Alternative weighting schemes for fine-tuned extended similarity index calculations

Kenneth López Pérez[1,§], Anita Rácz[2,§*], Dávid Bajusz[3,§], Camila Gonzalez[1], Károly Héberger[2], Ramón Alain Miranda-Quintana[1,*]

[1] *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States*

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

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

§These authors contributed equally to this work.

*Corresponding authors: quintana@chem.ufl.edu, racz.anita@ttk.mta.hu

This paper is intended for the Conferentia Chemometrica 2023 special issue.

ABSTRACT

Extended similarity indices (i.e. generalization of pairwise similarity) have recently gained importance because of their simplicity, fast computation and superiority in tasks like diversity picking. However, they operate with several meta parameters that should be optimized. Earlier, we extended the binary similarity indices to 'discrete non-binary' and 'continuous' data; now we continue with introducing and comparing multiple weighting functions. As a case study, the similarity of CYP enzyme inhibitors (4016 molecules after curation) was characterized by their extended similarities, based on 2D descriptors, MACCS and Morgan fingerprints. A statistical workflow based on sum of ranking differences (SRD) and analysis of variance (ANOVA) was used for finding the optimal weight function(s). Overall, the best weighting function is the fraction ("frac"), while optimal extended similarity indices were also found, and their differences are revealed across different data sets. We intend this work to be a guideline for users of extended similarity indices regarding the various weighting options available. Source code for the calculations is available at https://github.com/mqcomplab/MultipleComparisons.

INTRODUCTION

The application of molecular fingerprints (binary strings of molecular features with 0/1 denoting absence/presence) and similarity measures has been the backbone of cheminformatics for decades. Ligand similarity searches have constituted a high-throughput alternative for computational drug design and virtual screening,[1,2] and they have remained tightly integrated even into today's AI-enhanced, structure-based workflows,[3] as well as generative models.[4] After our 2015 work that provided a statistical basis for the decade-long collective habit of preferring the Tanimoto coefficient,[5] we have thoroughly investigated a large number of alternatives collected from diverse scientific fields by Todeschini et al.,[6] regarding their use in metabolomics,[7] molecular modeling[8] and even food science.[9] In certain cases, we could establish a perfect consistency between two or more metrics even with analytical methods.[10]

As cheminformatics and related fields move gradually into the domain of big data, e.g. by handling molecular datasets in and above the billion regime,[11,12] a deeply rooted bottleneck of similarity metrics gets revealed. By their original definition, similarity measures are calculated between exactly two entities, resulting in a disadvantageous, quadratic scaling of computational demand if we want to characterize (cluster, etc.) a large dataset by pairwise similarity calculations. To rectify this, we have recently developed and introduced an extension of the mathematical framework of similarity calculations, allowing for calculating a single similarity value for an arbitrarily large group of objects.[13,14] Without reiterating the whole framework here, we briefly note that, for binary fingerprints, this extension is based on the generalization of the terms

- a: the number of coincident 1′s (common *on* bits)
- d: the number of coincident 0′s (common *off* bits)
- and $b+c$: the number of 0-1 or 1-0 pairs

to the following, more general terms:

- 1-similarity counters: number of bit positions where 1's occur over a coincidence threshold γ
- 0-similarity counters: number of bit positions where 0's occur over a coincidence threshold γ
- dissimilarity counters: number of bit positions where neither 1's, nor 0's occur over the coincidence threshold γ.

An important feature of the resulting *extended* or *n-ary similarity metrics* is that they realize a true generalization of the well-known pairwise similarities, as they provide the same results in the $n=2$ case as the "traditional" pairwise definitions. As a further extension of the framework, we have defined *extended many-item*[15] and *extended continuous similarities*[16] to cases of discrete non-binary and continuous data, respectively. Most importantly, the usage of extended similarities displays a much more advantageous, linear scaling of computational demand vs. library size, during common tasks such as diversity selection or clustering.[14] Recently, the various flavors of extended similarity metrics have been implemented into several workflows, including data visualization,[17] activity cliff- detection,[18] or molecular dynamics trajectory sampling.[19]

Besides their many advantages, extended similarities require the optimization of several parameters that are absent from the traditional pairwise definitions. One such parameter is the definition of the coincidence threshold: when comparing five objects and having a bit position with three co-occurring 1's and two 0's, should we consider that a 1-similarity counter, or a dissimilarity counter? Similarly, should we distinguish the "strength" of similarity counters by assigning a larger weight to a bit position with five co-occurring 1's vs. the bit position detailed above (three 1's and two 0's)? While we have thoroughly investigated the first question in our earlier reports, there is much left to be explored regarding the use of different weighting schemes and their effect on the outcome of extended similarity calculations.

Here, we will introduce multiple weighting schemes and compare their usage in a common scenario of differentiating two groups of active vs. inactive molecules against the important drug (anti)target, CYP 2C9.[20] For continuity, we use the same dataset as in our recent work,[16] but use different molecular representations to cover the binary and continuous extended similarity metrics.

METHODS

Extended similarity

The key insight behind the extended similarity indices is that a condensed vector of the whole dataset is enough to quantify the similarity of the set. That is, given N molecules represented by vectors with M components, we will arrange them in a **N** × **M** matrix. Then, we just need to calculate the sum of each of the columns of this matrix, thus generating a vector $\Sigma = [\sigma_1, \sigma_2, ..., \sigma_M]$. In order to recover the information about the 1-, 0-, and dissimilarity-indicators (*a*, *d*, *b* + *c*) we need to calculate the indicator $\Delta = |2\sigma_k - N|$, which quantifies the "agreement" between elements in the *k*th column, and a coincidence threshold, γ, that indicates up to which point we count the column as contributing to the similarity or dissimilarity of the set. However, even with this classification we still need to distinguish between cases with partial coincidence of elements in a column. For that, we need to define weight functions, that penalize the partial coincidence over the similar columns, f_s, and the dissimilar ones f_d. These functions only need to obey very general conditions, namely:

$$\begin{aligned} & f_s(N) = 1 \\ & x > y \Rightarrow f_s(x) > f_s(y) \end{aligned} \tag{1}$$

$$\begin{aligned} & f_d(N \bmod 2) = 1 \\ & x > y \Rightarrow f_s(x) < f_s(y) \end{aligned} \tag{2}$$

Until now, all the extended similarity applications have only used a very particular form of these functions, with a simple linear dependency:

$$f_s(x) = \frac{x}{N}$$
$$f_d(x) = 1 - \frac{x - N \bmod 2}{N} \tag{3}$$

Here, we explore the effect of changing these weights in non-linear ways, as summarized in Table 1.

Table 1. Weighting schemes introduced and compared in this work (for reference, we include the original, fractional weighting scheme, labeled "Frac" from here on).

Order	Notation	f_s	f_d
1	Frac	$\frac{x}{N}$	$1-\frac{x - N \bmod 2}{N}$
2	Poly_2	$\left(\frac{x}{N}\right)^2$	$1-\left(\frac{x - N \bmod 2}{N}\right)^2$
-2	Poly_M2	$\left(\frac{x}{N}\right)^{-2}$	$1-\left(\frac{x - N \bmod 2}{N}\right)^{-2}$
4	Poly_4	$\left(\frac{x}{N}\right)^4$	$1-\left(\frac{x - N \bmod 2}{N}\right)^4$
-4	Poly_M4	$\left(\frac{x}{N}\right)^{-4}$	$1-\left(\frac{x - N \bmod 2}{N}\right)^{-4}$
8	Poly_8	$\left(\frac{x}{N}\right)^8$	$1-\left(\frac{x - N \bmod 2}{N}\right)^8$
-8	Poly_M8	$\left(\frac{x}{N}\right)^{-8}$	$1-\left(\frac{x - N \bmod 2}{N}\right)^{-8}$

16	Poly_16	$\left(\frac{x}{N}\right)^{16}$	$1-\left(\frac{x-N \bmod 2}{N}\right)^{16}$
-16	Poly_M16	$\left(\frac{x}{N}\right)^{-16}$	$1-\left(\frac{x-N \bmod 2}{N}\right)^{-16}$

Dataset

A large dataset of cytochrome P450 (CYP) 2C9 ligands from Pubchem Bioassay (AID 1851) was used to compare the different weighting schemes,[21] to provide continuity with our earlier work.[16] Cytochrome P450 enzymes (CYP) are important mediators of drug metabolism, therefore generally important anti-targets in drug design: consequently, many datapoints of CYP bioactivity for deposited into public databases, and CYP enzymes are likewise popular targets of QSAR and machine learning studies.[22] In total, 12,161 molecules were applied after data curation: 4016 inhibitors with a potency of 10 μM or better (actives) and 8145 inactive species. MACCS[23] and Morgan fingerprints (radius: 4, length: 1024)[24] were generated with RDKit,[25] while the Dragon 7 software was used for the calculation of 2D descriptors.[26,27] Highly correlated variables (above 0.997) and constant variables were excluded from the sets.[28] The details and descriptions of the different descriptor sets can be found in the DRAGON software manual.

RESULTS AND DISCUSSION

First, similarity calculations were carried out for the three separate datasets (2D descriptors, MACCS and Morgan fingerprint) in the case of active and total groups. Our major assumption was that the better similarity metrics can provide higher similarity values for the actives and bigger differences between the actives and total similarity values. Some restrictions had to be made before the evaluation of the similarity values in each case studies. Based on previous assumptions, we have used only the non-weighted version of extended similarity values for the

further evaluations. While in the case of 2D descriptor dataset, coincidence threshold limit was set to minimum to avoid the combinatorial "explosion" in the further ANOVA evaluations, in the case of MACCS and Morgan fingerprints, similarity values were calculated in each coincidence threshold limits. SRD was used for the comparison of the similarity metrics in the case of each weighting function for the active group. The results of SRD were channeled into a factorial ANOVA, in which the weighting functions (9) and similarity metrics (16) were used as factors along with the different datasets (3). The workflow of the three datasets will be discussed in the next sections.

In the case of 2D descriptors for the active set, the basic SRD input dataset contained 16 similarity metrics in the columns, and 14 different 2D descriptor sets in the rows. The dimensions of the input dataset were the same for each weighting functions. Thus, the SRD calculations were carried out for the nine different weighting functions and the cross-validated SRD results were summed up for the further analysis.

In the case of both MACCS and Morgan fingerprint datasets for actives, the similarity metrics were in the columns, and the similarity values for each coincidence threshold limits were in the rows. It resulted in 9 rows and 16 columns. Like in the case of 2D descriptors set, in total nine SRD calculations were carried out with the same settings and further analyzed by ANOVA.

As SRD is in % scale for each case study, we have merged together the results for the factorial ANOVA.

In the factorial ANOVA analysis, our aim was to find the differences and best solutions between the weighting functions for the different dataset alternatives, such as continuous (2D descriptors) and binary type vectors (fingerprints). Therefore, at first in **Figure 1** we show the comparison of the weighting functions for the SRD results of the three datasets.

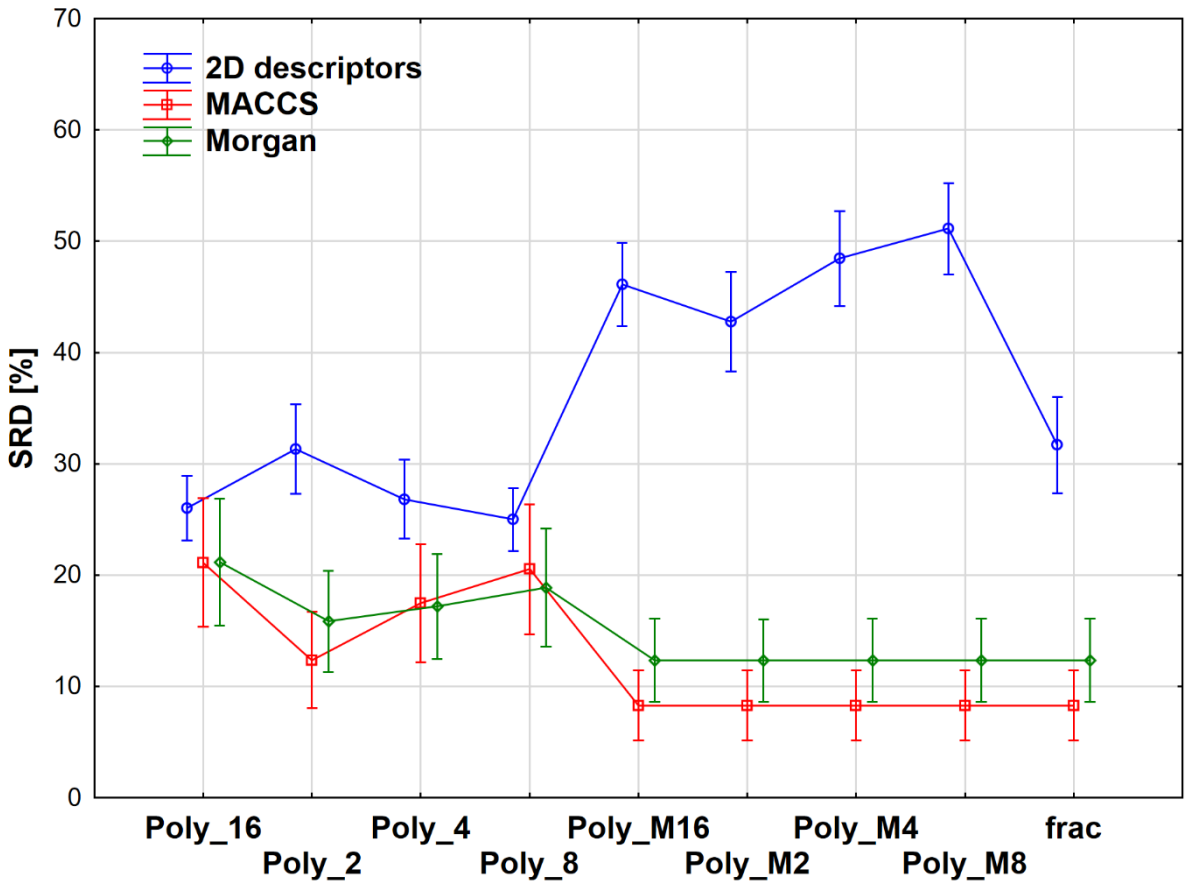

Figure 1. The SRD values (%) are plotted based on the weighting functions.

As it is shown, the best weighting function is depending on the used dataset type. Significant differences were detected in the factor of the weighting functions (at alpha = 0.05). While MACCS had the lowest (best) SRD values in total, the binary datasets show almost the same pattern. Thus, the positive orders of the weighting functions are optimal for the continuous descriptor type sets (especially Poly_8 and Poly_16), while the negative orders and the fraction weighting functions are better for the binary datasets.

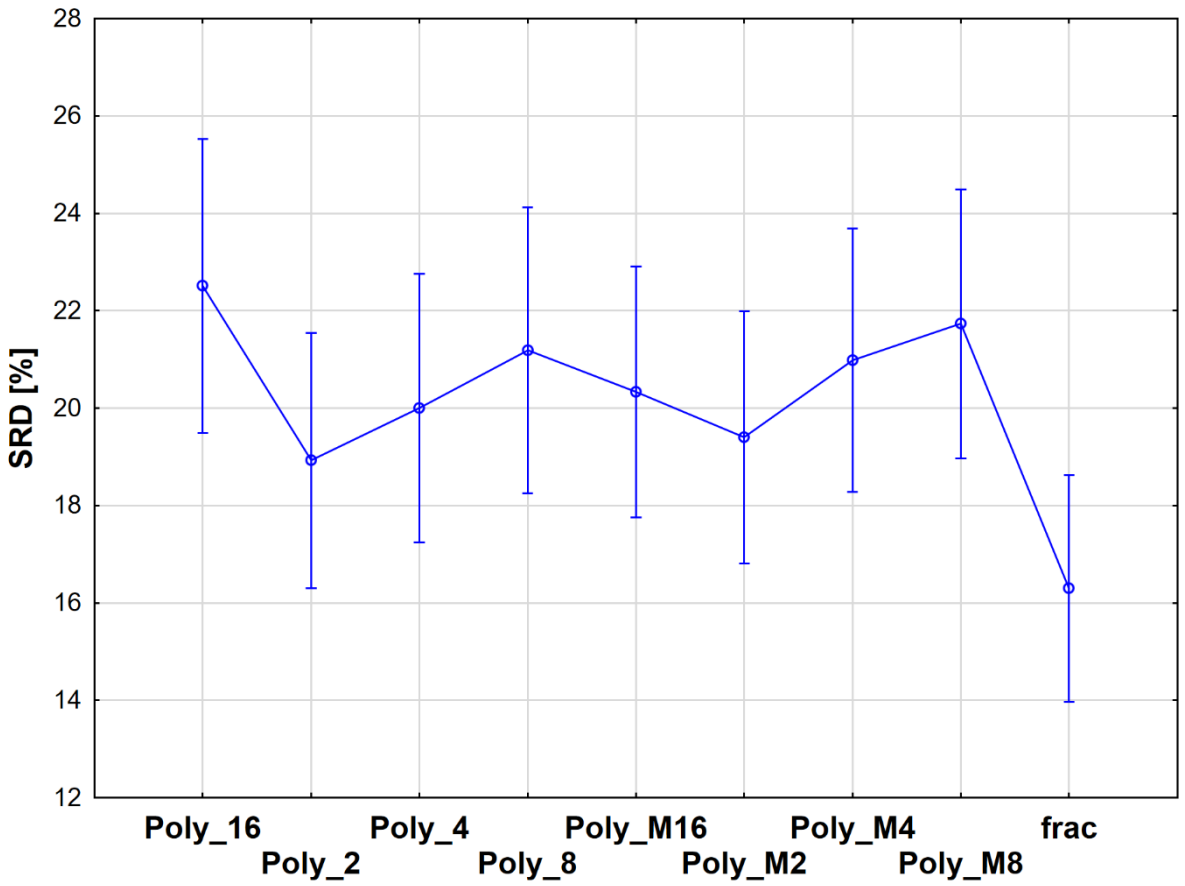

Figure 2. SRD values (%) based on the weighting functions.

If we compare the SRD values for the actives together for the three datasets in the sense of weighting functions, we can see the most optimal ones overall in **Figure 2**. The fractional weighting function looks like the best option, but it is biased towards the binary datasets. Although the confidence intervals (95%) are wider in this case, the weighting functions have still significant impact on the outcome.

From the similarity metrics point of view, we can safely say, based on **Figure 3**, that the current metrics are generally worse for continuous sets, than the binary ones. The extended similarity index denoted by eJa0 is close to random or even reverse ranking in the case of binary datasets, and interestingly eBUB produces dramatically different SRD scores for MACCS and Morgan

fingerprints. However, the patterns for MACCS and Morgan datasets are still very similar. SRD scores for many extended similarity indices are the same or closely resemble the hypothetical best ranking. These datasets are in concordance with the previous results (cf. Fig. 5 in ref.[13]) that eCT1 and eCT2 indices are amongst the best ones for all the three datasets. We can extend this now with eAC, eRT and eSM, which are still suitable for the two binary case studies and acceptable for the continuous dataset as well. In the case of 2D descriptors, the list can be extended with eJa and eSS2, too.

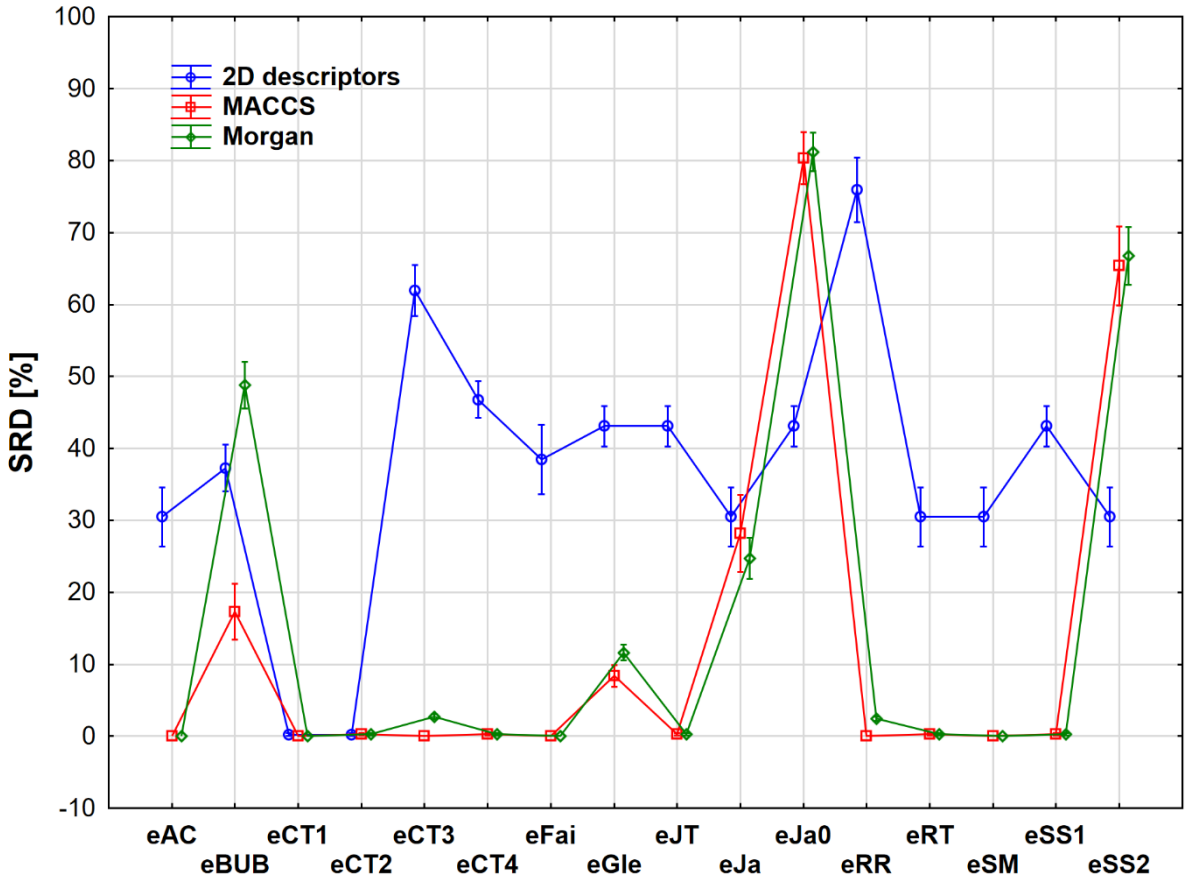

Figure 3. SRD values based on the similarity metrics and the dataset types.

Next, we have analyzed the most frequent weighting functions and similarity indices with regard to the differences between the "active" and "total" group similarities for each dataset. **Figure 4** shows the frequencies of the weighting functions occurring above the cut-off limit of 0.10, while **Figure 5** shows the most frequent similarity metrics above this cut-off limit.

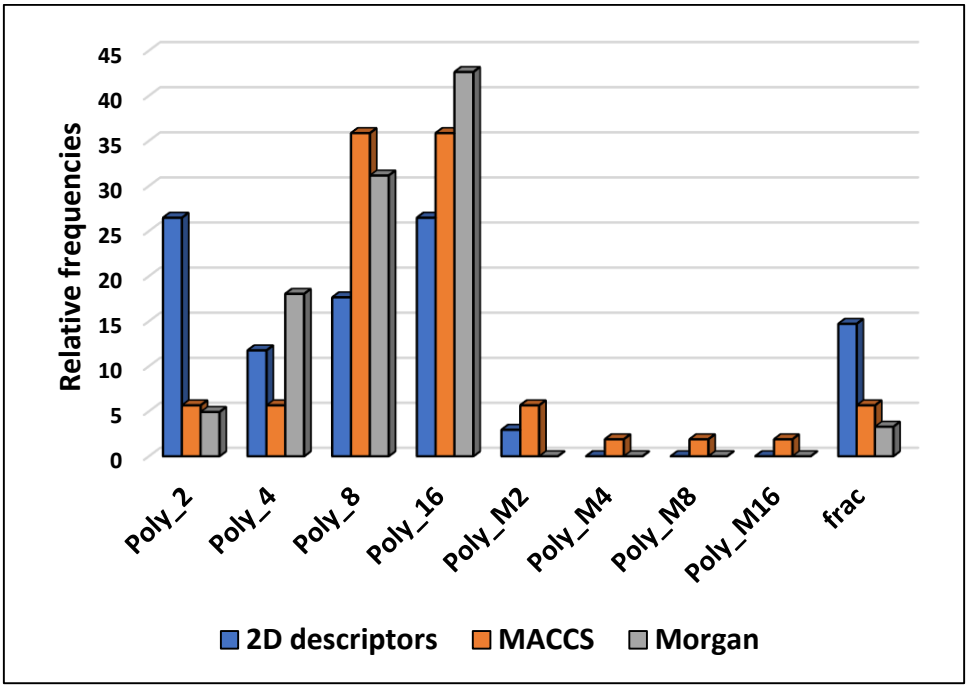

Figure 4. Relative frequencies of the weighting functions above 0.10 similarity values in the case of Active-Total data.

As shown, the top weighting functions in this case were the Poly_8 and Poly_16 variants. These two were the most frequent in all case studies regarding the Active-Total similarity values. Poly_8 and especially Poly_16 can be good options if the aim of the study is to maximize the difference between the active set and the whole database (actives and inactives). On the other hand, the Poly_2 weighting function is also a good choice for continuous descriptors as input data.

We have evaluated the frequencies of the similarity metrics above the 0.10 cut-off value in the same way. **Figure 5** shows the distribution of the similarity metrics in the three case studies for the Active-Total values. The result shows that eCT4 can be an optimal choice for the similarity calculations especially in the case of binary datasets like MACCS or Morgan, where the aim is to determine the differences between the actives and the whole dataset of molecules. On the other hand, the eGle, eJT, eJa and eSS1 indices are also good options for continuous variables in the input matrix.

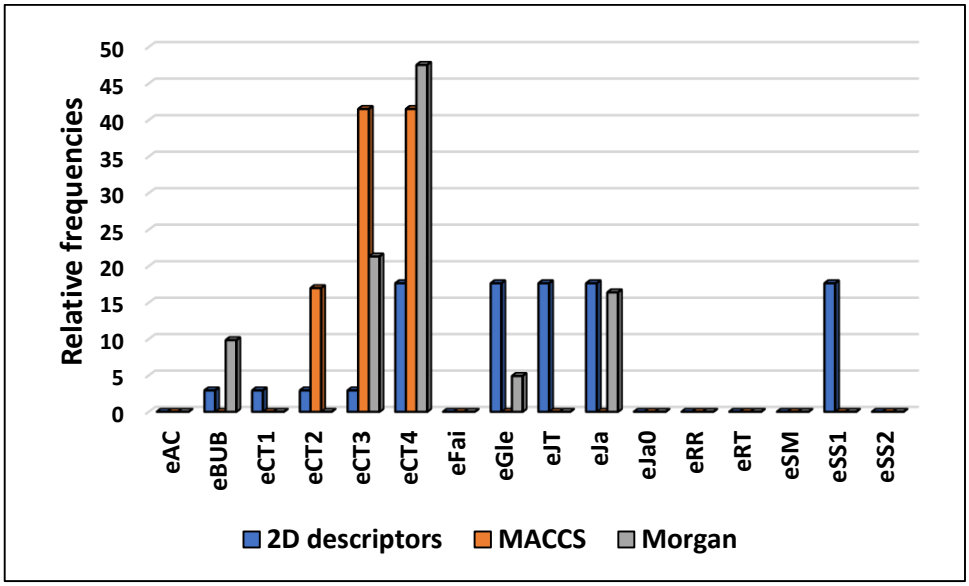

Figure 5. Relative frequencies of the similarity indices above 0.10 cut-off limit for the Active-Total similarity values of the three case studies.

CONCLUSION

We have compared several weighting functions in the case of extended similarity metrics based on sum of ranking differences (SRD scores) combined with factorial ANOVA. The superiority of positive-power weighting functions was clear for the continuous descriptors, while negative-power alternative (and fraction) are better for binary variables in the input dataset. Interestingly, in the evaluation of Active-Total values (differences in the similarities of the active molecules vs. the whole dataset), positive weighting functions, especially Poly_8 and Poly_16 were amongst the best for all the case studies. Thus, weights can be optimized based on the aim of the application. From the extended similarity metrics point of view, we have seen that the most optimal ones are input variable dependent, while eCT1 and eCT2 are good choices for the similarity calculation of the active sets in any case. On the other hand, eCT3 and eCT4 are very promising for binary case studies in both Actives and Active-Total evaluation, while eJa can be an optimal index for the continuous descriptors in both Actives and Active-Total evaluation. As the used weighting functions and similarity metrics had significant impact on the outcome,

selecting the optimal ones in a case specific manner is of outmost importance in the similarity calculations stage of any cheminformatics related research. Source code for the calculations is available at https://github.com/mqcomplab/MultipleComparisons

ACKNOWLEDGEMENTS

The authors received funding from the National Research Development and Innovation Office of Hungary under grant no. K134260 (KH) and FK146063 (D.B.), as well as the Hungarian Academy of Sciences: János Bolyai Research Scholarship (A.R.). The work of A.R. was supported by the ÚNKP-23-5 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. RAMQ and KLP thank support from the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM150620.

REFERENCES

1. Willett P. Similarity-based approaches to virtual screening. *Biochemical Society Transactions*. 2003;(31):603-606.

2. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today*. 2006;**11**(23-24):1046-1053. doi:10.1016/j.drudis.2006.10.005.

3. Tropsha A, Popov KI, Wellnitz J, Maxfield T. Hit Discovery using Docking ENriched by GEnerative Modeling (HIDDEN GEM): A Novel Computational Workflow for Accelerated Virtual Screening of Ultra-large Chemical Libraries. *Molecular Informatics*. 2024;**43**(1):e202300207. doi:10.1002/minf.202300207.

4. Tibo A, He J, Janet JP, Nittinger E, Engkvist O. Exhaustive local chemical space exploration using a transformer model. October 2023. doi:10.26434/chemrxiv-2023-v25xb.

5. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*. 2015;**7**:20. doi:10.1186/s13321-015-0069-3.

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. *Journal of Chemical Information and Modeling*. 2012;**52**(11):2884-2901. doi:10.1021/ci300261r.

7. Rácz A, Andrić F, Bajusz D, Héberger K. Binary similarity measures for fingerprint analysis of qualitative metabolomic profiles. *Metabolomics*. 2018;**14**(3):29. doi:10.1007/s11306-018-1327-y.

8. Rácz A, Bajusz D, Héberger K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J Cheminform*. 2018;**10**(1):48. doi:10.1186/s13321-018-0302-y.

9. Gere A, Bajusz D, Biró B, Rácz A. Discrimination Ability of Assessors in Check-All-That-Apply Tests: Method and Product Development. *Foods*. 2021;**10**(5):1123. doi:10.3390/foods10051123.

10. Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K. Differential Consistency Analysis: Which Similarity Measures can be Applied in Drug Discovery? *Mol Inf*. 2021;**40**(7):2060017. doi:10.1002/minf.202060017.

11. Bajusz D, Keserű GM. Maximizing the integration of virtual and experimental screening in hit discovery. *Expert Opinion on Drug Discovery*. 2022;**17**(6):629-640. doi:10.1080/17460441.2022.2085685.

12. Silva MMP da, Guedes IA, Custódio FL, Krempser E, Dardenne LE. Deep Learning Strategies for Enhanced Molecular Docking and Virtual Screening. November 2023. doi:10.26434/chemrxiv-2023-zfv87-v2.

13. Miranda-Quintana RA, 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. doi:10.1186/s13321-021-00505-3.

14. Miranda-Quintana RA, 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. doi:10.1186/s13321-021-00504-4.

15. Bajusz D, Miranda-Quintana RA, Rácz A, Héberger K. Extended many-item similarity indices for sets of nucleotide and protein sequences. *Computational and Structural Biotechnology Journal*. 2021;**19**:3628-3639. doi:10.1016/j.csbj.2021.06.021.

16. Rácz A, Dunn TB, Bajusz D, Kim TD, Miranda-Quintana RA, Héberger K. Extended continuous similarity indices: theory and application for QSAR descriptor selection. *J Comput Aided Mol Des*. 2022;**36**(3):157-173. doi:10.1007/s10822-022-00444-7.

17. López-Pérez K, López-López E, Medina-Franco JL, Miranda-Quintana RA. Sampling and Mapping Chemical Space with Extended Similarity Indices. *Molecules*. 2023;**28**(17):6333. doi:10.3390/molecules28176333.

18. Dunn TB, López-López E, Kim TD, Medina-Franco JL, Miranda-Quintana RA. Exploring activity landscapes with extended similarity: is Tanimoto enough? *Molecular Informatics*. 2023;**42**(7):2300056. doi:10.1002/minf.202300056.

19. Rácz A, Mihalovits LM, Bajusz D, Héberger K, Miranda-Quintana RA. Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices. *J Chem Inf Model*. 2022;**62**(14):3415-3425. doi:10.1021/acs.jcim.2c00433.

20. Rácz A, Keserű GM. Large-scale evaluation of cytochrome P450 2C9 mediated drug interaction potential with machine learning-based consensus modeling. *J Comput Aided Mol Des*. 2020;**34**(8):831-839. doi:10.1007/s10822-020-00308-y.

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

22. Rácz A, Bajusz D, Miranda-Quintana RA, Héberger K. Machine learning models for classification tasks related to drug safety. *Mol Divers*. 2021;**25**(3):1409-1424. doi:10.1007/s11030-021-10239-x.

23. Bajusz D, Rácz A, Héberger K. 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*. Oxford: Elsevier; 2017:329-378. doi:10.1016/B978-0-12-409547-2.12345-5.

24. Rogers D, Hahn M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*. 2010;**50**(5):742-754. doi:10.1021/ci100050t.

25. RDKit: Open-Source Cheminformatics Software. http://rdkit.org/. Accessed May 2, 2016.

26. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. *MATCH Communications in Mathematical and in Computer Chemistry*. 2006;**56**:237-248.

27. Dragon 7.0, Kode Cheminformatics. Dragon 70, Kode Cheminformatics.

28. Rácz A, Bajusz D, Héberger K. Intercorrelation Limits in Molecular Descriptor Preselection for QSAR/QSPR. *Mol Inf*. 2019;**38**(8-9):1800154. doi:10.1002/minf.201800154.