iSIM: Instant Similarity

Kenneth López-Pérez,[1] Taewon D. Kim,[1] Ramón Alain Miranda-Quintana[1*]

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

[*] Email: quintana@chem.ufl.edu

ABSTRACT

The quantification of molecular similarity has been present since the beginning of cheminformatics. Although several similarity indices and molecular representations have been reported, all of them ultimately reduce to the calculation of molecular similarities of only two objects at a time. Hence, to get the average similarity of a set of molecules, all the pairwise comparisons need to be computed, which demands a quadratic scaling in the number of computational resources. Here we propose an exact alternative to this problem: *i*SIM (Instant Similarity). *i*SIM performs comparisons of multiple molecules at the same time and yields the *same* value as the average pairwise comparisons of molecules represented with binary fingerprints and real-value descriptors. In this work, we introduce the mathematical framework and several applications of *i*SIM in chemical sampling, visualization, diversity selection, and clustering.

1. Introduction

Molecular fingerprints are one the most common representations of compounds in cheminformatics. The simplest version of fingerprints are binary vectors, where the presence of a structural feature is represented by a 1 and the absence by a 0.[1] Another popular representation are molecular descriptors, which correspond to useful numbers that encode information about a molecule; commonly they could be calculated from graph theory, quantum chemistry, topological or experimental methods, to mention some sources.[2] Despite their apparent differences, both descriptors and fingerprints can be used to calculate the similarity between two molecules. From a mathematical point of view, a similarity index is a metric that measures how "related" are two

points in a chemical space.[3] Multiple similarity measurements have been reviewed and analyzed, with the well-known Jaccard-Tanimoto coefficient (JT)[4,5] being the usual go-to in the cheminformatics community.[6] The main point of calculating similarity measurements lies on the *"molecular similarity principle"*: similar molecules have similar properties/activities.[7] This powerful idea is at the core of virtual screening[8–11], hit selection[12], QSAR/QSPR modeling[13,14], many chemical space exploration methods[15,16], activity landscape description[17,18], diversity selection[19], clustering[20,21], and many more applications.

The common way of quantifying similarity is by comparing two molecules. If one wants the similarity/diversity of a library, the typical way of doing so would be calculating the average similarity of all the possible comparisons in the library, which is a computationally costly O(N^2) step. Motivated to solve this problem, our group recently developed the concept of extended similarity.[22,23] Extended similarity performs the comparison of all the molecules in a set at the same time and yields a similarity metric for the whole set. Briefly, for a matrix of size $N \times M$, where M is the size of the fingerprint or number of molecular descriptors and N the number of molecules in a set, the first step is to sum the elements column wise, $\sum = [\sigma_1, \sigma_2, \ldots, \sigma_M]$. Each column sum, σ_k, can be used to classify the column when it is compared to a threshold in the following way: i) if $2\sigma_k - N > \gamma$ it will be a 1-similarity column, ii) if $N - 2\sigma_k > \gamma$ it will be a 0-similarity column, iii) otherwise it will count as dissimilarity. Then, using $\Delta_{\sigma_k} = |2\sigma_k - N|$ as independent variable a weighting function should be used to consider partial similarity and dissimilarity, the function should be positive and increasing. Now the variables of any similarity metric can be transformed using the sum of the weighted or non-weighted counters. The major advantage of extended similarity is that it calculates a similarity metric for the whole set much more efficiently than by using the traditional pairwise comparisons, with this calculation now scaling as O*(N)*. [22,23]

Extended similarity has been applied to several cheminformatics problems like diversity selection, [23,24] molecular dynamics simulations,[25,26] library diversity studies,[27–29] activity cliffs,[30] descriptor selection for QSAR/QSPR model,[31] fingerprint evaluations,[32] and chemical space visualization.[33] Despite the advantages, linear scaling and simultaneous multiple comparisons, there are certain some drawbacks like the need of a coincidence threshold analysis to determine the best similarity/dissimilarity separation and a different numeric value than the pairwise comparisons. Those limitations inspired this work where we show the mathematical framework,

analysis, and cheminformatics applications of *i*SIM, an "instantaneous" similarity measurement for binary fingerprints and molecular descriptors that yields virtually the same value as the average pairwise similarity comparisons in a linear scaling with the number of observations.

2. Theory

2.1 Binary representations

Comparisons over molecular fingerprints are based on three key indicators: the number of times there is a coincidence of two "on" bits between the fingerprints (denoted by a), the number of times there is a coincidence of two "off" bits between the fingerprints (denoted by d), and the mismatches between the fingerprints, when one bit is "on" and the other is "off" (denoted by $b + c$). With these ingredients one can propose a plethora of similarity indices, which could be interpreted as such as long as they are monotonically increasing functions of a and d, and monotonically decreasing functions of $b + c$. Here, we will be concerned mainly with the Russel-Rao (RR)[34], Jaccard-Tanimoto (JT)[4,5], and Sokal-Michener (SM)[35] indices:

$$\mathrm{RR} = \frac{a}{a+d+b+c} \tag{1}$$

$$\mathrm{JT} = \frac{a}{a+b+c} \tag{2}$$

$$\mathrm{SM} = \frac{a+d}{a+d+b+c} \tag{3}$$

(Notice that, trivially, $\mathrm{RR} \le \mathrm{JT} \le \mathrm{SM}$.)

The very definition of, say, a, seems to imply that when we have N molecules, as we need to consider the $\binom{N}{2} = \frac{N(N-1)}{2}$ distinct pairs to check the coincidence or not of on bits. However, it is possible to access the same information in far fewer operations. The first step is to arrange all the fingerprints in a matrix, with each fingerprint corresponding to a row. Then, we just need to find the sum of each column, which generates a vector $K = [k_1 k_2 ... k_m]$, with element k_q corresponding to the sum of the q^{th} column, and m indicating the length of the fingerprints. The key insight is to note that the k's are all that we need to calculate the number of times we will have coincidence or not of any type of bits. For instance, there will be $\binom{k_q}{2} = \frac{k_q(k_q-1)}{2}$ instances in

which two on bits will coincide in column q. Likewise, there will be $\binom{N-k_q}{2}=\frac{\left(N-k_q\right)\left(N-k_q-1\right)}{2}$ coincidences of off bits. Finally, the number of mismatches is $k_q\left(N-k_q\right)$. It is natural then to make the following identification (with the sums running over all bit positions):

$$a \to \sum_{q=1}^{m} \frac{k_q\left(k_q-1\right)}{2} \tag{4}$$

$$d \to \sum_{q=1}^{m} \frac{\left(N-k_q\right)\left(N-k_q-1\right)}{2} \tag{5}$$

$$b+c \to \sum_{q=1}^{m} k_q\left(N-k_q\right) \tag{6}$$

With this, we have everything in place to define instantaneous similarity (*i*SIM) versions of the previously discussed indices, *i*RR, *i*JT, and *i*SM, as:

$$\begin{aligned} i\mathrm{RR} &= \frac{\sum_{q=1}^{m} \frac{k_q\left(k_q-1\right)}{2}}{\sum_{q=1}^{m}\left\{\frac{k_q\left(k_q-1\right)}{2}+\frac{\left(N-k_q\right)\left(N-k_q-1\right)}{2}+k_q\left(N-k_q\right)\right\}} \\ &= \frac{\sum_{q=1}^{m} k_q\left(k_q-1\right)}{mN\left(N-1\right)} \end{aligned} \tag{7}$$

$$i\mathrm{JT} = \frac{\sum_{q=1}^{m} \frac{k_q\left(k_q-1\right)}{2}}{\sum_{q=1}^{m}\left\{\frac{k_q\left(k_q-1\right)}{2}+k_q\left(N-k_q\right)\right\}} \tag{8}$$

$$
\begin{aligned}
i\mathrm{SM} &= \frac{\sum_{q=1}^{m}\left\{\frac{k_q\left(k_q-1\right)}{2}+\frac{\left(N-k_q\right)\left(N-k_q-1\right)}{2}\right\}}{\sum_{q=1}^{m}\left\{\frac{k_q\left(k_q-1\right)}{2}+\frac{\left(N-k_q\right)\left(N-k_q-1\right)}{2}+k_q\left(N-k_q\right)\right\}} \\
&= \frac{\sum_{q=1}^{m}\left\{k_q\left(k_q-1\right)+\left(N-k_q\right)\left(N-k_q-1\right)\right\}}{mN\left(N-1\right)}
\end{aligned}
\tag{9}
$$

The case of *i*RR and *i*SM is special, because the denominators in Eqs. (1) and (3) are always constant, since $a+d+b+c=m$, the number of digits in the fingerprints (a fact that we explicitly use in the 2nd, simpler, form of the *i*RR and *i*SM indices shown above). Then, we can interpret Eqs. (7) and (9) as effectively combining the RR and SM similarities over each independent bit positions. Given the constant-denominator characteristic, it is then easy to see that the *i*SIM version of these indices will provide the exact average of all the pairwise RR and SM values over the given set, but at only O(*N*) cost. *i*JT, on the other hand, will not in general give exactly the same value as the average of the pairwise Tanimoto calculations. Once again, the key is that the JT denominator is not the same for arbitrary pairs of fingerprints. In this case, we can interpret *i*JT as an O(*N*) mediant approximation[36,37] to the O(N^2) average. Despite this simplification, as shown below, *i*JT still provides superb estimates for the pairwise average over a varied set of conditions.

2.2 Real-value representations

The previous results are promising, so it is certainly desirable to extend them to more general types of molecular representations. Here, we show how this can be done for vectors of real values. The key insight is to use inner products between the molecular "vectors" instead of the more limited *a*, *d*, and *b* + *c* indicators used in the binary case. To do this we will focus on the case where the *i*th molecule, $X^{(i)}$, is represented by a vector of descriptors $\left|X^{(i)}\right\rangle=\left[x_1^{(i)},x_2^{(i)},...,x_m^{(i)}\right]$. Without losing any generality, these vectors are considered to be normalized: $\forall i,q:0\le x_q^{(i)}\le 1$. The main motivation behind the focus on normalized descriptors is that we can then easily define the "flipped" representation of a molecule, $\left|\tilde{X}^{(i)}\right\rangle$, as the real-value equivalent of flipping the bits of a binary representation, that is: $\left|\tilde{X}^{(i)}\right\rangle=\left[1-x_1^{(i)},1-x_2^{(i)},...,1-x_m^{(i)}\right]$. In terms of inner products, the previously analyzed indices can be written as:

$$\mathrm{RR} = \frac{\left\langle X^{(i)} \middle| X^{(j)} \right\rangle}{m} \tag{10}$$

$$\mathrm{JT} = \frac{\left\langle X^{(i)} \middle| X^{(j)} \right\rangle}{\left\langle X^{(i)} \middle| X^{(i)} \right\rangle + \left\langle X^{(j)} \middle| X^{(j)} \right\rangle - \left\langle X^{(i)} \middle| X^{(j)} \right\rangle} \tag{11}$$

$$\mathrm{SM} = \frac{\left\langle X^{(i)} \middle| X^{(j)} \right\rangle + \left\langle \tilde{X}^{(i)} \middle| \tilde{X}^{(j)} \right\rangle}{m} \tag{12}$$

Notice that, for simplicity, we have directly used the fact that the denominators of the RR and SM indices are constant and equal to the total length of the molecular vectors, *m*.

Once again, the way of writing Eqs. (10)-(12) seems to suggest that calculating the average of all the RR, JT, or SM comparisons demands O(N^2). However, we can actually calculate the sum of all the involved inner products in O(N) (albeit, with a larger overhead, compared to the binary case).

First, for the inner products between the molecular representations, we have:

$$\left\langle X^{(i)} \middle| X^{(j)} \right\rangle = \sum_{q=1}^{m} x_q^{(i)} x_q^{(j)} \tag{13}$$

Then, for the relevant inner products appearing in Eqs. (10)-(12):

$$\begin{aligned} \sum_{i<j} \left\langle X^{(i)} \middle| X^{(j)} \right\rangle &= \sum_{i<j} \sum_{q=1}^{m} x_q^{(i)} x_q^{(j)} \\ &= \frac{1}{2} \sum_{q=1}^{m} \left\{ \left(\sum_i x_q^{(i)} \right)^2 - \sum_i \left(x_q^{(i)} \right)^2 \right\} \end{aligned} \tag{14}$$

$$\begin{aligned} \sum_{i<j} \left\langle \tilde{X}^{(i)} \middle| \tilde{X}^{(j)} \right\rangle &= \sum_{i<j} \sum_{q=1}^{m} \left[1 - x_q^{(i)} \right] \left[1 - x_q^{(j)} \right] \\ &= \frac{1}{2} \sum_{q=1}^{m} \left\{ \left(\sum_i \left[1 - x_q^{(i)} \right] \right)^2 - \sum_i \left[1 - x_q^{(i)} \right]^2 \right\} \end{aligned} \tag{15}$$

From these expressions, it is clear that we can follow a similar route to the one taken for the binary input. First, we need to arrange all the molecular vectors in a matrix X. Then, we need to generate some related matrices: a) The "flipped" matrix $\tilde{X} = 1 - X$, b) the Hadamard (element-wise) squares of these matrices, X^2, $\tilde{X}^2$. That is, if the element in row i and column q in X is given by $x_q^{(i)}$, then the corresponding elements of matrices $\tilde{X}, X^2, \tilde{X}^2$ will be

$1-x_q^{(i)}, \left(x_q^{(i)}\right)^2, \left[1-x_q^{(i)}\right]^2$, respectively. It is important to remark that since we are only taking element-wise products, generating these auxiliary matrices will only demand O(N) operations. Then, the sum of the columns of matrices X and $\tilde{X}$ gives vectors with components $\sum_i x_q^{(i)}, \sum_i \left[1-x_q^{(i)}\right]$, respectively. On the other hand, the sum of the columns for the Hadamard squares gives the factors $\sum_i \left(x_q^{(i)}\right)^2, \sum_i \left[1-x_q^{(i)}\right]^2$. These are all the ingredients necessary to calculate the real-value *i*SIM similarity indices:

$$i\mathrm{RR} = \frac{\sum_{q=1}^{m}\left\{\left(\sum_i x_q^{(i)}\right)^2 - \sum_i \left(x_q^{(i)}\right)^2\right\}}{mN\left(N-1\right)} \tag{16}$$

$$i\mathrm{JT} = \frac{\frac{1}{2}\sum_{q=1}^{m}\left\{\left(\sum_i x_q^{(i)}\right)^2 - \sum_i \left(x_q^{(i)}\right)^2\right\}}{\left(N-1\right)\sum_{q=1}^{m}\sum_i \left(x_q^{(i)}\right)^2 - \frac{1}{2}\sum_{q=1}^{m}\left\{\left(\sum_i x_q^{(i)}\right)^2 - \sum_i \left(x_q^{(i)}\right)^2\right\}} \tag{17}$$

$$i\mathrm{SM} = \frac{\sum_{q=1}^{m}\left\{\left(\sum_i x_q^{(i)}\right)^2 - \sum_i \left(x_q^{(i)}\right)^2\right\} + \sum_{q=1}^{m}\left\{\left(\sum_i \left[1-x_q^{(i)}\right]\right)^2 - \sum_i \left[1-x_q^{(i)}\right]^2\right\}}{mN\left(N-1\right)} \tag{18}$$

Once again, *i*RR and *i*SM provide the same exact results as the average of all the pairwise comparisons, due to the convenient constant denominators. For *i*JT, this is just a median-like approximation but, as it will be illustrated below with different numerical tests, Eq. (17) provides an excellent approximation to the O(N^2) result.

3. Systems

10,000 random datasets were generated, with the number of fingerprints ranged from 100 to 1000 and the size of the fingerprints ranged from 166 to 2048. For the binary case, to ensure that datasets covered the complete range of the similarity indexes domains, each dataset was randomly biased to have different proportion of ones and zeros.

For testing on real libraries, 30 CHEMBL curated datasets by van Tilborg et al.[38] were used. Three binary fingerprint types were generated using RDKit[39]: RDKIT[39] (m = 2048), MACCS[40] (m = 167)

and ECFP4[41] (m = 2014). All the real and discrete descriptors offered by the RDKit Descriptors[39] module was computed, descriptors with calculation errors or nan values were dropped for a total of 208 descriptors (full list, SI). Min max normalization was used prior to *i*SIM calculations. The code used in this manuscript can be found at: https://github.com/mqcomplab/iSIM

4. Numerical Results

4.1 Average similarity

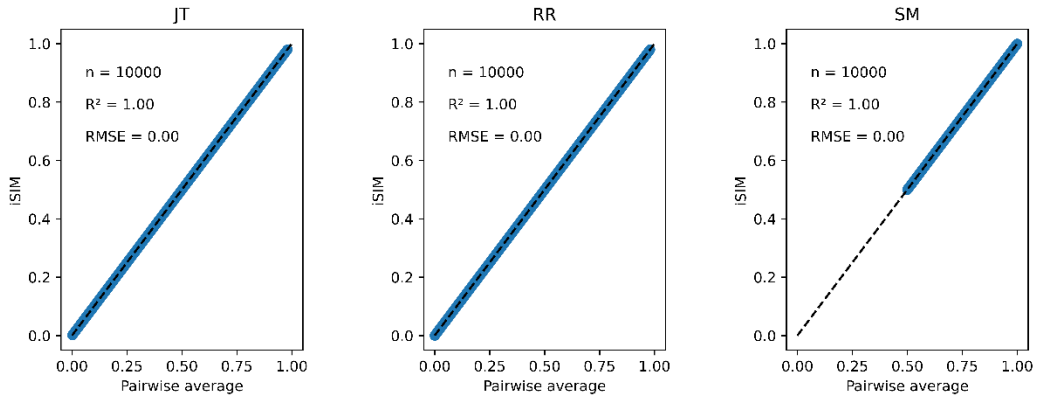


Figure 1: *i*SIM vs pairwise results for 10,000 randomly generated libraries. Molecules represented with binary fingerprints.

Our first tests were oriented towards checking the correspondence between the *i*SIM results and the average of the pairwise comparisons over a large number of libraries. For this, we used the 10,000 randomly generated libraries described in section 3. As can be seen in Fig. 1, the *i*SIM results perfectly reproduce the more computationally demanding standard comparisons. In our previous contributions, we had only focused on the relation between the previously extended similarity results and the pairwise metrics as far as the ability of the extended indices to preserve the ranking of the comparisons (see, for example, Fig. 7 in Ref. 23). The test presented in Fig. 1 is much more demanding, because we are comparing the similarity values obtained from both approaches. As noted in the Theory section, we expected the *i*RR and *i*SM results to be (analytically) identical to the pairwise averages. Even more remarkable, we see that *i*JT provides a superb estimate for the O(N^2) averages. This behavior is also observed over real datasets. In Fig. 2 we show a similar comparison, but now over 30 CHEMBL libraries, each represented with three

different types of fingerprints. Figure S1 and S2 include the same comparisons with more similarity indexes formulas that *i*SIM can be applied to.

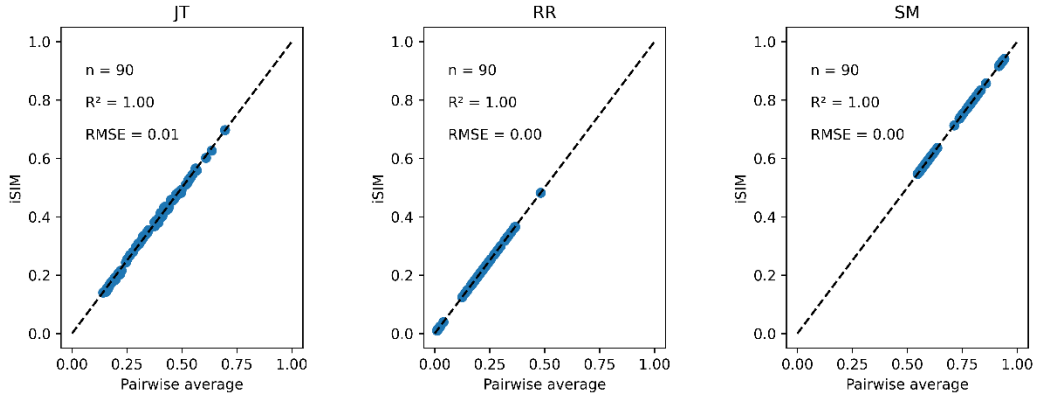


Figure 2: *i*SIM vs pairwise results for 30 CHEMBL libraries (binary input). Molecules represented with binary MACCS, RDKit, and ECFP4 (binary) fingerprints.

Figs. 3 and 4 present the equivalent results, but for molecules represented with (normalized) descriptors (e.g., real values). Once again, *i*RR and *i*SM show a perfect agreement both for the randomly generated and for the real data. The median approximation in *i*JT is also remarkably robust over real data, essentially operating at close level as for the binary fingerprints.

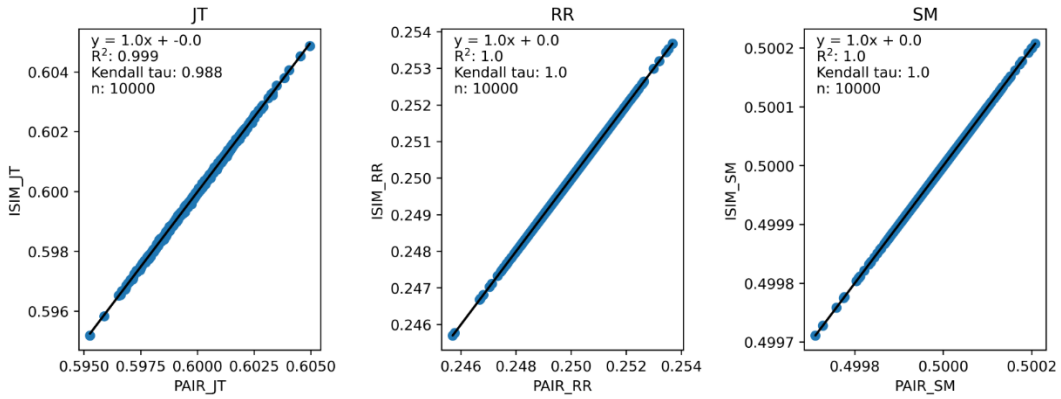


Figure 3: *i*SIM vs pairwise results for 10,000 randomly generated libraries (real input). Molecules represented with random generated fingerprints with real normalized variables.

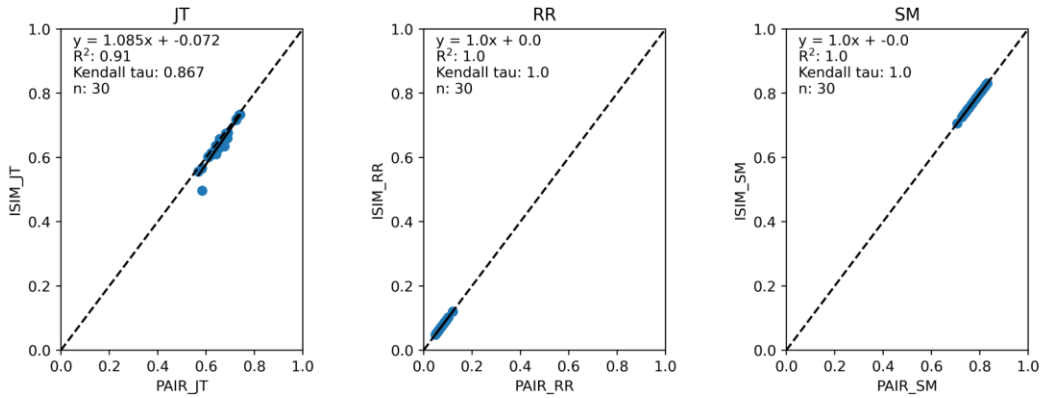


Figure 4: *i*SIM vs pairwise results for 30 CHEMBL libraries (real input). Molecules represented with 208 RDKIT real normalized descriptors.

4.2 Local analysis of molecular libraries

4.2.1 Complementary similarity

Complementary similarity calculations can also be applied with *i*SIM, as they were previously applied using extended similarity. One molecule is taken out of the set, and *i*SIM is calculated on the remaining compounds, in this way low values will correspond to molecules that inhabit high density regions in chemical space. Conversely, high complementary similarity corresponds to molecules from low density regions, thus overall, least similar to the rest of the set. This tool enables a ranking of molecules on how similar they are to the rest of the set, the most similar molecule, the medoid, has the lowest complementary similarity and on the end of the ranking we will have the outlier. [cite] As example in Fig. 5, medoid and outlier molecules from a dataset can be identified doing the complementary similarity ranking. Since we have a ranking of the molecules, the medoid and outlier cutoff can be flexible depending on the user needs, this gives an opportunity of visualization of relevant structures for the set. The information contained in the complementary similarity ranking has proven to be very valuable in stratifying the data as a pre-processing step in clustering,[25] as well as a way to quickly sample different regions of chemical space.[33]

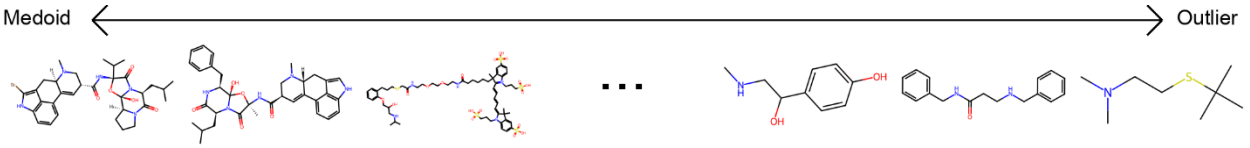


Figure 5: Structures of the CHEMBL214 database ranked by increasing complementary similarity using RDKIT fingerprints and iRR similarity index. Structures shown correspond to the top (medoids) and bottom (outliers) three molecules.

4.2.2 Diversity selection

To further expand on the applicability of *i*SIM, we focused on the classical cheminformatics task of sampling a given library in the most diverse way possible: the diversity picking problem. Just like the extended indices before, *i*SIM naturally leads to a diversity selection algorithm (*i*SIMDiv):

a) Pick a molecule and add it to the Selected set. (This is usually done at random, but in order to increase the reproducibility of our results, in all cases we start from the medoid of the set.)

b) At every step, pick the molecule that will result in the lowest *i*SIM for the Selected set.

As shown in Fig. 6A, this simple recipe leads to more diverse sub-sets than the popular MaxMin diversity selection algorithm.[42,43] There, we tested the performance of these algorithms over the CHEMBL214 library, corresponding to the 5-HT1a receptor. [cite] We selected a library with 3,317 molecules (represented using RDKIT fingerprints), and we monitored the process of selecting up to the 10% most diverse compounds. (The general trends observed for this library were corroborated for other libraries, similarity indices, and fingerprint types, see the SI.) If we quantify the chemical diversity of the selected set as inversely related to the average of the pairwise similarities of the molecules in the selected sub-set (the "y axis" in Fig. 5A), we see that *i*SIM (with the *i*RR metric), at worst, finds sets that are as diverse as those found by MaxMin with the standard pairwise RR. This happens at the very early stages, when we have only picked a handful of molecules, but then quickly the *i*SIM results become more diverse. This is no surprise since, by definition, *i*SIM is constructed to reproduce the average of the pairwise comparisons. Hence, the *i*SIMDiv algorithm is directly maximizing this measure of chemical diversity.

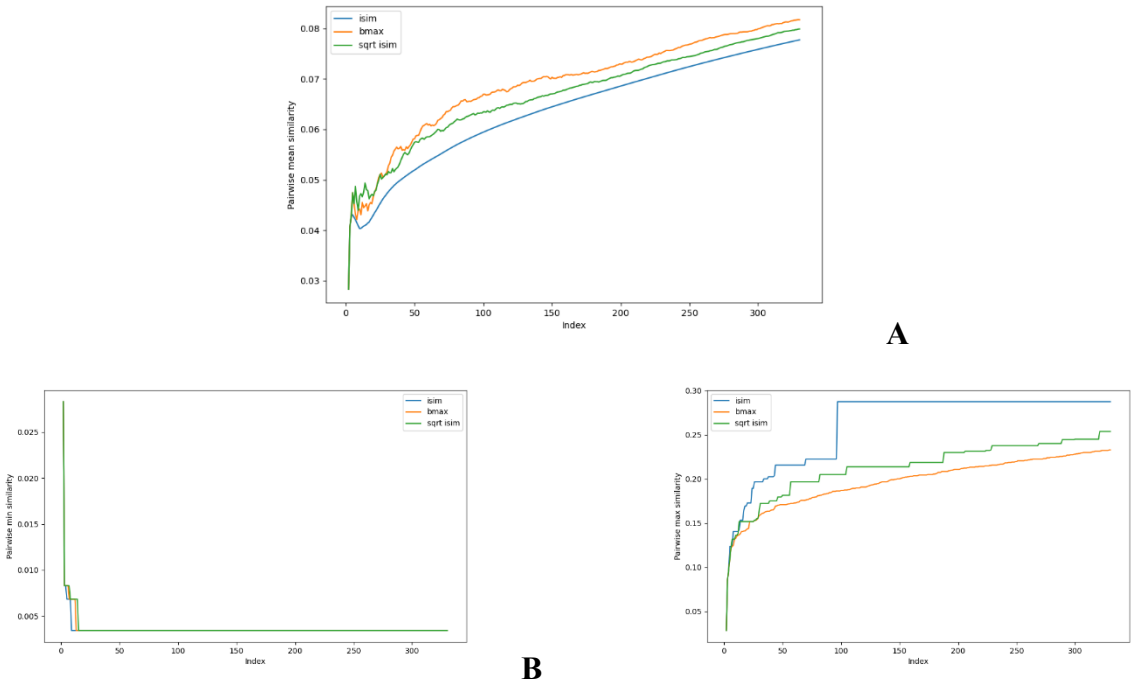


Figure 6: MaxMin (bmax, yellow), *i*RR (isim, blue), and sqrt_*i*RR (sqrt_isim, green) results for the diversity sampling of the CHEMBL214 dataset represented with RDKIT fingerprints: A) pairwise similarity of the Selected set, B) minimum similarity between elements of the Selected set, C) maximum similarity between elements of the selected set.

If at the "global" or "coarse" level of the selected set it is clear that *i*SIMDiv produces more diver sets, it is also interesting to study the "local" relations among the selected molecules. For instance, as shown in Fig. 6B, *i*SIMDiv is the algorithm that first finds a pair of "orthogonal" molecules in the data, that is, a pair of molecules with 0 similarity between them. On the other hand, we also see in Fig. 6C that *i*SIMDiv tends to produce selected sets where the closest pair of molecules is more similar to each other than the closest pair of molecules selected by MaxMin. This is in line with the properties of MaxMin, since this method tries to maximize the minimum distance between the new added molecule and those already selected. As a way to bridge the local gap between MaxMin and *i*SIMDiv we propose a version of *i*SIM that attempts to minimize not the sum of similarities, but the sum of the *square roots* of the similarities. This sqrt_*i*SIM can be easily calculated at the same cost as *i*SIM: for any *i*SIM variant, after calculating the sums of the columns of the molecular representations and generating the analogues of the a, d, and $b + c$

indicators, we take their square roots and we use those in the same expression for the similarity indices. For example, in the case of *i*RR, we would be minimizing:

$$\text{sqrt\_}i\text{RR} = \sqrt{\frac{2}{m}} \frac{1}{N(N-1)} \sum_{q=1}^{m} \sqrt{k_q(k_q-1)} \tag{19}$$

As can be seen in Fig. 6C, minimizing this new objective function results in selected sets that are much locally closer to MaxMin, in the sense of having almost maximally dissimilar pairs of closest molecules. However, as reflected in Fig. 6A, this new sampling strategy also produces sub-sets that are more globally diverse than MaxMin (albeit, not as diverse as those generated by *i*SIMDiv). In other words, by making changes to the objective function calculated within the *i*SIM framework, we can control the global and local properties of the sampled sets. Plots showing same trends on the chemical diversity selection method for more databases, fingerprint representations and similarity indexes are included in the SI.

Another way of modifying the *i*SIM objective diversity metric that allows a faster diversity selection is what we called *i*SIMRevDiv: *i*SIM reversed diversity selection. In this algorithm we start with all the points, and we iterate to find the molecule that, if removed, the remaining set will result the in the lowest similarity value. This process is then repeated until the number of desired molecules is reached. *i*SIMRevDiv will be extremely useful in cases where more than 50% of the set wants to be picked. Figures 7 show the *i*SIM and computing time comparison between the *i*SIMDiv and *i*SIMRev methods for the CHEMBL214 database represented with RDKIT fingerprints and using *i*RR metric. Figure 7A shows how when the diversity selection is started from the outlier, both forward and reversed iSIM diversity selection methods will yield the same average pairwise similarity results, which enables the user to use any of the two methods depending on the data percentage that wants to be picked. Figure 7B shows computing times, and further supports the idea that for selections over 50% of the data the *i*SIMDivRev will be less computationally costly with the same high-quality results.

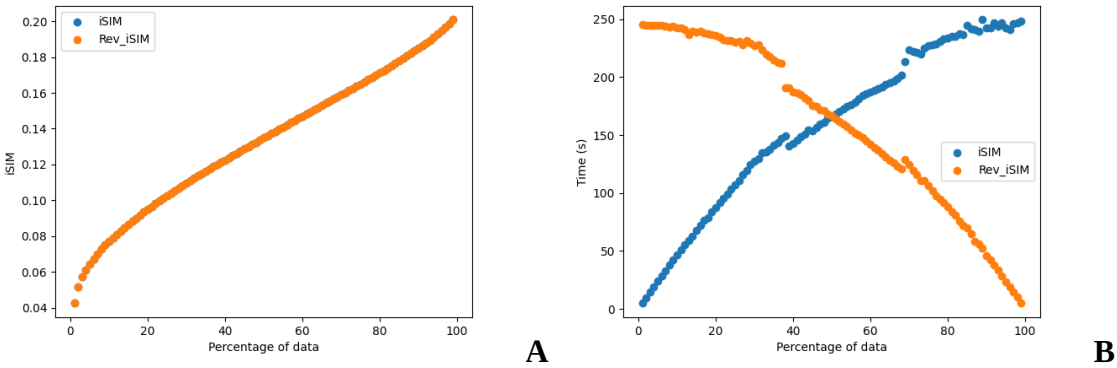


Figure 7: A) *i*SIMDiv and *i*SIMRevDiv selections for different data percentages (1-99%, in 1% steps) for the CHEMBL214 dataset represented with RDKIT fingerprints and selected by *i*RR index. B) Computing time variation of the diversity selection methods with the data percentage selected.

To further analyze the comparison of the *i*SIM diversity selection both forward and reversed ways, random fingerprints dataset were generated, and the results were consistent with the ones previously observed (see Figure S6). The fact that diversity selection can be done in a reversed way, puts *i*SIM diversity selection a foot in front of typical algorithms like MaxMin were the computation of a pairwise similarity is required.

With the aim of comparing visually all diversity selection methods proposed in this work, Principal Component Analysis (PCA)[44,45] and t-Distributed Stochastic Neighbor Embedding (t-SNE)[46] plots were generated identifying the 10% most diverse subset according to each of the methods. For the PCA in Figure 8 it is appreciated how *i*SIMDiv and *i*SIMRevDiv cover more of the two-dimensional space of the first principal components scores than the commonly used MaxMin algorithm and sqrt_*i*SIMDiv, which is consistent with our previous results. On Figure 9, the plots are very similar, which enforces that the proposed diversity selection methods are at the level of quality of the MaxMin.

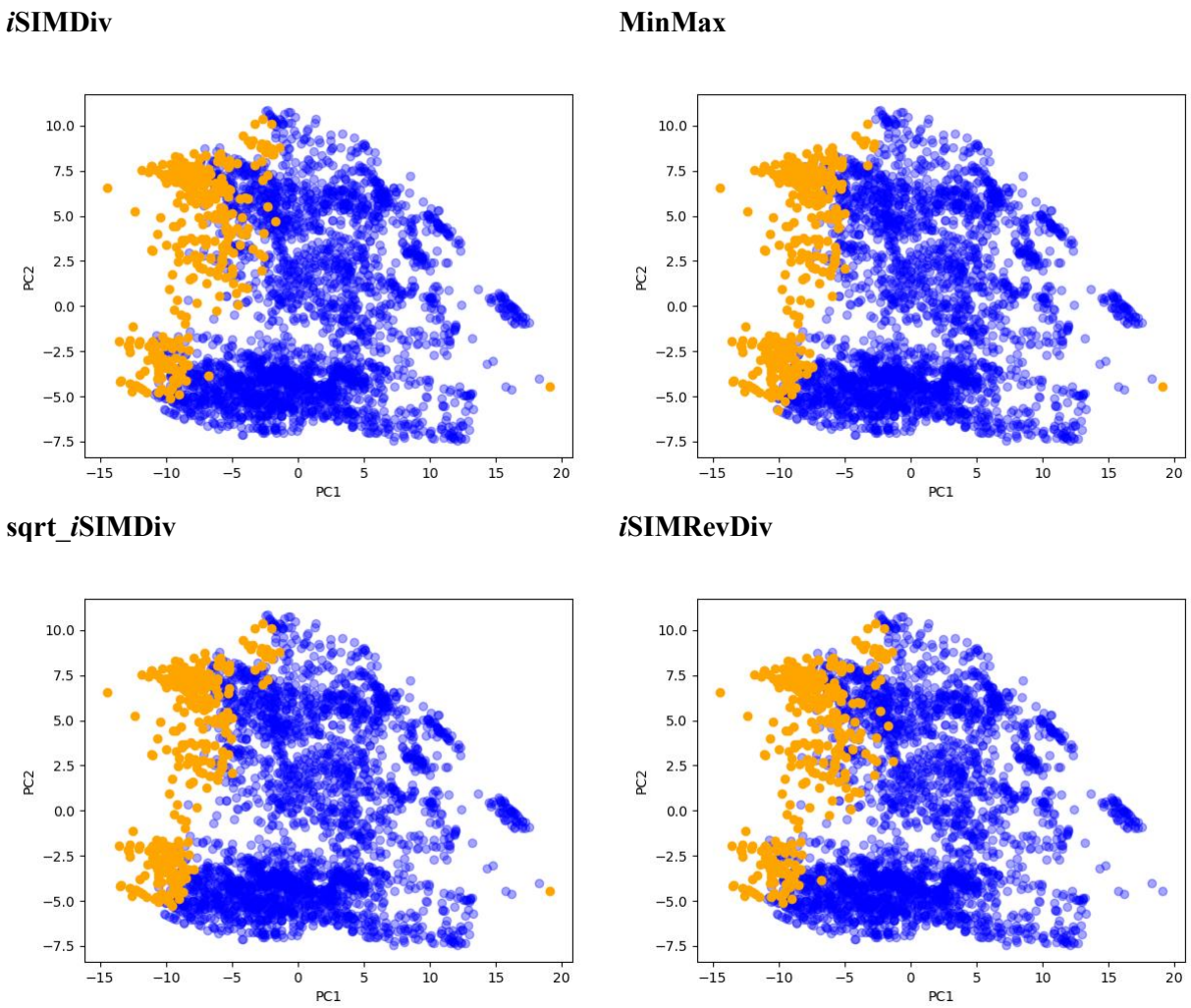


Figure 8: Principal component analysis scoring plots of the two first components for the CHEMBL214 dataset represented by RDKIT binary fingerprints by diversity selection methods. *i*SIM related methods use *i*RR similarity index as metric.

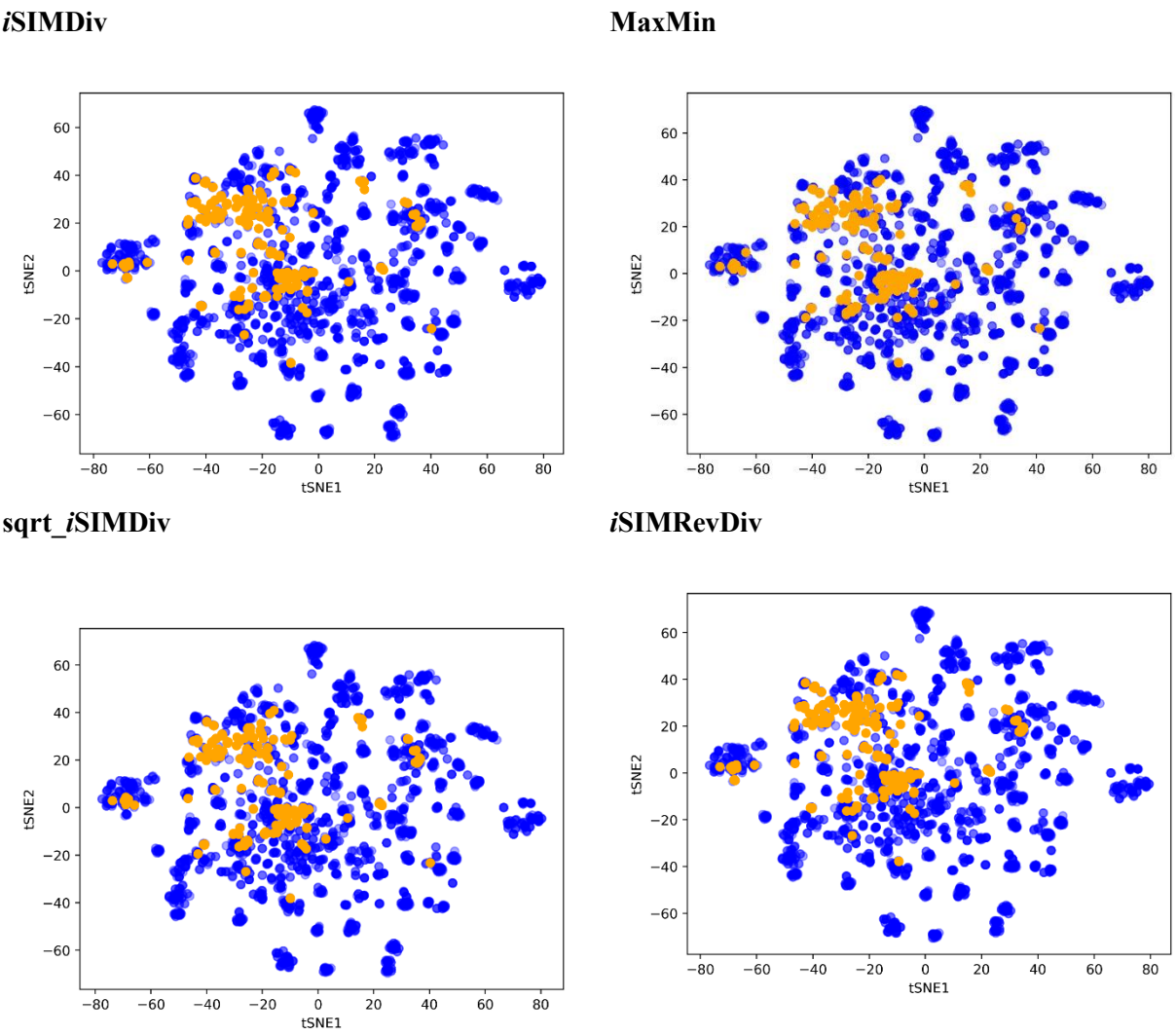


Figure 9: t-SNE plots for the CHEMBL214 dataset represented by RDKIT binary fingerprints by diversity selection methods. *i*SIM related methods use *i*RR similarity index as metric.

4.3 Clustering

As a final proof-of-principle demonstration of the versatility of the *i*SIM framework, we look at the clustering of molecular libraries. While there are many ways in which the notion of comparing multiple elements at the same time could be applied to clustering problems, perhaps the most natural one is in the context of hierarchical agglomerative (HA) algorithms. Note that *i*SIM can be used as a linkage criterion in the sense that at any given point we can choose to combine the two sets that produce the largest *i*SIM value for their union. In more mathematical terms, given sets $c_1, c_2, ..., c_K$, we combine clusters i, j such that: $i, j = \arg\max_{p,q} i\mathrm{SIM}\left(c_p \bigcup c_q\right)$.

This is the criterion that we used in Fig. 10 to cluster the CHEMBL214 (n = 3317) and CHEMBL2835 (n = 615) libraries (using iSM and MACCS fingerprints). We can also use the computed iSIM values to determine the optimum number of clusters in the data. If we follow the evolution of $i\text{SIM}_k$ (the iSIM of the cluster formed in the k^{th} step) we see that this quantity will tend to decrease with increasing k, but it will tend to reach some "stability" when an optimum separation of the data is achieved. In other words, we look for the largest value of k for which the quantity $\left|i\text{SIM}_{k+1} - i\text{SIM}_k\right|$ is as close to 0 as possible.

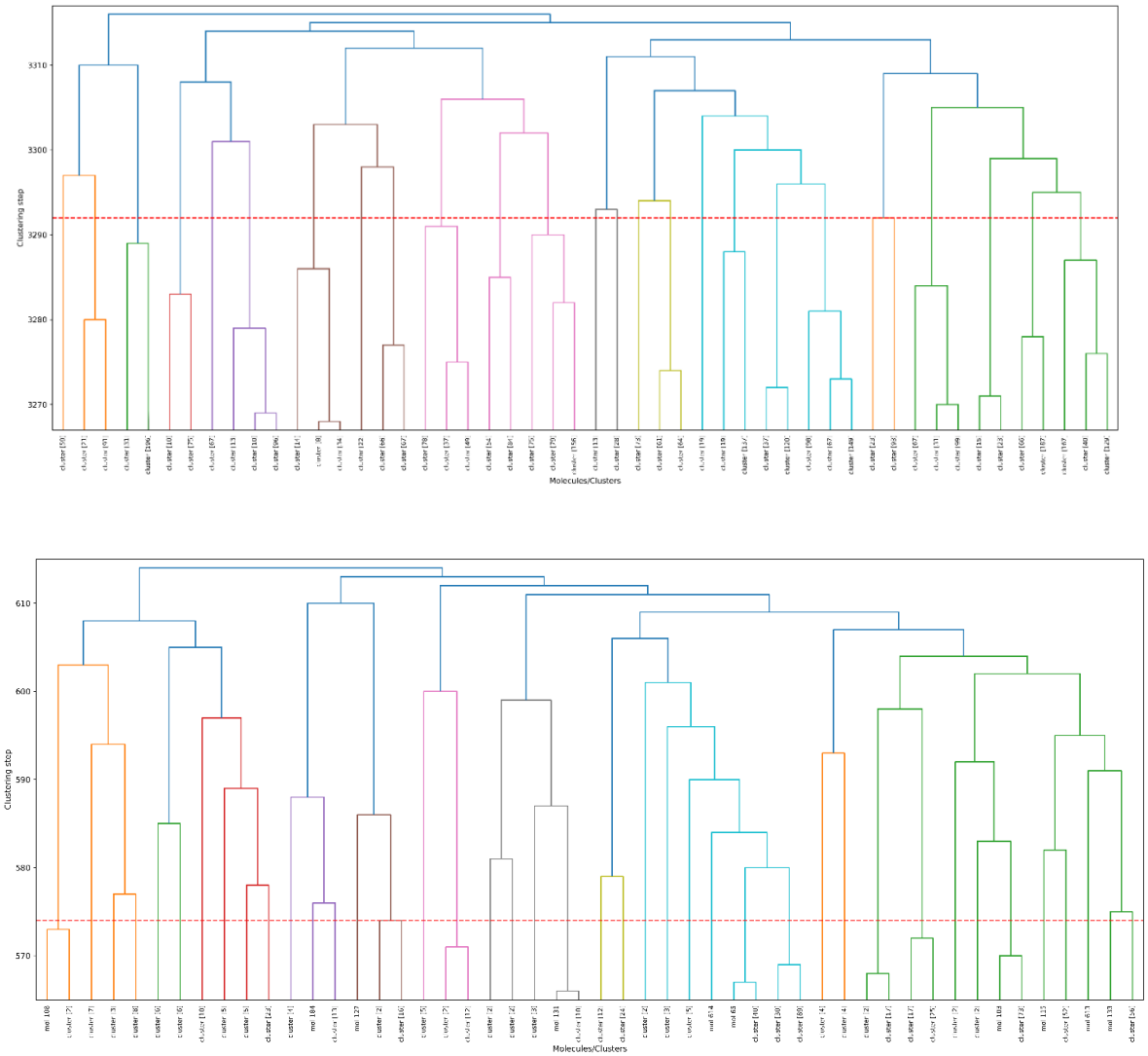
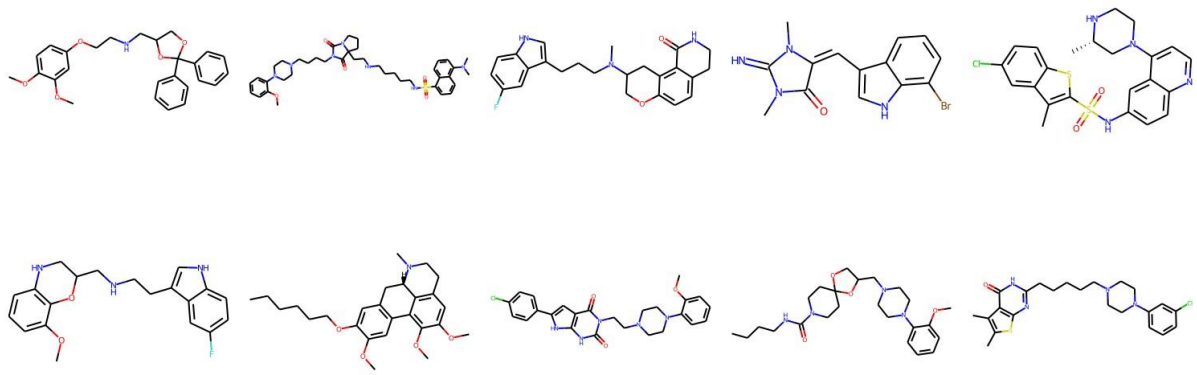


Figure 10: Dendrograms from hierarchical clustering of molecules in the CHEMBL214 (top) and CHEMBL2835 (down) libraries using iSM on MACCS fingerprints. Number of elements in each

cluster indicated in brackets. Coloring corresponds to the final 10 clusters. Dashed red line represents the optimal number of clusters cut-off (25 for CHEMBL214, 41 for CHEMBL2835).

Finally, clustering can be used to navigate through the molecular library, identifying representative structures associated to different basins in chemical space. For example, in Fig. 11 we show the medoids of the CHEMBL214 and CHEMBL2835 libraries in the case in which one selects 10 clusters in each of them. Note how our clustering is able to identify well-defined regions of chemical space that correspond to distinct scaffolds and functional groups. These structures, however, should not be mistaken for the most diverse structures in the original library. (A common practice in some fields tends to identify the cluster centroids with a diverse representation of the set.) For instance, if we calculate the *i*SIM for the set of medoids when one has a number of clusters equal to the 10% of the total number of points, we get 0.766 and 0.810 for CHEMBL214 and CHEMBL2835, respectively, which is far from a maximally-diverse set. That is, if the *i*SIMDiv and MaxMin tend to sample the data by increasing chemical diversity, the sampling through the medoids of the clusters offers a more "uniform" picture of the original set.



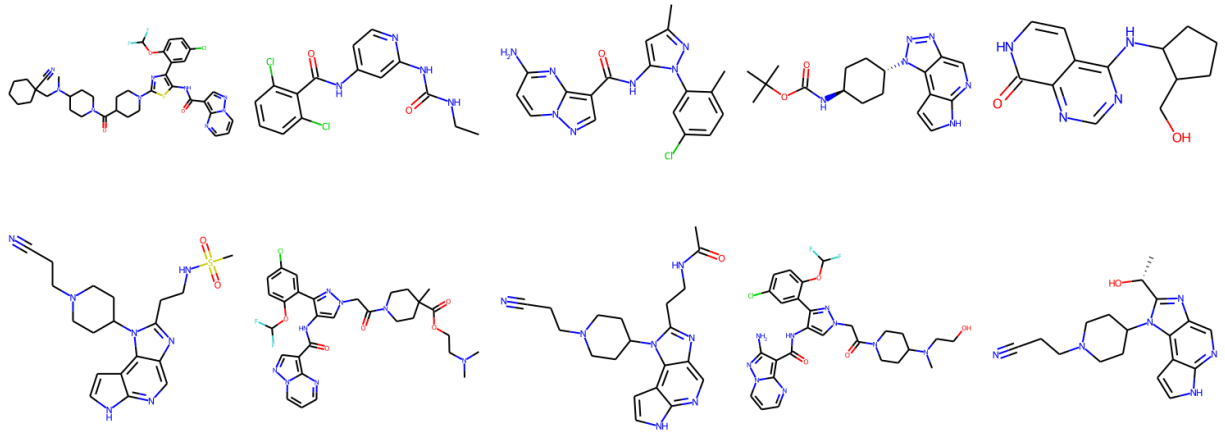


Figure 11: Medoids of each of the 10 colored clusters in the CHEMBL214 (top) and CHEMBL2835 (down) libraries using *i*SM on MACCS fingerprints.

5. Conclusions

*i*SIM has the ability of performing the comparisons of multiple objects at the same time, either if they are represented by binary fingerprints or real-value descriptors. The analytical mathematical operations behind *i*SIM, and the evidence from randomly generated data and real molecule libraries, show that the same exact value of average pairwise comparisons can be achieved for similarity indexes with the denominator equal to the length of the fingerprint, like RR and SM. In cases where the denominator is not equal to the length of the fingerprint, like JT, *i*SIM still provides an exceptional approximation to the pairwise comparison average, highlighting the robustness of the mediant approximation theorem. This brings the two key advantages of *i*SIM: the much more attractive linear scaling O(*N*) compared to the traditional pairwise indices, and the greater simplicity (no need to define coincidence thresholds and weight functions) compared to our previous extended similarity indices.

We showed that *i*SIM can be used to calculate complementary similarity of each of the molecules in the library and a ranking can be done to identify and visualize molecules as part of high-density or low-density regions. Different diversity selection methods using the proposed framework can be done depending on the necessity. *i*SIMDiv and *i*SIMRevDiv methods were developed to have two alternatives that output the same diversity results but differ in computing times depending on the percentage of data to select, adapting to the user's necessities. The *i*SIM metric can also be modified depending on the diversity selection is wanted to be globally or locally

coerced, which can be done taking the square root of the *i*SIM counters to select data that will have a lower maximum pairwise similarity. Remarkably, all the proposed diversity selection methods have the same or better quality as the commonly used MaxMin. Another application of our work is hierarchical clustering, as we can use *i*SIM as clustering objective function to be maximized when combining two molecules/clusters. The change in *i*SIM for the new cluster per clustering step can also be used as a metric to determine the optimal number of clusters. Overall, *i*SIM provides a flexible and easy-to-use framework to analyze molecular libraries, but that could be easily adapted to any problems that use comparisons between objects (metabolomics, MD simulations, etc.).

Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM150620.

References

(1) Fernández-de Gortari, E.; García-Jacas, C. R.; Martinez-Mayorga, K.; Medina-Franco, J. L. Database Fingerprint (DFP): An Approach to Represent Molecular Databases. *J Cheminform* **2017**, *9* (1), 9. https://doi.org/10.1186/s13321-017-0195-1.

(2) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley, 2000. https://doi.org/10.1002/9783527613106.

(3) Gugler, S.; Reiher, M. Quantum Chemical Roots of Machine-Learning Molecular Similarity Descriptors. *J Chem Theory Comput* **2022**, *18* (11), 6670–6689. https://doi.org/10.1021/acs.jctc.2c00718.

(4) Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytologist* **1912**, *11* (2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

(5) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science (1979)* **1960**, *132* (3434), 1115–1118. https://doi.org/10.1126/science.132.3434.1115.

(6) 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. https://doi.org/10.1021/ci300261r.

(7) Johnson, M. A.; Maggiora, G. M.; others. Concepts and Applications of Molecular Similarity. *(No Title)* **1990**.

(8) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J Chem Inf Model* **2012**, *52* (5), 1103–1113. https://doi.org/10.1021/ci300030u.

(9) Zoete, V.; Daina, A.; Bovigny, C.; Michielin, O. SwissSimilarity: A Web Tool for Low to Ultra High Throughput Ligand-Based Virtual Screening. *J Chem Inf Model* **2016**, *56* (8), 1399–1404. https://doi.org/10.1021/acs.jcim.6b00174.

(10) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discov Today* **2007**, *12* (5–6), 225–233. https://doi.org/10.1016/j.drudis.2007.01.011.

(11) Cohen, J. M.; Rice, J. W.; Lewandowski, T. A. Expanding the Toolbox: Hazard-Screening Methods and Tools for Identifying Safer Chemicals in Green Product Design. *ACS Sustain Chem Eng* **2018**, *6* (2), 1941–1950. https://doi.org/10.1021/acssuschemeng.7b03368.

(12) Posner, B. A.; Xi, H.; Mills, J. E. J. Enhanced HTS Hit Selection via a Local Hit Rate Analysis. *J Chem Inf Model* **2009**, *49* (10), 2202–2210. https://doi.org/10.1021/ci900113d.

(13) Ning, X.; Rangwala, H.; Karypis, G. Multi-Assay-Based Structure−Activity Relationship Models: Improving Structure−Activity Relationship Models by Incorporating Activity Information from Related Targets. *J Chem Inf Model* **2009**, *49* (11), 2444–2456. https://doi.org/10.1021/ci900182q.

(14) Helgee, E. A.; Carlsson, L.; Boyer, S.; Norinder, U. Evaluation of Quantitative Structure−Activity Relationship Modeling Strategies: Local and Global Models. *J Chem Inf Model* **2010**, *50* (4), 677–689. https://doi.org/10.1021/ci900471e.

(15) van Hoorn, W. P.; Bell, A. S. Searching Chemical Space with the Bayesian Idea Generator. *J Chem Inf Model* **2009**, *49* (10), 2211–2220. https://doi.org/10.1021/ci900072g.

(16) Buonfiglio, R.; Engkvist, O.; Várkonyi, P.; Henz, A.; Vikeved, E.; Backlund, A.; Kogej, T. Investigating Pharmacological Similarity by Charting Chemical Space. *J Chem Inf Model* **2015**, *55* (11), 2375–2390. https://doi.org/10.1021/acs.jcim.5b00375.

(17) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J Med Chem* **2012**, *55* (7), 2932–2942. https://doi.org/10.1021/jm201706b.

(18) Krein, M. P.; Sukumar, N. Exploration of the Topology of Chemical Spaces with Network Measures. *J Phys Chem A* **2011**, *115* (45), 12905–12918. https://doi.org/10.1021/jp204022u.

(19) Huggins, D. J.; Venkitaraman, A. R.; Spring, D. R. Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem Biol* **2011**, *6* (3), 208–217. https://doi.org/10.1021/cb100420r.

(20) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J Chem Inf Comput Sci* **1994**, *34* (5), 1094–1102. https://doi.org/10.1021/ci00021a011.

(21) Kovács, P.; Tran, F.; Hanbury, A.; Madsen, G. K. H. Similarity Clustering for Representative Sets of Inorganic Solids for Density Functional Testing. *J Chem Theory Comput* **2022**, *18* (1), 441–447. https://doi.org/10.1021/acs.jctc.1c00536.

(22) 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†. *J Cheminform* **2021**, *13* (1), 32. https://doi.org/10.1186/s13321-021-00505-3.

(23) 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* (1), 33. https://doi.org/10.1186/s13321-021-00504-4.

(24) Verhellen, J. Graph-Based Molecular Pareto Optimisation. *Chem Sci* **2022**, *13* (25), 7526–7535. https://doi.org/10.1039/D2SC00821A.

(25) Chang, L.; Perez, A.; Miranda-Quintana, R. A. Improving the Analysis of Biological Ensembles through Extended Similarity Measures. *Physical Chemistry Chemical Physics* **2022**, *24* (1), 444–451. https://doi.org/10.1039/D1CP04019G.

(26) 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* (14), 3415–3425. https://doi.org/10.1021/acs.jcim.2c00433.

(27) 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* (9), 2186–2201. https://doi.org/10.1021/acs.jcim.1c01013.

(28) 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* (13), 4042–4055. https://doi.org/10.1021/acs.jcim.3c00520.

(29) 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* (6), 2100285. https://doi.org/10.1002/minf.202100285.

(30) 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* (7). https://doi.org/10.1002/minf.202300056.

(31) 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* (3), 157–173. https://doi.org/10.1007/s10822-022-00444-7.

(32) Redžepović, I.; Furtula, B. Chemical Similarity of Molecules with Physiological Response. *Mol Divers* **2023**, *27* (4), 1603–1612. https://doi.org/10.1007/s11030-022-10514-5.

(33) 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* (17), 6333. https://doi.org/10.3390/molecules28176333.

(34) Russell, P. F.; Rao, T. R.; others. On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras. *J Malar Inst India* **1940**, *3* (1).

(35) Sokal, R. R.; Michener, C. D. University of Kansas. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas science bulletin. University of Kansas* **1958**.

(36) Guthery, S. B. *A Motif of Mathematics*; Docent Press, 2011.

(37) Tou, E. R. The Farey Sequence: From Fractions to Fractals. *Math Horizons* **2017**, *24* (3), 8–11. https://doi.org/10.4169/mathhorizons.24.3.8.

(38) 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. https://doi.org/10.1021/acs.jcim.2c01073.

(39) RDKit: Open-Source Cheminformatics. Https://Www.Rdkit.Org.

(40) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(41) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J Chem Doc* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.

(42) Parreño, F.; Álvarez-Valdés, R.; Martí, R. Measuring Diversity. A Review and an Empirical Analysis. *Eur J Oper Res* **2021**, *289* (2), 515–532. https://doi.org/10.1016/j.ejor.2020.07.053.

(43) Kuo, C.; Glover, F.; Dhir, K. S. Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming*. *Decision Sciences* **1993**, *24* (6), 1171–1185. https://doi.org/10.1111/j.1540-5915.1993.tb00509.x.

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

(45) Ivosev, G.; Burton, L.; Bonner, R. Dimensionality Reduction and Visualization in Principal Component Analysis. *Anal Chem* **2008**, *80* (13), 4933–4944. https://doi.org/10.1021/ac800110w.

(46) der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *Journal of machine learning research* **2008**, *9* (11).