For the special issue in honor of Gerry Maggiora

Extended Continuous Similarity Indices–Theory and Application for QSAR Descriptor Selection

Anita Rácz[1‡], Timothy B. Dunn[2‡], Dávid Bajusz[3], Taewon D. Kim[2], Ramón Alain Miranda-Quintana[2,4*], Károly Héberger[1*]

[1] Plasma Chemistry Research Group, Research Centre for Natural Sciences, Magyar tudósok krt. 2, 1117 Budapest, Hungary

[2] Department of Chemistry, University of Florida, Gainesville, FL 32611, USA

[3] Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Magyar tudósok krt. 2, 1117 Budapest, Hungary

[4] Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA

* Corresponding authors.

R. A. Miranda-Quintana (quintana@chem.ufl.edu) K. Héberger (heberger.karoly@ttk.hu)

‡ These authors have contributed equally.

For the special issue in honor of Gerry Maggiora

Abstract

Extended (or *n*-ary) similarity indices have been recently proposed to extend the comparative analysis of binary strings. Going beyond the traditional notion of pairwise comparisons, these novel indices allow comparing any number of objects at the same time. This results in a remarkable efficiency gain with respect to other approaches, since now we can compare N molecules in O(N) instead of the common quadratic O(N^2) timescale. This favorable scaling has motivated the application of these indices to diversity selection, clustering, phylogenetic analysis, chemical space visualization, and post-processing of molecular dynamics simulations. However, the current formulation of the *n*-ary indices is limited to vectors with binary or categorical inputs. Here, we present the further generalization of this formalism so it can be applied to numerical data, *i.e*. to vectors with continuous components. We discuss several ways to achieve this extension and present their analytical properties. As a practical example, we apply this formalism to the problem of feature selection in QSAR and prove that the extended continuous similarity indices provide a convenient way to discern between several sets of descriptors.

1. Introduction

Similarity and distance measures are cornerstones of a vast range of methodologies in the fields of molecular modeling, drug design and cheminformatics [1, 2]. In some common examples, their binary implementations are used to quantify the similarity of binary molecular fingerprints (with the Tanimoto coefficient unquestionably being the most popular one) [3], while their continuous implementations constitute the basics of clustering algorithms [4]. The applications of molecular similarity (as expressed by pairwise similarity calculations between binary fingerprints) in ligand-based virtual screening were thoroughly explored by the groups of Jürgen Bajorath [5, 6],Peter Willett [7, 8], and many others, with a large body of works from the latter group dedicated to data fusion practices [9, 10]. Binary similarity measures from many sub-fields were collected by Todeschini and colleagues [11], and further analyzed by our group to select ideal candidates for specific applications in metabolomics [12] and molecular design [13] studies. We have also shown that two similarity measures can be consistent with each other in a surprisingly high percentage of cases, even when they are poorly correlated [14].

Recently, we have introduced several methodological frameworks to extend the usage of similarity measures beyond the common cases mentioned above. Most importantly, we have demonstrated that the mathematical expansion of the core concepts of similarity measures can provide a way to quantify the similarity of an arbitrary number of objects at the same time. We first showed this on binary (molecular) fingerprints: the resulting similarity measures were termed extended (or *n*-ary) similarity measures [15]. They employ the core concept of similarity and dissimilarity counters, which have replaced the *a*, *b*, *c* and *d* terms that are commonly applied in the well-known, pairwise definitions of the similarity measures to describe the number of bit positions where two fingerprints have co-occurring one (*a*) or zero (*d*) bits, or a one bit that is exclusive to either of the fingerprints (*b* and *c*). In our framework, the 1-similarity, 0-similarity, and dissimilarity counters express the number of bit positions where the number of co-occurring one (or zero) bits is above, or below, a pre-defined coincidence threshold, respectively. For pairwise comparisons, these generalizations naturally revert to the well-known definitions of the classical, pairwise similarity measures. We have shown that the new methodology is not only computationally efficient, scaling as *O(n)* with the number of compared objects *n*, but it can be successfully applied for tasks such as diversity selection, clustering, as well as the visualization of large sections of chemical space [16–19]. A further generalization involved the extension of this framework to allow for more than two possible characters ($t = 2$) in an object (vector), opening the possibility to apply the extended similarity measures in bioinformatics, for the comparison of nucleotide ($t = 4$) or protein sequences ($t = 20$) [20]. We have termed these, even further generalized definitions extended many-item, or (t, n) similarity measures, to distinguish them from the above-mentioned, extended binary, or (2, n) similarity measures.

In this study, realizing the potential of further possible generalizations to extended similarity measures, we introduce extended continuous, or ($\mathbb{R}$, n) similarity measures, to provide a way to compare an arbitrary number of vectors with real values. This generalization will employ the same concepts as mentioned above, with novel formulas for determining the number of similarity and dissimilarity counters. As we will show in section 2.1, there are at least three ways to generalize the extended indices so they can handle continuous-valued vectors. All of these variants were implemented and compared in the Results section.

To demonstrate the utility of the new class of similarity metrics, we use them in descriptor selection. Quantitative Structure-Activity Relationships (QSAR) are one of the earliest and most important concepts in molecular design [21]. QSAR realizes a linear or non-linear regression between numerical descriptors of compound structure and experimentally determined or calculated physicochemical parameters and bioactivity. While multiple linear regression (MLR) has ruled the QSAR field for a long time as a classical regression algorithm, the last decades have seen the emergence of several new algorithms, many of them based on Machine Learning [22], including some interesting examples that are adapted from other fields [23]. In the meantime, new families of molecular descriptors were introduced: contemporary descriptor calculator software (such as Dragon) can generate thousands of continuous (and, discrete and binary) descriptors. Also, public bioactivity repositories such as ChEMBL [24] or PubChem Bioassay [25] allow access to large molecular datasets for the more thorough training of QSAR models. The increasing number of descriptors, more complex algorithms and larger training datasets are factors that drive up the computational demand of QSAR modeling: to mitigate this, it is common practice to apply one or more descriptor (feature) selection algorithms to reduce the input dataset of the modeling algorithm by pre-selecting the most meaningful descriptors to work with [26]. In turn, descriptor selection has its own computational cost as a limiting factor: less sophisticated (less demanding) algorithms will sample the descriptor space only superficially, while more sophisticated options, such as genetic algorithms, will be more time-consuming [27]. A thorough review of descriptor selection methods is given by Goodarzi *et al*. [28]. While we do not necessarily gain prediction accuracy from descriptor selection [29], a smaller number of descriptors will convey a significant speedup to QSAR modeling in most cases, especially if the descriptor selection approach is not laborious either.

Here, we apply the new class of extended continuous similarity metrics in a simple descriptor selection scenario. Using a large and relevant ADME-related dataset of cytochrome P450 (CYP) 2C9 inhibitors (actives) and inactive species, we calculate group-wise similarities based on several descriptor families to find the best ones at discriminating the group of actives from the total dataset. Therefore, we provide a novel, simple variable selection tool for QSAR/QSPR analyses. This idea can constitute the basis of more complex descriptor selection approaches with a more thorough exploration of the descriptor space to yield the set of descriptors that can optimally distinguish the actives from the total set of ligands.

2. Materials and methods

2.1 Extended Continuous Similarity Indices

There are several ways to extend the domain of definition of our *n*-ary indices such that they can be applied to quantify the similarity of an arbitrary number of vectors with continuous components. The strategies that we will consider here all start from a common point: scaling the input data between 0 and 1. In other words, we will work with vectors:

$$\begin{aligned} &V_1 = \left(x_{11}, x_{12}, ..., x_{1m}\right) \\ &V_2 = \left(x_{21}, x_{22}, ..., x_{2m}\right) \\ &... \\ &V_n = \left(x_{n1}, x_{n2}, ..., x_{nm}\right) \end{aligned} \tag{1}$$

where: $\forall i, j : 0 \le x_{ij} \le 1$.

Variant 1

The first way of quantifying the similarity of these vectors is to see how different the components are from the average of their column (*e.g.*, how distant is a feature from its average value). Hence, the first step is to calculate the vector of column-wise averages:

$$A = \left(a_1, a_2, ..., a_m\right) \tag{2}$$

where:

$$a_i = \frac{1}{n}\sum_{j=1}^{j=n} x_{ji} \tag{3}$$

We now have to subtract this average from the corresponding normalized elements (e.g., centering) and find the absolute of these differences:

$$\begin{aligned} &\left(\left|x_{11} - a_1\right|, \left|x_{12} - a_2\right|, ..., \left|x_{1m} - a_m\right|\right) \\ &\left(\left|x_{21} - a_1\right|, \left|x_{22} - a_2\right|, ..., \left|x_{2m} - a_m\right|\right) \\ &... \\ &\left(\left|x_{n1} - a_1\right|, \left|x_{n2} - a_2\right|, ..., \left|x_{nm} - a_m\right|\right) \end{aligned} \tag{4}$$

The next step is to sum all these differences across a given column and form a new vector with the results:

$$S=\left(s_1,s_2,...,s_m\right) \tag{5}$$

where:

$$s_i=\sum_{j=1}^{j=n}\left|x_{ji}-a_i\right| \tag{6}$$

Now, analogously to the original binary extended similarity indices, we need to define a new vector of "coincidences":

$$D=\left(\delta_1,\delta_2,...,\delta_m\right) \tag{7}$$

where:

$$\delta_i=\sqrt{n^2-2ns_i} \tag{8}$$

This is directly related to our previous works [15, 16]. The key insight is that if the i^{th} column of the normalized data has k 1's and $n-k$ 0's, then $\delta_i=\Delta_{n(k)}=\left|2k-n\right|$, that is, the indicator we use in our original paper to quantify the coincidence. The main difference is that the simpler $\left|2k-n\right|$ expression is only defined over strings of 1's and 0's, while Eq. (8) is defined over real numbers in the [0, 1] interval.

Having established this connection, we can now follow a similar route as in the binary case. First, we defined a coincidence threshold, γ, and if $\delta_i>\gamma$ then we use $f_s\left(\delta_i\right)$ to estimate the similarity, and if $\delta_i\le\gamma$ then we use $f_d\left(\delta_i\right)$ to calculate the dissimilarity. By analogy of the 1- and 0-similarities of the binary case, we can distinguish between "high-content similarities" (where the column average is higher) and "low-content similarities" (where the column average is lower):

If *n* is odd:

δ_i will be a "high-content similarity" if $\delta_i>\gamma$ and $a_i\ge\dfrac{\left(n-n\bmod 2\right)\big/2+n\bmod 2}{n}$, moreover,

δ_i will be a "low-content similarity" if $\delta_i>\gamma$ and $a_i\le\dfrac{\left(n-n\bmod 2\right)\big/2}{n}$.

If *n* is even:

δ_i will be a "high-content similarity" if $\delta_i>\gamma$ and $a_i\ge 0.5$, and δ_i will be a "low-content similarity" if $\delta_i>\gamma$ and $a_i<0.5$.

This procedure extends many of the notions of the binary case in a natural way. However, while the notion of quantifying similarity by measuring the distance to the mean is a common one, we should be aware that a high similarity in this case implies that the components of vector S (Eq. (5)) can be very close to zero. This means that if we do not use a high enough coincidence threshold, we have the risk of identifying all the columns as corresponding to low-content similarities. The problem with this is that several indices will be ill-defined (*e.g.* they will involve division by zero), because their denominators only include high-content similarities and dissimilarity counters. For instance, taking the more common case of the traditional binary similarity indices (and the standard convention that a, $b + c$, and d represent the number of common "on" bits, the mismatches, and common number of "off" bits, respectively), this situation will be equivalent to saying that $a = b + c = 0$. Hence, indices without d in their denominator (like Jaccard-Tanimoto, Baroni-Urbani-Buser, *etc*., see Table 1) could not be calculated. Once again, this would not be a problem if we select a large enough coincidence threshold. Nonetheless, the potential issues that could be caused by this prevalence of 0-similarities motivate us to explore another variant of extended continuous indices.

Variant 2

As noted above, the raison d'être for this new approach is to increase the number of high-content (as opposed to low-content) similarities. Here we also measure similarity according to the distance from the mean, so we also need to calculate the column-average vector (Eq. (2)), and we need to form the matrix given in Eq. (4). The key difference is that now we carry out an additional transformation of this matrix before calculating the similarities, namely, we work instead with a new matrix defined by:

$$\begin{gathered}\left(1-\left|x_{11}-a_1\right|,1-\left|x_{12}-a_2\right|,...,1-\left|x_{1m}-a_m\right|\right)\\ \left(1-\left|x_{21}-a_1\right|,1-\left|x_{22}-a_2\right|,...,1-\left|x_{2m}-a_m\right|\right)\\ ...\\ \left(1-\left|x_{n1}-a_1\right|,1-\left|x_{n2}-a_2\right|,...,1-\left|x_{nm}-a_m\right|\right)\end{gathered} \qquad (9)$$

The rationale behind this is quite simple: in Eq. (4) a high similarity will correspond to a very small element, while in Eq. (9) a high similarity will correspond to an element that is close to 1.

From here we proceed as usual, first calculating the vector of column sums:

$$\overline{S} = \left(\overline{s}_1, \overline{s}_2, ..., \overline{s}_m\right) \tag{10}$$

with:

$$\overline{s}_i = \sum_{j=1}^{j=n}\left(1 - \left|x_{ji} - a_i\right|\right) = n - \sum_{j=1}^{j=n}\left|x_{ji} - a_i\right| \tag{11}$$

In this case, we will follow a simpler recipe to determine the character of the counters:

$$\begin{aligned} &2\overline{s}_i - n > \gamma \rightarrow \text{high - content} \\ &n - 2\overline{s}_i > \gamma \rightarrow \text{low - content} \\ &\left|2\overline{s}_i - n\right| \leq \gamma \rightarrow \text{dissimilarity} \end{aligned} \tag{12}$$

From purely theoretical arguments, we should expect this variant to be better than the previous one, if anything because it will lead to indices that can be calculated regardless of the coincidence threshold selected. Nonetheless, it still measures similarity taking the mean as a reference, so it seems desirable to explore yet another option, which measures similarity directly from the normalized values.

Variant 3

Starting from the scaled data (Eq. (1)), we only need to calculate the sums along each column:

$$\sigma_i = \sum_j x_{ji} \tag{13}$$

Then, we use these numbers to assign the type of counters, analogously to what we did in variant 2:

$$\begin{aligned} &2\sigma_i - n > \gamma \rightarrow \text{high - content} \\ &n - 2\sigma_i > \gamma \rightarrow \text{low - content} \\ &\left|2\sigma_i - n\right| \leq \gamma \rightarrow \text{dissimilarity} \end{aligned} \tag{14}$$

Notice that this method is essentially equivalent to the original binary case (the analogy is clear if we notice that if all the x_{ij} are either 0 or 1, then $\sigma_i = k_i$ in our original notation).

This variant has two potential advantages: its simplicity, and the ability of looking at the data from a different point of view (since it does not rely on the calculation of the average). However, the latter can bring a potential problem: by not referring to an average and using the raw normalized values to directly calculate the similarity, this variant should be more prone to depend on the scaling (normalization) procedure. This can lead to a pathological behavior, since a normalization method that gives very small values for the x_{ij} will lead to an input that suffers from the overly abundance of low-content similarities of variant 1. This will once again imply that we will not be able to calculate all indices, unless we use a very high coincidence threshold.

Having reached this point in any of the previous variants, we can easily classify each column as contributing to the high-content, low-content similarity, or dissimilarity between the compared objects. Notice that, as in the binary case, the minimum possible value for γ in all these cases is also equal to $n \bmod 2$. Once we have classified all the counters, the process to calculate the similarity indices is exactly the same as in the binary case (see Table 1 for the list of all the expressions). Notice that here we can also decide whether to include or not weight functions in the denominator of the indices, leading to the weighted (w) or non-weighted (nw) flavors, respectively.

The formulae of *n*-ary continuous indices are enumerated in Table 1 (notice that the cJa and cJa0 differ in that in the latter we do not differentiate between high-content and low-content similarities). That is, notice how the original formulation of some of these indices (e.g., the asymmetric indices, like Gle, Ja, etc.) distinguished between the high- and low-content similarities, assigning a more important role to the latter. As we showed in our original paper, we can generalize these indices by replacing every occurrence of the high-content similarity by the sum of the high- and low-content similarities, which leads to more symmetrical expressions (and novel ways to quantify similarity).

Table 1. Formulae and notations of the extended continuous similarity indices.

Additive indices				
Label	**Type**[a]	**Notation**[b]	**Name**	**Equation**
cAC	cAC_hc	cACw	continuous Austin-Colwell	$s_{cAC(hc\_wd)} = \frac{2}{\pi}\arcsin\sqrt{\frac{\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}+\sum_{lc-s} f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}}}$
		cACnw		$s_{cAC(hc\_d)} = \frac{2}{\pi}\arcsin\sqrt{\frac{\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}+\sum_{lc-s} f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)}+\sum_d C_{n(k)}}}$
cBUB	cBUB_hc	cBUBw	continuous Baroni-Urbani-Buser	$s_{cBUB(hc\_wd)} = \frac{\sqrt{\left[\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}\right]\left[\sum_{lc-s} f_s(\Delta_{n(k)})C_{n(k)}\right]}+\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}}{\left\{\sqrt{\left[\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}\right]\left[\sum_{lc-s} f_s(\Delta_{n(k)})C_{n(k)}\right]}+\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right\}}$

		cBUBnw		$s_{cBUB(hc\_d)} = \frac{\sqrt{[\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)}][\sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}]} + \sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{\left\{\sqrt{[\sum_{hc-s} C_{n(k)}][\sum_{lc-s} C_{n(k)}]} + \sum_{hs-s} C_{n(k)} + \sum_d C_{n(k)}\right\}}$
cCT1	cCT1_hc	cCT1w	continuous Consoni-Todeschini (1)	$s_{cCT1(hc\_wd)} = \frac{\ln(1+\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)} + \sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)})}{\ln(1+\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)})}$
		cCT1nw		$s_{cCT1(hc\_d)} = \frac{\ln(1+\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)} + \sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)})}{\ln(1+\sum_s C_{n(k)} + \sum_d C_{n(k)})}$
cCT2	cCT2_hc	cCT2w	continuous Consoni-Todeschini (2)	$s_{cCT2(hc\_wd)} = \frac{\ln(1+\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}) - \ln(1+\sum_d f_d(\Delta_{n(k)}) C_{n(k)})}{\ln(1+\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)})}$
		cCT2nw		$s_{cCT2(hc\_d)} = \frac{\ln(1+\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}) - \ln(1+\sum_d f_d(\Delta_{n(k)}) C_{n(k)})}{\ln(1+\sum_s C_{n(k)} + \sum_d C_{n(k)})}$
cFai	cFai_hc	cFaiw	continuous Faith	$s_{cFai(hc\_wd)} = \frac{\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)} + 0.5 \sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}$
		cFainw		$s_{cFai(hc\_d)} = \frac{\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)} + 0.5 \sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s C_{n(k)} + \sum_d C_{n(k)}}$
		cGKnw		$s_{cGK(hc\_d)} = \frac{2\min(\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)}, \sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}) - \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}{2\min(\sum_{hc-s} C_{n(k)}, \sum_{lc-s} C_{n(k)}) + \sum_d C_{n(k)}}$
		cHDnw		$s_{cHD(hc\_d)} = \frac{1}{2}\left(\frac{\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_{hc-s} C_{n(k)} + \sum_d C_{n(k)}} + \frac{\sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_{lc-s} C_{n(k)} + \sum_d C_{n(k)}}\right)$
cRT	cRT_hc	cRTw	continuous Rogers-Tanimoto	$s_{cRT(hc\_wd)} = \frac{\sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + 2 \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}$
		cRTnw		$s_{cRT(hc\_d)} = \frac{\sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s C_{n(k)} + 2 \sum_d C_{n(k)}}$
cRG	cRG_hc	cRGw	continuous Rogot-Goldberg	$s_{cRG(hc\_wd)} = \frac{\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{2 \sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}} + \frac{\sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{2 \sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}$
		cRGnw		$s_{cRG(hc\_d)} = \frac{\sum_{hc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{2 \sum_{hc-s} C_{n(k)} + \sum_d C_{n(k)}} + \frac{\sum_{lc-s} f_s(\Delta_{n(k)}) C_{n(k)}}{2 \sum_{lc-s} C_{n(k)} + \sum_d C_{n(k)}}$
cSM	cSM_hc	cSMw	continuous Simple matching, Sokal-Michener	$s_{cSM(hc\_wd)} = \frac{\sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}$
		cSMnw		$s_{cSM(hc\_d)} = \frac{\sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{\sum_s C_{n(k)} + \sum_d C_{n(k)}}$
cSS2	cSS2_hc	cSS2w	continuous Sokal-Sneath (2)	$s_{cSS2(hc\_wd)} = \frac{2 \sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{2 \sum_s f_s(\Delta_{n(k)}) C_{n(k)} + \sum_d f_d(\Delta_{n(k)}) C_{n(k)}}$
		cSS2nw		$s_{cSS2(hc\_wd)} = \frac{2 \sum_s f_s(\Delta_{n(k)}) C_{n(k)}}{2 \sum_s C_{n(k)} + \sum_d C_{n(k)}}$

Asymmetric indices				
Label	**Type**	**Notation**	**Name**	**Equation**
cCT3	cCT3_hc	cCT3w	continuous Consoni-Todeschini (3)	$s_{cCT3(hc\_wd)} = \frac{\ln\left(1+\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}\right)}$
		cCT3nw		$s_{cCT3(hc\_d)} = \frac{\ln\left(1+\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_s C_{n(k)}+\sum_d C_{n(k)}\right)}$
	cCT3_lc	cCT3lcw		$s_{cCT3(lc\_wd)} = \frac{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}\right)}$
		cCT3lcnw		$s_{cCT3(lc\_d)} = \frac{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_s C_{n(k)}+\sum_d C_{n(k)}\right)}$
cCT4	cCT4_hc	cCT4w	continuous Consoni-Todeschini (4)	$s_{cCT4(hc\_wd)} = \frac{\ln\left(1+\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}\right)}$
		cCT4nw		$s_{cCT4(hc\_d)} = \frac{\ln\left(1+\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_{hc-s} C_{n(k)}+\sum_d C_{n(k)}\right)}$
	cCT4_lc	cCT4lcw		$s_{cCT4(lc\_wd)} = \frac{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}\right)}$
		cCT4lcnw		$s_{cCT4(lc\_d)} = \frac{\ln\left(1+\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}\right)}{\ln\left(1+\sum_s C_{n(k)}+\sum_d C_{n(k)}\right)}$
cGle	cGle_hc	cGlew	continuous Gleason	$s_{cGle(hc\_wd)} = \frac{2\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{2\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}}$
		cGlenw		$s_{cGle(hc\_d)} = \frac{2\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{2\sum_{hc-s} C_{n(k)}+\sum_d C_{n(k)}}$
	cGle_lc	cGlelcw		$s_{cGle(lc\_wd)} = \frac{2\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{2\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}}$
		cGlelcnw		$s_{cGle(lc\_d)} = \frac{2\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{2\sum_s C_{n(k)}+\sum_d C_{n(k)}}$
cJa	cJa_hc	cJaw	continuous Jaccard	$s_{cJa(hc\_wd)} = \frac{3\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{3\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}}$
		cJanw		$s_{cJa(hc\_d)} = \frac{3\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{3\sum_{hc-s} C_{n(k)}+\sum_d C_{n(k)}}$
	cJa_lc	cJalcw		$s_{cJa(lc\_wd)} = \frac{3\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{3\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}}$
		cJalcnw		$s_{cJa(lc\_d)} = \frac{3\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{3\sum_s C_{n(k)}+\sum_d C_{n(k)}}$
cRR	cRR_hc	cRRw	continuous Russel-Rao	$s_{cRR(hc\_wd)} = \frac{\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}}$
		cRRnw		$s_{cRR(hc\_d)} = \frac{\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{\sum_s C_{n(k)}+\sum_d C_{n(k)}}$
	cRR_lc	cRRlcw		$s_{cRR(lc\_wd)} = \frac{\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}+\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k)}}$
		cRRlcnw		$s_{cRR(lc\_d)} = \frac{\sum_s f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{\sum_s C_{n(k)}+\sum_d C_{n(k)}}$
cSS1	cSS1_hc	cSSw	continuous Sokal-Sneath (1)	$s_{cSS1(hc\_wd)} = \frac{\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)} + 2\sum_d f_d\left(\Delta_{n(k)}\right)C_{n(k}}$
		cSSnw		$s_{cSS1(hc\_d)} = \frac{\sum_{hc-s} f_s\left(\Delta_{n(k)}\right)C_{n(k)}}{\sum_{hc-s} C_{n(k)}+2\sum_d C_{n(k)}}$

	cSS1_lc	cSSlcw		$s_{cSS1(lc\_wd)} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)} + 2\sum_d f_d(\Delta_{n(k)})C_{n(k)}}$
		cSSlcnw		$s_{cSS1(lc\_d)} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)} + 2\sum_d C_{n(k)}}$
cJT	cJT_hc	cJTw	continuous Jaccard-Tanimoto	$s_{cJT(hc\_wd)} = \frac{\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}}{\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)} + \sum_d f_d(\Delta_{n(k)})C_{n(k)}}$
		cJTnw		$s_{cJT(hc\_d)} = \frac{\sum_{hc-s} f_s(\Delta_{n(k)})C_{n(k)}}{\sum_{hc-s} C_{n(k)} + \sum_d C_{n(k)}}$
	cJT_lc	cJTlcw		$s_{cJT(lc\_wd)} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)} + \sum_d f_d(\Delta_{n(k)})C_{n(k)}}$
		cJTlcnw		$s_{cJT(lc\_d)} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)} + \sum_d C_{n(k)}}$

[a] *hc*: high-content similarity; *lc*: low content similarity

[b] *w*: weighted, *nw*: non-weighted

2.2 Dataset and descriptors

A large dataset of cytochrome P450 (CYP) 2C9 ligands from Pubchem Bioassay (AID 1851) was used as a case study to highlight the applicability of the *n*-ary indices for continuous variables [30]. Cytochrome P450 enzymes are important mediators of drug metabolism, therefore they are widely studied in the field of QSAR/QSPR: many compounds were evaluated against this enzyme family and they are recurring targets in machine learning classification studies as well [31]. In total, 12161 molecules were applied after the data curation and preparation step. The dataset contained 4016 inhibitors with a potency of 10 µM or better (actives) and 8145 inactive species. Dragon 7 software was used for the calculation of 2D descriptors [32, 33]. **Table 2** shows the 19 different 2D descriptor groups, which were calculated in the study (the groups are predefined by the applied software). We have applied the same numbering system for the descriptor sets as it was used in the Dragon software. The excluded numbers (13-20, 26-27) are connected to 3D descriptors. Highly correlated variables (above 0.997) and constant variables were excluded from the sets [34]. The details and descriptions of the different descriptor sets can be found in the DRAGON software manual.

Table 2. The applied 2D descriptor packages with the number of descriptors

Dragon number	2D Descriptor	Size
1	Constitutional	45
2	Ring descriptors	32

3	Topological indices	72
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	50
7	2D matrix-based descriptors	436
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P-VSA-like descriptors	55
11	ETA indices	21
12	Edge adjacency indices	324
21	Functional groups count	128
22	Atom-centred fragments	98
23	Atom-type E-state indices	88
24	CATS 2D	145
25	2D Atom Pairs	746
28	Molecular properties	14
29	Drug-like indices	12

2.3 Statistical analysis

First, we had to normalize the descriptor sets before the calculation of the continuous *n*-ary indices. Two different methods were used for this step: rank transformation and mean scaling. The equations are the following:

$$y_{mean}\left(x_i\right)=\frac{x_i-\min X}{\max X-\min X} \qquad (15)$$

and

$$y_{rank}\left(x_i\right)=\frac{\operatorname{rank}\left(x_i\right)-1}{\max \operatorname{rank}\left(X\right)-1} \qquad (16)$$

After the normalization of the dataset, 16 different continuous *n*-ary indices were calculated for the 19 descriptor sets. We have calculated the *n*-ary indices for the active and inactive groups, as well as the total dataset, corresponding to three different levels of similarity. We assume the active group to be more coherent – based on earlier examples from our research group, where a small number of descriptors was sufficient to define simple multicriteria optimization rules for kinase [35] and GPCR ligands [36], distinguishing them from a larger set of commercially available compounds. By comparison, the inactive set should display a lesser degree of similarity, while the total dataset (containing both the active and inactive sets) should be the most diffuse, *i.e.* less similar overall. A further level of comparison was introduced by calculating the absolute differences between the similarities of the active group vs. the total dataset (from here on, denoted as |active-total| values). Here, a larger difference corresponds to more discriminatory power of the given descriptor set and similarity metric. The datasets – with the 16 continuous *n*-ary metrics in the columns and the descriptor sets in the rows – were evaluated and compared with factorial ANOVA and the multi criteria decision making tool, sum of ranking differences (SRD) [37]. The SRD procedure is not a simple extension of the Spearman footrule to equal numbers (ties) in the input matrix [38], but contains two validation steps: i) comparison of ranks with random numbers (CRRN) [39], and ii) cross-validation [40]. It is a generally applicable multicriteria decision making tool [41], whose applications were demonstrated in a wide range of fields from food chemistry [42] to medical applications [43], as well as politics [44] and sports [45]. The sum of ranking differences (SRDs) is calculated as the city block (Manhattan) distance (d_{kj}) between the rank values of the gold standard and the rank values of the original data. In the calculation process, always the columns of the dataset are compared to the reference column. Sum of ranking differences (SRD) helped to compare and rank the descriptor sets and the *n*-ary continuous indices. SRD was carried out separately for the similarities of the active and inactive sets, as well as the absolute differences between the actives and the total dataset (|active-total| values). In all cases, the maximum values were used as the reference column. When the novel similarity metrics were compared, the dataset contained those in the columns and the descriptor sets in the rows, while in the comparison of the descriptor sets, the mentioned dataset was transposed. It is important to note, that in every SRD calculation, the variables with smaller SRD values are the better ones (these are closer to the reference). The scaled (between 0 and 100) and cross-validated SRD values were applied for the final evaluation by ANOVA.

Factorial ANOVA analysis is dedicated to compare the group averages according to the different factors. For the original datasets (containing the 16 similarity metrics for the active, inactive and the complete dataset of molecules), we have used several factors such as the *n*-ary indices (16 levels), the molecular descriptor sets (19), the different threshold limits (0.05 – 0.95 fraction of the total size of the set, with steps of 0.05) and the applied groups of molecules (active, inactive, total). For the final comparison of the descriptors based on their SRD values, we have used i) the descriptor sets, and ii) the actives, inactives and |active-total| groups as factors.

3. Results and discussion

The calculated *n*-ary continuous indices were used for descriptor (variable) selection in the case study of a large dataset of CYP 2C9 inhibitors and inactives. Moreover, the 16 different continuous similarity measures (weighted and non-weighted versions) were compared and ranked to find the most optimal ones for the task. We have calculated the similarity measures for three different sets of the dataset: actives, inactives and the complete dataset (total). As the optimal coincidence threshold limit (γ) is case-dependent, in each variant (1, 2, 3) of the similarity calculation, a coincidence threshold analysis was carried out to select the best threshold limit. In the next step, the most important descriptor sets, and the optimal similarity measures have been selected based on the continuous similarity values for the "active", "inactive" and "total" groups. In the optimal situation, the best similarity measures should return bigger similarity values for the group of active ligands, somewhat smaller similarities for the inactive ligands, and the lowest similarity for the most diffuse "total" group. An additional parameter, the absolute difference between the similarity of "active" and "total" groups was calculated to select and rank the examined descriptors and similarity indices with the SRD analysis, based on their ability to distinguish the active group within the total dataset. The whole process was carried out for the three different continuous *n*-ary similarity calculation variants; thus, we could compare their efficiencies for the task and finally select the most applicable one. Figure 1 shows the mentioned workflow of the study in an illustrative way.

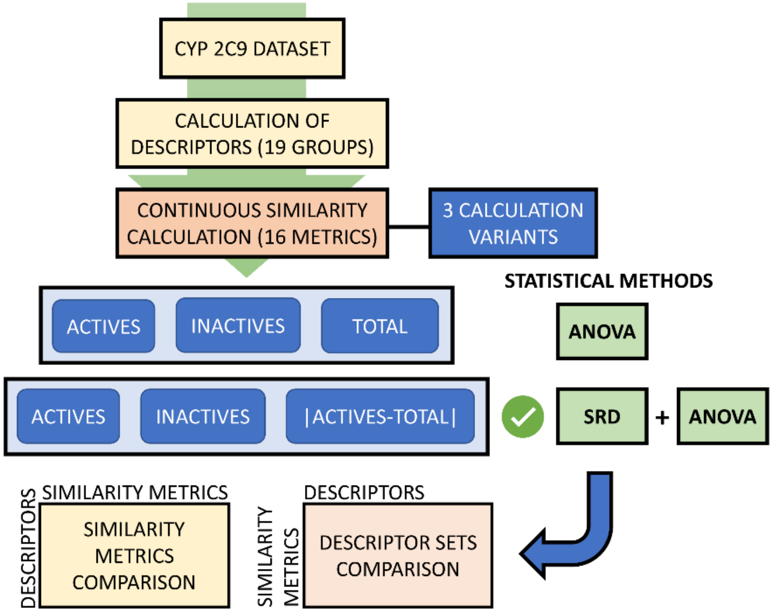

Figure 1. The applied workflow of the study, emphasizing the most important aspects of the analysis.

3.1 Variant 1

As we have two different normalization procedures for the descriptors: rank and mean normalization; as well as weighted and non-weighted versions for the continuous similarity calculations, the coincidence thresholds were compared for all the four cases. **Figure 2** shows the dependence of the similarity values on the applied threshold limits in the *x* axis. The similarity of the group of molecules: “actives”, “inactives” and “total” are compared in the factorial ANOVA plot. It is very clear that the weighted and non-weighted versions have the same shape or pattern, but the range of the similarity values are different. Naturally, we can say that in the optimal case, the groups are separated, especially the actives from the total. For the non-weighted version, this separation is slightly better based on the covered similarity range, while the use of rank normalization of the descriptors clearly gives us better results. We have selected 0.70 as the coincidence threshold limit for the further SRD analysis, based on the non-weighted and rank scaled version of the plots. In this case the active, inactive, and total groups are the farthest from each other.

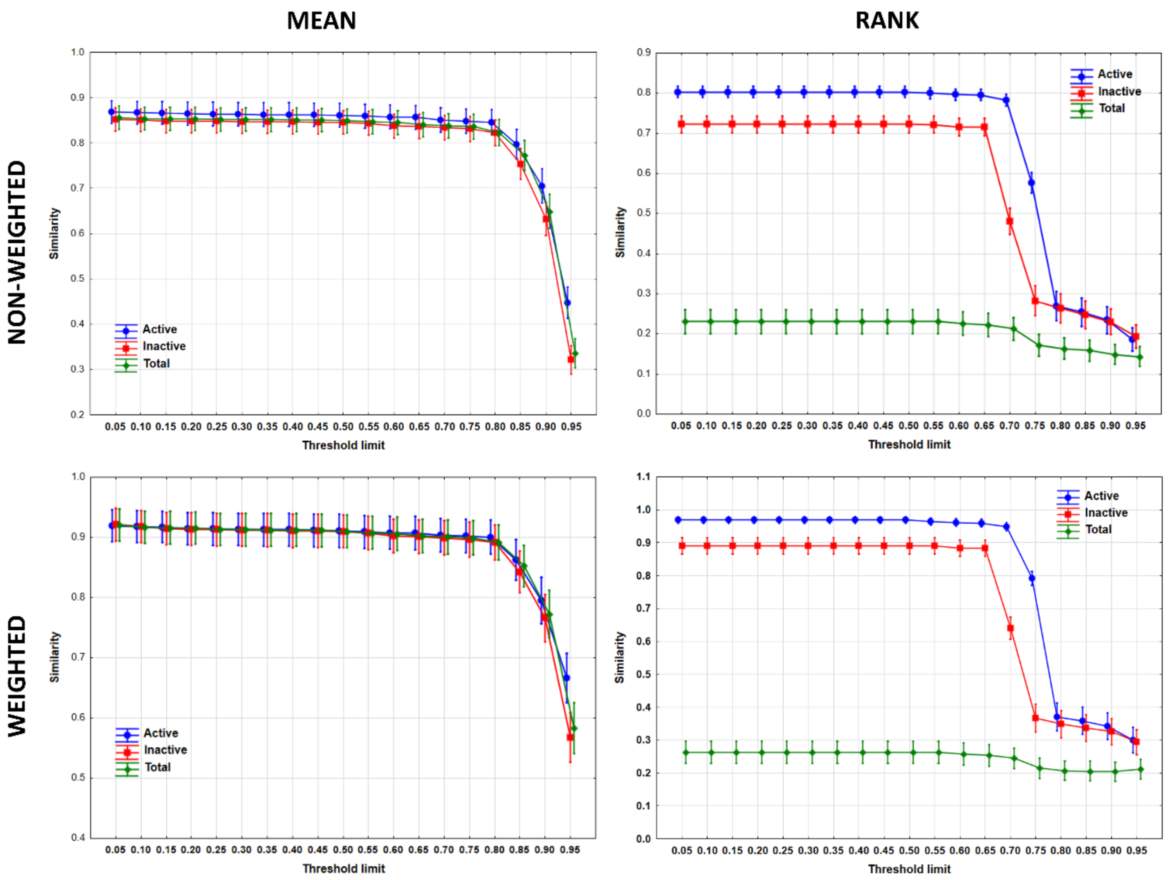

Figure 2. Threshold dependence for similarity values in the weighted and non-weighted variants of the continuous indices. First row: non-weighted version; second row: weighted version; first column: mean normalization; second column: rank normalization. Active molecules: blue line; Inactive molecules: red line; total dataset: green line. Similarity values are plotted against the different coincidence threshold limits.

The continuous *n*-ary similarity measures were also compared with factorial ANOVA. Molecule groups were selected as the second factor in this case as well. **Figure S1** in the supplementary information shows the result of the factorial ANOVA, where the similarity values are plotted against the different similarity measures. The same pattern can be noticed as in the case of the threshold limit selection. Again, the rank normalized version coupled with the non-weighted similarity calculation provides a much better result. Since it would be hard to select the most proper measures based on the ANOVA plot, the rank-normalized and non-weighted results were used for the SRD analysis for further evaluation. **Figure 3** shows the result of the SRD analysis, where the scaled SRD values were used for factorial ANOVA, instead of the original similarity values. The SRD analyses were carried out for the active, inactive and the additionally calculated |active-total| similarity values separately. This latter parameter is relevant because the bigger the difference between the similarities of the active group and the total dataset, the better the final model could be. In the SRD analyses, the maximum values were used as the reference. It means that those similarity measures, which had higher values for the different groups of molecules (or the difference between the actives and total set), are ranked better. In other words, the best similarity indices should be the most sensitive in finding the similarities amongst the actives and providing bigger differences in similarity between the actives and the total dataset. The result of the three SRD analyses were merged for the final ANOVA analysis. It is justified because the SRD values are scaled to the same range in each case. The smaller the SRD values, the better the similarity measure. We must make note of a difference between the use of the original similarity values and the calculated SRD values in the ANOVA analysis. For the original similarity values, all of the results with the various coincidence thresholds were used. For the SRD analysis, only the optimal threshold limit with non-weighted similarities and rank normalization was used, based on the conclusions from the ANOVA of the similarity values.

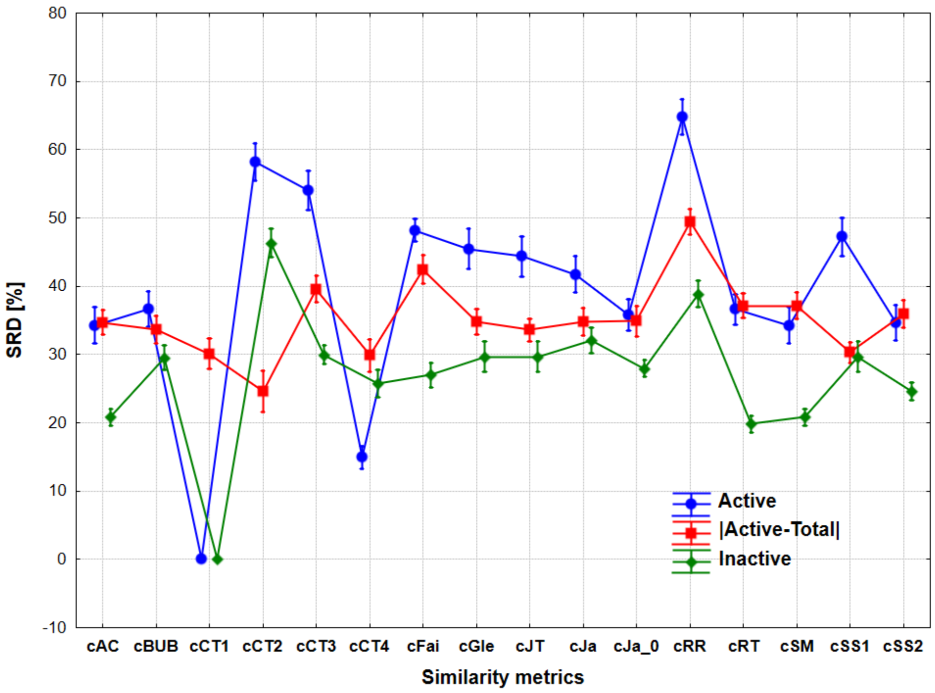

Figure 3. SRD values [%] to the gold standard for the active and inactive sets, and |active-total| values. The coincidence threshold was determined by variant 1. The continuous extended, individual similarity measures are plotted on the X axis (for their formulae, see Table 1). The similarity of the active group is marked with a blue line, the similarity of the inactive group is marked with a green line and the absolute difference between the active and the total group is marked with a red line.

The active, inactive and the |active-total| versions had different behaviors. As cCT1 has the best ranks in the active and inactive cases and still good SRD values for the |active-total| case, we can recommend that measure as the best one for variant 1. The cRR similarity measure can be considered the worst one based on the SRD values of the three cases.

Similarly, the molecular descriptor sets have been compared based on the original similarity values and the SRD values as well. The factorial ANOVA of the original values can be found in the Supplementary information as Figure S2. The original similarity values showed that the first half of the descriptor sets have better discrimination between the similarities of the groups. The SRD analysis of the active, inactive and |active-total| similarity sets provided extra information about the best descriptor sets. **Figure 4** shows the factorial ANOVA result of the three cases. Descriptor sets 3 and 8 have the smallest SRD values in all the three cases together, although the results are not consistent: where the difference between the active and total is bigger (thus the SRD value is smaller), the inactive group has a worse result. (Descriptor set 3 contains the topological indices, while descriptor set 8 contains the 2D autocorrelation descriptors.) Many descriptor sets cannot rank the |active-total| better than random, *e.g.*, No. 1 and 21-25. Some of the descriptor sets evaluate the active and |active-total| very similarly, *e.g.*, No. 3, 8, 11 and 28. The actives are found to be the most similar according to sets No. 1, 21, 22, 23.

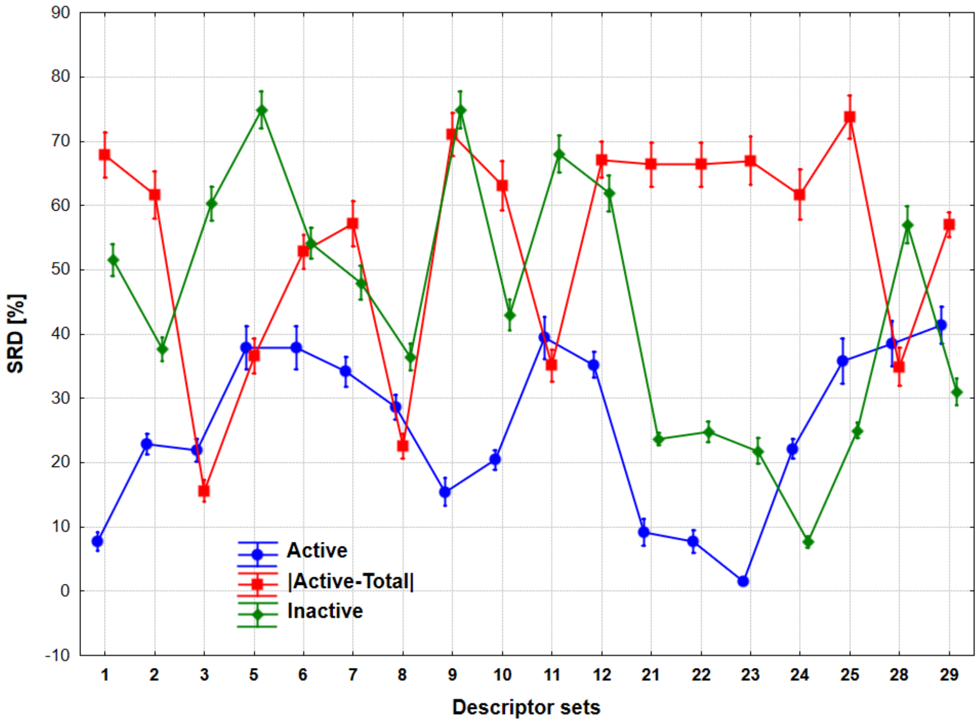

Figure 4. Factorial ANOVA of the SRD values [%] as a function of descriptor sets (Table 1) in the case of variant 1. The different descriptor sets are plotted in the X axis. The similarity of the active group is marked with a blue line, the similarity of the inactive group is marked with a green line and the absolute difference between the active and the total group is marked with a red line.

3.2 Variant 2

In the case of variant 2, the same process was carried out as in the case of variant 1. First, we have compared the coincidence thresholds with the different pretreatments (rank/mean, weighted/non-weighted). **Figure 5** shows the factorial ANOVA of the original similarity values. With the mean transformation, the curve has a long plateau part, then it drops quickly, while in the case of rank normalization, the curve has an inflexion point. In this point, at 0.5-0.55, the similarity of the three groups (active, inactive, total) are the farthest: large similarity for the active set, and small similarity for the inactive and total sets. Thus, we have selected 0.50 as the coincidence threshold limit for the further SRD analysis.

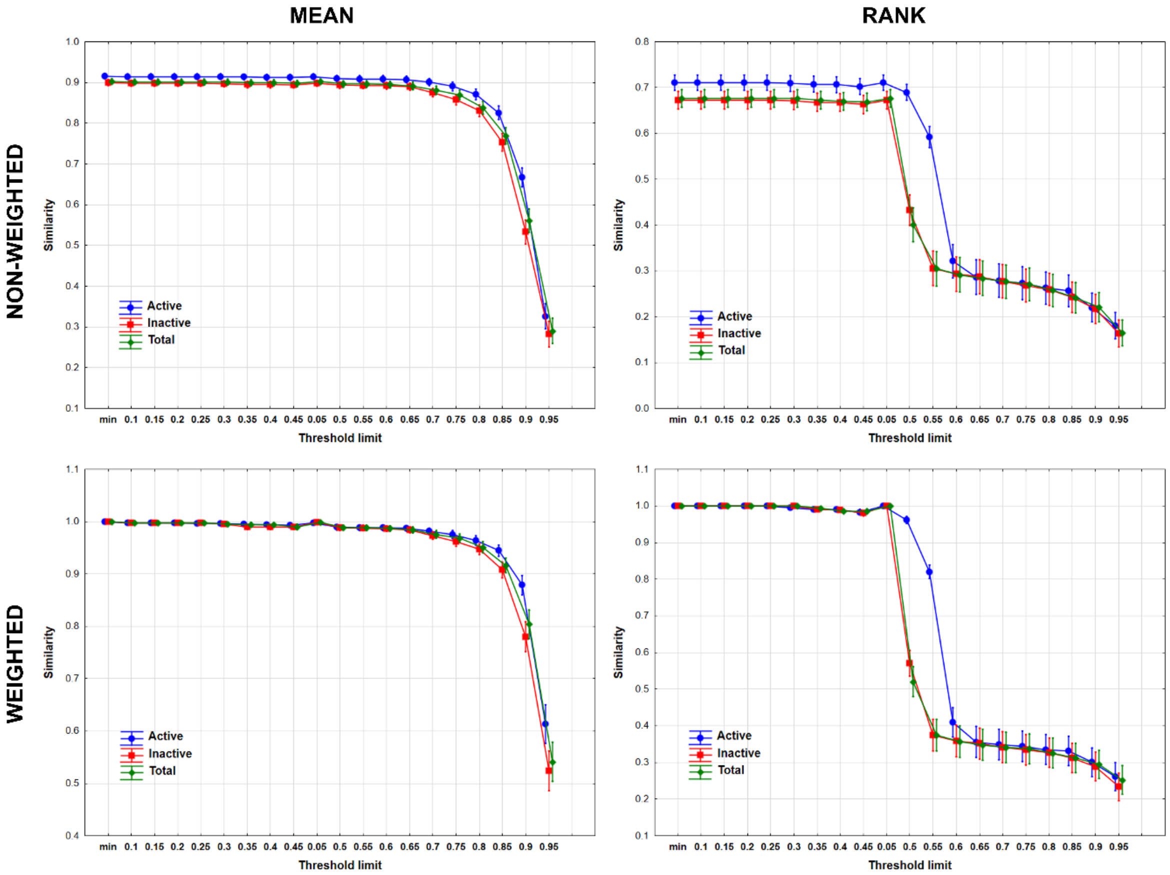

Figure 5. The factorial ANOVA for the four different scenarios of the similarity calculations in the case of variant 2. First row: non-weighted version; second row: weighted version; first column: mean normalization; second column: rank normalization. Active molecules: blue line; inactive molecules: red line; total dataset: green line. Similarity values are plotted against the different coincidence threshold limits.

The continuous *n*-ary similarity measures are compared in Figure S3 in the Supplementary information in the four pretreatment scenarios (with all the threshold limits), but we have also tested how the optimal threshold limit affects the result. In **Figure 6A**, where we use only the 0.50 threshold limit data, the lines of the different groups are further from each other compared to Figure S3. However, it would be still hard to find the most optimal similarity measure based on this figure, because most of them are in the same range and the lines are at about the same distance from each other. In the optimal case, the similarity metric should provide higher similarity within the group of actives and smaller similarity within the total dataset: this holds for all metrics. **Figure 6B** shows that SRD values can more easily select the most prominent continuous measures. As in this case, the smaller the SRD value, the better the applied metric, here we can highlight the cCT1, cCT3 and cCT4 metrics, because they have the smallest SRD values consistently in all the three cases (active, inactive, |active-total|).

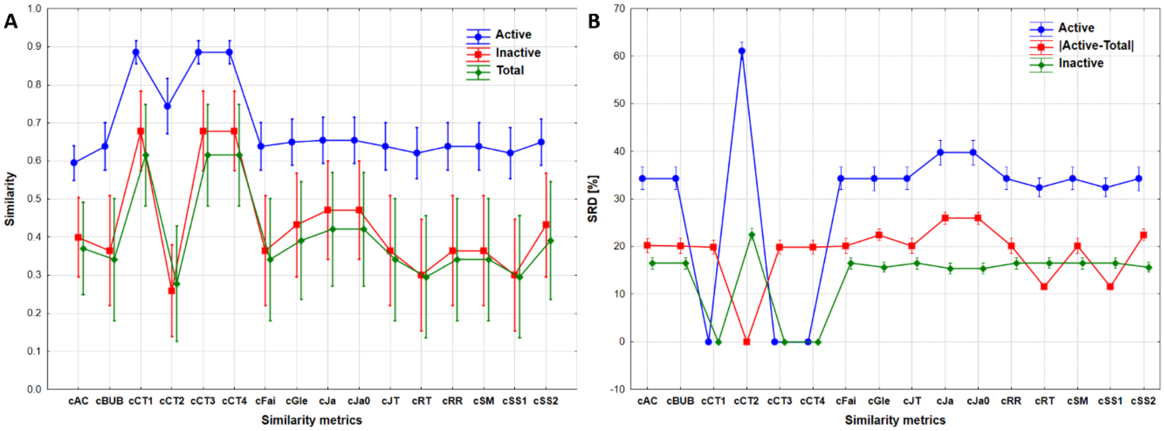

Figure 6. The factorial ANOVA of the original similarity values (A) and the scaled SRD values (B) with the continuous similarity measures and molecule groups as factors. (For the formulae of the similarity metrics, see Table 1.) The similarity of the active group is marked with blue line, the similarity of the inactive group is marked with green line and the absolute difference between the active and the total group is marked with red line.

In the case of descriptor set selection, the same analyses have been made. Figure S4 in the Supplementary information shows the factorial ANOVA with the descriptor sets and molecule groups as factors for the four preprocessing scenarios. **Figure 7** presents the results focusing only to the optimal threshold limit 0.50 based on the original data and the scaled SRD values.

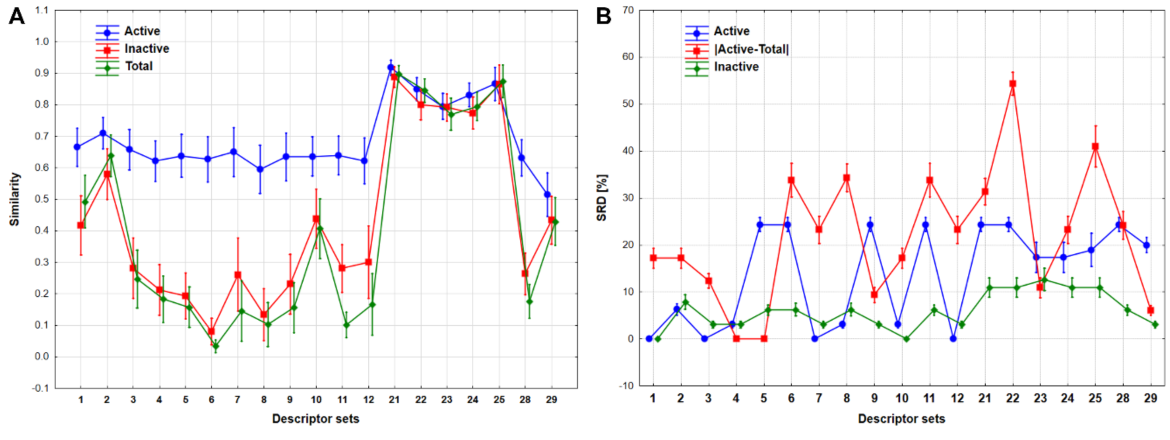

Figure 7. The factorial ANOVA of the original similarity values (A) and the scaled SRD values (B) with the descriptor sets and molecule groups as factors. The similarity of the active group is marked with a blue line, the similarity of the inactive group is marked with a green line and the absolute difference between the active and the total group is marked with a red line.

The line of the active group is further away from the others with the use of the optimal threshold limit, but we can safely say (based on **Figure 7A)** that descriptor sets 21-25 have no discriminative power between the active and the other groups, which is not advantageous for their use in QSAR models. In the descriptor selection phase, those descriptor sets can be more important, which are capable to find the active molecules that are more similar to each other than the whole dataset. Based on **Figure 7B**, descriptor sets 3 and 4 have remarkably good SRD values in all the three cases (active, inactive and |active-total|). These descriptor sets are the topological indices (3) and the walk and path counts (4), which together contain 118 descriptors. Moreover, all the mean SRD values of the descriptor sets tend to be closer to zero compared to the variant 1 results, which is a favorable feature in the case of variant 2.

3.3 Variant 3

In the case of variant 3, the calculation is much simpler and less robust than the other two variants. This resulted in different plots compared to the others, such as in the case of the optimal threshold limit determination. **Figure 8** shows that mean normalization is not able to select any threshold limit, but in the case of rank normalization the group similarities are separated better, but only in the beginning of the plot. The “typical” curve shape that we had in the other two cases, is now missing. In the case of mean, a linearly or slightly convex decreasing curve can be seen, while in the case of rank transformation the curve plateaus off at the end. Therefore, we have decided to use the “min” threshold limit, which is the minimum coincidence threshold possible, calculated as *n*mod2. In this case, based on the rank transformed data, the three group similarities can be separated better. The SRD analyses were carried out with the “min” coincidence threshold data.

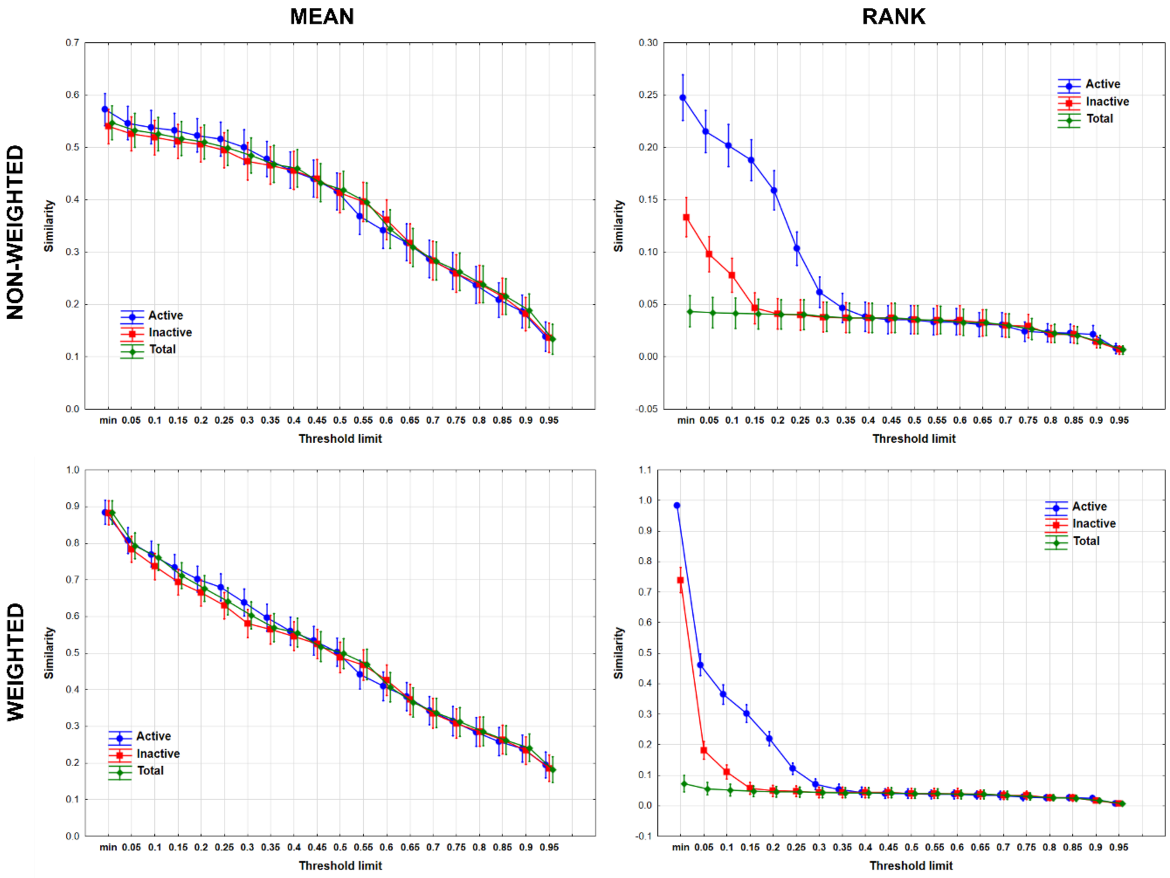

Figure 8. The factorial ANOVA for the four different scenarios of the similarity calculations in the case of variant 3. First row: non-weighted version; second row: weighted version; first column: mean normalization; second column: rank normalization. Active molecules: blue line; Inactive molecules: red line; total dataset: green line. Similarity values are plotted against the different coincidence threshold limits.

The continuous *n*-ary similarity measures were also compared. The Supplementary information contains the factorial ANOVA for the four cases together as Figure S5. Here we show the result of the factorial ANOVA based on the scaled SRD values in **Figure 9**. It is still true, that the original similarity values cannot be used for the selection of the most optimal similarity measure. SRD analysis with the selected coincidence threshold limit ("min") data could extend the results and provide a more consistent picture about the comparison of the indices. **Figure 9** clearly shows that cCT1 can be selected as the most optimal continuous similarity measure. On the other hand, all the cCT$_i$ measures are somewhat better than the others, especially in returning higher similarities for the active set.

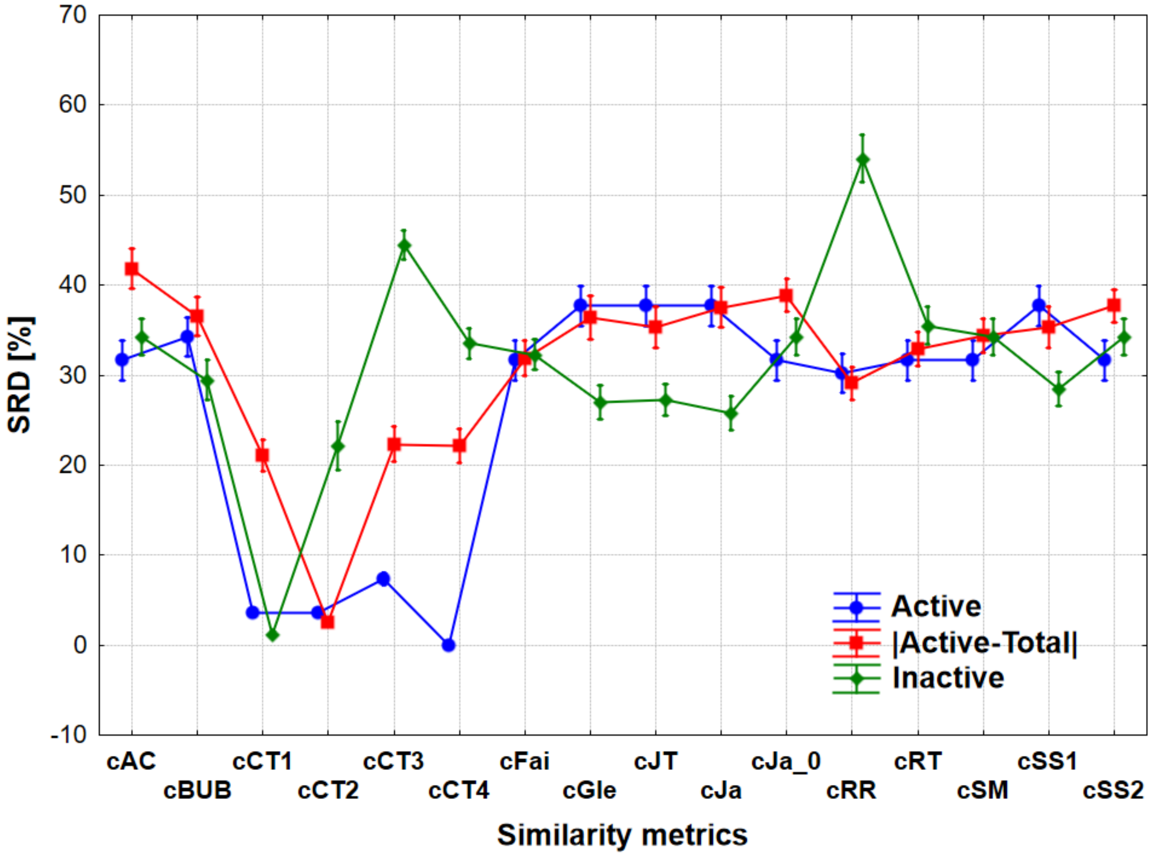

Figure 9. The result of factorial ANOVA based on the SRD values [%] in the case of variant 3. The continuous similarity measures are plotted in the X axis (for their formulae, see Table 1). The similarity of the active group is marked with a blue line, the similarity of the inactive group is marked with a green line and the absolute difference between the active and the total group is marked with a red line.

The molecular descriptor sets were compared with the same workflow as in the case of variant 1 and 2. The results of the factorial ANOVA based on the original similarity values with the four different pretreatment scenarios are shown in the Supplementary information as Figure S6. Finally, the results of the factorial ANOVA based on the scaled SRD values are shown in **Figure 10**. In Figure S6, the similarity values calculated with this variant have no discriminative power. Even the SRD analysis could not select the best sets properly, because it was not sensitive enough. However, the inactive molecules can be ranked worse for almost all descriptor sets, with two definite but diverse exceptions (No. 12 and 29) As the variant 3 is a simpler and less robust version of the calculation, it could not provide a definite selection for the task.

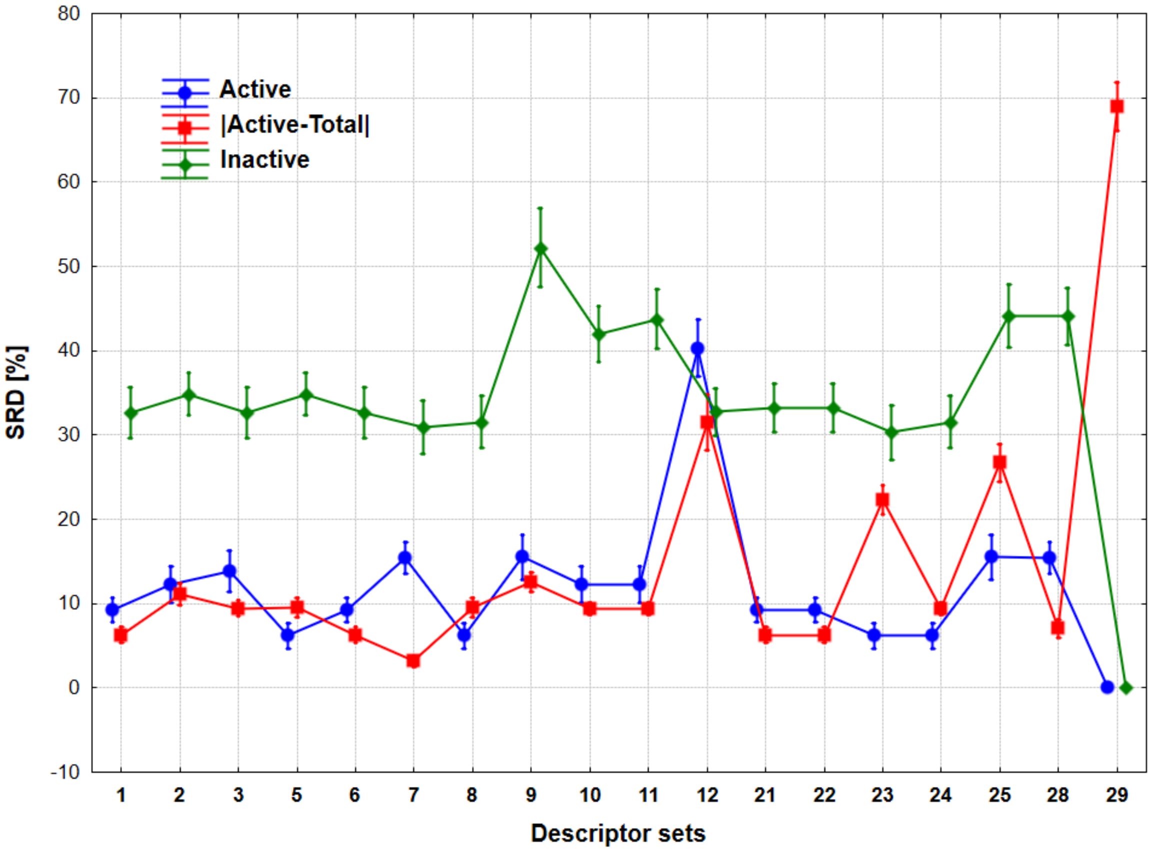

Figure 10. Factorial ANOVA of SRD values [%] in the case of variant 3. The different descriptor sets are plotted in the X axis. The similarity of the active group is marked with blue line, the similarity of the inactive group is marked with green line and the absolute difference between the active and the total group is marked with red line.

Conclusion

We have generalized our recently introduced *n*-ary similarity indices such that they can be now applied to vectors with continuous components. This greatly expands the domain of applicability of the extended similarity framework, which can now be applied to the selection of molecular descriptors for QSAR/QSPR modeling. We proposed three ways to calculate the extended (or *n*-ary) continuous similarity indices, depending on the way of defining the similarity between the different elements to be compared. We also considered how different factors impact the characteristics of these indices, including the way of normalizing the data, and the inclusion or omission of weight factors in the denominators of the similarity indices. A case study of a publicly available dataset of CYP 2C9 inhibitors (actives) and inactives was used for comparing the various possible similarity metrics and coincidence thresholds (cutoff values to determine whether a certain variable/descriptor contributes to the similarity or dissimilarity of the given dataset).

The first variant for the calculation of extended continuous similarities is based on how different the elements of an array (in this case, a column vector) are from their average. This is an intuitive measure that can be easily related to the original *n*-ary formalism for binary fingerprints, but it has some important disadvantages. For instance, indices without low-content similarity counters in the denominator could be ill-defined for relatively small values of the coincidence threshold. Overall, for the descriptor selection case study, cCT1 showed the best ranks in the active and inactive cases and still good SRD values for the |active-total| case, so we can recommend this measure as the best one.

The second variant attempts to remedy the issues of Variant 1 by converting the low-content similarities to high-content similarities, but it also quantifies similarity by measuring how distant are the different components to the corresponding column average. Here, the cCT1, cCT3 and cCT4 indices have the smallest SRD values consistently in all the three cases (active, inactive, |active-total|). Descriptor sets 21-25 have no discriminative power between the active and the other groups. On the other hand, descriptor sets 3 (topological indices) and 4 (walk path counts) have remarkably good SRD values in all the three cases (together, these sets contain 118 descriptors).

Finally, the third variant takes a different approach to measuring the similarity between the elements of a set, by directly assessing how related are the components in each column of the normalized matrix (just like it is done to calculate the counters in the binary case). Now, cCT1 can once again be selected as the most optimal continuous similarity measure. More generally, all the cCTi measures (*i*=1, …4) are somewhat better than the others, especially in returning higher similarities for the active set. However, this variant places almost all descriptor sets in the same position, so it is not as clear to give a precise indication of the best conditions for this option.

Overall, this work bridges the missing gap in the applicability of extended similarity indices, which can now handle more general types of input. While we have shown here different ways in which one can handle continuous inputs, Variant 2 seems to be the more robust of these options, mainly because the original similarity indices used in cheminformatics tend to favor high-content-similarities (1-similarity in the binary case). This means that using this variant we will have access to a more diverse toolkit of extended similarity measures. We have shown that the extended similarity metrics with the use of ANOVA and SRD methods can be successfully applied for the selection of continuous molecular descriptor sets, but this formalism opens the way for other applications, including the analysis of three-dimensional structures and conformations of biological ensembles, since we could directly represent them via their coordinates in real space. We are currently exploring this line of research, by studying the different conformations obtained via Molecular Dynamics simulations. These results will be presented elsewhere in due course.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Data availability statement

Data is available from the authors upon reasonable request. Python code for calculating the extended similarity metrics is freely available at: https://github.com/ramirandaq/MultipleComparisons

For the special issue in honor of Gerry Maggiora

Funding

- National Research, Development and Innovation Office of Hungary (OTKA), contracts K_20 134260 (KH) and PD_20 134416 (AR)
- University of Florida: startup grant: RAMQ
- Hungarian Academy of Sciences: János Bolyai Research Scholarship: DB
- Ministry for Innovation and Technology of Hungary: ÚNKP-21-5 New National Excellence Program: DB

References

1. Bajusz D, Rácz A, Héberger K (2017) Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In: Chackalamannil S, Rotella DP, Ward SE (eds) Comprehensive Medicinal Chemistry III. Elsevier, Oxford, pp 329–378
2. Bender A, Glen RC (2004) Molecular similarity: A key technique in molecular informatics. Org. Biomol. Chem. 2:3204–3218
3. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7:20. https://doi.org/10.1186/s13321-015-0069-3
4. Saxena A, Prasad M, Gupta A, et al (2017) A review of clustering techniques and developments. Neurocomputing 267:664–681. https://doi.org/10.1016/J.NEUCOM.2017.06.053
5. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50:205–216. https://doi.org/10.1021/ci900419k
6. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 12:225–33. https://doi.org/10.1016/j.drudis.2007.01.011
7. Willett P (2009) Similarity methods in chemoinformatics. Annu Rev Inf Sci Technol 43:1–117. https://doi.org/10.1002/aris.2009.1440430108
8. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov

Today 11:1046–53. https://doi.org/10.1016/j.drudis.2006.10.005

9. Willett P (2013) Fusing similarity rankings in ligand-based virtual screening. Comput Struct Biotechnol J 5:e201302002. https://doi.org/10.5936/csbj.201302002

10. Willett P (2013) Combination of similarity rankings using data fusion. J Chem Inf Model 53:1–10. https://doi.org/10.1021/ci300547g

11. Todeschini R, Consonni V, Xiang H, et al (2012) Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. J Chem Inf Model 52:2884–2901. https://doi.org/10.1021/ci300261r

12. Rácz A, Andrić F, Bajusz D, Héberger K (2018) Binary similarity measures for fingerprint analysis of qualitative metabolomic profiles. Metabolomics 14:. https://doi.org/10.1007/s11306-018-1327-y

13. Rácz A, Bajusz D, Héberger K (2018) Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. J Cheminform 10:48. https://doi.org/10.1186/s13321-018-0302-y

14. Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K (2021) Differential Consistency Analysis: Which Similarity Measures can be Applied in Drug Discovery? Mol Inform 40:2060017. https://doi.org/10.1002/minf.202060017

15. Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K (2021) Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics. J Cheminform 13:32. https://doi.org/10.1186/s13321-021-00505-3

16. Miranda-Quintana RA, Rácz A, Bajusz D, Héberger K (2021) Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. J Cheminform 13:33. https://doi.org/10.1186/s13321-021-00504-4

17. Dunn TB, Seabra GM, Kim TD, et al (2021) Diversity and Chemical Library Networks of Large Data Sets. J Chem Inf Model. https://doi.org/10.1021/ACS.JCIM.1C01013

18. Chang L, Perez A, Miranda-Quintana RA (2021) Improving the analysis of biological ensembles through extended similarity measures. bioRxiv. https://doi.org/10.1101/2021.08.08.455555

19. Flores-Padilla A, Eurídice Juárez-Mercado] K, Naveja JJ, et al (2021)

Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. ChemRxiv. https://doi.org/10.33774/CHEMRXIV-2021-0PQ98

20. Bajusz D, Miranda-Quintana RA, Rácz A, Héberger K (2021) Extended many-item similarity indices for sets of nucleotide and protein sequences. Comput Struct Biotechnol J 19:3628–3639. https://doi.org/10.1016/j.csbj.2021.06.021

21. Cherkasov A, Muratov EN, Fourches D, et al (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57:4977–5010. https://doi.org/10.1021/jm4004285

22. Piir G, Kahn I, García-Sosa AT, et al (2018) Best Practices for QSAR Model Reporting: Physical and Chemical Properties, Ecotoxicity, Environmental Fate, Human Health, and Toxicokinetics Endpoints. Environ Health Perspect 126:126001. https://doi.org/10.1289/EHP3264

23. Algamal ZY, Qasim MK, Lee MH, Mohammad Ali HT (2020) High-dimensional QSAR/QSPR classification modeling based on improving pigeon optimization algorithm. Chemom Intell Lab Syst 206:104170. https://doi.org/10.1016/J.CHEMOLAB.2020.104170

24. Gaulton A, Bellis LJ, Bento AP, et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100-7. https://doi.org/10.1093/nar/gkr777

25. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) Chapter 12 – PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Annual Reports in Computational Chemistry. pp 217–241

26. Andersen CM, Bro R (2010) Variable selection in regression-a tutorial. J Chemom 24:728–737. https://doi.org/10.1002/cem.1360

27. Leardi R (2007) Genetic algorithms in chemistry. J Chromatogr A 1158:226–233. https://doi.org/10.1016/J.CHROMA.2007.04.025

28. Goodarzi M, Dejaegher B, Heyden Y Vander (2012) Feature Selection Methods in QSAR Studies. J AOAC Int 95:636–651. https://doi.org/10.5740/JAOACINT.SGE_GOODARZI

29. Eklund M, Norinder U, Boyer S, Carlsson L (2014) Choosing Feature Selection and Learning Algorithms in QSAR. J Chem Inf Model 54:837–843.

https://doi.org/10.1021/CI400573C

30. National Center for Biotechnology Information. PubChem Database. Source=NCGC, AID=1851

31. Rácz A, Bajusz D, Miranda-Quintana RA, Héberger K (2021) Machine learning models for classification tasks related to drug safety. Mol Divers 25:1409–1424. https://doi.org/10.1007/s11030-021-10239-x

32. Mauri A, Consonni V, Pavan M, Todeschini R (2006) DRAGON SOFTWARE: AN EASY APPROACH TO MOLECULAR DESCRIPTOR CALCULATIONS. MATCH Commun Math Comput Chem 56:237–248

33. (2018) Dragon 7.0, Kode Cheminformatics. Dragon 70, Kode Cheminformatics

34. Rácz A, Bajusz D, Héberger K (2019) Intercorrelation Limits in Molecular Descriptor Preselection for QSAR/QSPR. Mol Inform 38:1800154. https://doi.org/10.1002/minf.201800154

35. Bajusz D, Ferenczy GG, Keserű GM (2015) Property-based characterization of kinase-like ligand space for library design and virtual screening. Med Chem Commun 6:1898–1904. https://doi.org/10.1039/C5MD00253B

36. Kelemen AA, Ferenczy GG, Keserű GM (2015) A desirability function-based scoring scheme for selecting fragment-like class A aminergic GPCR ligands. J Comput Aided Mol Des 29:59–66. https://doi.org/10.1007/s10822-014-9804-5

37. Héberger K (2010) Sum of ranking differences compares methods or models fairly. TrAC Trends Anal Chem 29:101–109. https://doi.org/10.1016/j.trac.2009.09.009

38. Sipos L, Gere A, Popp J, Kovács S (2018) A novel ranking distance measure combining Cayley and Spearman footrule metrics. J Chemom 32:e3011. https://doi.org/10.1002/cem.3011

39. Héberger K, Kollár-Hunek K (2011) Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. J Chemom 25:151–158. https://doi.org/10.1002/cem.1320

40. Héberger K, Kollár-Hunek K (2019) Comparison of validation variants by sum of ranking differences and ANOVA. J Chemom 33:e3104. https://doi.org/10.1002/CEM.3104

41. Lourenco JM, Lebensztajn L (2018) Post-Pareto Optimality Analysis With Sum of

Ranking Differences. IEEE Trans Magn 54:1–10. https://doi.org/10.1109/TMAG.2018.2836327

42. Gere A, Rácz A, Bajusz D, Héberger K (2021) Multicriteria decision making for evergreen problems in food science by sum of ranking differences. Food Chem 344:128617. https://doi.org/10.1016/j.foodchem.2020.128617

43. Saratxaga CL, Bote J, Ortega-Morán JF, et al (2021) Characterization of Optical Coherence Tomography Images for Colon Lesion Differentiation under Deep Learning. Appl Sci 2021, Vol 11, Page 3119 11:3119. https://doi.org/10.3390/APP11073119

44. Sziklai BR (2021) Ranking institutions within a discipline: The steep mountain of academic excellence. J Informetr 15:101133. https://doi.org/10.1016/J.JOI.2021.101133

45. West C (2018) Statistics for Analysts Who Hate Statistics, Part VII: Sum of Ranking Differences (SRD). LCGC North Am 36:2–6