eSALI values vs. fingerprint types. A) Average over the libraries, coincidence thresholds, and similarity indices; B) γ -resolved averages.

4.2 Local ranking analysis

As remarked above, we can also use the extended similarity indices to explore structure-activity relations at the local level. We considered two approaches to study the correlation between selected subsets of structural and property similarity rankings: how many identical elements are in each subset (measured by the set Tanimoto or Jaccard index), and the pairwise relation between elements of each subset (calculated via cosine similarity). Fig. 4 presents the average behavior of these two measures, which show markedly different trends. On one hand (Fig. 4J), the set Tanimoto (Jaccard) values increase monotonically with the size of the subsets (represented by N, and taking values of 10%, 20%, …, up to 90% of the size of the original libraries). This is no surprise, because the more elements we consider, the more likely we are to have them replicated in the property and structural subsets, until we reach the obvious limit of Jaccard = 1 in the case in which we select 100% of the data. On the contrary, the cosine similarity (Fig. 4C) slightly decreases with increasing N, since by including more molecules in the subsets we are more likely to decrease their average pairwise similarity.

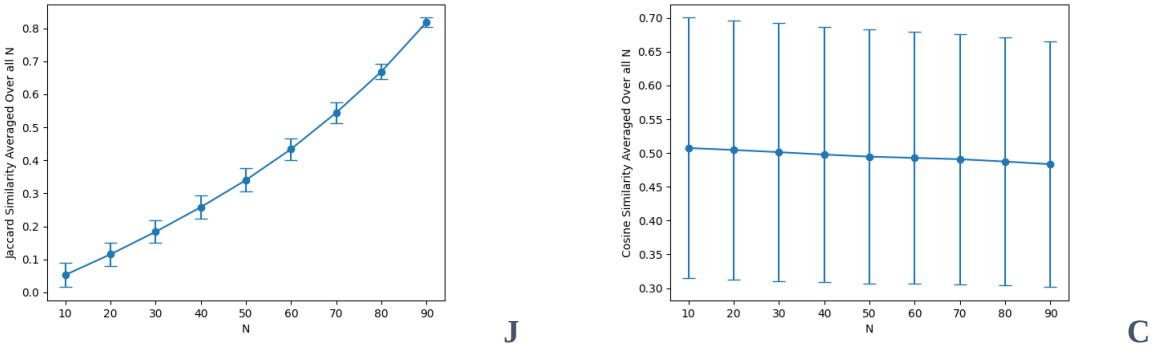

Figure 4: Average set Tanimoto (Jaccard, J) and cosine (C) similarity vs. the size of the selected subsets.

Reassuringly, the trends in both the set Tanimoto (Fig. 5J1) and cosine (Fig. 5C1) values are virtually independent of the coincidence threshold. This means that we do not have to pay too much attention to the selection of this hyperparameter to perform a medoid-based analysis. Note also (Figs. 5J2, 5C2) the previously discussed tendency of the set Tanimoto and cosine values to increase/decrease with N, respectively.

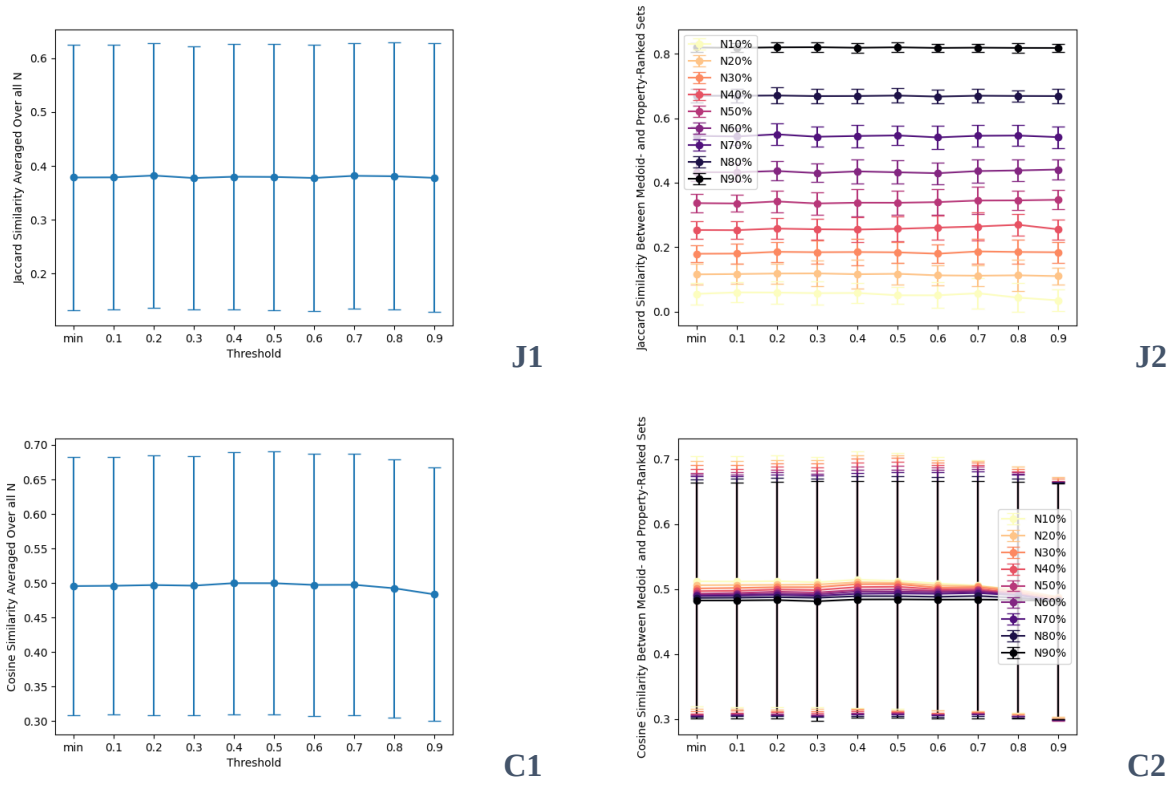

Figure 5: Average set Tanimoto (Jaccard, J) and cosine (C) similarity vs. coincidence threshold. 1) Average over the libraries, similarity indices, fingerprint type, and N; 2) N-resolved averages.

The dependency with the similarity indices (Fig. 6) is a more interesting. First, in the set Tanimoto case (Fig. 6J1), we see that all the indices have essentially exactly the same behavior (something that is also reflected in the N-resolved study, Fig. 6J2). That is, at least for the current selection of molecular libraries, properties, and similarity indices, none of the latter outperforms their counterparts as far as selecting exactly the same molecules from the property and structural rankings. If we consider the cosine analysis it is also certainly difficult to unambiguously say that a given index outperforms all the others. In the fully-averaged case (Fig. 6C1), we can only see some shy "local maxima", with the RR index slightly outperforming the rest, but only by the finest of margins. Unsurprisingly (and as it was also remarked in the global case considered in the previous section), the 1-similarity indices seem to do a relatively better job (in particular, JT and CT3). This tendency appears more clearly when we consider each value of N separately (Fig. 6C2), with once again RR appearing to be the "shy" winner. Hence, even if all the indices are essentially equivalent if we are interested in the identical coincidence of the property and structural rankings,

the 1-similarity indices seem to do a slightly better job at finding molecules in the two rankings that are more similar to each other (albeit not identical). As a word of caution, we warn the reader to not take these conclusions as a universal truth. While we expect the (small) relative advantage of the 1-similarity indices over the 0-similarity ones to be applicable to broad classes of libraries and/or properties, the same should not be expected of the preference for the RR index. First, this index only appears to outperform the others by a very tiny margin. Second, even this small advantage should not be extrapolated beyond the particular properties considered here. The key take-home message from this section, however, is that one can (efficiently) perform this same structure-property medoid-ranking analysis in order to determine, for a particular combination of library/property, which is the similarity that provides the best agreement with the experimental results.

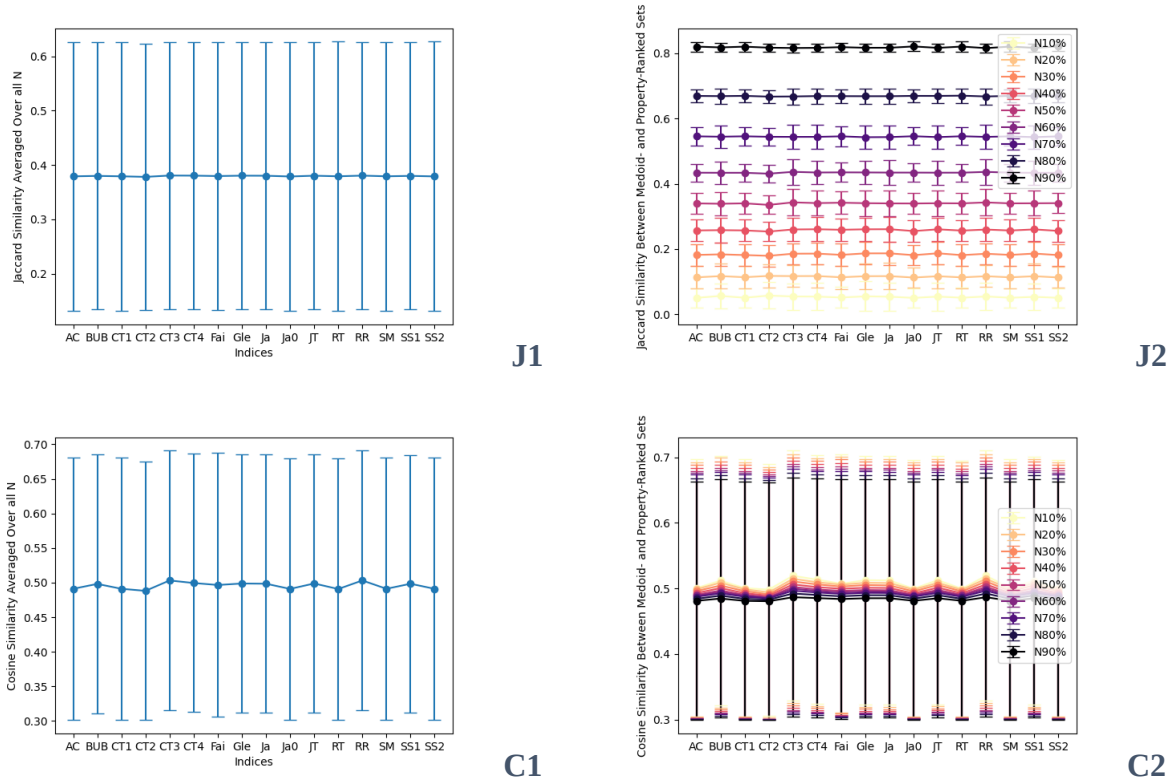

Figure 6: Average set Tanimoto (Jaccard, J) and cosine (C) similarity vs. similarity indices. 1) Average over the libraries, coincidence thresholds, similarity indices, fingerprint type, and N; 2) N-resolved averages.

As in the global case, the role of the molecular representation chosen also provides some very valuable insights. Once again, the exact equivalency between the property and structural rankings is virtually independent of the fingerprint type (see the Jaccard values in Figs. 7J1 and 7J2). However, the cosine analysis indicates that MACCS is the clear winner. It is surprising that this minimalistic representation can capture key structural features such as to provide a better proxy to identify molecules with related properties at a ratio consistently above RDKit (appearing now in a close 2nd place) and ECFP4 (once again with a performance markedly inferior to the other alternatives).

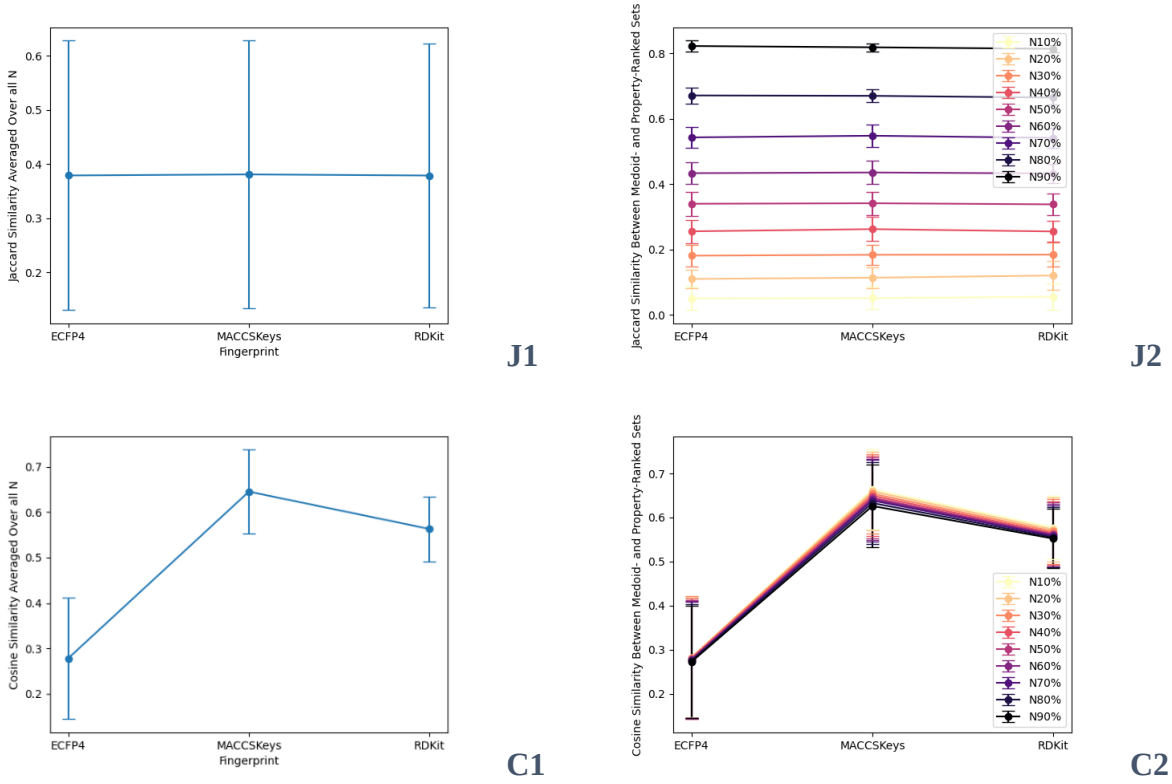

Figure 7: Average set Tanimoto (Jaccard, J) and cosine (C) similarity vs. fingerprint type. 1) Average over the libraries, coincidence thresholds, similarity indices, and N; 2) N-resolved averages.

4.3 Landscape studies using eSALI values

Figure 8 illustrate examples of average eSALI values calculated for different datasets (combining all the coincidence thresholds, similarity indices, and fingerprint types previously considered in

this work). eSALI values enable to quantify the landscapes' roughness (based on specific datasets) and at the same time, eSALI serves as a metric to determine the datasets "modelability". For example, dataset 6 shows a higher average eSALI value (in contrast with the rest of the datasets) that suggests a major roughness in their data, and lower model ability.

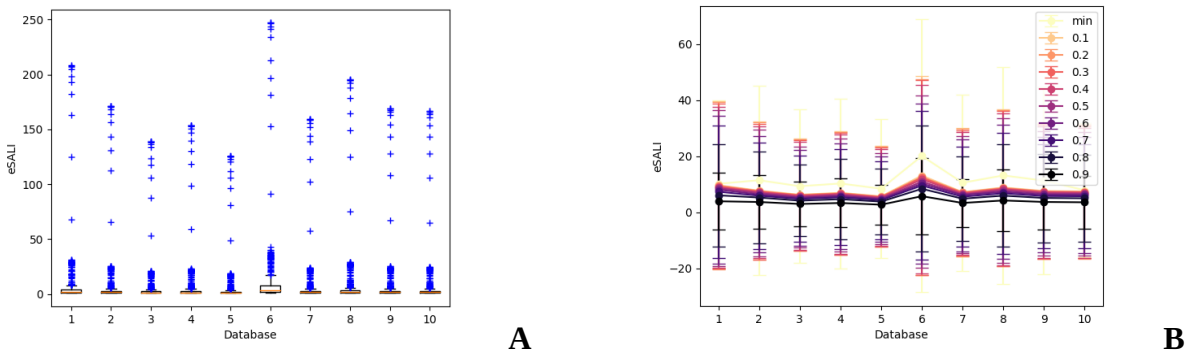

Figure 8: Average eSALI values of the tent datasets considered in this work. A) Average over the libraries, coincidence thresholds, similarity indices, and fingerprints types; B) γ-resolved averages.

Namely, eSALI index points to the presence of activity cliffs in data sets that interfere with the training and accuracy of machine learning models, as discussed recently.[11] Also, eSALI index is an interesting option to explore the roughness of large datasets, owing to its ease of calculation and very low computational cost.

One of the main perspectives of this work is that extended metrics (like eSALI index) could be applied to explore a plethora of properties related to the structure of compounds. For example, toxicity and side effects. Also, it is possible to use this index to develop consensus (or fused) similarity metrics that have demonstrated extensive applicability in drug design and discovery.[40, 41]

5. CONCLUSIONS

In this study, we showed the applicability of the *n*-ary indices to quantify the activity landscapes of ten compound data sets retrieved from the literature using MACCS keys, RDKit, and ECFP4 fingerprints: 16 extended similarity indices and 11 coincidence thresholds. A swift answer to the

title question on the efficacy of the Tanimoto is that, in most cases, this is a perfectly good choice. But the more nuanced study that we presented here shows that this is a characteristic shared by virtually all the similarity indices that put a higher importance on the coincidence of "on" bits rather than in the common absence of identical features between the molecules to be compared. Even more, we showed that there are situations in which the extensively used Tanimoto index can be (even if so slightly) surpassed by some of its lesser-known relatives, like the Russell-Rao index. It was also shown in this work that the medoid algorithm facilitates efficiently computing a structure-property medoid-ranking analysis to determine, for a particular combination of data set/property, which is the similarity coefficient that provides the best agreement with the experimental results as far as correlating structural motifs with nominal property values.

ACKNOWLEDGEMENTS

J. L. M.-F. thanks support from DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant IN201321. E.L.-L. is grateful to Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for the Ph.D. scholarship number, 762342 (No. CVU: 894234). R. A. M. Q. acknowledges funding from UF in the form of a startup grant and a UFII SEED Award.

References

(1) Maggiora, G. M. On outliers and activity cliffs - Why QSAR often disappoints. *J. Chem Inf. Model.* **2006**, *46* (4), 1535-1535, Editorial Material. DOI: 10.1021/ci060117s.

(2) Maggiora, G. M.; Medina-Franco, J. L.; Iqbal, J.; Vogt, M.; Bajorath, J. From Qualitative to Quantitative Analysis of Activity and Property Landscapes. *J. Chem Inf. Model.* **2020**, *60* (12), 5873-5880.

(3) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Perez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* **2014**, *19* (8), 1069-1080.

(4) Stumpfe, D.; Hu, H.; Bajorath, J. Advances in exploring activity cliffs. *J. Comput. Aided Mol. Des.* **2020**, *34*, 929-942.

(5) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving concept of activity cliffs. *ACS Omega* **2019**, *4*, 14360-14368.

(6) Bajorath, J. Representation and identification of activity cliffs. *Expert Opinion on Drug Discovery* **2017**, *12* (9), 879-883.

(7) Reutlinger, M.; Schneider, G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *Journal of Molecular Graphics & Modelling* **2012**, *34*, 108-117, Article. DOI: 10.1016/j.jmgm.2011.12.006.

(8) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (7), 2932-2942.

(9) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem Inf. Model.* **2008**, *48* (3), 646-658.

(10) Aldeghi, M.; Graff, D. E.; Frey, N.; Morrone, J. A.; Pyzer-Knapp, E. O.; Jordan, K. E.; Coley, C. W. Roughness of molecular property landscapes and its impact on modellability. *J. Chem Inf. Model.* **2022**, *62* (19), 4660-4671. DOI: 10.1021/acs.jcim.2c00903.

(11) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem Inf. Model.* **2022**, *62* (23), 5938-5951. DOI: 10.26434/chemrxiv-2022-mfq52.

(12) Todeschini, R.; Ballabio, D.; Consonni, V. Distances and Other Dissimilarity Measures in Chemometrics. In *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, John Wiley & Sons, Ltd., 2015.

(13) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem Inf. Model.* **2012**, *52* (11), 2884-2901.

(14) Miranda-Quintana, R. A.; Cruz-Rodes, R.; Codorniu-Hernandez, E.; Batista-Leyva, A. J. Formal theory of the comparative relations: its application to the study of quantum similarity and dissimilarity measures and indices. *J. Math. Chem.* **2010**, *47* (4), 1344-1365, Article. DOI: 10.1007/s10910-009-9658-6.

(15) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Differential Consistency Analysis: Which similarity measures can be applied in drug discovery? *Molecular Informatics* **2021**, *40* (7), 2060017.

(16) Miranda-Quintana, R. A.; Kim, T. D.; Heidar-Zadeh, F.; Ayers, P. W. On the Impossibility of Unambiguously Selecting the Best Model for Fitting Data. *J. Math. Chem.* **2019**, *57* (7), 1755-1769.

(17) Miranda-Quintana, R. A.; Bajusz, D.; Racz, A.; Heberger, K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics. *Journal of Cheminformatics* **2021**, *13*, 32. DOI: 10.1186/s13321-021-00505-3.

(18) Miranda-Quintana, R. A.; Racz, A.; Bajusz, D.; Heberger, K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. *Journal of Cheminformatics* **2021**, *13*, 33. DOI: 10.1186/s13321-021-00504-4.

(19) Bajusz, D.; Miranda-Quintana, R. A.; Racz, A.; Heberger, K. Extended many-item similarity indices for sets of nucleotide and protein sequences. *Comput. Struct. Biotechnol* **2021**, *19*, 3628-3639. DOI: 10.1016/j.csbj.2021.06.021.

(20) 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.

(21) Flores-Padilla, E. A.; Juarez-Mercado, K. E.; Naveja, J. J.; Kim, T. D.; Miranda-Quintana, R. A.; Medina-Franco, J. L. Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. *Molecular Informatics* **2022**, *41* (6), 2100285. DOI: 10.1002/minf.202100285.

(22) Racz, A.; Dunn, T. B.; Kim, T. D.; Miranda-Quintana, R. A.; Heberger, K. Extended continuous similarity indices: theory and application for QSAR descriptor selection. *J. Comput. Aided Mol. Des.* **2022**, *36* (3), 157. DOI: s10822-022-00444-7.

(23) Racz, A.; Mihalovits, L.; Bajusz, D.; Heberger, K.; Miranda-Quintana, R. A. Molecular dynamics simulations and diversity selection by extended continuous similarity indices. *J. Chem Inf. Model.* **2022**, *62* (14), 3415. DOI: 10.26434/chemrxiv-2022-t61ls.

(24) Chang, L.; Perez, A. Deciphering the Folding Mechanism of Proteins G and L and Their Mutants. *J. Am. Chem. Soc.* **2022**. DOI: 10.1021/jacs.2c04488.

(25) Verhellen, J. Graph-based molecular Pareto optimisation. *Chemical Science* **2022**, *13*, 7526-7535.

(26) Redzepovic, I.; Furtula, B. Chemical similarity of molecules with physiological response. In *Molecular Diversity*, 2022.

(27) Goel, H.; Yu, W.; MacKerell, A. D. hERG Blockade Prediction by Combining Site Identification by Ligand Competitive Saturation and Physicochemical Properties. *Chemistry* **2022**, *4* (3), 630-646.

(28) Dunn, T. B.; Seabra, G. M.; Kim, T. D.; Juarez-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* (9), 2186-2201. DOI: 10.1021/acs.jcim.1c01013.

(29) ChEMBL FTP Directory. In *ChEMBL*, 2022.

(30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Margariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data *Nucleic Acids Res.* **2019**, *47*, D930-D940.

(31) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: an update *Nucleic Acids Res.* **2014**, *42* (D1), D1083–D1090. DOI: 10.1093/nar/gkt1031.

(32) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107. DOI: 10.1093/nar/gkr777.

(33) Hahn, W. C.; Bader, J. S.; Braun, T. P.; Califano, A.; Clemons, P. A.; Druker, B. J.; Ewald, A. J.; Fu, H.; Jagu, S.; Kemp, C. J.; et al. An expanded universe of cancer targets. *Cell* **2021**, *184* (5), 1142-1155.

(34) He, J.-H.; Chen, L.-X.; Li, H. Progress in the discovery of naturally occurring anti-diabetic drugs and in the identification of their molecular targets. *Fitoterapia* **2019**, *134*, 270-289.

(35) Tomaselli, D.; Lucidi, A.; Rotili, D.; Mai, A. Epigenetic polypharmacology: A new frontier for epi-drug discovery. *Medical Research Reviws* **2020**, *40* (1), 190-244.

(36) Lopez-Lopez, E.; Cerda-Garcia-Rojas, C. M.; Medina-Franco, J. L. Tubulin Inhibitors: A Chemoinformatic Analysis Using Cell-Based Data *Molecules* **2021**, *26* (9), 2483.

(37) Lopez-Lopez, E.; Rabal, O.; Oyarzabal, J.; Medina-Franco, J. L. Towards the understanding of the activity of G9a inhibitors: an activity landscape and molecular modeling approach. *J. Comput. Aided Mol. Des.* **2020**, *34*, 659-669.

(38) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31-36.

(39) Landrum, G. *RDKit: Open-source cheminformatics; http://www.rdkit.org*. (accessed.

(40) Lopez-Lopez, E.; Medina-Franco, J. L. Towards Decoding Hepatotoxicity of Approved Drugs through Navigation of Multiverse and Consensus Chemical Spaces *Biomolecules* **2023**, *13* (1), 176.

(41) Lopez-Lopez, E.; Cerda-Garcia-Rojas, C. M.; Medina-Franco, J. L. Consensus Virtual Screening Protocol Towards the Identification of Small Molecules Interacting with the Colchicine Binding Site of the Tubulin-microtubule System. *Molecular Informatics* **2023**, *42* (1), 2200166.