Molecular Diversity (2023) 27:1603–1612
https://doi.org/10.1007/s11030-022-10514-5

ORIGINAL ARTICLE

Chemical similarity of molecules with physiological response

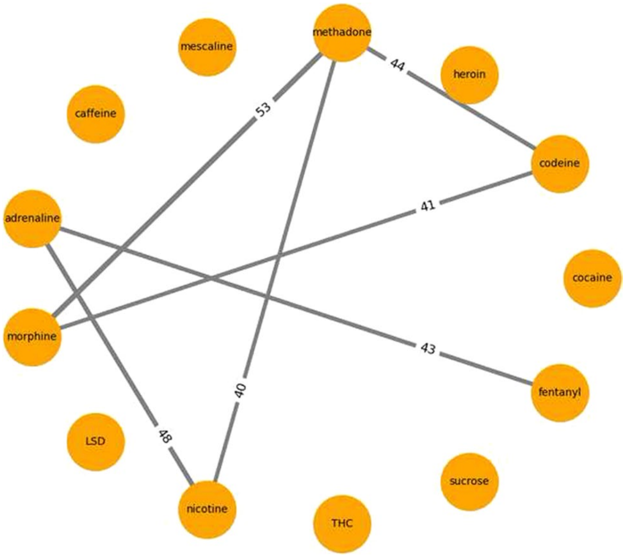
Izudin Redžepović[1] · Boris Furtula[1]

Received: 10 June 2022 / Accepted: 8 August 2022 / Published online: 17 August 2022

Abstract

Measuring the similarity among molecules is an important task in various chemically oriented problems. This elusive concept is hard to define and quantify. Moreover, the complexity of the problem is elevated by bifurcating the notion of molecular similarity to structural and chemical similarity. While the structural similarity of molecules is being extensively researched, the so-called chemical similarity is being mentioned scarcely. Here, we propose a way of converting the physico-chemical properties into molecular fingerprints. Then, using the apparatus of measuring the structural similarity, the chemical similarity can be assessed. The proof of a concept is demonstrated on a set of molecules that induce diverse physiological responses.

Graphical abstract



Keywords Binary fingerprints · Drugs · Log*P* · Melting point · pKa

✉ Izudin Redžepović
izudin.redzepovic@pmf.kg.ac.rs

[1] Department of Chemistry, Faculty of Science, University of Kragujevac, P. O. Box 60, 34000 Kragujevac, Serbia

Introduction

A naturally occurring result from observing two objects is the extent of similarity among them [1]. Although we deal with this elusive concept daily, it is challenging to properly define similarity or quantify it because of its multidimensional character. The similarity between molecules plays a key role in cheminformatics [2, 3]. Furthermore, it is also significant in modern drug design and the material industry [4–6]. For example, Krasowski and coworkers have used the similarity approach to assess the structural similarity for a wide spectrum of clinically important drugs to the target molecules of DOA/Tox screening tests [7]. Also, the authors in [8] have shown that compounds producing cross-reactivity in steroid hormone immunoassays have a high degree of structural similarity to the target hormone. Additionally, a similarity approach was used to design novel zeolite materials for separations based on adsorption [9]. Such a big interest in molecular similarity may be especially accounted to the one renowned postulate, i.e., to similarity property principle (SPP) [10]. Namely, this viewpoint of the structure–property relationship implies that compounds with high structural similarity are likely to have similar physico-chemical properties. From this statement, it is obvious that SPP has a simple and logical foundation. However, a relationship between structural features and a physico-chemical property (or some type of activity) of the corresponding molecule is complex and it is not always so apparent. Therefore, there are stumbling stones within SPP, such as activity cliffs [11–14].

To measure the similarity of two molecules, a quantification of molecular structure is a necessary step. For this reason, diverse ways of encoding structural information have been proposed. Presently, an enormous number of molecular descriptors is available, with an increasing tendency [15, 16]. Types of descriptors range from simple, such as counting descriptors, up to complex quantum-chemical molecular descriptors [17]. The molecular descriptors are usually categorized according to the dimensionality of molecular representation needed for the calculation of a descriptor. Thus, there are 1D, 2D, 3D, and higher-dimension descriptors. Besides similarity investigations, molecular descriptors have also found important applications in QSPR/QSAR studies [18, 19]. A special place in the similarity-related calculations is reserved for structural fingerprints [20, 21]. In their simplest form, these are numerical strings constructed by zeros and ones. More precisely, one in a bit-string represents the presence of a certain structural feature in a molecule, while zero signifies its absence, in this way producing molecule-specific linear bit patterns. With this, the molecular structure is converted into a binary vector that may be manipulated with. This type of reasoning was used to create a novel method for representing and analyzing 3D protein–ligand binding interactions. In other words, a binary fingerprint was constructed to model intermolecular connections, where 1 denotes a certain bond between protein and ligand and 0 assigns the lack of a bond. These are called interaction fingerprints [22, 23]. One of the most popular structural fingerprints is the extended-connectivity circular fingerprint [24]. Even though this descriptor was not originally developed as a binary sequence, if necessary, it can be transformed into a binary analog.

The similarity of two molecules is usually perceived as the amount of coherency between their structural features. However, there is another aspect of molecular similarity, usually referred to as chemical similarity. The most obvious manifestation of the chemical similarity between molecules is in the case of compounds that exert similar activities but are structurally quite different. Moreover, the opposite situation is more frequent, i.e., when structurally similar molecules do not show similar physico-chemical properties or activities [25, 26]. For example, Boström and others have found that there is a significant probability that minor modifications on one ligand, in a pair of structurally similar ligands, will produce high changes in the binding sites, hence, the changes in their activities [27]. The main obstacle regarding chemical similarity is its quantification. While structural similarity has been heavily studied, the chemical similarity is poorly understood. Several attempts were made to examine the chemical similarity of some molecules. One of them was made by Xenides and coworkers, who have applied the information theory approach to generate clusters of chemically similar compounds [28].

In this study, a novel method was introduced to quantitatively determine the chemical similarity of molecules. More precisely, a plain binary fingerprint of a molecule was developed by encoding its physico-chemical properties. Within the present paper, we examine the chemical similarity of compounds depicted in Fig. 1. This set consists of 13 molecules that induce diverse physiological responses and was also studied in paper [28]. These compounds cause pleasant, euphoric, and analgesic effects. Several published papers have shown that some of these molecules are producing similar effects, or they are acting as antagonists [29–32]. Since these molecules are well-known and widespread, and some of them are consumed daily, it is of utmost interest to examine relationships among them, i.e., to quantify their chemical similarity. In the following section, we are going to expose a procedure that enables this.

Computational methodology

Construction of a fingerprint

The very first step in constructing binary fingerprints for assessing the chemical similarity of compounds is to provide several physico-chemical properties for underlying molecules. The more diverse properties are supplied, the better the chemical description of a molecule is conceived. Due to the limitation of available experimental data, this might be a tricky task. Therefore, experimental values may be replaced with the properties provided by, e.g., quantum chemistry computations at a sufficiently high level of theory. For our set of compounds, melting point (MP), log*P*, and pKa experimental values were used for this purpose and are collected in Table 1. These values were retrieved from PubChem [33] and DrugBank [34] chemistry data repositories. A reason for using these sets of physico-chemical properties is that these experimentally determined properties were available for all thirteen molecules under the consideration. Moreover, both log*P* and pKa are known as high-quality indicators of physiological effects. In order to avoid fingerprint dependance on dataset size, in this approach we do not apply standardization of the physico-chemical properties. The advantage of this is twofold, the resulting molecular fingerprint remains the same within any dataset of compounds with the use of the same physico-chemical properties arranged in the same order, and the developed fingerprint is highly informative. Regarding the latter, this means that the obtained fingerprint is not sparse, which might be the case when min–max scaling is applied, for example. Therefore, we have developed fingerprints based on the physico-chemical properties in the following manner.

To encode as much chemical information as possible, values from Table 1 are rounded up to two decimals and then multiplied by 10^2. In this way, experimental values are converted into integers without losing valuable information. Then, for each property five digits are reserved for encoding into a string, since our values are no bigger than five digits. If the obtained integer has less than 5 digits, then one or more zeros are added at the beginning of an experimental value to get a string of five digits. By completing this step, all used values are encoded into a numerical string made of five digits. Further construction of molecular fingerprints based on physico-chemical properties demands the transformation of these integer strings into binary strings. This step is done using the binary coded decimal (BCD) system. Namely, in this type of encoding every digit in a five-digits-string (even zero) is replaced by the corresponding 4-bits-long binary code (see Table 2). This conversion is transforming the experimental value into a binary string with a length of 20 bits.

An additional bit is added at the beginning of a string to encode the sign of an experimental value. Zero denotes positive, while one stands for the negative sign. In this way, every physico-chemical property from Table 1 is transformed into a 21-bits-long binary code. Finally, by merging obtained strings for MP, log*P*, and pKa, in this order, the molecular fingerprint based on the physico-chemical properties, with the length of 63 bits, is constructed. In Fig. 2 this procedure is depicted in the case of the molecule of cocaine.

The authors of the paper [35] have studied similarity coefficients that are usually utilized in cheminformatics investigations. They have found that some of the examined metrics exhibit better characteristics than others. Namely, out of all coefficients that yield the similarity results within [0,1]-range, Jaccard (*Ja*) [36], Jaccard-Tanimoto (*JT*) [37], Gleason (*Gle*) [38], Sokal-Sneath (*SS*) [39], and Consonni-Todeschini (*CT*) [40] have shown satisfactory performance in similarity calculations related to real and simulated cheminformatics binary data. They are defined as follows:

$$Ja = \frac{3a}{3a + b + c} \tag{1}$$

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

$$Gle = \frac{2a}{2a + b + c} \tag{3}$$

$$SS = \frac{a}{a + 2b + 2c} \tag{4}$$

$$CT = \frac{ln(1 + a)}{ln(1 + a + b + c)} \tag{5}$$

In Eqs. (1)–(5) a is the frequency of bits 1 that fingerprints of molecules A and B have in common, b is the frequency of bits 1 present in fingerprint A but not in B, and c is the frequency of bits 1 found in fingerprint B but not in A. As can be seen, most of these indices differ only in the weights that they give to some parts of the fingerprints during comparative analysis. Namely, *JA* and *Gle* coefficients highlight the same features in fingerprints, while *SS* emphasizes their differences. To analyze the chemical similarity of compounds depicted in Fig. 1, we have employed *Ja*, *JT*, *Gle*, *SS*, and *CT* asymmetric similarity indices to measure the coherence between fingerprints based on physico-chemical properties. For all these computations, a Python script was coded with an implementation of the RDKit cheminformatics package [41]. In addition, we have calculated the extended versions of our similarity indices. The extended similarity indices allow simultaneous comparison of more than two molecules at the same time. These similarity coefficients are entirely general, and they do not depend on the fingerprints used [42, 43]. In the original paper, their features were investigated by sum of ranking differences (a statistical method that we also use here to get better insight into our

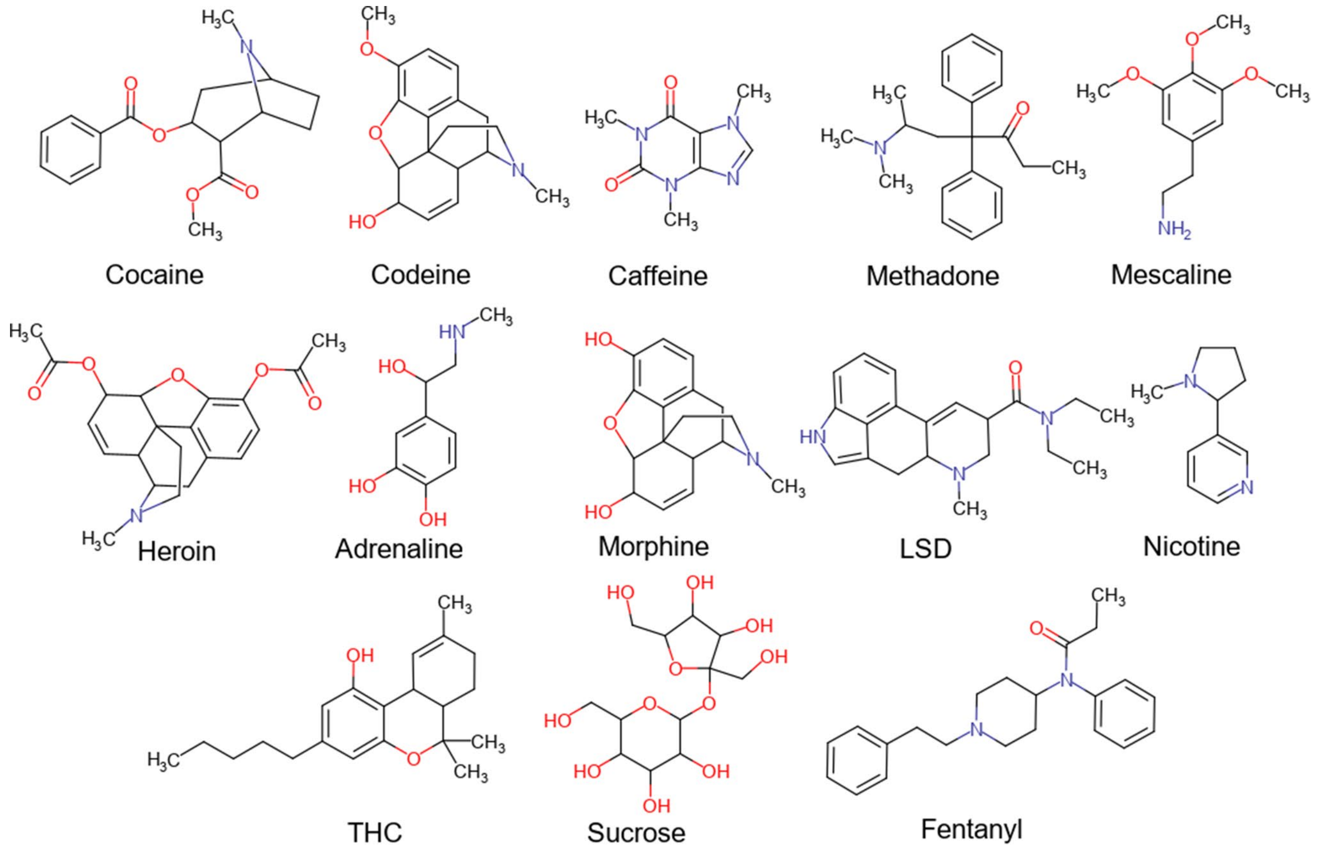


Fig. 1 The compounds with physiological effects

Table 1 The experimental physico-chemical properties that are used to construct binary fingerprints

Molecule	MP (°C)	log*P*	pKa
Cocaine	98.00	2.30	8.61
Codeine	155.00	1.39	8.20
Caffeine	236.00	− 0.07	14.00
Methadone	235.00	3.93	9.20
Mescaline	35.50	0.78	9.56
Heroin	173.00	1.58	7.95
Adrenaline	211.50	− 1.37	8.59
Morphine	255.00	0.87	8.21
LSD	82.50	2.95	7.80
Nicotine	− 79.00	1.17	8.50
THC	200.00	5.65	10.60
Sucrose	185.50	− 3.70	12.60
Fentanyl	83.50	4.05	8.99

results) and ANOVA. The definition of the extended form of *Ja*, *JT*, *Gle*, *SS*, and *CT* and the corresponding Python scripts for their calculation are freely available at: https://github.com/ramirandaq/MultipleComparisons.

Table 2 The BCD system is used to encode digits

Binary code	Digit
0000	0
0001	1
0010	2
0011	3
0100	4
0101	5
0110	6
0111	7
1000	8
1001	9

Sum of ranking differences (SRD)

The SRD is a novel general-purpose statistical procedure to compare models, methods, analytical techniques, etc. [44]. So far, it has been successfully used on many different problems, e.g., for the correct split of training and test sets in QSAR, column selection in supercritical fluid chromatography, and analysis of chromatographic retention data [45–47]. Here, we use this tool to compare the results of the similarity obtained by different metrics. This technique is available as MS Office Excel macro at http://aki.ttk.hu/srd/. In the input matrix, the objects (in the present case molecules) are arranged in rows and the variables (models or methods, in the present case similarity coefficients) are arranged in the columns. The results are ranked for each method (similarity coefficient) to the ranking of experimental or reference values. If the standard value is not available, like in this case, then the mean value for all methods (similarity indices) may be used. With the scaling of SRD values between 0 and 100, it is possible to compare these values to different methods/models. The full description of SRD calculation and its validation may be found elsewhere [44, 48]. In general, the closer the SRD value is to zero (i.e., the closer is the ranking to the golden standard), the better is the method. The proximity of SRD values indicates the similarity of the methods, thus in our case, the similar performance of tested similarity coefficients.

Results and discussion

The developed plain binary fingerprints based on physico-chemical properties of compounds depicted in Fig. 1 have been mutually compared using *Ja*, *JT*, *Gle*, *SS*, and *CT* similarity coefficients and by their extended versions. The calculated similarities, in percentages, are given in Figs. 1S-5S in Supporting Information as heatmaps, while Table 3 summarizes obtained results. As one may see, all five applied metrics yielded comparable results, that is, the same trends have been identified, especially in the case of extended indices. This is expected considering the closeness of their definitions. With an average value of 57%, the highest similarities are obtained by the *CT* coefficient, while the *SS* index gives the lowest mean value (14%). In the case of extended indices, the chemical similarity of four indices is 50%, while the extended *CT* index shows a similarity of 63%. The *Ja* and *CT* indices yielded comparable results by both approaches, standard pairwise and extended, while for the other measures higher similarities are obtained by their extended versions. As for the data variation, most of the similarity values demonstrate comparable scattering. The highest standard deviation is obtained for the *Ja* index, while the lowest data dispersion gives *SS* coefficient. It was found that, on average, the chemically most similar compound to other molecules is adrenaline. Its mean chemical similarities by *Ja*, *JT*, *Gle*, *SS*, and *CT* are 58%, 32.7%, 48.5%, 19.8%, and 66.5%, respectively. On the other hand, with 39.8%, 18.9%, 30.9%, 10.7%, and 50.7% of similarity, LSD is the least similar to the other compounds. Also, comparable to LSD, the low similarity is obtained for THC by *SS* and *CT* indices.

All five similarity coefficients have found that morphine and methadone are chemically the most similar compounds, while cocaine and caffeine show the lowest chemical similarity. Namely, the obtained values for *Ja*, *JT*, *Gle*, *SS*, and *CT* for morphine-methadone similarity are 77%, 53%, 69%, 36%, and 80%, respectively, while these values for cocaine-caffeine similarity amount 14%, 5%, 10%, 3%, and 23%, respectively. Such a high chemical similarity between methadone and morphine is in accordance with the experimental findings, where these two opioids are found to have similar physiological responses. Moreover, both have an analgesic effect, and they are used as substitution pain killers [49]. This finding supports the assumption that molecules with similar physico-chemical properties should stimulate similar physiological reactions. On the other hand, such similarity between methadone and morphine, within this set of molecules, is quite surprising, considering the high structural similarity of morphine and codeine. Namely, it is reasonable to expect that morphine and codeine show the highest

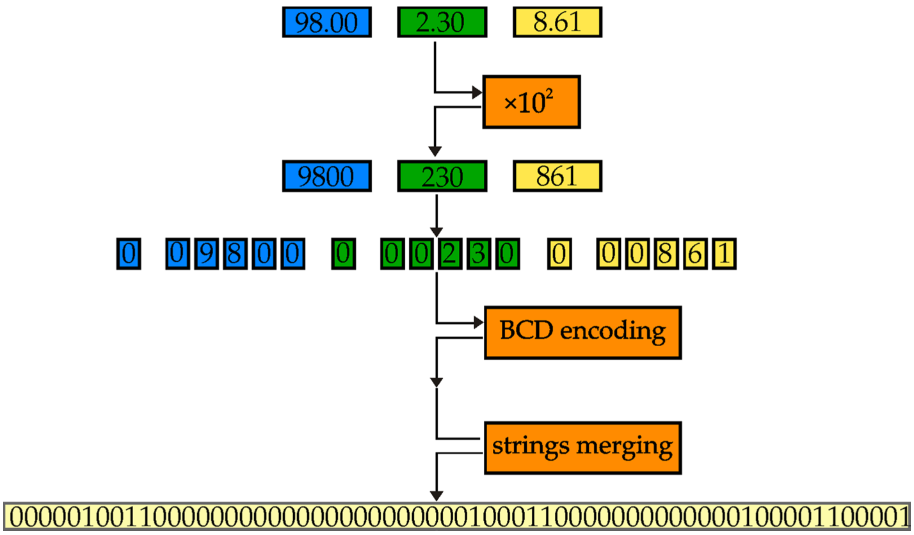


Fig. 2 The procedure of constructing binary 63-bits-long fingerprint based on physico-chemical properties of cocaine molecule

chemical similarity since their structures differ in only one methyl group. However, the chemical similarity of these two compounds ranks as the fifth highest, among all similarities, by all coefficients and it amounts to 68%, 41%, 58%, 26%, and 72% according to *Ja*, *JT*, *Gle*, *SS*, and *CT* index, respectively. Such result may be attributed to the big differences in MP and log*P*, while their pKa values disagree by only 0.01. Even though both molecules cause analgesic effects in the human body, it was found that the magnitude and lasting of these effects significantly differ [50].

The other two molecules that also exhibited high chemical similarity (ranked as the second highest) are adrenaline and nicotine. Both compounds are known as euphoric feeling inducers and their similarities calculated by *Ja*, *JT*, *Gle*, *SS*, and *CT* are 73%, 48%, 65%, 31%, and 78%, respectively. Even though these two molecules share some structural features, like an aromatic ring and a nitrogen atom, structural coherence between adrenaline and mescaline, for example, is more noticeable. However, their chemical similarity is ~ 9% lower on average, compared to adrenaline-nicotine similarity. On the other hand, the effects of nicotine on the heart and systemic blood pressure are almost identical to those of adrenaline [51].

The lowest chemical similarity was detected for cocaine and caffeine by all five metrics, as stated above. For example, the *SS* gives only 3% of similarity between these two molecules. Such a finding is expected considering big differences in all three physico-chemical properties (Table 1). Also, they belong to different types of drugs regarding the sensation they cause in our body, i.e., cocaine is classified as a “hard” drug, while caffeine is marked as a “soft” drug.

As we previously mentioned, the same trends are identified by all coefficients and the differences come only from the amount of computed similarity. Since the Jaccard-Tanimoto coefficient is the most used index in the cheminformatics community, therefore, we decided to present the similarity assessments obtained by this metric. In Fig. 3 chemical similarity of our compounds is depicted. The graph is constructed to reflect the similarity of molecules in the cases where it amounts to ≥ 40%. Namely, an edge is established between two nodes (molecules) only if their chemical similarity is ≥ 40%. This high threshold is chosen to show molecules with a high chemical similarity since the average similarity calculated by *JT* is 25% (Table 3). At the level of 40%, around 50% of molecules are connected, and two clusters of similar molecules are observed. The first group consists of methadone, morphine, and codeine that exhibit a high chemical similarity among themselves. The second, a loosely connected group includes adrenaline, nicotine, and fentanyl. These two clusters of molecules are related through the connection between methadone and nicotine.

In Table 4 the correlation coefficients (*R*) between similarity results calculated by *Ja*, *JT*, *Gle*, *SS*, and *CT* are presented. The similarities obtained by these coefficients are highly correlated. The highest linear correlation is observed between *Ja* and *Gle* with $R = 0.9981$, while the lowest correlation, $R = 0.9545$, is between *SS* and *CT*. Such a good correlation between *Ja* and *Gle* comes from the fact that these indices differ only by the weights they set on the same features, i.e., on the bits 1 present in the same place in two fingerprints.

To get better insight into obtained similarity results, we have employed the SRD statistical procedure. As described in Sect. 2.2, this tool enables us to compare similarity coefficients. Since the reference value is not available, the average value for all five indices has been used as an “ideal” standard for each molecule, for ranking purposes. The calculated SRD values, of every similarity coefficient, are presented in Table 3. The *Ja*, *JT*, *Gle*, and *SS* indices yielded the same SRD values, while the SRD for the *CT* was 92. This finding shows that our coefficients are useful for the similarity assessment of fingerprints based on physico-chemical properties, especially the first four indices since they produced SRD values that are close to zero. Also, these results reveal

Table 3 The results of chemical similarity analysis of compounds from Fig. 1 by five different metrics

Coefficient	Min	Max	Mean	*s*	SRD	Scaled SRD	Ext. indices
Ja	14	77	48	14	6	0.1972	50
JT	5	53	25	10	6	0.1972	50
Gle	10	69	38	13	6	0.1972	50
SS	3	36	14	7	6	0.1972	50
CT	23	80	57	11	92	3.0243	63

The minimum (min), maximum (max), mean, and standard deviation (*s*) values in %, were obtained for each similarity coefficient. The SRD stands for the sum of ranking differences value, the scaled SRD data are scaled between 0 and 100 and they are given in %. The ext. indices column shows similarity among all molecules (in %), which is calculated by the extended forms of similarity measures [42, 43] that *Ja*, *JT*, *Gle*, and *SS* show similar performance, compared to *CT*, which is in accordance with their definitions but cannot be concluded from the previous results. It is interesting to note that *Ja*, *JT*, *Gle*, and *SS* indices produce different rankings in the case of interaction fingerprints in virtual screening scenarios [23]. The validation of the SRD procedure has been carried out by performing the comparison of ranks with 78 random numbers (CRRN). This is a randomization test that gives a distribution of SRD values with randomized ranks. Based on this validation, it can be concluded whether the SRD value characterizing a coefficient overlap with the use of random numbers (in that case, the coefficient is statistically not distinguishable from randomly assigned ranks). The obtained results are depicted in Fig. 4 with a magnified view.

As can be seen, the scaled SRD of similarity coefficients is extremely low, compared to random SRD. Most importantly, there is no overlap between the left side (real numbers) and the right side (random numbers). The location of the scaled SRD for similarity coefficients (located between 0 and 4) is far from the SRD for random numbers (located between 50 and 81). It can be concluded that the probability that the real variables are random is negligible.

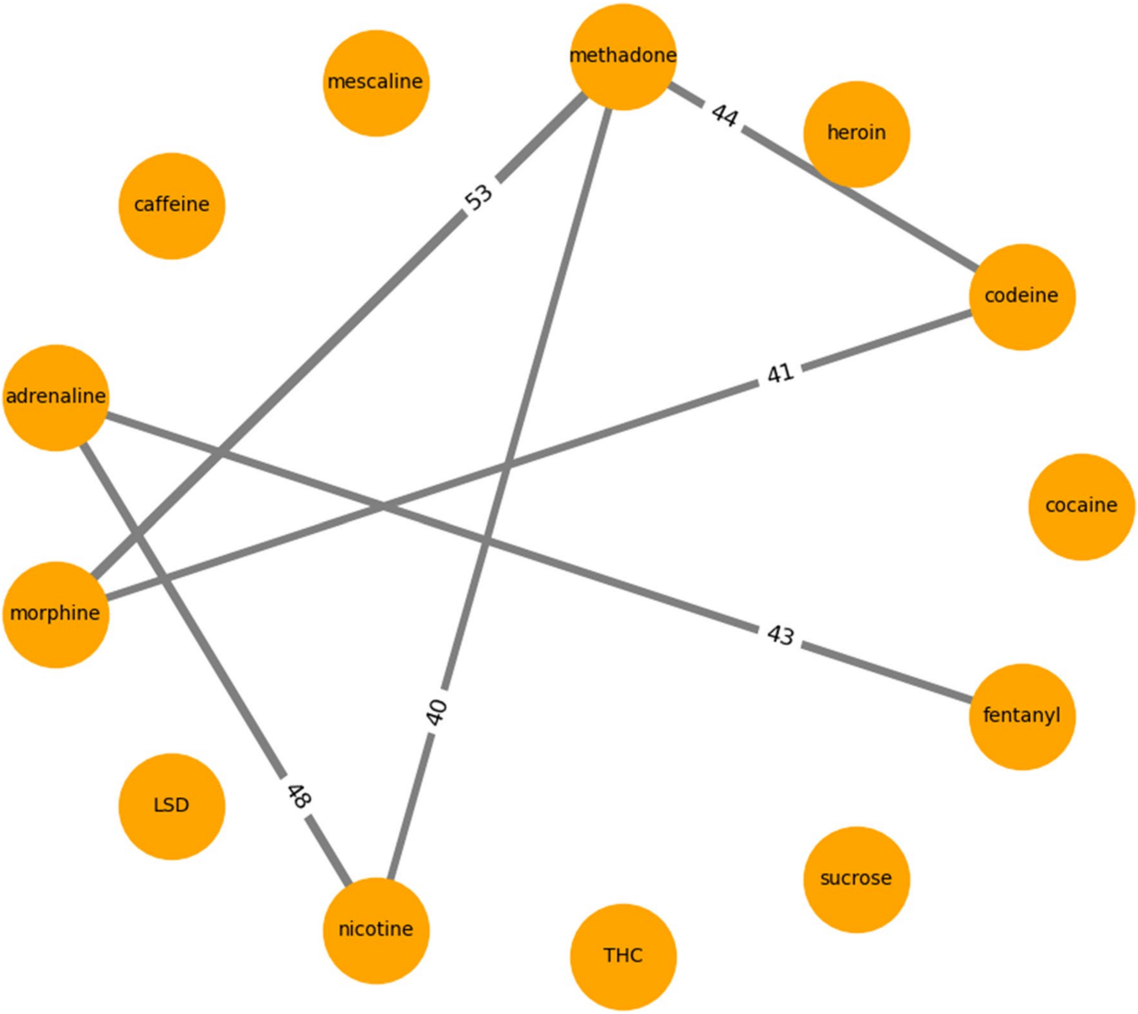


Fig. 3 The chemical similarity of compounds calculated by the Jaccard-Tanimoto similarity coefficient. Note that the graph is constructed to reflect the similarity of molecules in the cases where their amounts ≥ 40%. The edges are labeled with the percentage of similarity between two compounds

Table 4 The Pearson correlation coefficients between similarity values computed by five different metrics

	Ja	*JT*	*Gle*	*SS*	*CT*
Ja	1.0000				
JT	0.9879	1.0000			
Gle	0.9981	0.9955	1.0000		
SS	0.9741	0.9973	0.9860	1.0000	
CT	0.9931	0.9717	0.9872	0.9545	1.0000

Conclusion

Chemical similarity is an important aspect of similarity between two molecules. Here, we have examined the chemical similarity of 13 compounds with the physiological response. To do so, we have developed plain binary fingerprints based on physico-chemical properties, i.e., on melting point, log*P*, and pKa. The Jaccard, Jaccard-Tanimoto, Gleason, Sokal-Sneath, and the Consonni-Todeschini similarity coefficients have been used to calculate the similarity of fingerprints. It was found that adrenaline on average is the most similar to other molecules, while LSD and THC are the least similar to other compounds. All five similarity coefficients have found that morphine and methadone are chemically the most similar compounds, while cocaine and caffeine show the lowest chemical similarity. The sum of ranking differences statistical procedure has shown that applied similarity indices can be successfully used for similarity analysis of developed binary fingerprints. The advantage of the applied methodology is that it summarizes the information on the physico-chemical features in a simple and straightforward way, which enables the calculation of the chemical similarity of the compounds. Therefore, this approach is a useful tool that can provide information on the amount of chemical similarity of molecules only using several available physico-chemical properties.

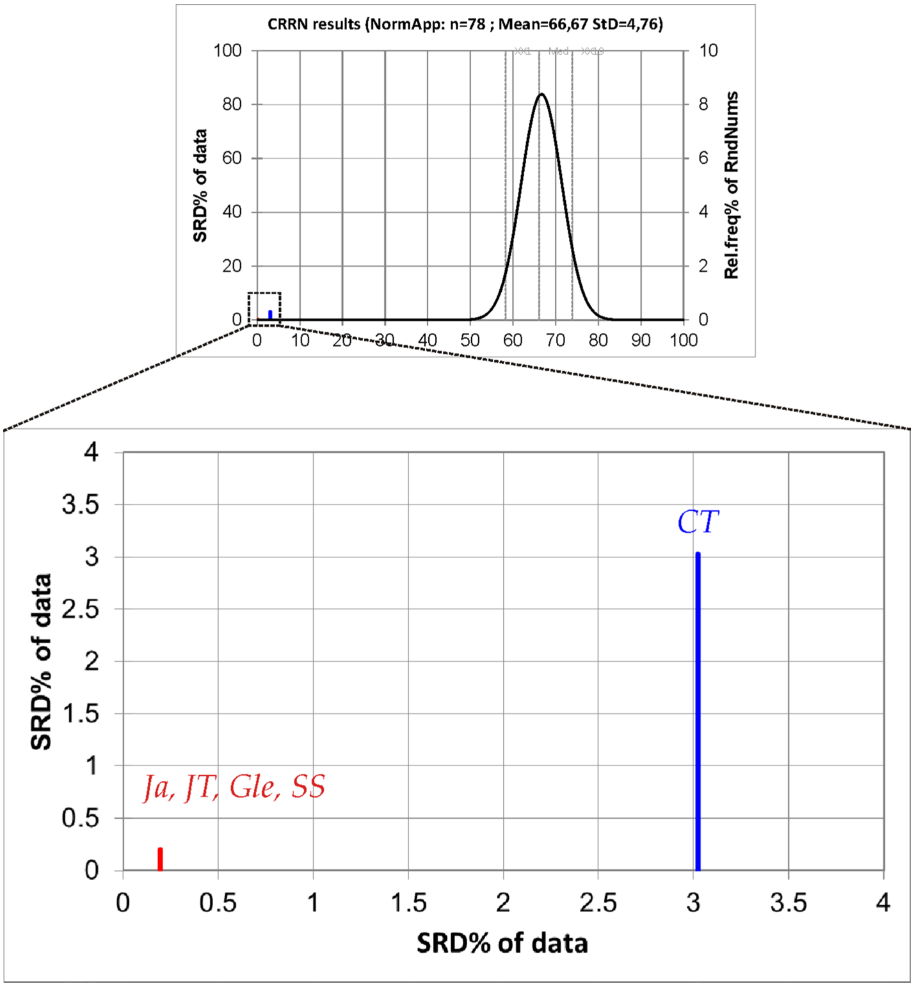


Fig. 4 *X* and left *Y* axes: The percentage of scaled SRD for different similarity coefficients (scaled between 0 and 100, i.e., put on the same scale as the random numbers). The scaled SRD for *Ja*, *JT*, *Gle*, and *SS* is 0.1972% (red) and for *CT* is 3.0243% (blue). Right *Y*-axis: The frequencies of random SRD are plotted (the black curve corresponds to random SRD distribution)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-022-10514-5.

Acknowledgements This work was supported by the Serbian Ministry of Education, Science and Technological Development (Grant No. 451-03-68/2022-14/200122).

Data availability The data used in this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Tversky A (1977) Features of similarity. Psychol Rev 84:327–352. https://doi.org/10.1037/0033-295X.84.4.327
2. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2:3204–3218. https://doi.org/10.1039/B409813G
3. Maldonado AG, Doucet JP, Petitjean M, Fan B-T (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10:39–79. https://doi.org/10.1007/s11030-006-8697-1
4. Dean PM (1995) Defining molecular similarity and complementarity for drug design. In: Dean PM (ed) Molecular similarity in drug design. Springer, Dordrecht, pp 1–23. https://doi.org/10.1007/978-94-011-1350-2_1
5. Coley CW, Rogers L, Green WH, Jensen KF (2017) Computer-assisted retrosynthesis based on molecular similarity. ACS Cent Sci 3:1237–1245. https://doi.org/10.1021/acscentsci.7b00355
6. Liu Y, Cao Y, Lai W, Yu T, Ma Y, Ge Z (2021) A strategy for predicting the crystal structure of energetic N-oxides based on molecular similarity and electrostatic matching. CrystEngComm 23:714–723. https://doi.org/10.1039/D0CE01501F
7. Krasowski MD, Pizon AF, Siam MG, Giannoutsos S, Iyer M, Ekins S (2009) Using molecular similarity to highlight the challenges of routine immunoassay-based drug of abuse/toxicology screening in emergency medicine. BMC Emerg Med 9:5. https://doi.org/10.1186/1471-227X-9-5
8. Krasowski MD, Drees D, Morris CS, Maakestad J, Blau JL, Ekins S (2014) Cross-reactivity of steroid hormone immunoassays: clinical significance and two-dimensional molecular similarity prediction. BMC Clin Pathol 14:13. https://doi.org/10.1186/1472-6890-14-33
9. Martin RL, Willems TF, Lin L-C, Kim J, Swisher JA, Smit B, Haranczyk M (2012) Similarity-driven discovery of zeolite materials for adsorption-based separations. ChemPhysChem 13:3595–3597. https://doi.org/10.1002/cphc.201200554
10. Rouvray D (1990) The evolution of the concept of molecular similarity. In: Johnson MA, Maggiora GM (eds) Concepts and applications of molecular similarity. Wile, New York. ISBN: 978-0-471-62175-1
11. Maggiora GM (2006) On outliers and activity cliffs – why QSAR often disappoints. J Chem Inf Model 46:1535. https://doi.org/10.1021/ci060117s
12. Guha R, Van Drie JH (2008) Structure−activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48:646–658. https://doi.org/10.1021/ci7004093
13. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry: miniperspective. J Chem Inf Model 55:2932–2942. https://doi.org/10.1021/jm201706b
14. Medina-Franco JL (2013) Activity cliffs: facts or artifacts? Chem Biol Drug Des 81:553–556. https://doi.org/10.1111/cbdd.12115
15. Xue L, Bajorath J (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb Chem High Throughput Screen 3:363–372. https://doi.org/10.2174/1386207003331454
16. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim. https://doi.org/10.1002/9783527613106
17. Engel T, Gasteiger J (2018) Applied chemoinformatics: achievements and future opportunities. Wiley, Weinheim
18. Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 96:1027–1044. https://doi.org/10.1021/cr950202r
19. Dearden JC, Cronin MTD, Kaiser KLE (2009) How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ Res 20:241–266. https://doi.org/10.1080/10629360902949567
20. Willett P (2011) Similarity searching using 2D structural fingerprints. In: Bajorath J (ed) Chemoinformatics and computational chemical biology. Humana, Totowa, pp 133–158. https://doi.org/10.1007/978-1-60761-839-3_5
21. O'Boyle NM, Sayle RA (2016) Comparing structural fingerprints using a literature-based similarity benchmark. J Cheminform 8:6. https://doi.org/10.1186/s13321-016-0148-0
22. Deng Z, Chuaqui C, Singh J (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein−ligand binding interactions. J Med Chem 47:337–344. https://doi.org/10.1021/jm030331x
23. 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
24. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
25. Kubinyi H (1998) Similarity and dissimilarity: a medicinal chemist's view. Perspect Drug Discov Des 9:225–252. https://doi.org/10.1023/A:1027221424359
26. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? J Med Chem 45:4350–4358. https://doi.org/10.1021/jm020155c
27. Boström J, Hogner A, Schmitt S (2006) Do structurally similar ligands bind in a similar fashion? J Med Chem 49:6716–6725. https://doi.org/10.1021/jm060167o
28. Xenides D, Fostiropoulou D, Vlachos DS (2020) A metric space approach on the molecular vs. chemical similarity of some analgesic and euphoric compounds. MATCH Commun Math Comput Chem 83:261–284
29. Kaiko RF, Kanner R, Foley KM, Wallenstein SL, Canel AM, Rogers AG, Houde RW (1987) Cocaine and morphine interaction in acute and chronic cancer pain. Pain 31:35–45. https://doi.org/10.1016/0304-3959(87)90004-2
30. Van Soeren M, Mohr T, Kjaer M, Graham TE (1996) Acute effects of caffeine ingestion at rest in humans with impaired epinephrine responses. J Appl Physiol 80:999–1005. https://doi.org/10.1152/jappl.1996.80.3.999
31. Parrott AC (2015) Why all stimulant drugs are damaging to recreational users: an empirical overview and psychobiological explanation. Hum Psychopharmacol 30:213–224. https://doi.org/10.1002/hup.2468
32. Graziane NM, Sun S, Wright WJ, Jang D, Liu Z, Huang YH, Nestler EJ, Wang YT, Schlüter OM, Dong Y (2016) Opposing mechanisms mediate morphine-and cocaine-induced generation

of silent synapses. Nat Neurosci 19:915–925. https://doi.org/10.1038/nn.4313

33. PubChem, https://pubchem.ncbi.nlm.nih.gov/, Accessed 10 Dec 2021
34. DrugBank, https://go.drugbank.com/, Accessed 10 Dec 2021
35. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P (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
36. Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytol 11:37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x
37. Rogers DJ, Tanimoto TT (1960) A computer program for classifying plants. Science 132:1115–1118. https://doi.org/10.1126/science.132.3434.1115
38. Gleason HA (1920) Some applications of the quadrat method. Bull Torrey Bot Club 47:21–33. https://doi.org/10.2307/2480223
39. Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. W. H. Freeman and Co., London
40. Consonni V, Todeschini R (2012) New similarity coefficients for binary data. MATCH Commun Math Comput Chem 68:581–592
41. RDKit: Open-source cheminformatics, http://www.rdkit.org.
42. Miranda-Quintana Alain R, 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
43. Miranda-Quintana Alain R, Bajusz D, Rácz A, 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
44. Héberger K (2010) Sum of ranking differences compares methods or models fairly. Trends Anal Chem 29:101–109. https://doi.org/10.1016/j.trac.2009.09.009
45. Rácz A, Bajusz D, Héberger K (2015) Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters. SAR QSAR Environ Res 26:683–700. https://doi.org/10.1080/1062936X.2015.1084647
46. West C, Khalikova MA, Lesellier E, Héberger K (2015) Sum of ranking differences to rank stationary phases used in packed column supercritical fluid chromatography. J Chromatogr A 1409:241–250. https://doi.org/10.1016/j.chroma.2015.07.071
47. Vastag G, Apostolov S, Perišić-Janjić N, Matijević B (2013) Multivariate analysis of chromatographic retention data and lipophilicity of phenylacetamide derivatives. Anal Chim Acta 767:44–49. https://doi.org/10.1016/j.aca.2013.01.002
48. 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
49. Moreira de Barros GA, Baradelli R, Rodrigues DG, Toffoletto O, Domingues FS, Gayoso MV, Lopes A, Afiune JB, Guimarães GMN (2021) Use of methadone as an alternative to morphine for chronic pain management: a noninferiority retrospective observational study. PAIN Rep 6:e979. https://doi.org/10.1097/PR9.0000000000000979
50. Goldsack C, Scuplak SM, Smith M (1996) A double-blind comparison of codeine and morphine for postoperative analgesia following intracranial surgery. Anaesthesia 51:1029–1032. https://doi.org/10.1111/j.1365-2044.1996.tb14997.x
51. Dixon WE, Hoyle JC (1929) Studies in the pulmonary circulation: II. The action of adrenaline and nicotine. J Physiol 67:77–86. https://doi.org/10.1113/jphysiol.1929.sp002554

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.