Maximum Common Property: A New Approach for Molecular Similarity

Aurelio Antelo Collado
University of Informatics Science https://orcid.org/0000-0001-7532-0736
Ramón Carrasco-Velar
University of Informatics Science https://orcid.org/0000-0003-1318-6687
Nicolás García-Pedrajas
University of Cordoba https://orcid.org/0000-0002-4488-6849
Gonzalo Cerruela-García (gcerruela@uco.es)
Cordoba University https://orcid.org/0000-0001-9140-3347

Research article

Keywords: Maximum Common Property, Electrotopographic State Index, Molecular similarity, Tanimoto function, Maximum Common Structure

Posted Date: April 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-22241/v1

RESEARCH

Maximum Common Property: A New Approach for Molecular Similarity

Aurelio Antelo Collado[1], Ramón Carrasco Velar[2*], Nicolás García-Pedrajas[3] and Gonzalo Cerruela-García[4]

Abstract

The maximum common property similarity (MCPhd) method is presented using descriptors as a new approach to determine the similarity between two chemical compounds or molecular graphs. This method uses the concept of maximum common property arising from the concept of maximum common substructure and is based on the electrotopographic state index for atoms. A new algorithm to quantify the similarity values of chemical structures based on the presented maximum common property concept is also developed in this paper. To verify the validity of this approach, the similarity of a sample of compounds with antimalarial activity is calculated and compared with the results obtained by the small molecule subgraph detector (SMSD) method. The results obtained by the MCPhd method differ significantly from those obtained by the SMSD method, improving the quantification of the similarity. A major advantage of the proposed method is that it helps to understand the analogy or proximity between physicochemical properties of the molecular fragments or subgraphs compared with the biological response or biological activity. In this new approach, more than one property can be potentially used. The method can be considered a hybrid procedure because it combines descriptor and the fragment approaches.

Keywords: Maximum Common Property; Electrotopographic State Index; Molecular similarity; Tanimoto function; Maximum Common Structure

*Correspondence: rcarrasco@uci.cu
[2]University of Informatics Science, Havana, Cuba, ORCID: https://orcid.org/0000-0003-1318-6687
Full list of author information is available at the end of the article
†Equal contributor

Introduction

Molecular similarity is one of the most explored and employed concepts in cheminformatics (chemical informatics or chemoinformatics) [1]. Moreover, it is currently one of the central subjects in medicinal chemistry research [1, 2]. Molecular similarity can be evaluated using different approaches, which can be classified into two principal categories: those based on descriptors and those based on substructures [3]. To estimate similarity among molecules, it is necessary to identify those structural or chemical/physical properties that are useful to correlate and then predict the relationships among them.

Similarity calculations based on molecular descriptors use fingerprint representations [3, 4]. These representations can be codified both by topological or topographic descriptors. Topological descriptors are the most popular because the 2D representation of molecules is computationally less difficult to work with than the 3D representation [1].

This work proposes a different approach in contrast with what is rigorously known as molecular similarity or chemical similarity [1]. The descriptor and the method of reduction of the graph used contain both structural and chemical-physical information. Thus, the approach allows evaluations and comparisons to be made by accounting for not only the structure but also other properties associated with the electrostatic nature of the molecule or fragment. The methods of structural similarity in 2D are more popular and simple. However, when working with only the topology of the molecules, most of the information associated with the spatial distribution is lost, except in the molecules that are essentially flat. As opposed to 2D methods, 3D methods consider that the properties of molecules tend to be strongly associated with the spatial distribution of their atoms [5, 6]. On the other hand, the configuration of minimum energy in a molecular structure with more than one degree of rotational freedom does not condition that this configuration is responsible for the biological response.

This issue causes a dilemma for researchers: losing all three-dimensional information for the sake of simplicity in the calculations or complicating the calculations and possibly delaying the results. The possibility of obtaining large data sets is an unquestionable reality. In that case, the eventual distortion of the 3D results due to not adjusting to the required conformation must be compensated by the increase in the number of compounds. However, such voluminous processing is not currently an impediment in terms of computational cost [7, 8].

Another concept that has been used for more than two decades is the scaffold and, more recently, scaffold hopping. These concepts allow the reduction of the molecule by eliminating R-substituents from the nucleus supposedly responsible for the activity in series of compounds in the first case, and in the other case, they allow the scaffold to be determined and enable comparisons to be made between structurally different compounds [9]. In other words, this approach bears a certain similarity to the proposed method since both seek to identify structurally different compounds that may show similar biological activity.

For these reasons, the proposed similarity method is based on the molecular description with a 3D descriptor that has structural information and on the polarity of the molecular graph or its fragments defined by a chemical graph reduction method.

Furthermore, molecular similarity based on substructure allows obtaining the molecular fragment or common subgraph among pairs of compounds [10, 11]. Several similarity methods have been developed based on a group of algorithms aimed at obtaining the largest common subgraph among a pair of compounds, the maximum common subgraph (MCS) [12, 13]. To quantify the molecular similarity, this method uses the Tanimoto coefficient (Tc_{MCS}) [14, 15].

In this work, we introduce a new concept called maximum common property (MCPhd), inspired by MCS, to quantify the similarity based on substructure, using the electrotopographic state index for atoms ($Sstate_{3D}$) [17], which was developed from its parent electrotopology defined by Kier and Hall [16] from the connectivity matrix of the hydrogen-depleted chemical graph as an atomic descriptor.

The rest of the paper is organized in sections as follows: Materials and Methods describes the dataset and molecular codification, the general procedure and the proposed MCPhd algorithm; Results and Discussion describes the experimental results; and finally, Conclusions presents a summary of this work.

Materials and Methods

Sample used

We employed a set of 4-aminobicyclo[2.2.2]octan-2-yl 4-aminobutanoates (Table 1) reported by Weis *et al.* [18] and evaluated compounds against the multiresistant K-1 strain of Plasmodium falciparum.

Codification of structures

The electrotopographic state index for atoms [17] was used to codify chemical structures. This index is defined by Eq. (1).

$$Sstate_{3D} = I_i + \triangle I_{ij} \tag{1}$$

where $Sstate_{3D}$ is the calculated value of the atom i in the corresponding molecule and I_i is the intrinsic value of the atom i calculated with Eq. (2).

$$I_i = \left[(^2/_N) 2_v + 1 \right] / \delta \tag{2}$$

where N is the principal quantum number of atom I, δ^v is the number of valence electrons in the molecular skeleton (Z^v-h) and δ is the number of σ electrons in the skeleton ($\sigma - h$). For each atom of the molecular skeleton, δ^v is the number of valence electrons, σ is the number of electrons in σ orbitals and h is the number of hydrogen atoms bonded.

$\triangle I_{ij}$ represents the disturbance of the atoms of the environment, which is calculated by Eq. (3).

$$\triangle I_{ij} = \sum (I_i + I_j)/r_{ij}^2 \tag{3}$$

where the sum is over the difference of the intrinsic values of atom i with respect to each one of the other atoms in the molecule and r_{ij}^2 is the Euclidean distance between the analyzed atoms.

Graph reduction

The reduction of the chemical graph is carried out by the method described by Carrasco *et al.* [19], where the descriptor centers (CDs), rings of different orders (Rn), clusters of order 3 and 4 (C3 and C4, respectively), heteroatoms such as halogens, amino, etc. (X), and terminal groups such as methyl (M_3), methylene (M_2) and methyne (M) are defined. Examples of these parameters are shown in Fig. 1. This graph reduction procedure, named CALEDE, is inspired by the procedure developed by Avindon *et al.* [20], where each CD is assigned the total value of $Sstate_{3D}$, quantified as the sum of the value of $Sstate_{3Di}$ of each atom that conforms to it.

Definition of the Maximum Common Property

The maximum common property (MCPhd) between two fully connected and complete (not hydrogen-depleted) G_1 and G_2 chemical graphs is defined as the maximum similarity in the chemical-physical properties represented by the index $Sstate_{3D}$, which exists between subgraphs g_1 and g_2 of the molecular graphs G_1 and G_2, respectively. Both g_1 and g_2 represent the link of at least two CDs that are at a Euclidean distance dE(CD_1,CD_2) from their corresponding centers of mass from pairs of CDs.

To quantify the value of similarity between two compounds using the concept of the maximum common property (MCPhd), the calculation of the similarity of two compounds is assumed using the Tanimoto function or coefficient on the basis of the maximum common substructure called Tc_{MCS} [14, 15]. The Tc_{MCS} for two molecules A and B is defined as:

$$Tc_{MCS} = \frac{|MCS(A,B)|_b}{|A|_b + |B|_b + |MCS(A,B)|_b} \quad (4)$$

where $|A|_b$ is the number of links of A, $|B|_b$ is the number of links of B and $|MCS(A,B)|_b$ is the number of links of the MCS of A and B. If the concept MCPhd is replaced in Eq. (4), it yields:

$$Tc_{MCPhd} = \frac{|MCPhd(A,B)|_b}{|A|_b + |B|_b + |MCPhd(A,B)|_b} \quad (5)$$

where $|A|_b$ is the number of heavy atoms of A, $|B|_b$ the number of heavy atoms of B and $|MCPhd(A,B)|_b$ the smallest number of heavy atoms among the fragments with the highest MCP between A and B.

The Proposed MCPhd Algorithm

Figure 2 shows the algorithm used for the calculation of similarity. The algorithm uses the following parameters: (G_1 and G_2) two compounds or molecules, (u) the similarity threshold, (f) the similarity coefficient and (i) the index used to quantify the similarity. First, we obtain the subgraphs (f_1 and f_2) that have a maximum common property value quantified by the index based on the parameters and similarity coefficient. These subgraphs are obtained by performing the following steps:

1 The index (i) entered as a parameter is calculated for each atom in each G1 and G2 graph using the Chemical Development Kit (CDK) library [22]. Lines 1 and 2 of the algorithm are shown in Fig. 2.

2 The graphs (G_1 and G_2) on CDs are reduced, and the total index value of each one is obtained. Lines 3 and 4 of the algorithm are shown in Fig. 2.

3 The similarity matrix between the CDs obtained from the graphs (G_1 and G_2) is constructed using the similarity coefficient introduced as a parameter, along with the distance matrix between the CDs of each graph (G_1 and G_2) using the Euclidean distance. Line 5 of the algorithm is shown in Fig. 2.

4 The CDs from each graph (G_1 and G_2) that meet the condition that the similarity value must be higher than the similarity threshold (u), entered as a parameter, are selected. Line 5 of the algorithm is shown in Fig. 2.

5 Finally, the CDs of each graph G_1 and G_2 are selected at a distance of 0.15, using the Canberra distance coefficient [23] and the distance matrices of the graphs. For each pair of CDs selected, a list is created where the pairs of CDs that are at a distance less than or equal to 0.15 are stored. Finally, the largest list is selected, and if there are several lists of the same size, one is selected, and all its CDs are repeated in other lists. Line 5 of the algorithm is shown in Fig. 2

Then, with the subgraphs (f_1 and f_2) and graphs (G_1 and G_2) obtained, the values of the variables needed to quantify the similarity are obtained using the similarity coefficient (u) for discrete data entered as a parameter. Variable c is assigned the smallest number of heavy atoms belonging to the subgraphs (f_1 and f_2), while variables a and b are assigned the number of heavy atoms belonging to each graph (G_1 and G_2), respectively. Finally, these values are substituted in the similarity function to obtain the quantification of the similarity of the graphs (G_1 and G_2). Lines 6 to 16 of the algorithm are shown in Fig. 2.

The use of the algorithm is exemplified below using the molecules 6k and 6c present in the dataset as shown in Fig. 3 and 4, respectively. We use 5 parameters (G_1, G_2, u, f, i) for its operation, where G_1 and G_2 are the molecular graphs 6k and 6c respectively, (i) is the index ($Sstate_{3D}$), (u) is the similarity threshold, and (f) is the similarity function. For this example, we will use 0.95 and the modified Tanimoto coefficient (Tc_{MCPhd}) as the threshold and similarity function, respectively. Then, after assigning the parameters, the following steps are performed:

1 The $Sstate_{3D}$ index is calculated for each atom present in molecules 6k and 6c; these results are shown in Tables 2 and 3.

2 The 6k and 6c molecular graphs on CDs are reduced, and each is given the value of the total $Sstate_{3D}$ index. As shown in step A of Fig. 5, molecule 6k is reduced on the CDs ($R8_1$, $R5_2$, $R6_3$, $R6_4$, $R6_5$, $C3_6$, X_7), while molecule 6c is reduced on ($R6_1$, $R8_2$, $R6_3$, $R6_4$, $R6_5$, $C3_6$, M_7).
3 The similarity matrix between the CDs of each molecule 6k and 6c is constructed using the Tanimoto coefficient (Tc) for continuous data, together with the distance matrices between the CDs of each molecule (6k and 6c), as shown in step B of Fig. 5.
4 CDs are selected from each molecule (6k and 6c) that meet the condition that the similarity value is above the similarity threshold of 0.95. The CDs selected from molecules 6k and 6c are ($R8_1$, $R5_2$, $R6_3$, $R6_4$, $R6_5$ and $C3_6$) and ($R6_1$, $R8_2$, $R6_3$, $R6_4$, $R6_5$, $C3_6$ and $M3_7$), respectively, as shown in step C-a in Fig. 5. Furthermore, using the distance matrices of the graphs obtained in the previous step, for each pair of CDs, a list is constructed with the pairs of CDs that are at a Canberra distance less than or equal to 0.15, as shown in step C-b in Fig. 5.
5 From the lists of CD pairs obtained in the previous step, the following CDs are selected, namely, ($R8_1$, $R5_2$, $R6_4$ and $C3_6$) and ($R6_1$, $R8_2$, $R6_3$ and $C3_4$), corresponding to the lists (1, 3, 4 and 5) according to the larger size list with the same CDs in common.

Finally, the similarity value of the two molecules 6k and 6c is quantified using the modified Tanimoto coefficient (Tc_{MCPhd}), where the value of $|MCPhd(A,B)|_b$ is the lowest number of heavy bonds present between fragments f_1 and f_2, while the values of $|A|_b$ and $|B|_b$ are obtained from the number of heavy atoms present in molecules 6k and 6c, respectively. With these values, it is possible to quantify the similarity between molecules 6k and 6c. In step E of Fig. 5, it can be seen that the number of heavy atoms of fragments f_1 and f_2 is 23 and 24, respectively, so the value of $|MCPhd(A,B)|_b$ is 23, while the number of heavy atoms of molecules 6k and 6c is 36 and 37, respectively; that is, $|A|_b = 36$ and $|B|_b = 37$. Therefore, the calculated value of similarity between molecules 6k and 6c is 0.46.

Small Molecule Subgraph MCS approach

The Small Molecule Subgraph Detector (SMSD) algorithm differs from previous MCS algorithms in that it uses a combination of several algorithms to find the common maximum subset and filters the results in a way that is chemically relevant because it incorporates chemical knowledge (coincidence of atom type with information sensitive and insensitive to the bond) while searching for molecular similarity. In addition, the algorithm calculates the maximum subgraph common between two molecules (A and B) by combining the power of the VFLibMCS, MCSPlus and CDKMCS algorithms. These algorithms are used on a case-by-case basis, depending on the molecules under consideration for the common maximum subgraph search [24]. This algorithm is implemented in the SMSD tool available free of charge on the official site of the European Institute of Bioinformatics.

General Experimental Procedure

The experiments were carried out as shown in Fig. 6 based on a test of 36 compounds with a 2D structure, which have been tested experimentally in the study conducted by Weis *et al.* in 2014 [18]. The 3D structure of each compound was obtained through the Corina online service [25]. The 2D structures were used to calculate the molecular similarity (all against all) with the SMSD algorithm, while the 3D structures were processed to calculate the $Sstate_{3D}$ index for each atom and to reduce their graphs on CDs in order to apply the MCPhd algorithm to calculate the molecular similarity (all against all). Finally, the results obtained by both algorithms were compared through a screening process, which was evaluated by the percentage of success of finding structures with the same activity. The algorithms used were implemented using the JAVA language and were executed on an Intel(R) Core(TM) i7-7500U PC with 16 GB of RAM.

Results and Discussion

The molecular similarity methods used in this work, SMSD and MCPhd, use different approaches to quantify the similarity between two molecular graphs or molecules. Whereas SMSD uses graph isomorphism as a criterion, the similarity calculated with MCPhd is based on the criterion of analogy or proximity between the physicochemical properties of the molecular fragments or subgraphs that are compared, expressing these properties as an $Sstate_{3D}$ value.

This approach places MCPhd closer to the concepts of bioisosterism than SMSD. Bioisosterism means that two different molecules can provide similar biological responses if the structural aspects are phenomenologically accompanied by a physicochemical property associated with the biological response. This concept was coined by Friedman [26], extended by Burger [27] and recently used by Lassalas *et al.* [28] and Tahirova [29].

Using these two different approaches, different similarity values were obtained. For example, Table 4 shows

the results of the comparison with the remaining 35 molecules of the sample, with compounds 8c and 7j used as target elements since they had the minimum and maximum IC_{50} values, respectively.

To determine whether both methods yielded significantly equal results, the nonparametric statistical test is used for two independent Mann-Whitney [30] samples with a significance level of 5 %. The results of the Mann-Whitney U statistic were 243.00 and 98.00, with a value of p (bilateral asymptotic significance) of 0.00 and 0.00 for the most active compound (8c) and the least active compound (7j), respectively. It is then shown that the results obtained by both methods for both compounds were significantly different.

In addition, the results obtained by both methods had a low correlation; see Fig. 7 for the most active compound (8c) and Fig. 8 for the least active compound (7j).

To analyze the results from another perspective, a similarity function was developed for the dependent variable IC_{50} using the Tanimoto coefficient ($TcIC_{50}$, column in Table 4), and subsequently, the molecular similarities calculated by both methods were correlated with this new variable. As shown in Fig. 9, the similarity results obtained by the MCPhd method for the most active compound (8c) had a slope closer to that obtained with the $TcIC_{50}$ variable; in addition, the results correlated better with a correlation coefficient value of Pearson $r_{xy} = 0.84$ against $r_{xy} = 0.53$, as shown in Figs. 10 and 11, respectively.

Analogous behavior occurred with the similarity results obtained with the less active compound (7j). The slope for MCPhd was nearest to $TcIC_{50}$ than that for SMSD (Fig. 12), and r_{xy} for MCPhd vs $TcIC_{50}$ was greater than SMSD vs $TcIC_{50}$, as seen in Fig. 13 and 14.

To generalize these results, the similarity obtained with both methods of the rest of the 17 compounds selected as active by Baptista [31] was correlated against $TcIC_{50}$. The results showed (Table 5) that compared to the SMSD method, the MCPhd method improves the correlation coefficient in 65% of cases.

To perform a more exhaustive study comparing the molecular similarity results obtained by both methods, the following steps were performed: (1) The similarity is calculated with both methods for all compounds (one against all); (2) the results up to or equal to the similarity thresholds (0.90, 0.80 and 0.70) are selected for each method; and (3) in each method, the threshold with the highest percentage of success in finding structures with the same activity is selected as the best threshold, and its results are compared.

As a result, a threshold of 0.90 was selected for the SMSD method because 65% of structures with the same activity (active-active and inactive-inactive) were found out of 116 pairs; a threshold of 0.70 was selected for the MCPhd method because the percentage of finding structures with the same activity is 67% of 92 pairs found. Tables 6 and 7 show the results for the SMSD and MCPhd methods, respectively, which validate the selection.

If we analyze the results obtained with the best similarity threshold in each method, it can be inferred that the percentage of finding structures with the same activity (active-active and inactive-inactive) obtained using the MCPhd method (67%) was better than the results with SMSD (65%) by 2%. Analyzing only the active-active pairs, the increase was 11% (34% for the SMSD method against 45% for MCPhd). These results suggested, once again, that the MCPhd method improved the similarity results obtained by the SMSD method.

As a last criterion, the 41 (MCPhd) and 39 (SMSD) pairs of compounds classified in the active-active category shown in Tables 6 and 7 were compared for both methods. To do so, relationship graphs (Figs. 15 and 16) were drawn for the 17 compounds present in the 41 and 39 active-active pairs for each method.

Figures 15 and 16 show different behaviors in the relationships between the 17 compounds when the SMSD method was used compared to using the MCPhd method. Thus, for example, whereas SMSD classified compounds by separating them by the different families of bicyclo-butanoates (6, 7 and 8 in Table 1), the MCPhd method put almost all of them in the same cluster. This result implies that the MCPhd method enabled establishing similarity relationships between compounds even from different families. The reason is that in addition to the structural information content provided by the electrotopographic state index for atoms, it includes electrostatic information content.

Conclusions

This work proposed a new approach that uses the 3D structure of molecules with physical-chemical information to estimate the molecular similarity between chemical compounds. The method has been favorably compared with the standard SMSD method and shows

better performance in obtaining structures with the same activity using similarity cutoff values during the screening process. Furthermore, the proposal shows the ability to find similar compounds among different families. This strongly suggest the possibility of employing the MCPhd method for isosteric studies.

Finally, the proposal presented in this paper provides a promising method for extending this method to be used in the construction of QSAR models for molecular activity prediction.

Declarations:

Acknowledgements

Not applicable.

Author's contributions

Authors contributed equally to this work. All authors read and approved the final manuscript.

Funding

This work was supported in part by Project TIN2015-66108-P of the Spanish Ministry of Science and Innovation, by Project 1264182-F of the Andalusian Regional Government and by Project PP2019-Submod-1.2 of the Cordoba University.

Availability of data and materials

The data described in this article are available in Mendeley Data, doi: 10.17632/gnvwyddfnw.2.The algorithm implementation is freely available from the authors upon request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]University of Informatics Science, Havana, Cuba, ORCID: https://orcid.org/0000-0001-7532-0736. [2]University of Informatics Science, Havana, Cuba, ORCID: https://orcid.org/0000-0003-1318-6687. [3]University of Cordoba, Department of Computing and Numerical Analysis, Campus de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain, ORCID: https://orcid.org/0000-0002-4488-6849. [4]University of Cordoba, Department of Computing and Numerical Analysis, Campus de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain, ORCID: https://orcid.org/0000-0001-9140-3347.

References

1. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2013) Molecular Similarity In Medicinal Chemistry. J. Med. Chem 57:3186-3204. doi:10.1021/jm401411z
2. Kunimoto R, Vogt M, Bajorath J (2016) Maximum Common Substructure-Based Tversky Index: An Asymmetric Hybrid Similarity Measure. J. Comput Aided Mol Des 30:523-531. doi:10.1007/s10822-016-9935-y
3. Vogt M, Stumpfe D, Geppert H, Bajorath J (2010) Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, Or A Hopeless Endeavor? Guidelines For Virtual Screening. J. Med. Chem 12:5707-5715. doi:10.1021/jm100492z
4. Gardiner EJ, Holliday JD, O'dowd C, Willett P (2011) Effectiveness of 2D Fingerprints for Scaffold Hopping. Future Med. Chem 3:405-414. doi:10.4155/fmc.11.4
5. Good AC, Richards WG (1998) Explicit Calculation of 3D Molecular Similarity. Perspect. Drug Discovery Des. 9:321-338. doi:10.1023/A:1027280526177
6. Rush TS, Grant JA, Mosyak L, Nicholls A (2005) A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. J. Med. Chem. 48:1489-1495. doi:10.1021/jm040163o
7. Moffat K, Gillet VJ, Whittle M, Bravi G, Leach AR (2008) A Comparison of Field-Based Similarity Searching Methods: CatShape, FBSS, and ROCS. J. Chem. Inf. Model. 48:719-729. doi:10.1021/ci700130j
8. Tresadern G, Bemporad D (2010) Modeling Approaches for Ligand-Based 3D Similarity. Future Med. Chem. 2:1547-1561. doi:10.4155/fmc.10.244.
9. Hu Y, Stumpfe D, Bajorath J (2017) Recent advances in scaffold hopping. J. Med. Chem. 60:1238–1246. doi:10.1021/acs.jmedchem.6b01437
10. Kenny PW, Sadowski J (2005) Structure modification in chemical databases. Methods and Principles in Medicinal Chemistry. Wiley-Vch, Weinheim 23:271-285. doi:10.1002/3527603743.ch11
11. Hussain J, Rea C (2010) Computationally Efficient Algorithm To Identify Matched Molecular Pairs (Mmps) In Large Data Sets. J. Chem Inf Model 50:339-348. doi:10.1021/ci900450m
12. Duesbury E, Holliday JD, Willett P (2017) Maximum Common Subgraph Isomorphism Algorithms. Match Commun. Math. Comput. Chem 77:213-232.
13. Cerruela García G, Luque Ruiz I, Gómez-Nieto MÁ (2004) Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm. Journal of Chemical Information and Computer Sciences. 44:30–41. doi:10.1021/ci034167y
14. Maggiora GM, Shanmugasundaram V (2004) Molecular Similarity Measures. Methods Mol. Biol. 275:1-50. doi:10.1385/1-59259-802-1:001
15. Zhang B, Vogt M, Maggiora GM. Bajorath J (2015) Design Of Chemical Space Networks Using A Tanimoto Similarity Variant Based Upon Maximum Common Substructures. J. Comput Aided Mol Des 29:937-950. doi:10.1007/s10822-015-9872-1
16. Kier LB, Hall LH (1990) An Electrotopological-State Index for Atoms in Molecules. Pharm Res 7:801-807. doi:10.1023/A:1015952613760
17. Carrasco R (2003) New atomic and molecular descriptors. Applications. Editorial Universitaria 35-36.
18. Weis R, Seebacher W, Brun R, Kaiser M, Sat R, Faist J (2013) 4-Aminobicyclo[2.2.2]octan-2-yl 4-aminobutanoates with antiprotozoal activity. Monatsh Chem. doi:10.1007/s00706-013-1116-2.
19. Carrasco R, Prieto JO, Antelo A, Padrón JA, Cerruela G, Maceo ÁL, Alcolea R, Silva LG (2013) Hybrid Reduced Graph For SAR Studies. SAR and QSAR in Environmental Research 24:201-214. doi:10.1080/1062936X.2013.764926
20. Avidon VV, Pomerantsev IA, Golender VE, Rozenblit AB (1982) Structure-activity relationship oriented languages for chemical structure representation. J. Chem. Inf. Comp. Sci 22:207-214.
21. Antelo A, Paneque JL, Hernández MC, Ramón Carrasco R (2016) Molecular Similarity Using Hybrid Indices. Cuban Journal of Medical Informatics 8:487-498.
22. Willighagen EL, Mayfield JW, Alvarsson J et. al. (2017) The Chemistry Development Kit (CDK)V2.0: Atom Typing, Depiction, Molecular Formulas, And Substructure Searching. J Cheminf 9:33. doi:10.1186/s13321-017-022
23. Lance GN, Williams WT (1966) Computer programs for hierarchical polythetic classification ("similarity analysis"). Computer Journal, 9:60–64
24. Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM (2009) Small Molecule Subgraph Detector (SMSD) Toolkitl. J. Cheminform 1:12. doi:10.1186/1758-2946-1-12
25. Fast 3D Structure Generation with CORINA Classic (2020). https://www.mn-am.com/online_demos/corina_demo. Accessed 18 Feb 2020
26. Friedman HL (1951) Influence of Isosteric Replacements upon Biological Activity, National Academy of Sciences-National Research Council 206:295
27. Burger A (1991) Isosterism and bioisosterism in drug design in Progress in Drug Research. 37:287-371. doi:10.1007/978-3-0348-7139-6_7
28. Lassalas P, Oukoloff K, Makani V, James M, Tran V, Yao Y, Huang L, Vijayendran K, Monti L, Trojanowski JQ, Lee VM, Kozlowski MC, Smith III AB, Brunden KR, Ballatore C (2017) Evaluation of Oxetan-3-ol, Thietan-3-ol, and Derivatives Thereof as Bioisosteres of the Carboxylic Acid Functional Group. ACS Med. Chem. 8:864-868. doi:10.1021/acsmedchemlett.7b00212

29. Tahirova N, Poivet E, Xu L, Peterlin Z, Zou DJ, Firestein SS (2019) Bioisosterism reveals new structure-odor relationships, bioRxiv doi:10.1101/567701
30. Mann HB, Whitney DR.(1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, The Annals of Mathematical Statistics. vol. 18:50-60. doi:10.1214/aoms/1177730491
31. Baptista I, Camila Otero C, González S, Pertegás A, Galvez J, García R (2019) Aplicación de la topología molecular al análisis de la actividad antimalárica de 4-Aminobiciclo [2.2.2]Octan-2 il 4-Aminobutanoatos y sus análogos etanoatos y propanoatos. Nereis 11:51-65.

Figures

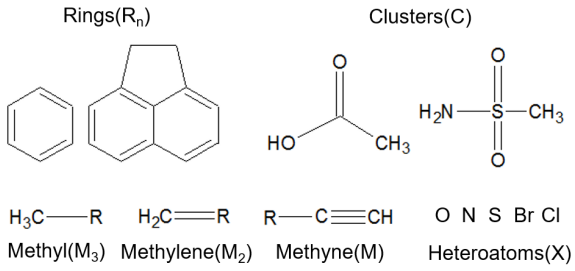


Figure 1 Examples of Descriptor Centers (CD) employed in the fragmentation of the chemical graphs.

```
Algorithm: MCPhd(G1, G2, u, f, i)
  input: Two graph G1 and G2, the similarity thershold (u), the
         coefficient of similarity (f) and index (i)
  exit: Two similar fragments (f1 and f2) and quantification of similarity
1.- G1.calculateTopographicIndices()
2.- G2.calculateTopographicIndices()
3.- reducedGraph graphG1 <- G1.getReducedGraph()
4.- reducedGraph graphG2 <- G2.getReducedGraph()
5.- frag(f1, f2, index) <- getFragmentoPMCCD(graphG1, graphG2, u, f, i)
6.- if f1 !=null and f2 !=null then
7.-     atomf1 <- getHeavyAtomsMCP(f1)
8.-     atomf2 <- getHeavyAtomsMCP(f2)
9.-     c <- min(atomf1, atomf2)
10.-    a <- getHeavyAtoms(G1)
11.-    b <- getHeavyAtoms(G2)
12.-    index <- c/(a+b-c)
13.-    frag(f1, f2, index) <- index
14.-    return frag(f1, f2, index)
15.- else
16.-    return frag(f1, f2, 0)
```

Figure 2 algorithm of the algorithm to calculate similarity.

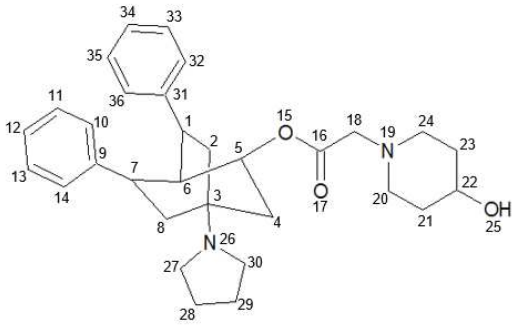


Figure 3 2D graph of the 6k molecule.

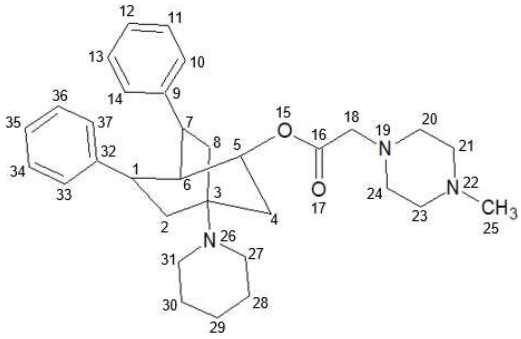


Figure 4 2D graph of the 6c molecule.

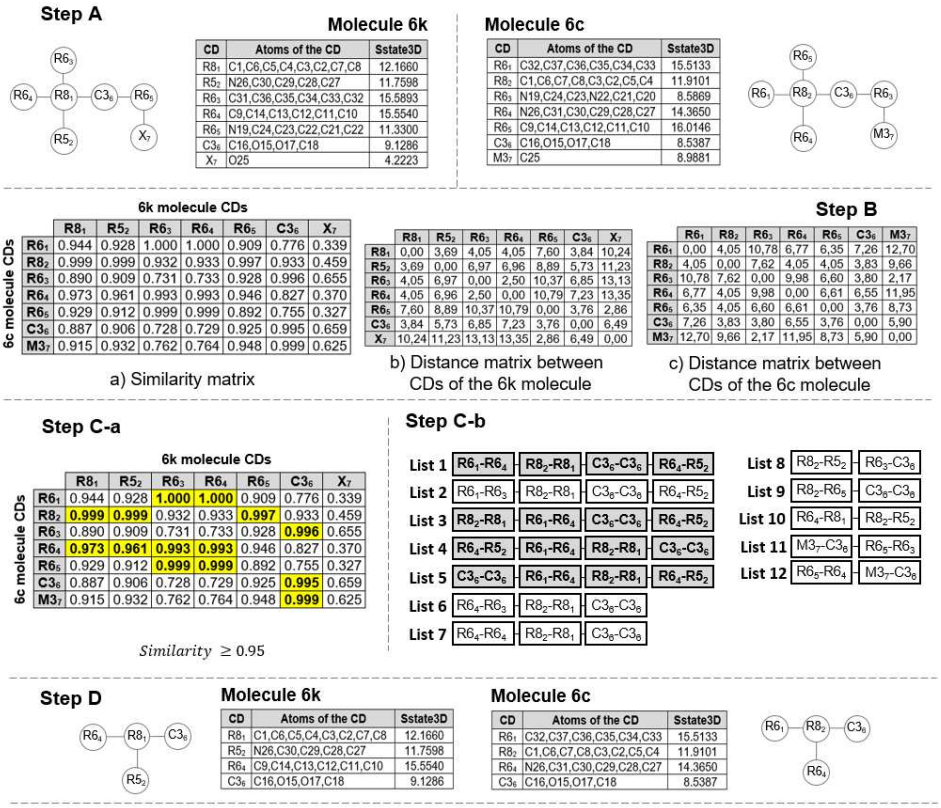


Figure 5 Example of applying the MCPhd algorithm to the 6k and 6c molecules belonging to the dataset.

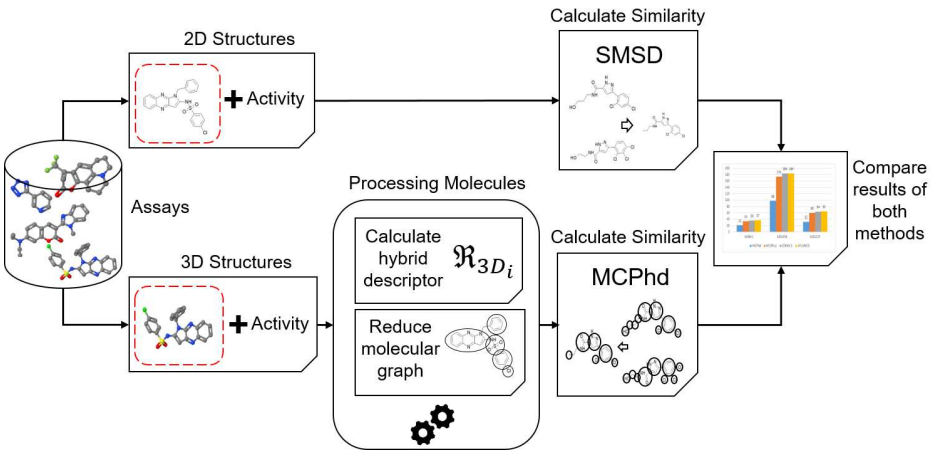


Figure 6 General Experimental procedure.

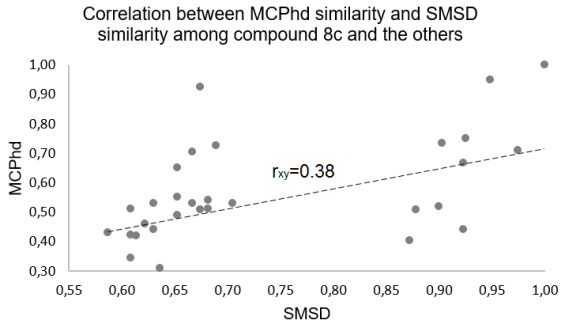


Figure 7 Correlation between MCPhd and SMSD similarity among compound 8c and the others.

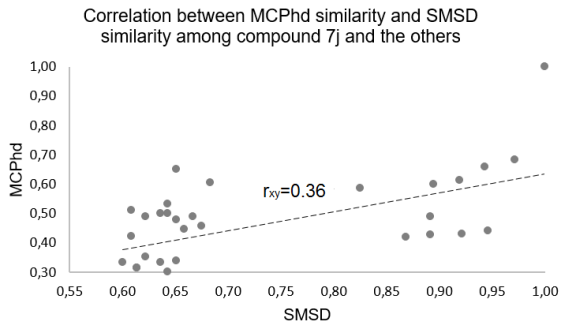


Figure 8 Correlation between MCPhd and SMSD similarity among compound 7j and the others..

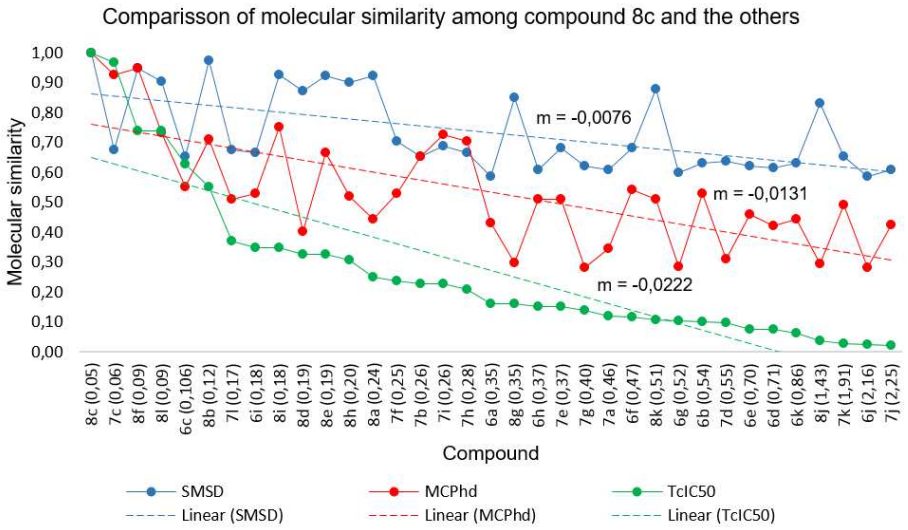


Figure 9 Comparison of molecular similarity among compound 8c and the others..

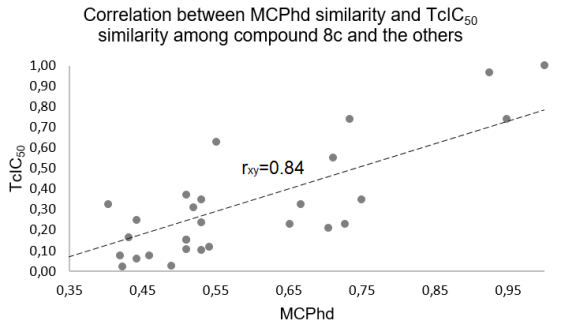


Figure 10 Correlation between MCPhd and $TcIC_{50}$ similarity among compound 8c and the others.

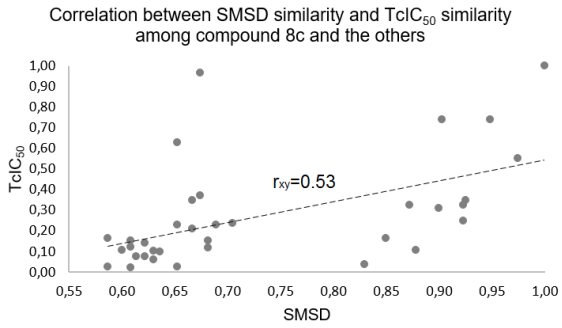


Figure 11 Correlation between SMSD and $TcIC_{50}$ similarity among compound 8c and the others.

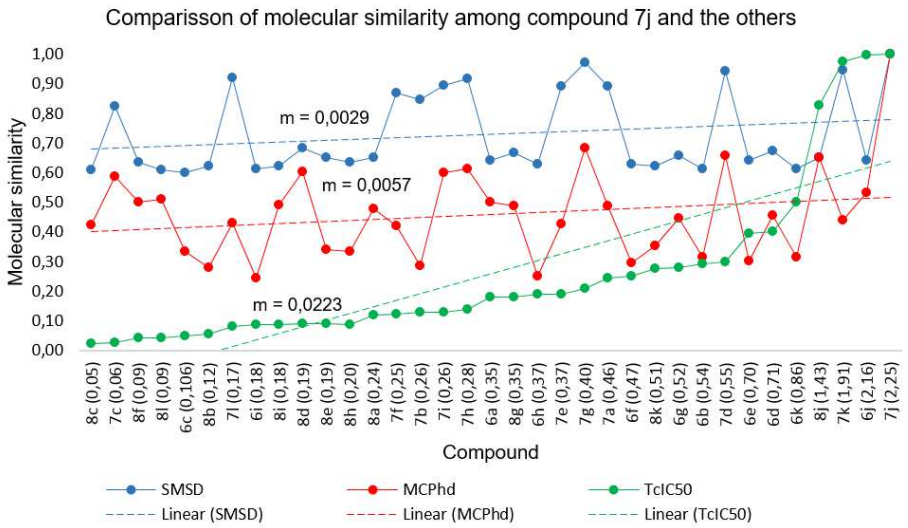


Figure 12 Comparison of molecular similarity among compound 7j and the others.

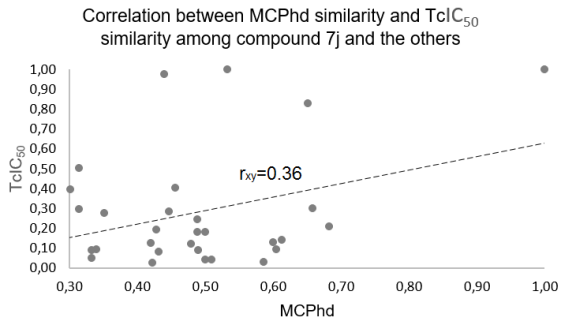


Figure 13 Correlation between MCPhd and $TcIC_{50}$ similarity among compound 7j and the others.

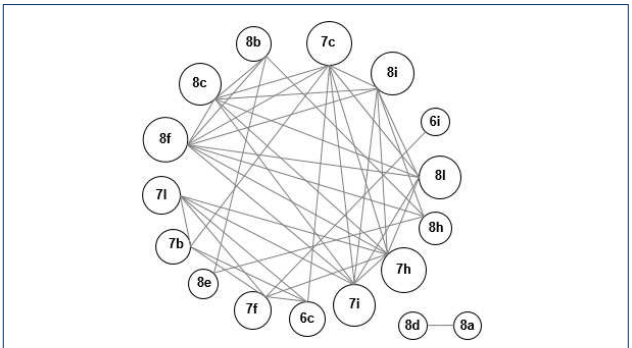


Figure 15 Relationship graph of active compounds obtained by the MCPhd method with a molecular similarity greater than 0.70.

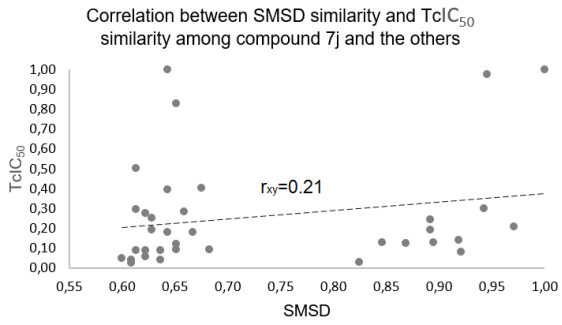


Figure 14 Correlation between SMSD and $TcIC_{50}$ similarity among compound 7j and the others.

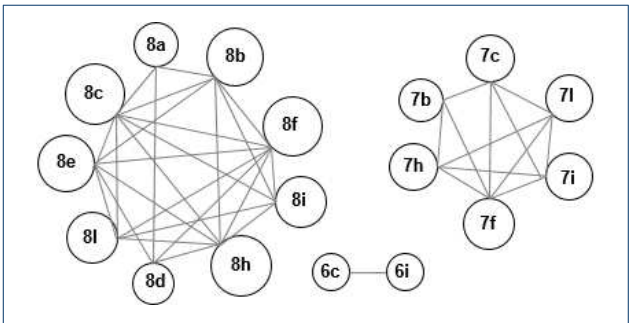
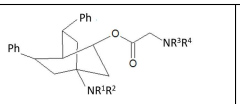
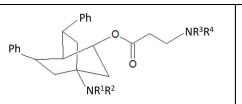
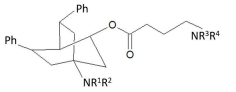
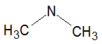
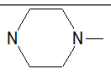
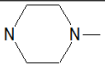
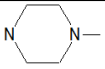
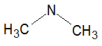
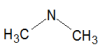
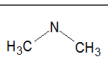
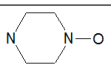
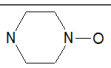
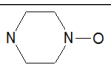


Figure 16 Relationship graph of active compounds obtained by the SMSD method with a molecular similarity greater than 0.90.

Tables

Table 1 Compounds set*

						
a	NR^1R^2		NR^3R^4		6a	IC_{50}= 0.35
					7a	IC_{50}= 0.46
					8a	IC_{50}= 0.24
b	NR^1R^2		NR^3R^4		6b	IC_{50}= 0.54
					7b	IC_{50}= 0.26
					8b	IC_{50}= 0.12
c	NR^1R^2		NR^3R^4		6c	IC_{50}= 0.106
					7c	IC_{50}= 0.06
					8c	IC_{50}= 0.05
d	NR^1R^2		NR^3R^4		6d	IC_{50}= 0.71
					7d	IC_{50}= 0.55
					8d	IC_{50}= 0.19
e	NR^1R^2		NR^3R^4		6e	IC_{50}= 0.70
					7e	IC_{50}= 0.37
					8e	IC_{50}= 0.19
f	NR^1R^2		NR^3R^4		6f	IC_{50}= 0.47
					7f	IC_{50}= 0.25
					8f	IC_{50}= 0.09
g	NR^1R^2		NR^3R^4		6g	IC_{50}= 0.52
					7g	IC_{50}= 0.40
					8g	IC_{50}= 0.35
h	NR^1R^2		NR^3R^4		6h	IC_{50}= 0.37
					7h	IC_{50}= 0.28
					8h	IC_{50}= 0.20
i	NR^1R^2		NR^3R^4		6i	IC_{50}= 0.18
					7i	IC_{50}= 0.26
					8i	IC_{50}= 0.18
j	NR^1R^2		NR^3R^4		6j	IC_{50}= 2.16
					7j	IC_{50}= 2.25
					8j	IC_{50}= 1.43
k	NR^1R^2		NR^3R^4		6k	IC_{50}= 0.86
					7k	IC_{50}= 1.91
					8k	IC_{50}= 0.51
l	NR^1R^2		NR^3R^4		6l	IC_{50}= nd
					7l	IC_{50}= 0.17
					8l	IC_{50}= 0.09

*Taken from reference Weiss

Table 2 Result of the $Sstate_{3D}$ calculation for each atom of the 6k molecule.

Molecule 6k					
Atom	**Number**	**Sstate3D**	**Atom**	**Number**	**Sstate3D**
C	1	0.27653	N	19	-1.4592
C	2	4.24474	C	20	3.38006
C	3	-2.11427	C	21	3.11542
C	4	4.25231	C	22	-0.12941
C	5	0.49396	C	23	3.09432
C	6	0.52371	C	24	3.32881
C	7	0.28273	O	25	4.22234
C	8	4.2063	N	26	-1.30847
C	9	-0.37931	C	27	3.52697
C	10	3.51969	C	28	2.9831
C	11	2.98731	C	29	3.01061
C	12	2.89649	C	30	3.54754
C	13	2.99502	C	31	-0.38433
C	14	3.5348	C	32	3.54001
O	15	-0.08644	C	33	2.99825
C	16	-0.10545	C	34	2.90069
O	17	5.5644	C	35	2.99544
C	18	3.75612	C	36	3.53921

Table 3 Result of the $Sstate_{3D}$ calculation for each atom of the 6c molecule.

Molecule 6c					
Atom	**Number**	**Sstate3D**	**Atom**	**Number**	**Sstate3D**
C	1	0.35527	C	20	3.314
C	2	4.17197	C	21	2.96595
C	3	-2.19195	N	22	-2.22668
C	4	4.25527	C	23	2.94306
C	5	0.37013	C	24	3.25612
C	6	0.50352	C	25	8.9881
C	7	0.21375	N	26	-1.46525
C	8	4.23211	C	27	3.52712
C	9	-0.24053	C	28	2.94719
C	10	3.50889	C	29	2.85191
C	11	3.00512	C	30	2.95634
C	12	2.94959	C	31	3.54763
C	13	3.09722	C	32	-0.13708
C	14	3.69435	C	33	3.3945
O	15	-0.48635	C	34	2.9161
C	16	-0.31876	C	35	2.85145
O	17	5.6053	C	36	2.96557
C	18	3.73854	C	37	3.52272
N	19	-1.66555			

Table 4 Molecular similarity values of the more active and inactive compounds with the rest of the sample.

Molecule 8c			Target	IC_{50}	Molecule 7j		
SMSD	**MCPhd**	**TcIC$_{50}$**			**SMSD**	**MCPhd**	**TcIC$_{50}$**
1.00	1.00	1.00	8c	0.05	0.61	0.42	0.02
0.67	0.93	0.97	7c	0.06	0.83	0.59	0.03
0.95	0.95	0.74	8f	0.09	0.64	0.50	0.04
0.90	0.73	0.74	8l	0.09	0.61	0.51	0.04
0.65	0.55	0.63	6c	0.106	0.60	0.33	0.05
0.97	0.71	0.55	8b	0.12	0.62	0.28	0.06
0.67	0.51	0.37	7l	0.17	0.92	0.43	0.08
0.67	0.53	0.35	6i	0.18	0.61	0.25	0.09
0.93	0.75	0.35	8i	0.18	0.62	0.49	0.09
0.87	0.40	0.33	8d	0.19	0.68	0.60	0.09
0.92	0.67	0.33	8e	0.19	0.65	0.34	0.09
0.90	0.52	0.31	8h	0.20	0.64	0.33	0.09
0.92	0.44	0.25	8a	0.24	0.65	0.48	0.12
0.70	0.53	0.24	7f	0.25	0.87	0.42	0.12
0.65	0.65	0.23	7b	0.26	0.85	0.29	0.13
0.69	0.73	0.23	7i	0.26	0.89	0.60	0.13
0.67	0.70	0.21	7h	0.28	0.92	0.61	0.14
0.59	0.43	0.16	6a	0.35	0.64	0.50	0.18
0.85	0.30	0.16	8g	0.35	0.67	0.49	0.18
0.61	0.51	0.15	6h	0.37	0.63	0.25	0.19
0.68	0.51	0.15	7e	0.37	0.89	0.43	0.19
0.62	0.28	0.14	7g	0.40	0.97	0.68	0.21
0.61	0.35	0.12	7a	0.46	0.89	0.49	0.24
0.68	0.54	0.12	6f	0.47	0.63	0.30	0.25
0.88	0.51	0.11	8k	0.51	0.62	0.35	0.27
0.60	0.29	0.11	6g	0.52	0.66	0.45	0.28
0.63	0.53	0.10	6b	0.54	0.61	0.31	0.29
0.64	0.31	0.10	7d	0.55	0.94	0.66	0.30
0.62	0.46	0.08	6e	0.70	0.64	0.30	0.40
0.61	0.42	0.08	6d	0.71	0.68	0.46	0.40
0.63	0.44	0.06	6k	0.86	0.61	0.31	0.50
0.83	0.29	0.04	8j	1.43	0.65	0.65	0.83
0.65	0.49	0.03	7k	1.91	0.95	0.44	0.97
0.59	0.28	0.02	6j	2.16	0.64	0.53	1.00
0.61	0.42	0.02	7j	2.25	1.00	1.00	1.00

Table 5 Correlation results between both methods for the active compounds in the sample.

Molecule	**IC_{50}**	**Correlation**		
		r_{xy} a*	**r_{xy} b***	**r_{xy} c***
8c	0.05	0.48	0.53	0.80
7c	0.06	0.40	0.07	0.72
8f	0.09	0.58	0.57	0.64
8l	0.09	0.56	0.51	0.60
6c	0.106	0.53	-0.13	0.10
8b	0.12	0.69	0.47	0.53
7l	0.17	0.62	0.07	0.17
6i	0.18	0.60	-0.19	0.14
8i	0.18	0.55	0.37	0.33
8d	0.19	0.74	0.25	0.02
8e	0.19	0.73	0.31	0.30
8h	0.20	0.76	0.26	0.35
8a	0.24	0.65	0.12	0.08
7f	0.25	0.52	0.15	0.17
7b	0.26	0.41	0.15	0.18
7i	0.26	0.51	0.14	0.09
7h	0.28	0.46	0.02	-0.13

a*- SMSD vs MCPhd, b*-SMSD vs TcIC$_{50}$ and c*-MCPhd vs TcIC$_{50}$

Table 6 Comparison of the observed and predicted by SMSD for several similarity thresholds.

Threshold	Real-Predicted	Pairs*	%	Predicted	Pairs*	%
0.90	Active-Active	39	34	Correct	75	65
	Inactive-Inactive	36	31			
	Active-Inactive	19	17	Incorrect	41	35
	Inactive-Active	22	19			
	Total	**116**	**100**	**Total**	**116**	**100**
0.80	Active-Active	52	28	Correct	106	57
	Inactive-Inactive	54	29			
	Active-Inactive	48	26	Incorrect	81	43
	Inactive-Active	33	18			
	Total	**187**	**100**	**Total**	**187**	**100**
0.70	Active-Active	63	30	Correct	121	57
	Inactive-Inactive	58	27			
	Active-Inactive	48	23	Incorrect	91	43
	Inactive-Active	43	20			
	Total	**212**	**100**	**Total**	**212**	**100**

Pairs*-Number of predicted pairs

Table 7 Comparison of the observed and predicted by MCPhd for several similarity thresholds.

Threshold	Real-Predicted	Pairs*	%	Predicted	Pairs*	%
0.90	Active-Active	14	33	Correct	26	62
	Inactive-Inactive	12	29			
	Active-Inactive	8	19	Incorrect	16	38
	Inactive-Active	8	19			
	Total	**42**	**100**	**Total**	**42**	**100**
0.80	Active-Active	16	36	Correct	28	62
	Inactive-Inactive	12	27			
	Active-Inactive	9	20	Incorrect	17	38
	Inactive-Active	8	18			
	Total	**45**	**100**	**Total**	**45**	**100**
0.70	Active-Active	41	45	Correct	62	67
	Inactive-Inactive	23	25			
	Active-Inactive	15	16	Incorrect	30	33
	Inactive-Active	15	16			
	Total	**92**	**100**	**Total**	**92**	**100**

Pairs*-Number of predicted pairs

Figures

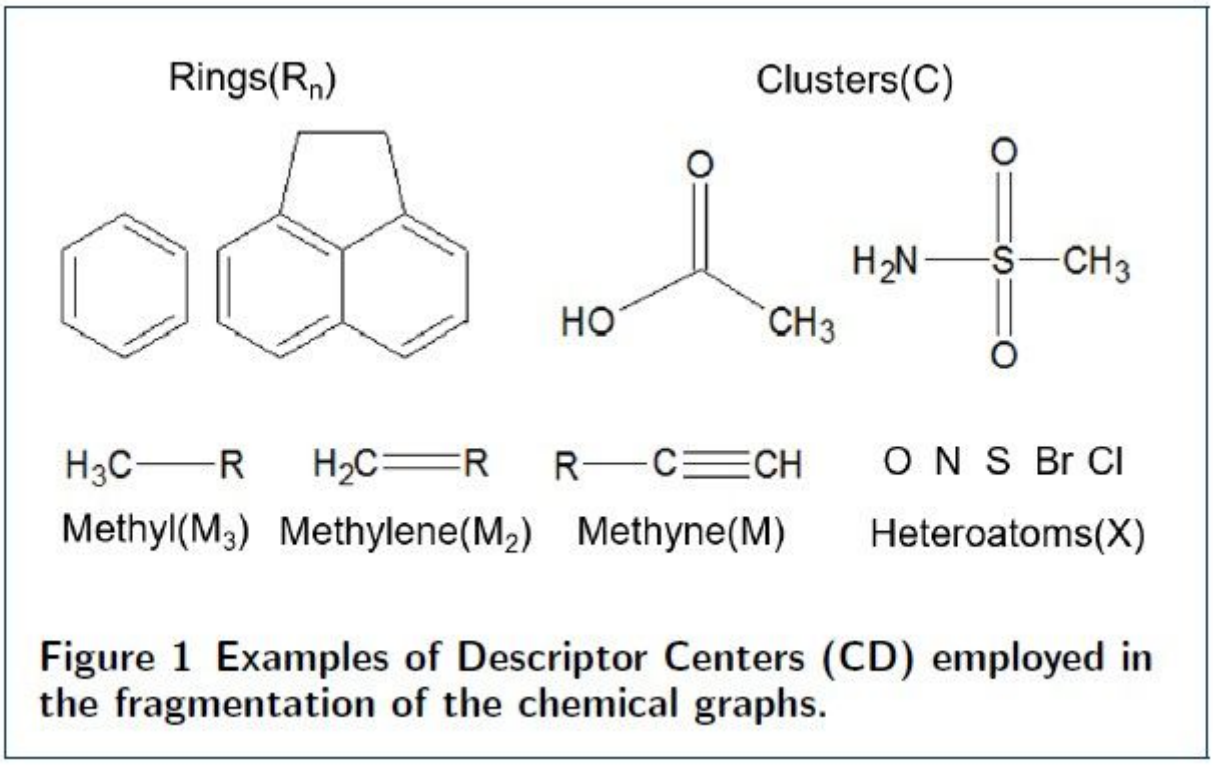
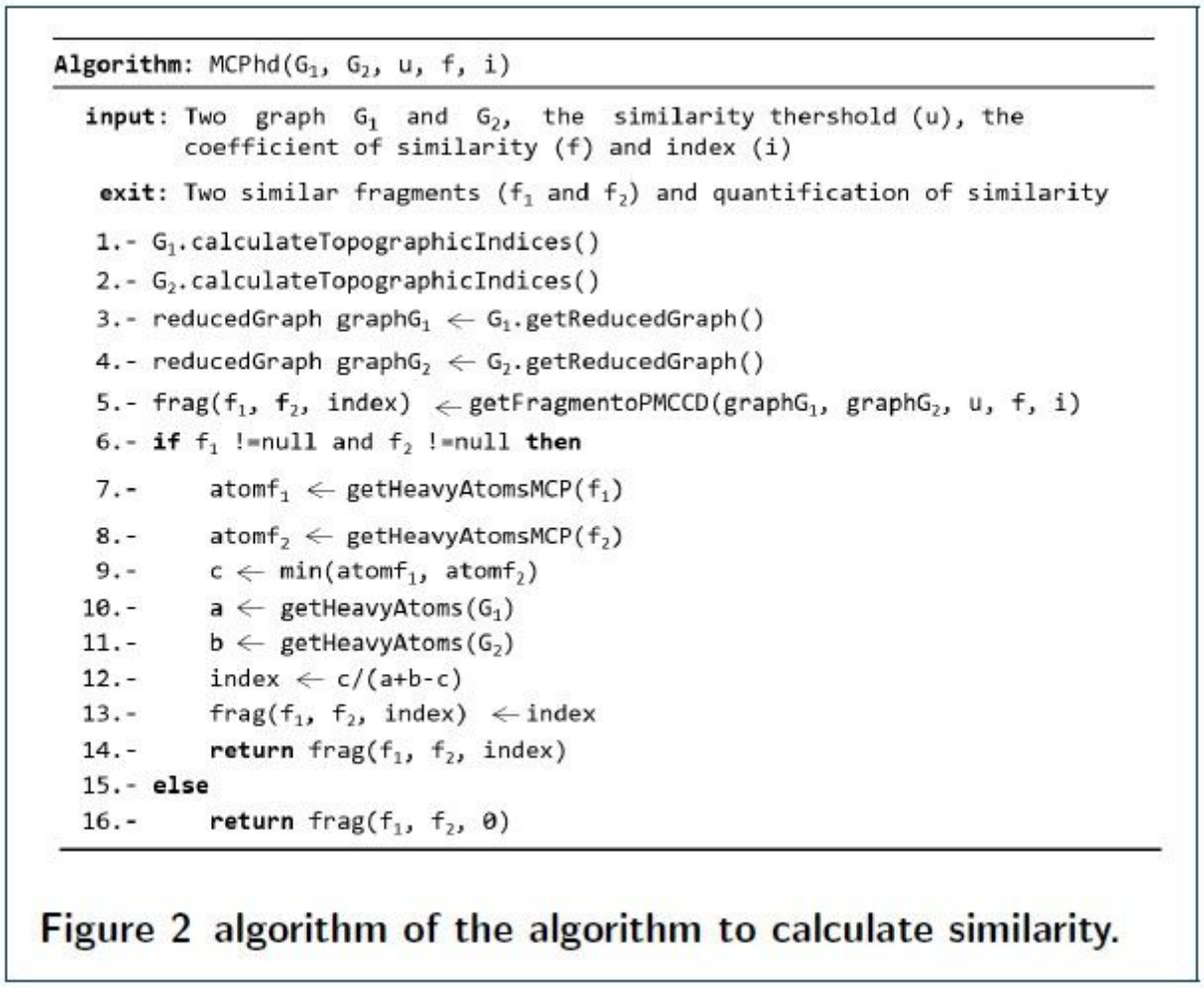


Figure 1 Examples of Descriptor Centers (CD) employed in the fragmentation of the chemical graphs.

Figure 1

Examples of Descriptor Centers (CD) employed in the fragmentation of the chemical graphs.



```
Algorithm: MCPhd(G1, G2, u, f, i)

  input: Two graph G1 and G2, the similarity thershold (u), the
         coefficient of similarity (f) and index (i)

   exit: Two similar fragments (f1 and f2) and quantification of similarity

 1.- G1.calculateTopographicIndices()
 2.- G2.calculateTopographicIndices()
 3.- reducedGraph graphG1 <- G1.getReducedGraph()
 4.- reducedGraph graphG2 <- G2.getReducedGraph()
 5.- frag(f1, f2, index) <- getFragmentoPMCCD(graphG1, graphG2, u, f, i)
 6.- if f1 !=null and f2 !=null then
 7.-     atomf1 <- getHeavyAtomsMCP(f1)
 8.-     atomf2 <- getHeavyAtomsMCP(f2)
 9.-     c <- min(atomf1, atomf2)
10.-     a <- getHeavyAtoms(G1)
11.-     b <- getHeavyAtoms(G2)
12.-     index <- c/(a+b-c)
13.-     frag(f1, f2, index) <- index
14.-     return frag(f1, f2, index)
15.- else
16.-     return frag(f1, f2, 0)
```

Figure 2 algorithm of the algorithm to calculate similarity.

Figure 2

algorithm of the algorithm to calculate similarity.

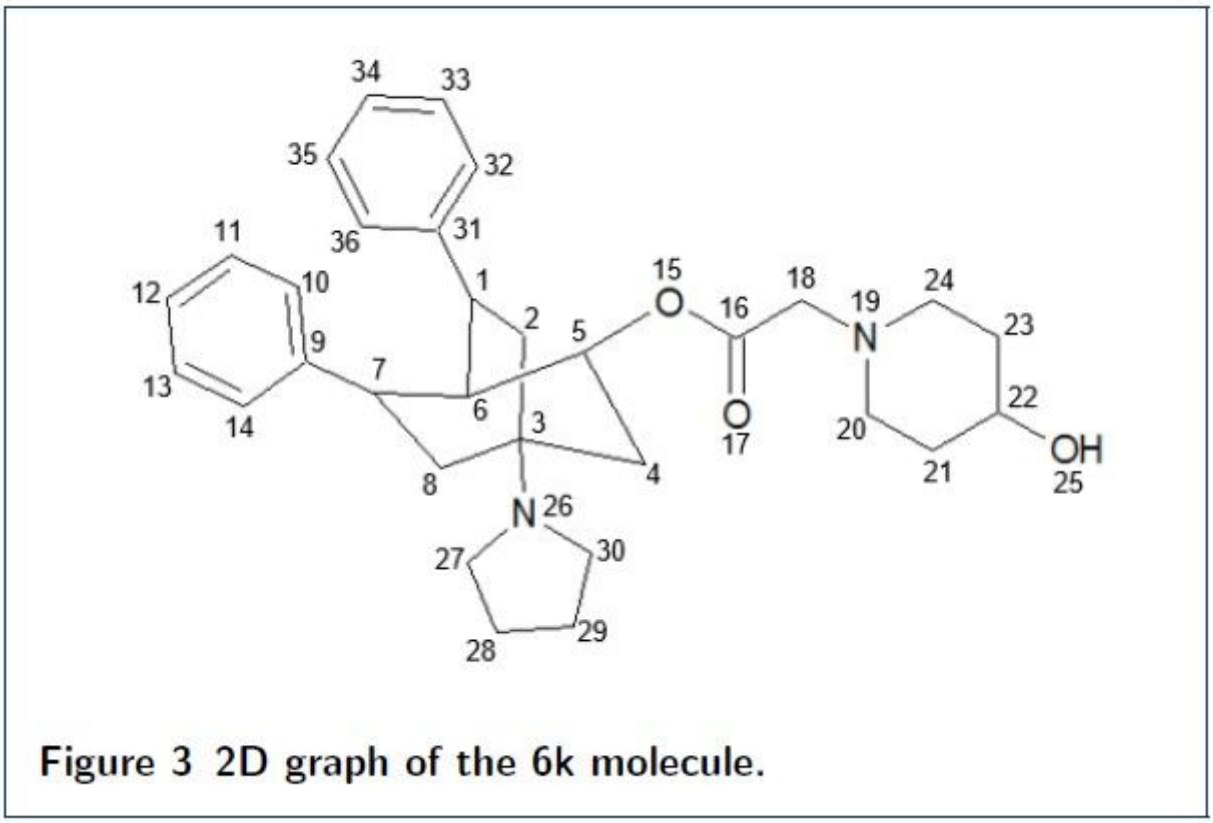


Figure 3 2D graph of the 6k molecule.

Figure 3

2D graph of the 6k molecule.

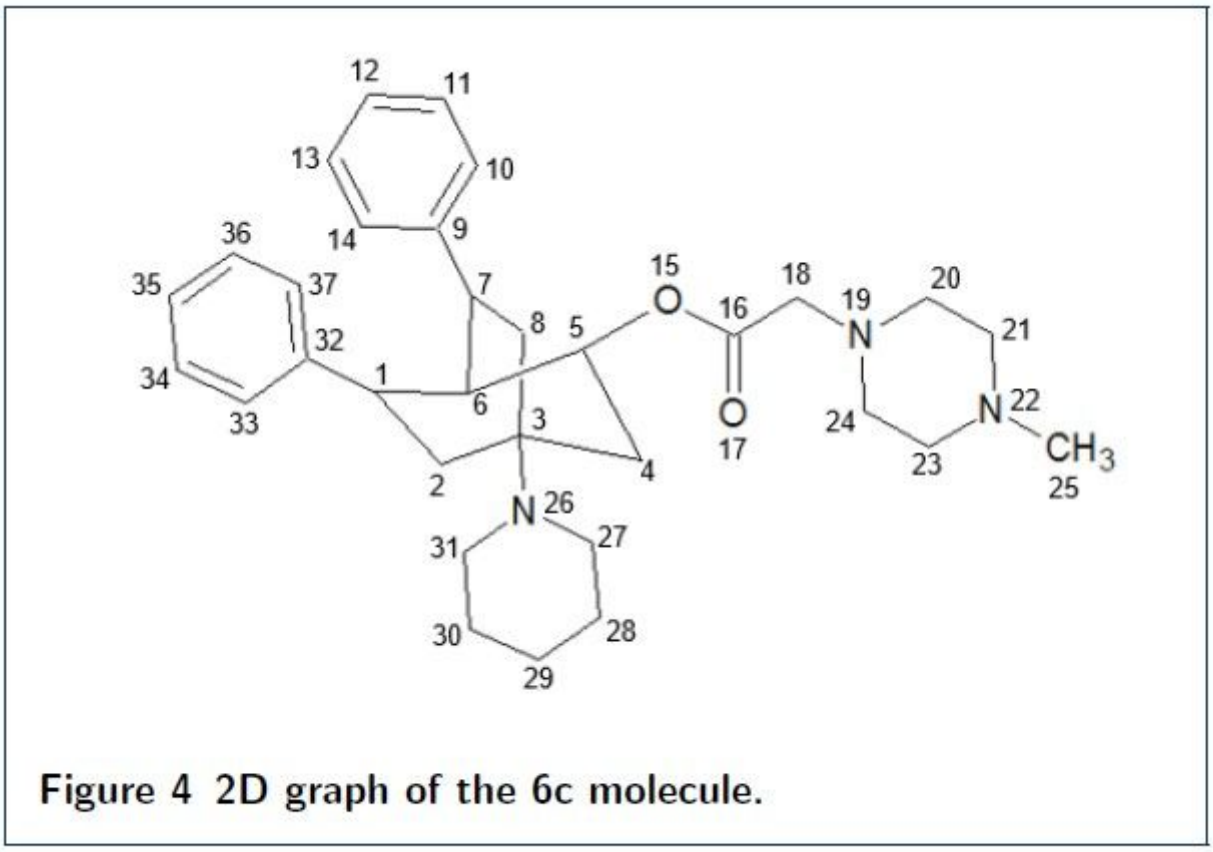


Figure 4 2D graph of the 6c molecule.

Figure 4

2D graph of the 6c molecule.

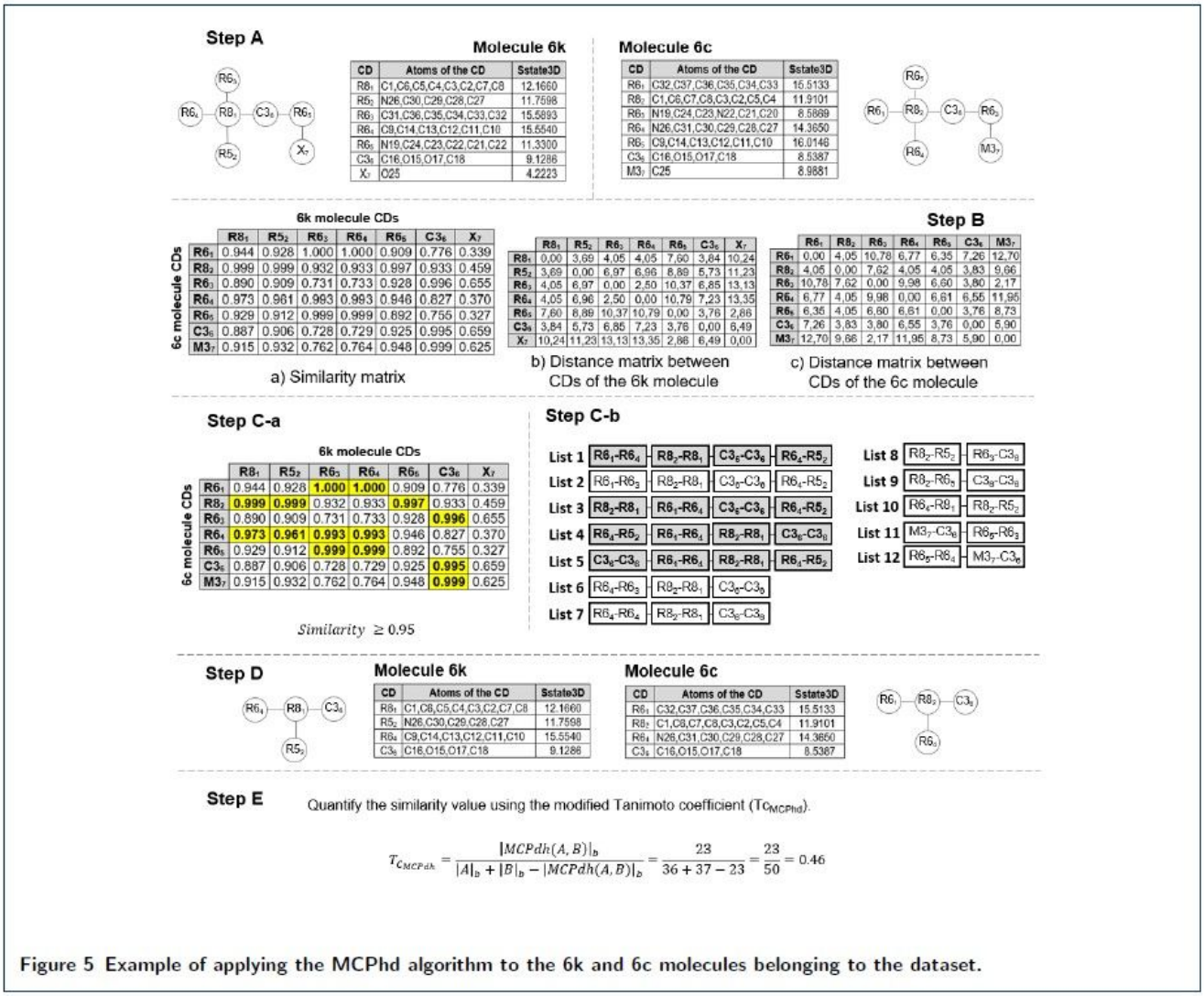


Figure 5 Example of applying the MCPhd algorithm to the 6k and 6c molecules belonging to the dataset.

Figure 5

Example of applying the MCPhd algorithm to the 6k and 6c molecules belonging to the dataset.

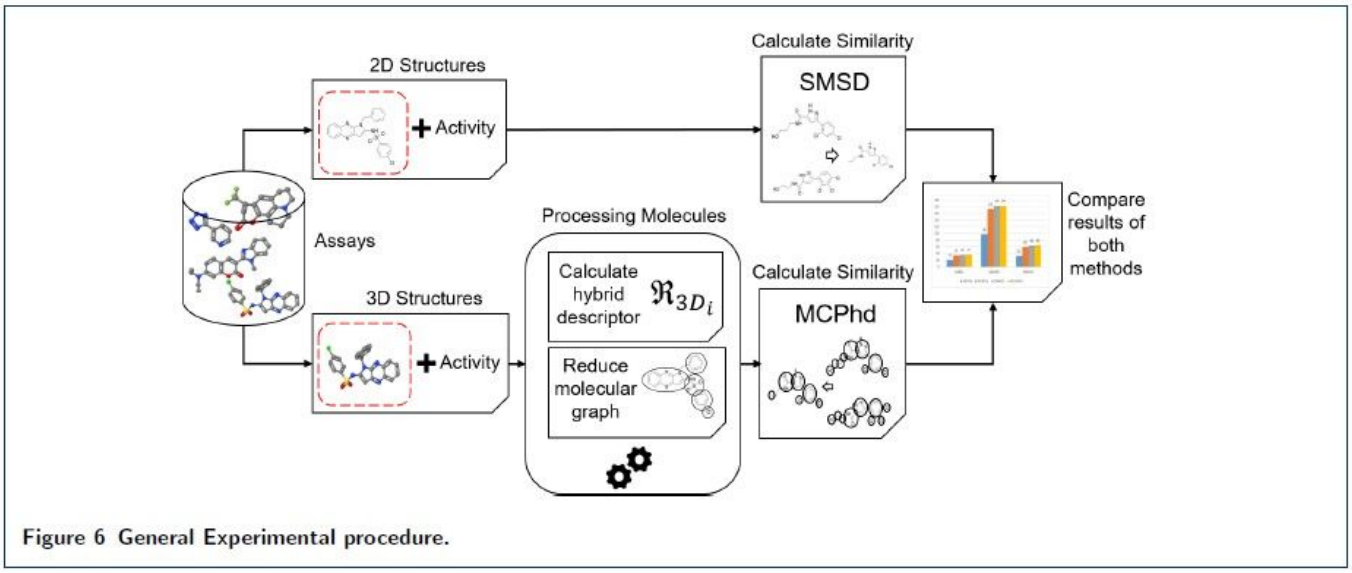


Figure 6 General Experimental procedure.

Figure 6

General Experimental procedure.

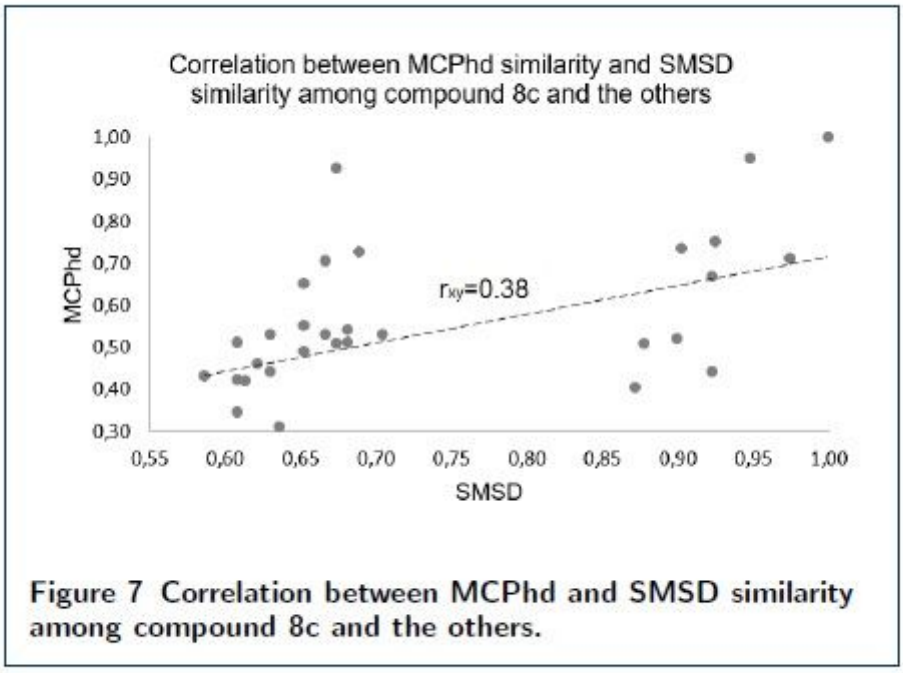


Figure 7 Correlation between MCPhd and SMSD similarity among compound 8c and the others.

Figure 7

Correlation between MCPhd and SMSD similarity among compound 8c and the others.

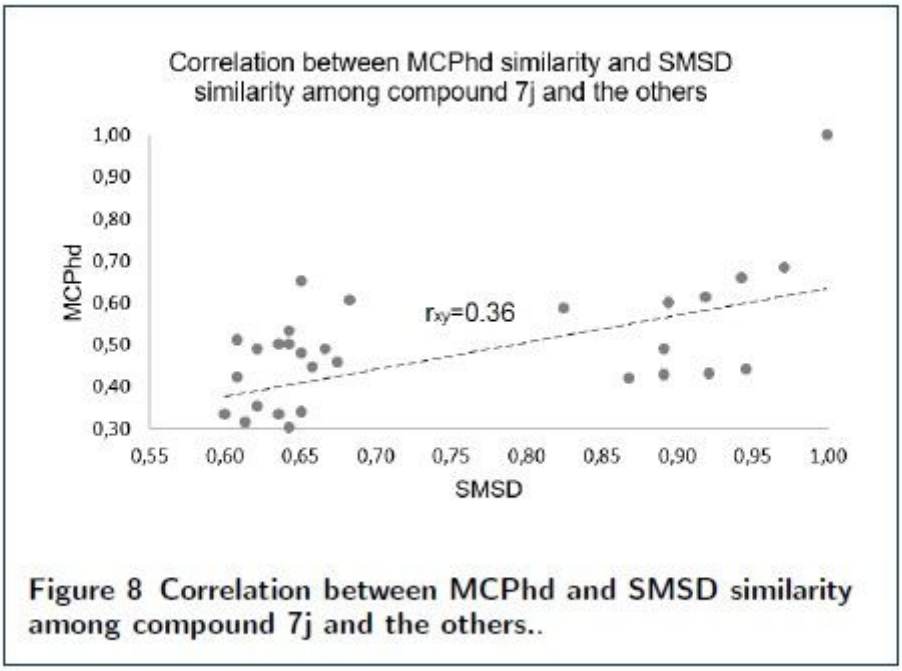


Figure 8 Correlation between MCPhd and SMSD similarity among compound 7j and the others..

Figure 8

Correlation between MCPhd and SMSD similarity among compound 7j and the others..

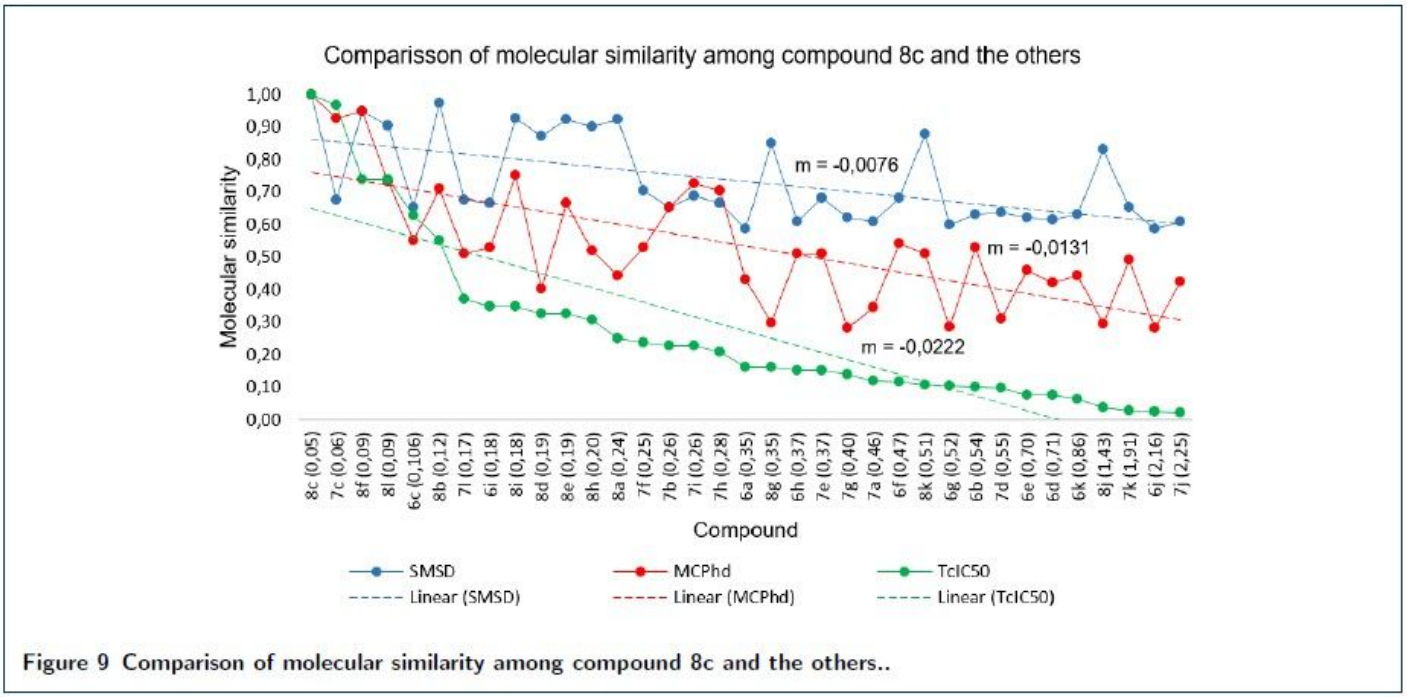


Figure 9 Comparison of molecular similarity among compound 8c and the others..

Figure 9

Comparison of molecular similarity among compound 8c and the others.

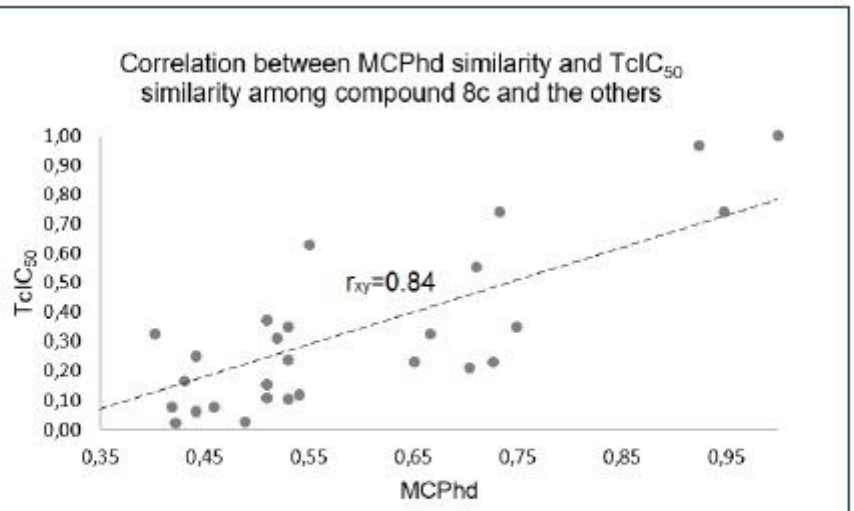


Figure 10 Correlation between MCPhd and $TcIC_{50}$ similarity among compound 8c and the others.

Figure 10

Correlation between MCPhd and T cIC50 similarity among compound 8c and the others.

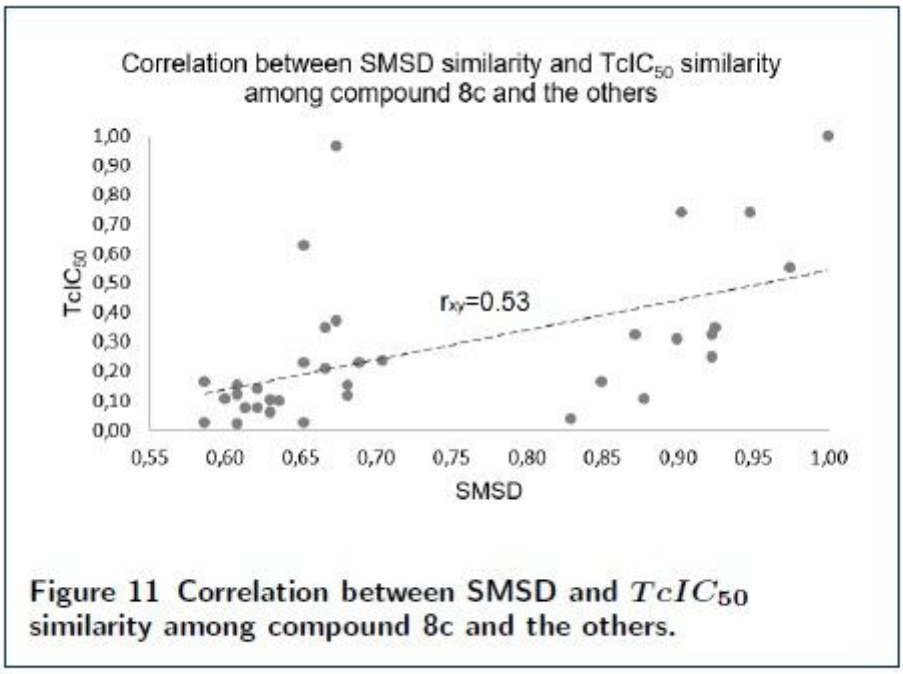


Figure 11 Correlation between SMSD and $TcIC_{50}$ similarity among compound 8c and the others.

Figure 11

Correlation between SMSD and T cIC50 similarity among compound 8c and the others.

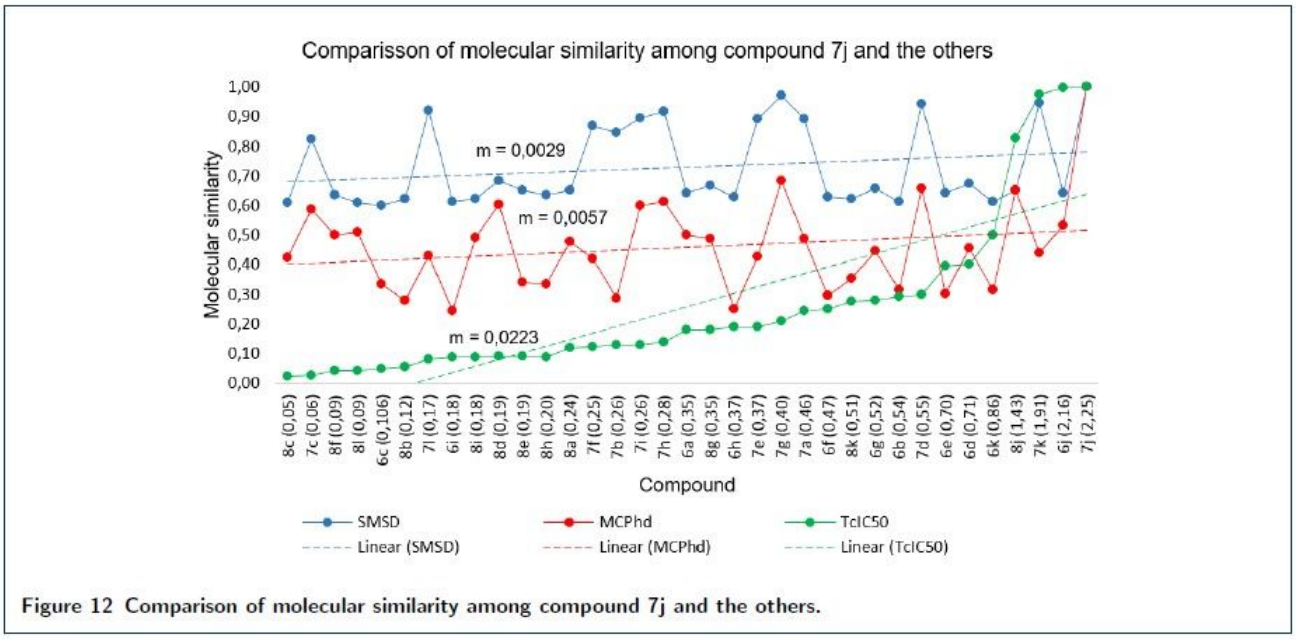


Figure 12 Comparison of molecular similarity among compound 7j and the others.

Figure 12

Comparison of molecular similarity among compound 7j and the others.

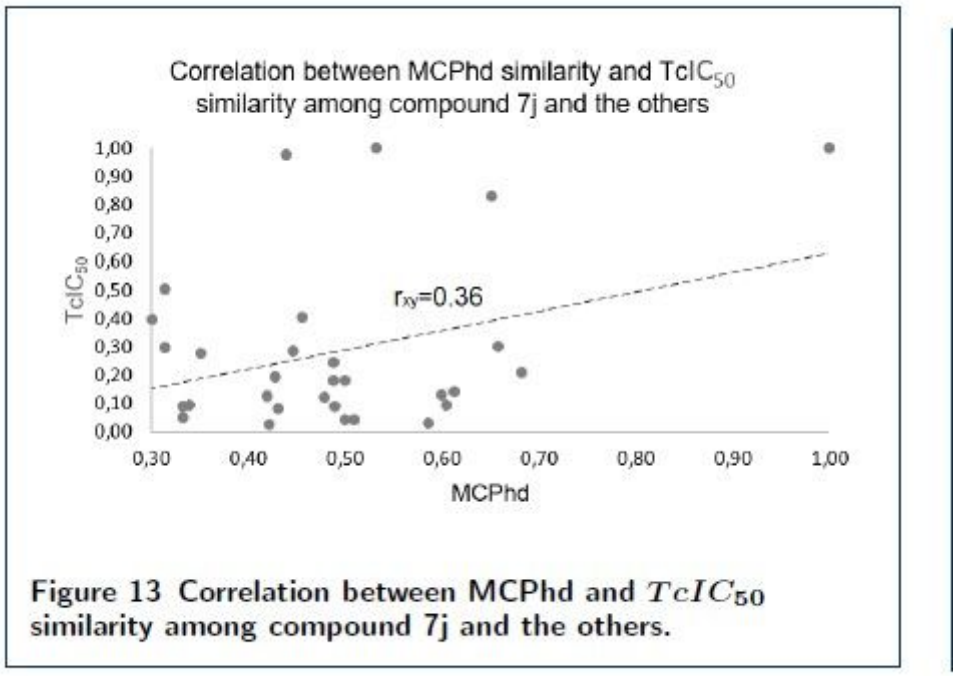


Figure 13 Correlation between MCPhd and $TcIC_{50}$ similarity among compound 7j and the others.

Figure 13

Correlation between MCPhd and T cIC50 similarity among compound 7j and the others.

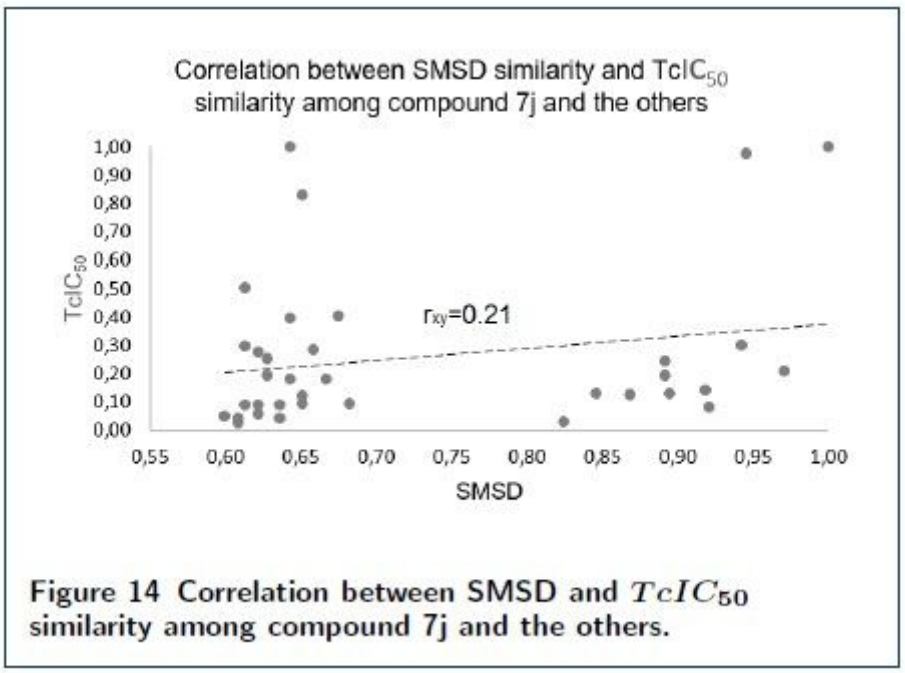


Figure 14 Correlation between SMSD and $TcIC_{50}$ similarity among compound 7j and the others.

Figure 14

Correlation between SMSD and T cIC50 similarity among compound 7j and the others.

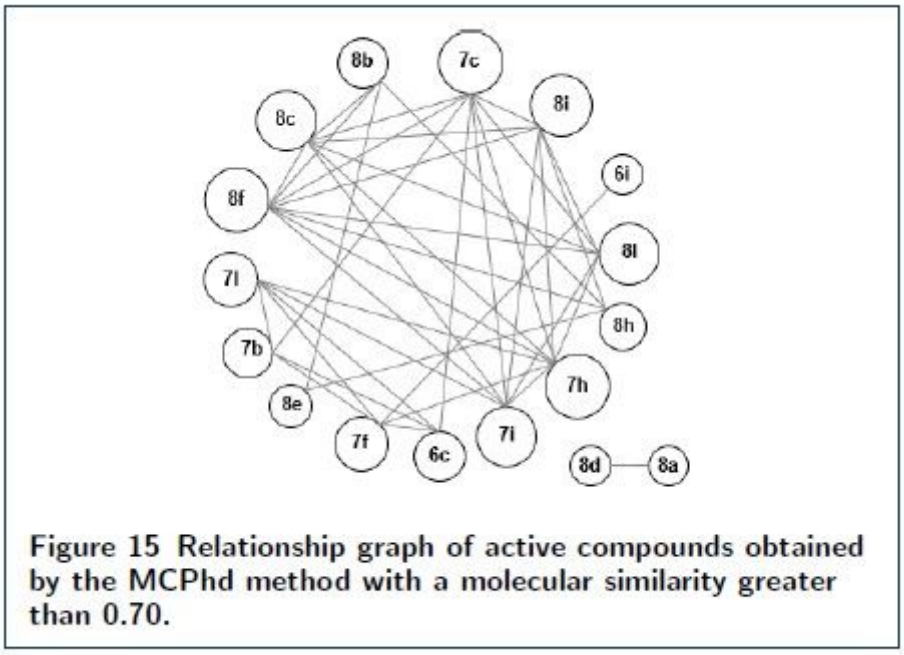


Figure 15 Relationship graph of active compounds obtained by the MCPhd method with a molecular similarity greater than 0.70.

Figure 15

Relationship graph of active compounds obtained by the MCPhd method with a molecular similarity greater than 0.70.

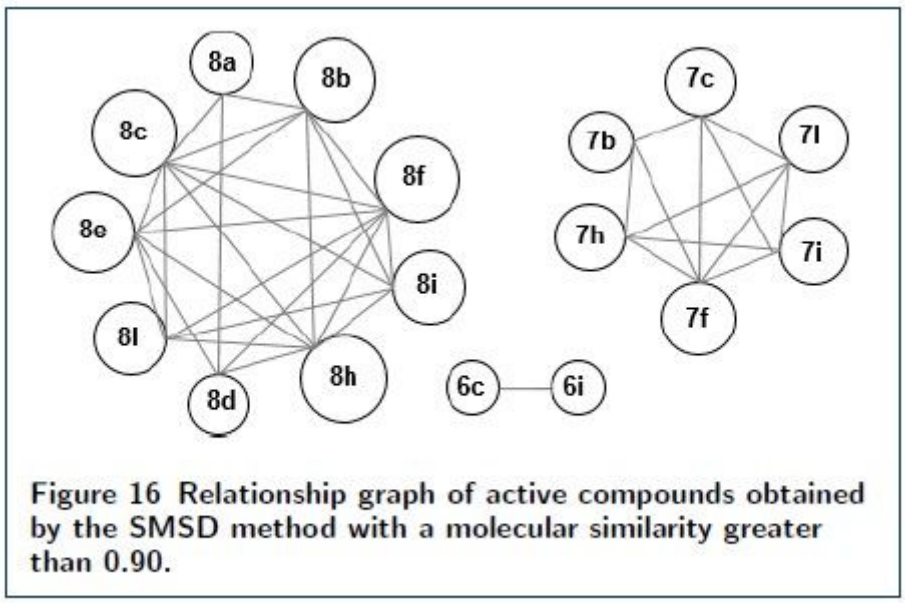


Figure 16 Relationship graph of active compounds obtained by the SMSD method with a molecular similarity greater than 0.90.

Figure 16

Relationship graph of active compounds obtained by the SMSD method with a molecular similarity greater than 0.90.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Dataset.zip