JOURNAL OF CHEMICAL INFORMATION AND MODELING
Article
pubs.acs.org/jcim

Noncontiguous Atom Matching Structural Similarity Function

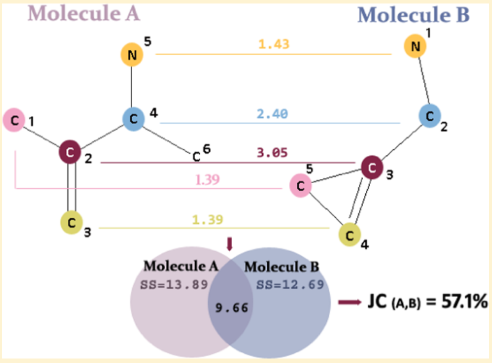
Ana L. Teixeira*,[†,‡] and Andre O. Falcao[†,||]

[†] LaSIGE, Faculty of Sciences, University of Lisbon, Campo Grande 1749-016 Lisbon, Portugal

[‡] CQB−Centro de Quimica e Bioquimica, Faculty of Sciences, University of Lisbon, Campo Grande 1749-016 Lisbon, Portugal

[||] Department of Informatics, Faculty of Sciences, University of Lisbon, Campo Grande 1749-016 Lisbon, Portugal

S *Supporting Information*



ABSTRACT: Measuring similarity between molecules is a fundamental problem in cheminformatics. Given that similar molecules tend to have similar physical, chemical, and biological properties, the notion of molecular similarity plays an important role in the exploration of molecular data sets, query-retrieval in molecular databases, and in structure−property/activity modeling. Various methods to define structural similarity between molecules are available in the literature, but so far none has been used with consistent and reliable results for all situations. We propose a new similarity method based on atom alignment for the analysis of structural similarity between molecules. This method is based on the comparison of the bonding profiles of atoms on comparable molecules, including features that are seldom found in other structural or graph matching approaches like chirality or double bond stereoisomerism. The similarity measure is then defined on the annotated molecular graph, based on an iterative directed graph similarity procedure and optimal atom alignment between atoms using a pairwise matching algorithm. With the proposed approach the similarities detected are more intuitively understood because similar atoms in the molecules are explicitly shown. This noncontiguous atom matching structural similarity method (NAMS) was tested and compared with one of the most widely used similarity methods (fingerprint-based similarity) using three difficult data sets with different characteristics. Despite having a higher computational cost, the method performed well being able to distinguish either different or very similar hydrocarbons that were indistinguishable using a fingerprint-based approach. NAMS also verified the similarity principle using a data set of structurally similar steroids with differences in the binding affinity to the corticosteroid binding globulin receptor by showing that pairs of steroids with a high degree of similarity (>80%) tend to have smaller differences in the absolute value of binding activity. Using a highly diverse set of compounds with information about the monoamine oxidase inhibition level, the method was also able to recover a significantly higher average fraction of active compounds when the seed is active for different cutoff threshold values of similarity. Particularly, for the cutoff threshold values of 86%, 93%, and 96.5%, NAMS was able to recover a fraction of actives of 0.57, 0.63, and 0.83, respectively, while the fingerprint-based approach was able to recover a fraction of actives of 0.41, 0.40, and 0.39, respectively. NAMS is made available freely for the whole community in a simple Web based tool as well as the Python source code at http://nams.lasige.di.fc.ul.pt/.

INTRODUCTION

Molecules are typical examples of unstructured data for which tasks such as searching, sorting, analyzing, and extracting knowledge are challenging. A molecule can have an arbitrary dimension, structure and composition, and moreover, there is not an univocal and unequivocal way of coding and comparing these molecules. Several computational tools have been developed over the years in pursuance of solving this issue. Fundamental observations that justify the amount of methods developed to compare molecules derive from the fact that similarity has a context,[1] and the representation of molecular structures implies information loss. Researchers have explored the concept of similarity between molecules which provides an important approach to search databases, predict properties of compounds, design structures with a predefined set of properties, and conduct structure-based drug design studies.[1−9] These studies are based on the "neighborhood" premise, which states that similar molecules usually have similar activities and properties.[1,2,10] The definition of similarity for molecules consists of comparing chemical structures, specifically representing the molecules and quantifying the similarity between them. Various methods to define structural similarity between molecules are available in the literature.[1,6] The most popular approaches to represent the structure of the molecules under comparison can be divided in three broad categories, approaches based on structural descriptors (two- and three-dimensional), molecular fragments, and graph matching (descriptor-independent methods).

Approaches Based on Structural Descriptors. Methods based on structural descriptors attempt to describe the information encoded in the molecular structure into a set of

Received: May 31, 2013
Published: September 17, 2013

dx.doi.org/10.1021/ci400324u | J. Chem. Inf. Model. 2013, 53, 2511−2524

numerical values and define some means for comparing them.[6] A large number of different descriptors can be used in similarity calculations and they differ in the complexity of the encoded information and in the computation time.[11] Molecular descriptors can be categorized according to their dimensionality or nature. According to the dimensionality of the encoded information, it is possible to define the following categories of descriptors: (1) one-dimensional (1D) descriptors which capture information that is slightly discriminative but fast to compute, such as the constitution of the structure and related properties (e.g., molecular weight, atom counts, and logarithm of the octanol/water partition coefficient); (2) two-dimensional (2D) descriptors which capture topological information, such as the degree of branching, overall shape and size (e.g., topological indexes, molecular walk counts, and shape indices); (3) three-dimensional (3D) descriptors which capture geometrical information, accounting for properties that depend on internal coordinates, internal orientation, conformation, or stereoisomerism (e.g., van der Waals volume, radial distribution function, and 3D molecule representation of structures based on electron diffraction (MoRSE) descriptors); (4) four-dimensional (4D) descriptors which capture electronic and quantum-chemical information usually derived from computationally expensive empirical schemes or molecular orbital calculations, accounting for reactivity, shape, and binding properties of a molecule (e.g., highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, hydrogen bonding, and polar interactions). These categories of descriptors are heterogeneous and not always mutually exclusive, originating the appearance of different schemes of categorization in the literature.

A descriptor positions each abstract molecular representation in the descriptor space. It is then possible to compare molecules, considering that the distance of the abstract molecular representations reflects their similarity in this specific descriptor space.[1] Depending on the context of the comparison the appropriate set of descriptors may change. For instance, in the property prediction context using a certain descriptor space, a set of structurally similar molecules may be also similar with respect to one property *A*, but completely dissimilar with respect to a property *B*.[1] Molecular similarity is a nonlinear problem for which there is not a set of descriptors or a similarity measure that correlates with every context of comparisons one can perform.[3,12] Moreover, published works have shown that compounds that are similar to known active molecules are themselves far less frequently active than one might expect.[6,13] Despite three-dimensional descriptors were expected to have a better performance than two-dimensional descriptors, the opposite trend has been verified in several studies.[5,14]

Approaches Based on Molecular Fragments. Molecular fragments are structural descriptors which indicate the presence of some structural fragments in molecules. For a given molecule, the fragments under consideration can be obtained using one of the following methods (1) a set of rules for their generation or (2) a predefined library of fragments.[8,15,16] An important class of molecular fragment descriptors are fingerprints. Fingerprints are bit strings that encode the presence or absence of topological patterns of atoms and bonds in a molecule. They can be divided into keyed or hashed approaches.[15,16] In keyed fingerprints, each bit position is associated with a specific group of descriptors, while in hashed fingerprints the values are mapped to overlapping bit segments. The length of bit strings can vary from hundreds of bit positions for structural fragment fingerprints to thousands of bits for connectivity fingerprints. Molecules can be efficiently stored and processed using fingerprints and in the context of searching chemical databases a molecule is considered to be structurally similar to the target if the calculated coefficient of similarity between the two fingerprints exceed a predefined threshold value. Fingerprints are one of the most widely used methodologies for global molecular similarity analysis, despite the fact that this molecular representation has some disadvantages such as information loss and a bias in the evaluation of molecular similarity due to differences in molecular complexity and size. There is information loss when representing molecules as fingerprints since, for example, binary fingerprints simply indicate the presence or absence of a given fragment rather than set bits for the number of matches and structural fragment fingerprints may not include in the library fragments that may be important for certain problems. Also, the average similarity appears to increase with the complexity and size of the query compound, since there is an higher bit density than for simpler molecules which will endorse a larger overlapping of the fingerprints.[1,8,15,17,18] Fingerprints can also be used to represent molecules in the context of property prediction or to efficiently filter out dissimilar structures from a data set, since quantifying two fingerprints as very dissimilar means, in principle, that the underlying structures are certainly dissimilar.[15]

Approaches Based on Graph Matching. A molecule can also be represented, using graph theory, as a labeled graph whose vertices correspond to the atoms and edges correspond to the covalent bonds. The representation of molecules using graphs has some advantages, namely, graphs are intuitive when representing a molecule since they are close to our understanding of a molecule and they have a solid mathematical background with different existing techniques to compare labeled graphs.[19] However, representing molecules as graphs raises an important issue, identical graphs do not necessarily represent identical structures and vice versa. This problem is originated by mesomeric structures (e.g., aromatic rings), stereochemistry (e.g., chirality), and tautomeric forms, among others.[19] The graph matching approach is descriptor-independent and often employs the concept of the maximum common subgraph (MCS).[19−21] A common substructure can be defined as a substructure present in two molecules with the same bonding profile and therefore the objective of an MCS algorithm is to find the common substructures with the largest number of atoms and bonds.[19] These searches tend to be time-consuming due to the NP-complete complexity of the subgraph isomorphism problem (a NP-complete problem cannot be solved in polynomial time);[22] however, many approximated heuristics have been proposed to overcome this complexity. These heuristics are based on techniques such as pruning the search tree of the exact algorithm,[21,23] greedy algorithms,[24−26] genetic algorithms,[27,28] and reduced representations of chemical graphs,[29−31] among others. This group of descriptor-independent methods benefits from improved sensitivity in relation to descriptor/fragment based similarity searches since they can find atom−atom, bond−bond, or atom−bond equivalences between query and target molecules.[19,32]

Quantifying the Degree of Similarity/Dissimilarity between Molecules. Following the selection of a molecular representation, to determine the numerical value of the similarity/dissimilarity of the molecules it is necessary to compare their abstract representation using a similarity coefficient/distance measure. This quantification can be obtained using simple distance measures such as Hamming or Euclidean, and association/similarity coefficients such as Tanimoto–Jaccard, Dice, or Cosine.[6,33] Distance measures consider a shared absence of fragments as evidence of similarity while association coefficients consider a shared presence of fragments as evidence of similarity, ignoring molecular features that are absent in both molecules. While the Tanimoto–Jaccard coefficient is the most popular similarity coefficient, there are several others that have been explored.[18,34] The question then is which coefficient performs better given the selected molecular representation.

Similarity is an abstract, problem-dependent, and subjective concept and its definition is, to a great extent, a semantic question. Similarity depends on comparative perception without a defined standard, in a certain degree "like beauty, it is in the eye of the beholder".[35] When judging, for example, the similarity of faces, some may consider that two faces are similar if they have a common complexion, while others would consider other facial characteristics such as the eyes, the nose, the ears, and the mouth. Because of this subjectivity it is difficult to develop methods for unambiguously quantifying the similarities of objects such as molecules. Moreover, in many situations, while two molecules are not similar, some of their parts are; the challenge is then the quantification of the degree of partial similarity between the given molecules. Some authors argue that all pattern recognition problems boil down to giving a quantitative interpretation of similarity between objects.[36] There are also studies that show that using the existing methods to quantify similarity is not always possible to take advantage of the similarity principle to predict properties/activities of molecular structures with good performance.[3,13]

Noncontiguous Atom Matching Structural Similarity. We propose a new atom alignment method for adequately quantifying the structural similarity between molecules with a high discriminative power of similar molecules. In general, to solve the global problem of quantifying the structural similarity between molecules, we decided to break it down into solvable different parts by reducing the molecule to atoms and compare atoms of different molecules in order to find the best alignment between them. These atoms should be considered not only by their intrinsic chemical characteristics but also according to their relation to the other atoms in the molecule. The similarities detected by an atom correspondence approach like the proposed one are consistent with the chemistry and structure of the molecules because it depends on the direct neighborhood of each atom as well as the overall topology of the molecule, becoming more intuitively understood because similar atoms in the molecules are explicitly shown. The relation between each atom and the whole molecule allows the consideration of important characteristics of the atoms and bonds such as the chirality and the double bond stereoisomerism, since these depend on the orientation and symmetry of the neighboring atoms in the space.[37] Although these characteristics are ignored by most 2D similarity methods, they are of great importance in many different fields since the molecular properties and biological effects of the stereoisomers are often significantly different.[38]

For the comparison of the bonding profiles of atoms on comparable molecules, we defined three main steps. First, a set of attributes should be selected in order to characterizes the molecule's atoms (e.g., atom type and chirality) and bonds (e.g., bond type and stereoisomerism) and all the topological relations between the atoms for the purpose of their comparison. Second, an adjustable weighting scheme that emphasizes certain characteristics and accounts for the differences with a penalty function, in accordance with the context of the problem, should be developed. Third, determine a value for a measure that represents the degree of similarity between the annotated molecular graph, based on a recursive concept of graph similarity and an optimal alignment between atoms using an heuristic approach.

In the following presentation, this noncontiguous atom matching structural similarity (NAMS) algorithm will be presented in detail, including the atom and bond matching functions, the alignment between the pair of molecules under comparison, and the quantification of the resemblance of the molecules. To clarify the application of the method, an illustrative example is also presented, and to demonstrate the granularity and effectiveness of this method, we present similarity analyses for three different data sets with different characteristics and compare the results against a similarity function based on path-based fingerprints.

METHODS

Concepts and Overview. The method proposed in this work is based on the concept of atom matching between two molecules, that is, for every atom of a molecule A find the atom of a molecule B that is more related to it, by scoring every possible atom comparison between A and B. If it is possible to define a score for each atom of A as related to B, it is then possible to get the best possible matchings by selecting the atom matchings that produce the best possible score, with the constraints that one atom of A may only be matched to one atom of B and vice versa. The main issue is then the definition of an atom scoring function. The present approach is noncontiguous as it may happen that matchings between atoms of different molecules may not reflect the contiguous atomic fragments of the other molecule.

Atoms are not isolated objects within molecules, their characteristics depend on (a) the bonds to other atoms and (b) the neighboring atoms. Yet the characteristics of these neighboring atoms depend as well on their bonds and their own neighboring atoms, and accordingly until all molecular bonds and atoms are exhausted within the molecule. The main idea is that each atom (α_{Ai}) of a given molecule A at position i can be represented by a corresponding graph that is centered in α_{Ai} and encompasses all the other bonds and atoms in the molecule. As such, the procedure for comparing atoms is essentially a procedure for comparing graphs, where each graph is a *view* of the full molecule from that atom. These graphs are directed graphs as there are different molecular graphs for each atom. However, the problem of comparing directed graphs is computationally expensive[22,39] even for moderately sized instances. Therefore, the following heuristic procedure was devised. Primarily it was adopted a simplified representation of the atomic directed graph. This representation encompasses a list of all the bonds of the molecule coupled to their topological distances to the atom α_{Ai} under observation. Each bond is represented by the start and end atoms as well as the covalent chemical bond between them (that is an *atom-bond-atom* tuple that, through the manuscript, we will designate as an *aba-bond*). Second, using this simplified representation, the procedure of comparing different atoms of different molecules becomes a localized problem of trying to match the best possible representation of an atom as related to its molecule, by matching each aba-bond pairings depending on their topological distances to the atoms being compared. The bond matching problem can be solved using a distance function and the same assignment algorithm suggested for atom matching. Finally, using the best possible alignment between the molecules under comparison as determined by the atom and bond matching functions, it is possible to produce a score that indicates the degree of superimposition between them.

The procedure described above can then be outlined in the following algorithm:

1. For each bond of each molecule discriminate each atom-bond-atom (an aba-bond) of the molecule, their structural characteristics and their respective topological distances as related to each atom in the molecule.
2. Use a distance-dependent bond matching function to compute a matching score between each aba-bond of each atom of each molecule and produce, for each two atoms being compared, a bond matching matrix.
3. Use an assignment algorithm to compute the best possible matching score between each possible pair of atoms, by matching the aba-bonds matrices of each atom being compared. The resulting matrix will have a score that quantifies how closely each atom of molecule *A* matches any other atom of molecule *B*.
4. Use the same assignment algorithm to assign each atom of molecule *A* to each atom of molecule *B*.
5. The similarity score between molecules *A* and *B* is then the sum of the similarities between the best possible atom alignments.
6. Compute a similarity coefficient by calculating the ratio between the molecule *A* and *B* superimposition and the sum of the self-superimposition of molecules *A* and *B*.

An important step of the algorithm is the aba-bonds comparison, as for each bond in the molecule, the start and end atoms are accounted, as well as different structural characteristics of the atoms and bonds. Namely, it is possible to use the nature of the atom with distinct atomic similarity functions as well as the nature of the bond (single, double, triple, aromatic), chain type (e.g., linear or cyclic), include/exclude hydrogen atoms and even include other specific characteristics of bonds or atoms that are dependent on the topology and geometry of the neighborhood, namely atomic chirality and cis–trans bond isomerism.[37]

In the following sections each step of the algorithm is detailed, clarifying the implementation decisions and illustrating with a simple example for comparing two small molecules. For presenting each part of the method a bottom-up approach will be followed, first describing the aba-bond matching procedures, then the atom matching using the scores produced for each atom, and finally the definition of a molecular similarity score.

Molecular Alignment by Bond Matching. Two desirable characteristics of a bond similarity function should be first that it produces a score based on bond characteristics, and second, includes a factor that, for the atoms under comparison, accounts for the respective topological distances. These characteristics and how the method makes use of them to produce a molecular alignment by bond matching will be described below.

Bond Similarity. A chemical bond matching function should involve the computation of a similarity function between the predefined set of characteristics of any two bonds. For directly matching aba-bonds a product function that multiplies the similarities between the set of characteristics of the bond that includes not only the bond itself, but also both end atoms was devised. By using a product function, each difference in the bond characteristics cumulatively decreases the end result, which asymptotically approaches 0, thus effectively working as a similarity function. Therefore if a generic aba-bond β can be defined by a tuple of P characteristics (h_m: $\beta = (h_1, h_2, ..., h_P)$, a way to compute a similarity function between bonds β^k and β^l is by cumulatively multiplying their paired attributes' similarity:

$$V_{\mathrm{nd}}(\beta^k, \beta^l) = \prod_{m=1}^{P} W_m(h_m^k, h_m^l) \tag{1}$$

where W_m is a function that outputs a value from 0 to 1.0, when comparing the same set of characteristics for two bonds. A resulting $W_m = 1.0$, means that the characteristics are exactly the same, whereas a value of $W_m < 1.0$ implies a difference.

aba-Bond Distance-Compensation Functions. The inclusion of the bond topological distances to atoms is fundamental for the bond matching system. Yet two characteristics are deemed important for adequately using topological distances to weight aba-bond matching: (a) aba-bonds that are closer to the atoms being matched should have a larger impact on the matching score than aba-bonds that are very far topologically from the atoms being compared; (b) it should be possible to pair two aba-bonds even if they appear in distinct levels, but such pairing should have a lower score in the final matching function. Several possibilities for such functions exist, but the following empirical and parametrized aba-bond distance compensation function that respects these requirements was devised:

$$V(\beta_{Ai}^k, \beta_{Bj}^l) = \frac{V_{\mathrm{nd}}(\beta_{Ai}^k, \beta_{Bj}^l)}{(|d_{Ai}^k - d_{Bj}^l| + \max(d_{Ai}^k, d_{Bj}^l) + 1.0))^{\mu}} \tag{2}$$

where d_{Ai}^k and d_{Bj}^l are the distances of aba-bonds β^k and β^l to atoms α_{Ai} and α_{Bj}, respectively. The parameter μ was used as a way to weight the importance of the bond distance to an atom, when $\mu \geq 0$. Lower values for the μ parameter disregard the importance of distance of the aba-bonds to the atoms being compared, thus focusing on a more functional and local matching, whereas larger values emphasize the importance of distance and add a more global matching of the structure pendant to the score.

aba-Bond Matching Function. As referred, all atoms within a molecule must be evaluated in relation to all the other atoms in that same molecule. Therefore, to compare two atoms of different molecules these relationships must be taken into account. Consequently, all the molecule bonds and their relations to the atom α_{Ai} are compared to all the bonds of the other molecule as related to the atom α_{Bj} with the constraint that at most one bond of a molecule can be associated to one bond of the other molecule. Thus

$$T(\alpha_{Ai}, \alpha_{Bj}) = \max \sum_{i=0}^{M_A} \sum_{j=0}^{M_B} b_{kl} V(\beta_{Ai}^k, \beta_{Bj}^l) \tag{3}$$

such that

$$\sum_{k=1}^{M_A} b_{kl} <= 1 \quad \forall\, j \leq M_B \tag{4}$$

$$\sum_{l=1}^{M_B} b_{kl} <= 1 \quad \forall\, i \leq M_A \tag{5}$$

Where M_A and M_B are the number of bonds present in molecules A and B respectively, b_{kl} is a parameter set to 1 if the bond β^k as related to the atom α_{Ai} (β^k_{Ai}) is matched to the bond β^l as related to the atom α_{Bj} (β^l_{Bj}) and 0 otherwise. Function $V(\beta^k_{Ai},\beta^l_{Bj})$ represents the similarity between bonds weighted by the respective topological distances of the aba-bond as described above (eqs 1 and 2). Equations 3−5 present a optimization problem that can be solved using the Kuhn−Munkres algorithm,[40,41] of known complexity $O(n^3)$ where $n = \max(M_A, M_B)$, for which there is a guaranteed optimal solution within polynomial time.

Molecular Alignment by Atom Matching. For any two molecules A and B, a global atomic matching is the best possible matching of all atoms of molecule A to all atoms of molecule B. Formally, for each atom α_{Ai} of molecule A, the purpose is to find the best matching atom α_{Bj} of molecule B. This approach requires the aba-bond matching function $T(\alpha_{Ai},\alpha_{Bj})$ presented above (eqs 3−5) to compute a matching score between any two atoms (α_{Ai} and α_{Bj}). Therefore, for two molecules A and B with a total number of atoms N_A and N_B respectively, the goal is to find the similarity ($S(A, B)$) between them, which represents the optimal matching between their atoms:

$$S(A, B) = \max \sum_{i=0}^{N_A} \sum_{j=0}^{N_B} a_{ij} T(\alpha_{Ai}, \alpha_{Bj}) \tag{6}$$

such that

$$\sum_{i=1}^{N_A} a_{ij} <= 1 \quad \forall\, j \leq N_B \tag{7}$$

$$\sum_{j=1}^{N_B} a_{ij} <= 1 \quad \forall\, i \leq N_A \tag{8}$$

where a_{ij} is a binary variable, set to 1 if atom i that belongs to molecule A (α_{Ai}) is matched to atom j belonging to molecule B (α_{Bj}) and 0 otherwise. The constraints 7 and 8 ensure that at most one atom of molecule A is matched against at most one atom of B and, respectively, that at most one atom of B is matched against at most one atom of A. The problem is then to find the set of values for a_{ij} that maximize the similarity score $S(A, B)$. With this approach, the subsequent formulation is once again a typical assignment problem that can be solved in polynomial time using the Kuhn−Munkres algorithm,[40,41] as described above for aba-bond matching, this time, however, for atom matching.

Translating the Molecular Alignment into a Structural Similarity Score. At this stage, the atoms from both molecules were matched and their optimal alignment scored. As the value computed measures the intersection (superimposition) between the molecules A and B ($S(A, B) = A \cap B$), it is possible to compute the similarity score between both molecules by using the self-scores produced by self-superimposition ($S(A, A)$ and $S(B, B)$). It is then appropriate to use the Jaccard−Tanimoto coefficient (JC), which represents the fraction of the common parts of both molecules relative to the union of all their parts:

$$\mathrm{JC}_{AB} = \frac{A \cap B}{A \cup B} = \frac{S(A, B)}{S(A, A) + S(B, B) - S(A, B)} \tag{9}$$

where $S(A, A)$ and $S(B, B)$ is the self-superimposition of molecules A and B, respectively. JC_{AB} is a similarity coefficient where a $\mathrm{JC}_{AB} = 1.0$ indicates total superimposition between both molecules, and thus identity, while a $\mathrm{JC}_{AB} = 0.0$ shows no point of intersection between both molecules.

Illustrative Example of the Application of NAMS. One simple example of the application of the presented method will help in clarifying its application to a concrete case. Two small molecules will be compared, namely molecule A, *3-methylbut-3-en-2-amine* (SMILES: CC(=C)C(N)C), and molecule B, *cyclopropenyl methanamine* (SMILES: NCC1=CC1) (Figure 1).

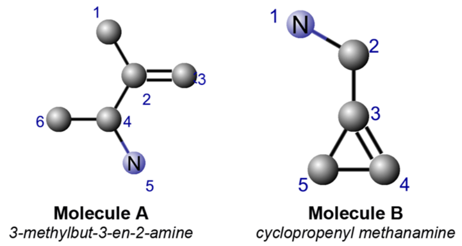


Figure 1. Chemical structures of the example molecules (A) 3-methylbut-3-en-2-amine (SMILES: CC(=C)C(N)C) and (B) cyclopropenyl methanamine (SMILES: NCC1=CC1). The canonical numbers of each atom are indicated near each atom symbol.

For simplification, only three characteristics of the aba-bond will be considered (h_m): left end-atom element, right end-atom element, and bond order. Considering then four different aba-bonds that occur in molecule A: (N—C), (C—N), (C—C), and (C=C). For this example, the respective W_m functions (degree of similarity between the set of characteristics of any two bonds) will be defined with very simple rules:

$$W_1 = \begin{cases} 0.1 & \text{if atom elements on left end of the bonds are different} \\ 1.0 & \text{if atom elements on left end of the bonds are equal} \end{cases} \tag{10}$$

$$W_2 = \begin{cases} 0.1 & \text{if atom elements on right end of the bonds are different} \\ 1.0 & \text{if atom elements on right end of the bonds are equal} \end{cases} \tag{11}$$

$$W_3 = \begin{cases} 0.8 & \text{if covalent bond orders are different} \\ 1.0 & \text{if covalent bond orders are the same} \end{cases} \tag{12}$$

With these rules, eq 1 is then used for computing the aba-bond topological distances. As a concrete example, V_{nd}((N—C), (C=C)) is calculated by iteratively applying the rules: W_1—as the left-end atom elements of each aba-bond differs (nitrogen ≠ carbon), apply a factor of 0.1 (rule 10); W_2—the right-end atom elements are the same (carbon = carbon), so the factor to apply is 1.0 (rule 11); W_3—as the bond orders (single covalent bond versus double covalent bond) are different, so the factor to apply is 0.8 (rule 12). For these two aba-bonds, V_{nd}((N—

C), (C═C)) = 0.1 × 0.8 × 1.0 = 0.08. Table 1 displays the similarity values between the four types of aba-bonds extant in molecule *A*.

Table 1. Sample Similarity Values (V_{nd}) between aba-Bonds Extant in Molecule *A* (3-Methylbut-3-en-2-amine)

V_{nd}	N—C	C—N	C—C	C═C
N—C	1.00	0.01	0.10	0.08
C—N	0.01	1.00	0.10	0.08
C—C	0.10	0.10	1.00	0.80
C═C	0.08	0.08	0.80	1.00

These values are the results of the comparison of different aba-bonds irrespective of their topological distances to each atom. To account for the topological distances for each calculated V_{nd} it is necessary to check each aba-bond as relative to each atom of the molecule. For the molecule *A* (Figure 1), starting at each of the atoms of the molecule, all possible aba-bond topological levels are displayed in Table 2.

Table 2. aba-Bonds Topological Distances (*d*) Starting from Each Atom α_i[a] of the Molecule *A* (3-Methylbut-3-en-2-amine)

	topological distance		
atom α_i	$d = 0$	$d = 1$	$d = 2$
C_1	$C_1{-}C_2$	$C_2{=}C_3$ $C_2{-}C_4$	$C_4{-}N_5$ $C_4{-}C_6$
C_2	$C_2{-}C_1$ $C_2{=}C_3$ $C_2{-}C_4$	$C_4{-}N_5$ $C_4{-}C_6$	
C_3	$C_3{=}C_2$	$C_2{-}C_1$ $C_2{-}C_4$	$C_4{-}N_5$ $C_4{-}C_6$
C_4	$C_4{-}C_2$ $C_4{-}N_5$ $C_4{-}C_6$	$C_2{-}C_1$ $C_2{-}C_3$	
N_5	$N_5{-}C_4$	$C_4{-}C_6$ $C_4{-}C_2$	$C_2{=}C_3$ $C_2{-}C_1$
C_6	$C_6{-}C_4$	$C_4{-}C_2$ $C_4{-}N_5$	$C_2{=}C_3$ $C_2{-}C_1$

[a] Where α is the atomic symbol of the atom in the position *i* as represented in Figure 1.

With the bond similarities determined, by calculating the $V_{\text{nd}}(\beta^k, \beta^l)$ coefficients using eq 1, it is then possible to compute the similarities between any two atoms, by applying the aba-bond distance-compensation function (eq 2), which weights, as detailed before, each bond similarity of two different atoms according to their topological distance. Equation 2 requires one user-defined parameter (μ) that only affects the denominator, which is also dependent on the relative topological distances of each aba-bond to the respective reference atom. As an example, the atom C_1 will be compared with the atom C_6 within the molecule *A*, where each of them has only three distinct topological levels (d = 0, 1, and 2) as represented in Table 2. Assuming that μ = 2.0, the denominators of eq 2 can be precalculated as presented in Table 3. Table 3 shows that for comparing aba-bonds, the smaller denominators are, as expected, the ones for aba-bonds that are closer to each other, producing higher aba-bond similarity scores, while there is an increase of the denominator as the topological distance between the aba-bonds grows, producing smaller aba-bond similarity scores. For instance, when comparing two aba-bonds that are both at $d = 1$, the similarity score of the bond decreases to 1/4, whereas if one bond is at $d = 0$ and the other at $d = 1$, the distance weighting factor is 1/9.

Table 3. Denominators of Equation 2 for Bond Comparison Based on the Topological Distances d_1^k and d_6^l of the aba-Bonds β^k and β^l to Atoms C_1 and C_6 of Molecule *A*, Considering μ = 2.0

d_1^k \ d_6^l	0	1	2
0	1	9	25
1	9	4	16
2	25	16	9

Equation 2 can now be applied to calculate the similarities between all possible bond matchings of any two given atoms, using the aba-bond similarities presented in Table 1 as numerator and the distance coefficients presented in Table 3 as denominator. Table 4 presents the results of aba-bond similarity matching between the atoms C_1 and C_6 of molecule *A* weighted by their topological distance. To finally compute the similarity between these two atoms, each aba-bond of each atom must be matched to the best possible aba-bond of the other atom. Applying the Kuhn–Munkres algorithm allows one to discover which aba-bond (if any) of C_1 best matches each aba-bond of C_6. On the same table, the highlighted values indicate the best possible assignments. The sum of the best possible alignment score between aba-bonds is the actual aba-bond similarity score (eqs 3–5) between these two atoms; thus, $T(C_1^A, C_6^A)$ = 1.000 + 0.250 + 0.063 + 0.111 + 0.063 = 1.487.

To compute the self-similarity score of molecule *A*, which represents the self-superimposition of the molecule, the procedure would be repeated for each atom and to obtain the final score for the atom matching use eqs 6–8. Yet to better illustrate the algorithm as a tool for inferring structural similarity between molecules, this procedure will be demonstrated for comparing two different molecules *A* and *B* (Figure 1). However, it must be stressed that molecule *B* includes a cyclic element, and therefore, another characteristic was defined to be considered by the W_m function beyond those already defined in rules 10–12:

$$W_4 = \begin{cases} 0.8 & \text{if one atom is within a cycle and the other} \\ & \quad \text{one is not} \\ 1.0 & \text{if both atoms are with in a cycle} \end{cases} \tag{13}$$

Adding this new rule 13 to the others 10–12, the end result of the atom matching matrix was calculated and is presented in Table 5. Similarly to the bond matching process, the Kuhn–Munkres algorithm was used to match the atoms of both molecules and produce an optimal atom-matching score to the system defined by eqs 6–8.

It is relevant to notice that the atom alignment has preserved the obvious characteristics of the molecules while adequately handling the cyclic element. Also, atom C_6 of molecule *A* was not paired, as no adequate matching for this atom was found on molecule *B*.

The total molecule similarity score is then $S(A, B)$ = 1.43 + 2.40 + 3.05 + 1.39 + 1.39 = 9.66. Using the same process, the self-similarity of molecules *A* and *B* was calculated as $S(A, A)$ = 13.89 and $S(B, B)$ = 12.69, respectively. These values allow us to use the eq 9 for attributing a final structural similarity

Table 4. aba-Bond Scores As Calculated by Equation 2 for the Atoms C_1 and C_6 of Molecule *A*, Considering Their Topological Distances ($d(C_1$, aba) and $d(C_6$, aba)) to the Respective Reference Atom and $\mu = 2.0$[a]

	aba-bond for C_6	$C_6 - C_4$	$C_4 - C_2$	$C_4 - N_5$	$C_2 - C_1$	$C_2 = C_3$
aba-bond for C_1	$d(C_6, aba)$ / $d(C_1, aba)$	0	1	1	2	2
$C_1 - C_2$	0	**1.000**	0.111	0.011	0.040	0.032
$C_2 = C_3$	1	0.089	0.200	0.020	0.050	**0.063**
$C_2 - C_4$	1	0.111	**0.250**	0.025	0.063	0.050
$C_4 - N_5$	2	0.004	0.006	**0.063**	0.011	0.009
$C_4 - C_6$	2	0.040	0.063	0.006	**0.111**	0.089

[a] The highlighted values represent the best possible alignment of the aba-bonds for the atoms C_1 and C_6 determined by eqs 6−8.

Table 5. Atom Similarity Scores between Molecules *A* and *B* Calculated by Equation 3[a]

Mol. *B* / Mol. *A*	N_1	C_2	C_3	C_4	C_5
C_1	0.49	1.44	1.38	1.35	**1.39**
C_2	0.41	1.40	**3.05**	2.00	2.00
C_3	0.46	1.20	1.21	**1.39**	1.19
C_4	0.43	**2.40**	2.20	1.89	2.09
N_5	**1.43**	0.46	0.43	0.57	0.53
C_6	0.53	1.36	1.51	1.19	1.19

[a] The highlighted values represent the best possible alignment of the atoms of molecule *A* and *B* determined by eqs 3−5.

coefficient between both molecules: $JC_{AB} = 9.66/(13.89 + 12.69 - 9.66) = 0.571$, or 57.1% similarity between both molecules. This value, even for this simplistically defined system, allows us to conclude that, even though an adequate atom matching between these two molecules was found, the structural similarity between both molecules is not very large, as expected by visual observation of their molecular graphs.

■ DATA

To assess NAMS, three data sets (A, B, and C) with different characteristics will be used and described below.

Data Set A—Hydrocarbons. The data set A comprises 100, randomly selected from the ThermInfo[42] data set, linear and cyclic hydrocarbon structures, i.e. compounds composed of carbon and hydrogen, sharing a key structural feature, the presence of stable carbon−carbon bonds. Although the compounds of this class are structurally related, they exhibit important differences that are not distinguishable when using and comparing molecules in terms of the presence or absence of molecular features.

A complete table with compound name, CASRN, SMILES, similarity matrix using fingerprints, and similarity matrix using NAMS for the 100 hydrocarbons is provided in Supporting Information 1.

Data Set B—Steroids and Their Binding Affinity to the Corticosteroid Binding Globulin (CBG) Receptor. The data set B comprises 31 steroids and their binding affinity (p*K*) to the corticosteroid binding globulin (CBG) receptor compiled by ref 43 and used in other studies regarding similarity calculations (e.g., ref 44) and activity prediction (e.g., refs 45−48). The CBG binding is expressed by an affinity constant (*K*), which is expressed as p*K* (equivalent to −log(*K*)). The more negative the p*K* value is, the higher the binding affinity. These steroids contain a characteristic arrangement that is composed of twenty carbon atoms bonded together in four fused rings (three cyclohexane and one cyclopentane rings), and they vary by the functional groups attached to four-ring core. For this data set, the aim is to compare the binding affinity based on the neighborhood principle.

A complete table with compound name, SMILES, binding affinity (p*K*), activity level, similarity matrix using fingerprints, and similarity matrix using NAMS for the 31 steroids is provided in Supporting Information 2.

Data Set C—Monoamine Oxidase (MAO) Inhibitor. The data set C comprises 1650 compounds considerably diverse with information about MAO inhibition level.[49,50] The activity is represented on a four-level scale: inactive compounds are represented as having activity 0, while the values 1, 2, and 3 correspond to increasing levels of activity. A preprocessing set was carried out where 5 molecules with unknown atoms were eliminated, 467 salts or clusters composed with more than one fragment were simplified in order to maintain only the main molecule (fragments such oxalic acid, sulfuric acid, sodium ion, among others were eliminated), molecules duplicated after fragment elimination (10 molecules) or with different affinity when clustered with other fragments (9 molecules) were eliminated. The final number of molecules is 1626 of which 288 are active (113 with activity level 1, 87 with activity level 2, and 88 with activity level 3) and 1338 are inactive. This data set has been previously used to assess how structurally similarity methods based on fingerprints relate with similar biological activity of molecules.[13] The data set has an high number of actives, about 17.4% larger than what is typically found in screening databases, since it contains a subset of compounds synthesized to follow up a lead. Martin et al.[13] assessed the adequacy and bias of the data set by determining the fraction of clusters of actives identified using different similarity thresholds (>0.85). Since the fraction of clusters containing active compounds increased with the increase of the threshold level, the authors[13] concluded that this data set is not misleading and adequate for research. For this data set, the aim is to retrieve compounds with similar activity level based on the similarity threshold.

■ IMPLEMENTATION

Binary fingerprints, computed from the presence or absence of molecular features are commonly compared using a similarity coefficient as a measure of similarity between structures, with the Tanimoto coefficient being the most widely used.[15] This is a particularly efficient, simple and among the most widely used methods in the case of two-dimensional or other easy to calculate descriptors due to their performance.[15,51,52] In addition, binary representations are suited to computer processing with a fast paced process. Binary fingerprints and similarity quantification using the Tanimoto coefficient have some drawbacks and limitations[1,15,33] such as the following: (1) They do not take into account bits that are off in both molecules. (2) They do not consider the frequency of detected

Table 6. Weights (W_m) of All Characteristics under Consideration for the Similarity Calculation for Each Data Set (A, B, and C)[a]

			atom and bond characteristics							
data set	μ	hydrogens	1	2	3	4	5	6	7	8
A	2.0	no	*b*	0.95	0.7	0.7	0.7	0.7	0.95	−0.2
B	2.0	yes	*b,c*	0.95	0.7	0.7	0.7	0.7	0.95	−0.2
C	2.0	no	*b*	0.9	0.8	0.7	0.8	0.8	0.95	−0.2

[a] Atom and bond characteristics under consideration: (1) the nature of the element; (2) whether the atom is chiral and its orientation; (3) whether the atom is part of at least one ring; (4) the bond order; (5) whether the bond is part of at least one ring; (6) whether the bond is aromatic; (7) whether a double bond has *E*−*Z* stereoisomerism; (8) a penalty function to account for unmatched atoms. [b] Atom substitution matrix that considers that each atom type is only fully similar to itself and completely different from all the others (Supporting Information 3 ASM = 0). [c] Atom substitution matrix with different scores computed according to their position on the periodic table (Supporting Information 3 ASM = 2).

fragments; therefore binary fingerprints tend to favor larger molecules in similarity due to bit saturation and smaller molecules in diversity selections. (3) In the cases where the bond path length exceeds the defined maximum, it is not possible to discriminate such fragments. (4) In the case of hashed fingerprints, it is possible that different fragments hash to the same bit and therefore there is information loss. (5) They do not include stereochemistry information which considerably tends to influence the properties/activities of the molecule. Nevertheless, there are several studies in the literature showing that fingerprint-based methods outperform other types of descriptors and similarity methods.[13,50,53−55] We can hence implement the developed structurally based similarity procedure (NAMS) in order to experimentally evaluate its capacity to compare molecules by examining the results for three different data sets with another implementation based on the widely used path-based fingerprints.

Fingerprint Implementation. The molecular similarity score is obtained by comparing binary path-based fingerprints (FP2) calculated by Openbabel[56] using the Tanimoto coefficient. The FP2 fingerprints identify all linear and ring fragments in the molecule with lengths varying from 1 to 7 and maps them onto a bit-string of length 1024 using a hash function (similar to the Daylight fingerprints). These fingerprints encode the substructures present in a molecule, which can be compared in order to obtain the proportion of substructures in common between the two molecules under consideration. As explained above, in this study the Tanimoto coefficient was used, since it is the most widely used similarity coefficient for comparing binary fingerprints by establishing a ratio between the number of chemical features that are common to both molecules compared to the number of chemical features that are present in both molecules.

NAMS Implementation. NAMS was implemented in Python (version 2.7). To process the chemical structures and extract the set of attributes to characterize the molecule's atoms, OpenBabel libraries (version 2.3.1) were used.[56] Although several code optimizations were used, this implementation was essentially designed for functionality and tool integration and not for computational performance.

The current implementation uses seven distinct bond characteristic parameters, necessary for the W_n functions of eq 1. These are (1) the nature of the atomic elements; (2) whether the atom is chiral and its orientation;[37] (3) whether the atom is part of at least one ring; (4) the bond order; (5) whether the bond is part of at least one ring; (6) whether the bond is aromatic; (7) whether a double bond has *E*−*Z* stereoisomerism.[37] For atom comparison and scoring, the approach followed was the use of *atom substitution matrices*. These are matrices where the atom differences are scored. Currently five different matrices were tested, differing in the distance scores given to comparing different atoms (Supporting Information 3). Finally, two parameters not specific of aba-bond characteristics are also user defined: namely, the μ parameter (eq 2) and a penalty parameter that accounts for unmatched atoms, when the atom counts of molecules differ.

The software also allows the user to specify whether hydrogen atoms should be accounted in the molecular comparison procedure. This has no effect in the use of the method, but it largely increases the computational cost.

Table 6 displays the parameters used for the similarity calculation for each data set (A, B, and C) analyzed.

Computational Efficiency. Being developed in a scripting language, the current prototype is not very efficient computationally. Yet, several optimizations were performed to allow its use in moderately sized databases. One of the most important optimizations involved identifying the critical phases of the algorithm. This was clearly the aba-bond matching procedure (eqs 3−5) that used the Kuhn−Munkres algorithm. The $O(n^3)$ complexity can be a significant factor when comparing large molecules, and therefore, an optional strategy was deployed that involved the use of a very fast greedy heuristic for aba-bond matching. This heuristic, in general, produces results similar to the Kuhn−Munkres algorithm, yet is up to four times faster in average. The results produced by the heuristic are only on occasion different from the ones reached from Kuhn−Munkres and have never accounted for differences above 3% in the final similarity score between two molecules.

A systematical test of NAMS over several databases in a common desktop PC (CPU Intel Core i3, running at 3.0 GHz with 4 GB of RAM) produced average computation times of 210 ms for comparing two molecules when using the Kuhn−Munkres algorithm. The heuristic processing times were on average 55 ms per pair of molecules compared. The heuristic approach was globally used for all the data sets in this study, however the Kuhn−Munkres was always used for the atom matching procedure (eqs 6−8).

■ RESULTS AND DISCUSSION

Assessing a similarity function is a challenging task, since similarity between two molecules is highly subjective, and even chemists are not consistent when comparing molecules.[57] For the assessment of NAMS against the molecular similarity calculated by comparing path-based fingerprints with the three data sets described above, three main points were considered important to evaluate: (1) identification of the most informative representation of molecular structures (avoid information loss); (2) fine granularity of the similarity score

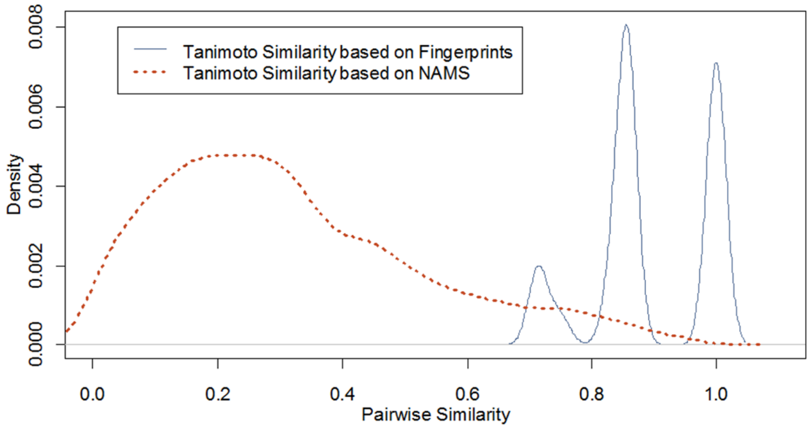
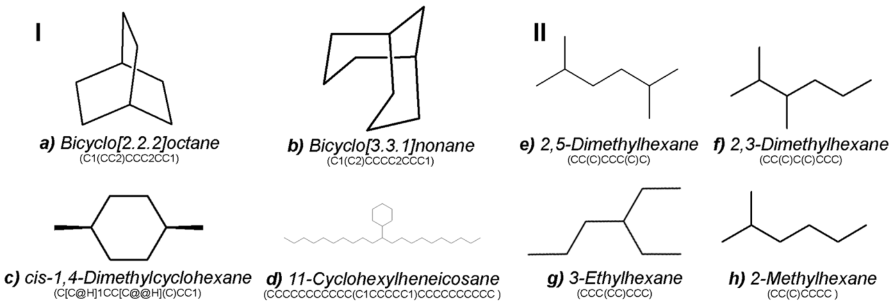


Figure 2. Distribution and variation of the pairwise similarity for a total of 4950 pairs (comparing the 100 hydrocarbons) using openbabel fingerprints and NAMS.



Chemical structure	a	b	c	d
a	1	1	1	1
b	0.72	1	1	1
c	0.38	0.42	1	1
d	0.12	0.16	0.22	1

Chemical structure	e	f	g	h
e	1	1	1	1
f	0.89	1	1	1
g	0.81	0.87	1	1
h	0.78	0.77	0.78	1

🟦 Tanimoto coefficient of similarity based on openbabel's fingerprints (FP2).
🟩 Tanimoto coefficient of similarity between molecules based on NAMS*.

Figure 3. Example of the pairwise similarity using fingerprint and NAMS, obtained for eight compounds extant in data set A. The upper part of the pairwise similarity scores matrix (blue background) was calculated using fingerprints while the lower part of the pairwise similarity scores matrix (green background) was calculated using NAMS. (I) Four considerably structurally different compounds: (a) bicyclo[2.2.2]octane (SMILES C1(CC2)CCC2CC1); (b) bicyclo[3.3.1]nonane (SMILES C1(C2)CCCC2CCC1); (c) cis-1,4-dimethylcyclohexane (SMILES C[C@H]1CC[C@@H](C)CC1); and (d) 11-cyclohexylheneicosane (SMILES CCCCCCCCCCC(C1CCCCC1)CCCCCCCCCC). (II) Four considerably structurally similar compounds: (e) 2,5-dimethylhexane (SMILES CC(C)CCC(C)C); (f) 2,3-dimethylhexane (SMILES CC(C)C(C)CCC); (g) 3-ethylhexane (SMILES CCC(CC)CCC); and (h) 2-methylhexane (SMILES CC(C)CCCC).

in order to be able to distinguish similar molecules; and (3) verify the molecular similarity principle[2] which states that structurally similar molecules tend to have similar properties (physical, chemical, or biological) more often than structurally dissimilar ones.

Discriminate Molecules with Repeated Substructures. Hydrocarbon fragments are present in most types of compounds, consequently a good similarity method should be able to distinguish hydrocarbons with similar structures. Therefore, the data set A was collected and the objective is to evaluate the distribution of the similarity scores between all pairs of molecules using path-fingerprints and NAMS. This data set presents an important challenge which is dealing with repeated substructures.

Figure 2 displays the distribution and variation of the pairwise similarity between the 100 hydrocarbons, totalizing 4950 different pairs of structures (excluding self-similarities) using fingerprints and NAMS. Figure 2 shows that using fingerprints to calculate the similarity between the pairs of molecules obtains a density curve with three distinct peaks: one with a mean value of 100% similarity which is the maximum value this distribution reaches, one with a mean value of 85.7% similarity which is slightly below the mean value of this distribution (89.1% ± 9.6%), and a smaller peak with a mean value of 71.4% similarity which is the minimum value of similarity obtained using fingerprints. On the other hand, using NAMS to calculate the similarity between the pairs of molecules (Figure 2) obtains a continuous density curve. This distribution ranges from 0 to 97.2% similarity with a mean value of 31.8% ± 21.2% similarity. Comparing with fingerprints for which only five different values of similarity between the 4950 pairs where obtained (71.4%, 75.0%, 83.3%, 85.7%, and 100%), using NAMS 859 different values of similarity are obtained showing its discriminative power. It is also interesting to mention that a similarity score between a pair of molecules of 100% was never obtained with NAMS, since one of its fundamental assumptions is that a molecule should only have a 100% similarity score when compared with itself.

Considering the question of effectiveness, i.e. being able to differentiate between molecules that are structurally different, Figure 3 gives an example of the similarity scores obtained using fingerprints and NAMS for four considerably structurally

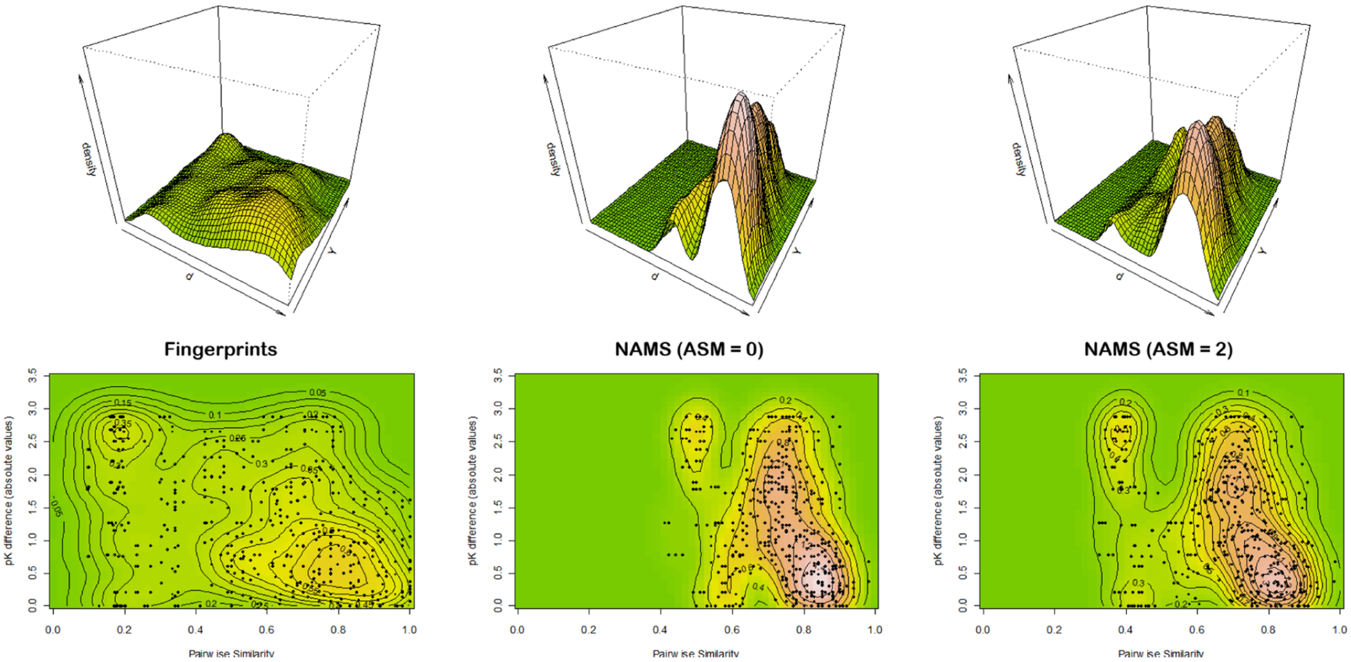


Figure 4. 2D kernel density estimator perspective and contour plots showing the distribution of the pairwise similarity between the 31 steroids, totalizing 465 different pairs of structures (excluding self-similarities), calculated using fingerprints and NAMS (using two different atom substitution matrices (Supporting Information 3 ASM = 0 and 2) and the corresponding difference in the pK absolute value.

different compounds (Figure 3I) and four considerably structurally similar compounds (Figure 3II). The similarity score using fingerprints for all the molecules in the example is 100%; therefore, there is no discriminative power between these structures. On the other hand, using NAMS to compare the considerably different structures (Figure 3I), it is possible to verify that for example the structure a is more similar to the structure b than any of the remaining structures, since both have two fused rings and similar size. Using NAMS to compare the similar structures (Figure 3II) it is possible to distinguish them in terms of shape and size. The structure h has 7 carbon atoms, while all the others have 8 carbon atoms; therefore, it is the structure with lowest similarity scores when compared with the others. The structures e and f are the most similar ones, since the only difference between them is the position of one methyl group. The structure g is more similar to the structure f because both have a substituent group in the position 3, a methyl in f, and an ethyl in g.

Discriminate Similar Molecules with Different Activity Levels. The data set B was chosen because although the structures have a similar structure, their activity level ranges from −5 to −7.881 with a mean value of −6.384 ± 1.082. Considering that the binding strength of a receptor−substrate complex strongly depends on the shape of the substrate, the aim is to analyze the difference in the binding affinity of each pair of steroids to the corticosteroid binding globulin (CBG) receptor solely based on their similarity.

Figure 4 displays the 2D kernel density estimator perspective and contour plots showing the distribution of the pairwise similarity between the 31 steroids, totalizing 465 different pairs of structures (excluding self-similarities), calculated using fingerprints and NAMS (with two different atom substitution matrices: the first one considers that each atom type is only fully similar to itself and completely different from all the others and the second one which considers different similarities between the atoms, based on their position in the periodic table (Supporting Information 3 ASM = 0 and 2)) and the corresponding difference in the pK absolute value. While the pairwise similarity using fingerprints has a wider distribution, using NAMS concentrates the similarity between 35 and 98%. Using NAMS, there are two main zones with an high density of pairs of steroids, one between 70% and 80% of similarity and another one with higher similarity values (>80%) and the difference in the absolute value of the binding affinity becomes smaller with the increase of similarity. There is an isolated island of pairs of steroids using NAMS with similarity values between 40 and 50% for the atom substitution matrix 0 or 35 and 45% for the atom substitution matrix two and high differences in the absolute value of the binding affinity.

The pairs of compounds in this island were further investigated, leading to the conclusion that three of the compounds were always present in these pairs, namely estradiol, estriol, and estrone (Figure 5). All of these compounds are estrogenic steroids, and although they share strong resemblance with the remaining steroids, there are some differences due to the aromatization process to convert anabolic steroids in estrogens. The "A" ring in the skeleton of a steroid (Figure 5a) is now aromatic, and it is a key functional group in all estrogens.

In general, it is possible to verify that although the structure of the 31 steroids is very similar, NAMS is able to discriminate them according to their pairwise similarity versus their difference in the binding activity level, since the density of points in Figure 5 is higher for higher similarity values versus lower difference in the binding activity level. NAMS was also able to discriminate the group of estrogenic steroids from the rest of the steroids. Using the fingerprints it is not possible to discriminate or relate the similarity between the molecules versus their difference in the binding activity level, since the distribution is wider and does not demonstrate any patterns.

Molecular Similarity for Inference. For data set C, the aim is to retrieve compounds with similar activity level based on

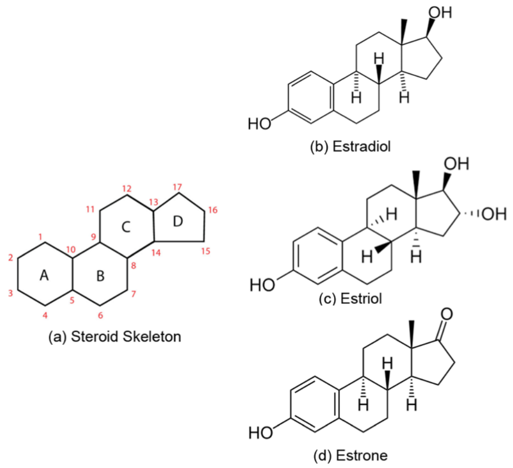


Figure 5. Basic skeleton of a steroid (a) and estrogenic steroids: (b) estradiol, (c) estriol, and (d) estrone.

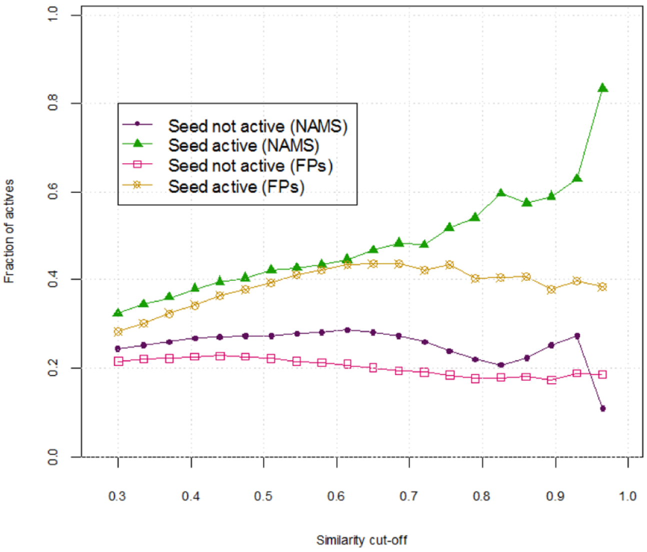


Figure 7. Fraction of active compounds (activity level = 1, 2, or 3) within those compounds similar to compounds that are active (activity level =1, 2, or 3) or inactive (activity level = 0) using fingerprints (FPs) and NAMS with different threshold cutoffs (starting from 30%).

the similarity threshold. For that purpose, using each compound as seed in each level of activity, the fraction of actives that are retrieved using fingerprints and NAMS to measure the similarity between the compounds with different threshold cutoffs are recorded.

Figure 6 shows the fraction of actives within those compounds similar to compounds of each level of activity (0, 1, 2, or 3), given the minimum threshold of similarity for the search using fingerprints (Figure 6a) or NAMS (Figure 6b). Figure 7 shows the fraction of actives (low, moderate, or high activity) within those compounds similar to compounds that are active or inactive using fingerprints and NAMS, given the minimum threshold of similarity for the search.

In general, using NAMS, the fraction of active compounds retrieved decreases as the similarity to an inactive is increased and increases as the similarity to an active is increased and the level of activity is higher. The same tendency is verified using fingerprints, except when the seed has a low activity level (level 1), following the curve for inactive seeds. However, the fraction of actives retrieved when the seed is active is higher when using NAMS for the same similarity cutoff, particularly to higher levels of similarity. The fraction of actives within similar compounds of high activity (level 3) using NAMS is similar to using fingerprints until reaching a cutoff level of 70% of similarity, from this point on the fraction of actives retrieved is higher and increases to 1 at a cutoff level of 96.5% of similarity.

These results support the similarity principle, which states that compounds similar to biologically active ones should also

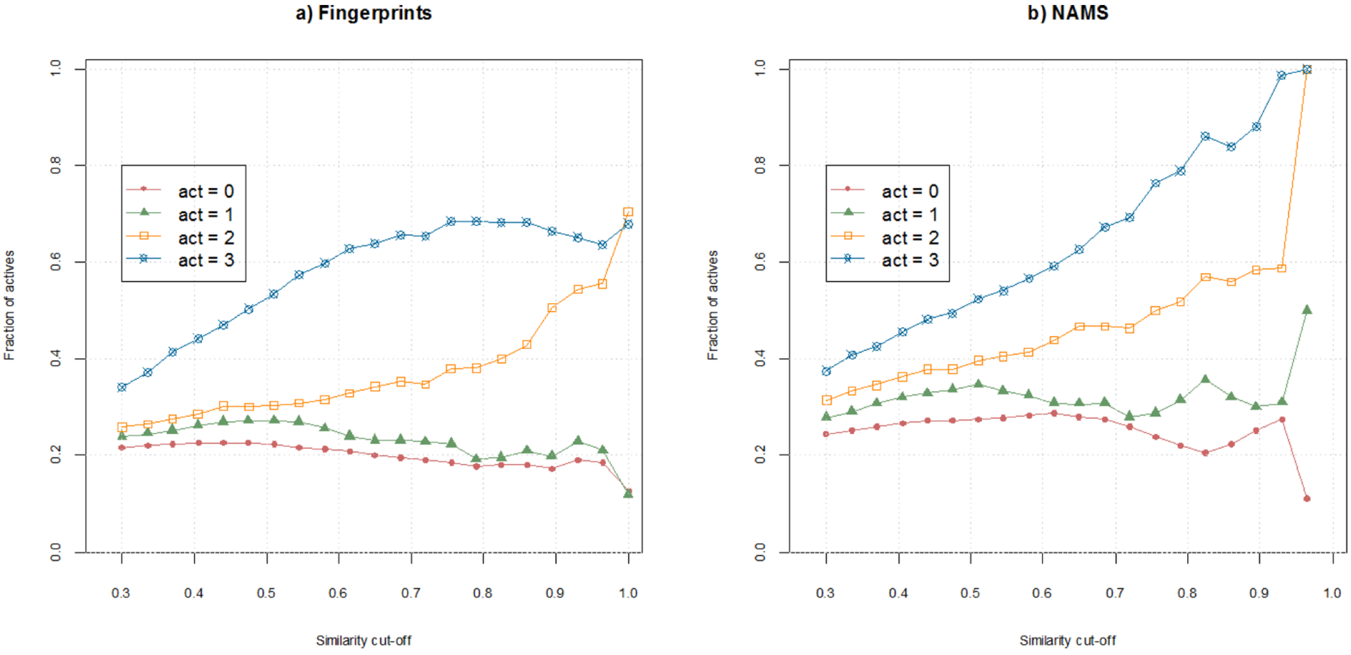
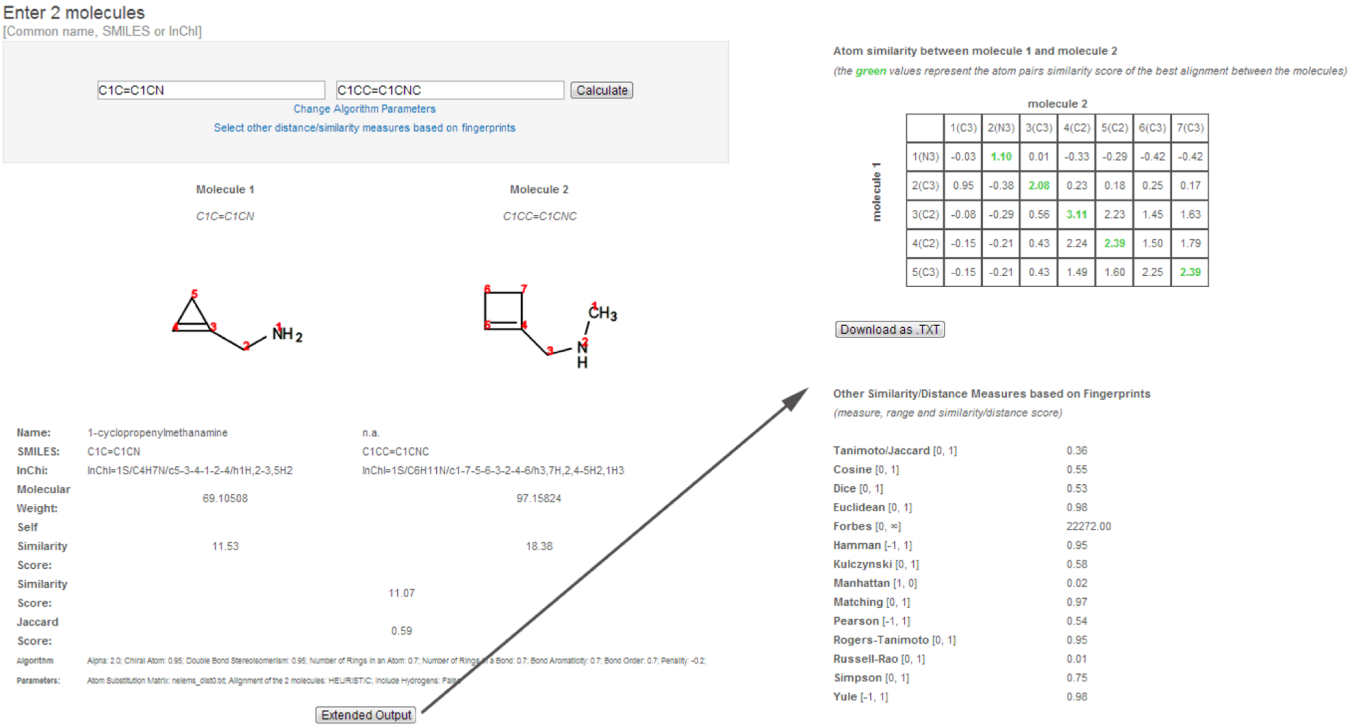


Figure 6. Fraction of active compounds in the data set that are similar to seeds with a certain level of activity (act = 0, act = 1, act = 2, act = 3) used for the similarity search with different threshold cutoffs (starting from 30%).



Enter 2 molecules
[Common name, SMILES or InChI]

C1C=C1CN | C1CC=C1CNC | Calculate

Change Algorithm Parameters

Select other distance/similarity measures based on fingerprints

Molecule 1: C1C=C1CN

Molecule 2: C1CC=C1CNC

	Molecule 1	Molecule 2
Name:	1-cyclopropenylmethanamine	n.a.
SMILES:	C1C=C1CN	C1CC=C1CNC
InChi:	InChI=1S/C4H7N/c5-3-4-1-2-4/h1H,2-3,5H2	InChI=1S/C6H11N/c1-7-5-6-3-2-4-6/h3,7H,2,4-5H2,1H3
Molecular Weight:	69.10508	97.15824
Self Similarity Score:	11.53	18.38
Similarity Score:	11.07	
Jaccard Score:	0.59	

Algorithm Parameters: Alpha: 2.0; Chiral Atom: 0.95; Double Bond Stereoisomerism: 0.95; Number of Rings in an Atom: 0.7; Number of Rings in a Bond: 0.7; Bond Aromaticity: 0.7; Bond Order: 0.7; Penalty: -0.2; Atom Substitution Matrix: nelems_dist0.txt; Alignment of the 2 molecules: HEURISTIC; Include Hydrogens: False

Extended Output

Atom similarity between molecule 1 and molecule 2

(the green values represent the atom pairs similarity score of the best alignment between the molecules)

molecule 1 \ molecule 2	1(C3)	2(N3)	3(C3)	4(C2)	5(C2)	6(C3)	7(C3)
1(N3)	-0.03	1.10	0.01	-0.33	-0.29	-0.42	-0.42
2(C3)	0.95	-0.38	2.08	0.23	0.18	0.25	0.17
3(C2)	-0.08	-0.29	0.56	3.11	2.23	1.45	1.63
4(C2)	-0.15	-0.21	0.43	2.24	2.39	1.50	1.79
5(C3)	-0.15	-0.21	0.43	1.49	1.60	2.25	2.39

Download as .TXT

Other Similarity/Distance Measures based on Fingerprints

(measure, range and similarity/distance score)

Tanimoto/Jaccard [0, 1]	0.36
Cosine [0, 1]	0.55
Dice [0, 1]	0.53
Euclidean [0, 1]	0.98
Forbes [0, ∞]	22272.00
Hamman [-1, 1]	0.95
Kulczynski [0, 1]	0.58
Manhattan [1, 0]	0.02
Matching [0, 1]	0.97
Pearson [-1, 1]	0.54
Rogers-Tanimoto [0, 1]	0.95
Russell-Rao [0, 1]	0.01
Simpson [0, 1]	0.75
Yule [-1, 1]	0.98

Figure 8. Screenshot of the NAMS Web-tool output when comparing two simple molecules: graphical representation of the molecules under comparison and molecule identifiers, self-similarity scores for each molecule, similarity between both molecules, atom similarity between both molecules with the best possible alignment highlighted in green, and finally other similarity scores using different similarity/distance coefficients based on fingerprints.

be active and vice versa, especially when the molecular comparison is highly discriminative.

Web Tool. A public and free Web tool for NAMS has been implemented. The objective was to produce an application simple to use with easily readable results that would allow the determination of the molecular similarity based on NAMS between a pair or a list of molecules. The user can input the molecules using their common name, SMILES string, or InChI identifier. The common name is resolved using the Chemical Identifier Resolver (http://cactus.nci.nih.gov/chemical/structure), directly called by the application. The user can also define several parameters for the atom and bond characteristics, already described above, that will influence the atom/bond matching similarity score. It is important to note that to avoid ambiguity NAMS only considers double bond stereoisomerism or atomic chirality as characteristics if the stereo information is correctly and explicitly written in the molecule identifier.[37]

The output produced consists of the molecule identifiers, a graphical representation of the molecular structure with the canonical numeration of atoms (generated using Openbabel), the similarity score between each pair of molecules, and a matrix of the atom similarity between both molecules, with the best possible alignment between the molecules highlighted (Figure 8). Furthermore, it is also possible to calculate the similarity score using different similarity/distance functions (e.g., Tanimoto, Cosine, Dice, Euclidean, etc.) based on fingerprints.

This Web tool was developed, mainly, in the PHP programming language. The application communicates with the Python code that uses Openbabel for converting the different representations of the molecule and determining the similarity score of the molecules in accordance with the described algorithm. This Web tool as well as the Python source code are freely available at http://nams.lasige.di.fc.ul.pt/.

CONCLUSIONS

In this work, we have defined a new noncontiguous atomic alignment method for the analysis of structural similarity between molecules. The atomic alignment approach often requires high computational cost; however, the similarities detected by the atom correspondence are more intuitively understood because similar atoms in the molecules are explicitly shown. This method is based on the comparison of atoms on comparable molecules taking into account their topological profiles. The similarity measure is defined on the annotated molecular graph, based on a recursive concept of graph similarity and an optimal alignment between atoms using a heuristic and a penalty function to account for the differences in both atoms/bonds characteristics and topological profiles. The stereoisomerism and chirality are also considered in this similarity function since they are of great importance in many different fields since the molecular properties and biological effects of the stereoisomers are often significantly different. Considering that all similarity functions have a context that both define and limit their use, all defined atomic/bonds characteristics have a corresponding weight that can be adjusted or even eliminated in accordance with the context of the problem. New characteristics are also rather easy to include in the method.

The number of parameters used by NAMS may seem a deterrent for its use, but the tests made suggest that, despite the fact that individual similarity scores do change, the similarity patterns are identical when comparing large databases. Also, the empirical tests over the three data sets presented strongly suggest that predefined default parameter values able to provide coherent results is attainable. Furthermore the nature of the method leads to its higher computational cost. We are currently in the process of rewriting the current implementation of NAMS in the C programming language and using triangulation hierarchies[58] in order to implement neighborhood search procedures.

NAMS was compared with one of the most widely used similarity methods (fingerprint-based similarity) using three data sets with different specificities. The method performed well and compared favorably to fingerprints for all three test cases. NAMS was able to distinguish either different or very similar hydrocarbons that were indistinguishable using a fingerprint-based approach and verifying the similarity principle using a data set of very similar steroids with differences in the binding affinity to the corticosteroid binding globulin receptor. The method was also able to recover a significantly higher average fraction of active compounds when searching a database of highly diverse set of molecules with information about the MAO inhibition level. For this set it was verified that the fraction of actives recovered per active seed searched, consistently increased with the similarity level, which further suggests that NAMS is actually capturing reliable structure−activity relationships.

It is nonetheless important to remember that, although structurally similar molecules are expected to exhibit similar properties, in some cases small changes in the structure of a molecule can bring thorough changes in some properties. Therefore, we cannot expect that in the context of property prediction there is a linear relationship between the molecular similarity of a pair of compounds and all the corresponding properties of that pair of molecules. A good similarity method is useful to construct a map of the chemical space; however, this is not enough to make good property/activity predictions. We are currently working on the development of tools able to analyze the chemical space defined by NAMS which may be able to recognize and make sense of structural patterns and their effects for property/activity prediction.

NAMS is made available freely for the whole community in a simple Web-based tool as well as the Python source code at http://nams.lasige.di.fc.ul.pt/.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Supporting Information 1: data set A—hydrocarbons. A complete table with compound name, CASRN, SMILES, similarity matrix using fingerprints, and similarity matrix using NAMS for the 100 hydrocarbons. Supporting Information 2: data set B—steroids and their binding affinity to the corticosteroid binding globulin receptor. A complete table with compound name, SMILES, binding affinity (pK), activity level, similarity matrix using fingerprints, and similarity matrix using NAMS for the 31 steroids (including the two atom substitution matrices (ASM = 0 and ASM = 1) used for the similarity calculation). Supporting Information 3: atom substitution matrices. Five different matrices used for comparing the nature of the atoms in terms of a distance score, varying from 0 to 1. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ateixeira@lasige.di.fc.ul.pt.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge Dr. Yvonne Martin, Dr. Robert Brown, and Abbott for the monoamine oxidase inhibitor dataset. The authors thank the Portuguese Fundação para a Ciência e Tecnologia for the Multiannual Funding Programme of the LaSIGE laboratory and doctoral grant SFRH/BD/64487/2009.

■ REFERENCES

(1) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem* **2004**, *2*, 3204−3218.

(2) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

(3) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discov.* **1998**, *9*−*11*, 225−252.

(4) Bajorath, J. Selected Concepts and Investigations in Compound Classi-cation, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245, DOI: 10.1021/ci0001482.

(5) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(6) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(7) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4183−4199, DOI: 10.1021/jm0582165.

(8) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225−233.

(9) Auer, J.; Bajorath, J. Molecular Similarity Concepts and Search Calculations. In *Bioinformatics*; Keith, J. M., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2008; Vol. *453*; pp 327−347.

(10) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059, DOI: 10.1021/jm960290n.

(11) Todeschini, R.; Consonni, V. In *Molecular Descriptors for Chemoinformatics*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, 2009.

(12) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119, DOI: 10.1021/ci800249s.

(13) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358, DOI: 10.1021/jm020155c.

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

(15) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386, DOI: 10.1021/ci970437z.

(16) Willett, P. Similarity-based virtual screening using 2D -ngerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(17) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity. *ChemMedChem* **2007**, *2*, 208−217.

(18) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of Coe-cients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Comb. Chem. High Throughput Screen* **2002**, *5*, 155−166.

(19) Ehrlich, H.-C.; Rarey, M. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 68−79.

(20) Barnard, J. M. Substructure searching methods: Old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538, DOI: 10.1021/ci00014a001.

(21) Raymond, J.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(22) Garey, R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*; W. H. Freeman: New York, 1979.

(23) Rahman, S.; Bashton, M.; Holliday, G.; Schrader, R.; Thornton, J. Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminf. [Online]* **2009**, *1*, No. 12, http://www.jcheminf.com/content/1/1/12 (accessed March, 2013).

(24) Hagadone, T. R. Molecular substructure similarity searching: e-cient retrieval in twodimensional structure databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521, DOI: 10.1021/ci00009a019.

(25) Berglund, A. E.; Head, R. D. PZIM: A Method for Similarity Searching Using Atom Environments and 2D Alignment. *J. Chem. Inf. Model.* **2010**, *50*, 1790−1795, DOI: 10.1021/ci1002075.

(26) Kawabata, T. Build-Up Algorithm for Atomic Correspondence between Chemical Structures. *J. Chem. Inf. Model.* **2011**, *51*, 1775−1787, DOI: 10.1021/ci2001023.

(27) Brown, R. D.; Jones, G.; Willett, P.; Glen, R. C. Matching two-dimensional chemical graphs using genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 63−70, DOI: 10.1021/ci00017a008.

(28) Wang, T.; Zhou, J. EMCSS: A New Method for Maximal Common Substructure Search. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 828−834, DOI: 10.1021/ci9601675.

(29) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic identi-cation of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639−643, DOI: 10.1021/ci00010a009.

(30) Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471−490.

(31) Batista, J.; Godden, J. W.; Bajorath, J. Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **2006**, *46*, 1937−1944, DOI: 10.1021/ci0601261.

(32) García, G. C.; Ruiz, I. L.; Gómez-Nieto, M. A. Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 30−41, DOI: 10.1021/ci034167y.

(33) Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Com-put. Sci.* **1998**, *38*, 983−996.

(34) Gillet, V. J.; Wild, D. J.; Willett, P.; Bradshaw, J. Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. *Comput. J.* **1998**, *41*, 547−558, DOI: 10.1093/comjnl/41.8.547.

(35) Bajorath, J. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Biomed Protocols; Humana Press: Totowa, NJ, 2004.

(36) Bronstein, A.; Bronstein, M.; Kimmel, R. *Numerical geometry of non-rigid shapes*; Springer: New York, 2008.

(37) Teixeira, A. L.; Leal, J. P.; Falcao, A. O. *Automated Identification and Classification of Stereochemistry: Chirality and Double Bond Stereoisomerism*; Technical Report, Department of Informatics, Faculty of Sciences, University of Lisbon: Lisbon, 2013; arXiv:1303.1724

(38) Islam, M.; Mahdi, J.; Bowen, I. Pharmacological Importance of Stereochemical Resolution of Enantiomeric Drugs. *Drug Safety* **1997**, *17*, 149−165.

(39) Köbler, J.; Schcöning, U.; Torán, J. *The Graph Isomorphism Problem: Its Structural Complexity*; Progress in Theoretical Computer Science Series; Birkhäuser: Boston, 1993.

(40) Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Log.* **1955**, *2*, 83−97.

(41) Munkres, J. Algorithms for the Assignment and Transportation Problems. *J. Soc. Ind. Appl. Math.* **1957**, *5*, 32−38.

(42) Teixeira, A. L.; Santos, R. C.; Simoes, J. A. M.; Leal, J. P.; Falcao, A. O. *ThermInfo: Collecting, Retrieving, and Estimating Reliable Thermochemical Data*; Technical Report, Department of Informatics, Faculty of Sciences, University of Lisbon: Lisbon, 2013; arXiv:1302.0710

(43) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular -eld analysis (CoMFA). 1. E-ect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(44) Good, A. C.; So, S. S.; Richards, W. G. Structure-activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

(45) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(46) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 521−534.

(47) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-Dimensional Quantitative Structure-Activity Relationships from Tuned Molecular Quantum Similarity Measures:â€. Prediction of the Corticosteroid-Binding Globulin Binding A-nity for a Steroid Family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.

(48) Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Perakyla, M. Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 607−613.

(49) *MAO Dataset, Pipeline Pilot v8.5*; Accelrys: San Diego CA, USA, 2008.

(50) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(51) Willett, P. Evaluation of Molecular Similarity and Molecular Diversity Methods Using Biological Activity Data. In *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Bajorath, J., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2004; Vol. 275; pp 51−63.

(52) Heikamp, K.; Bajorath, J. Large-Scale Similarity Search Pro-ling of ChEMBL Compound Data Sets. *J. Chem. Inf. Model.* **2011**, *51*, 1831−1839, DOI: 10.1021/ci200199u.

(53) Delaney, J. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Divers.* **1996**, *1*, 217−222.

(54) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(55) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(56) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminf. [Online]* **2011**, *3*, No. 33, http://www.jcheminf.com/content/3/1/33 (accessed March, 2013).

(57) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47*, 4891−4896.

(58) Jones, C. B.; Ware, J. M. Proximity Search with a Triangulated Spatial Model. *Comput. J.* **1998**, *41*, 71−83.