

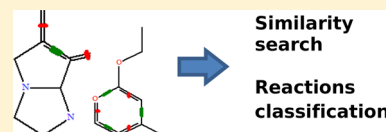
# Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches

Aurélie de Luca,<sup>†</sup> Dragos Horvath,<sup>†</sup> Gilles Marcou,<sup>†</sup> Vitaly Solov'ev,<sup>†,‡</sup> and Alexandre Varnek<sup>\*,†</sup>

<sup>†</sup>Laboratoire d'Infochimie, UMR7177 CNRS, Université de Strasbourg, 4 rue B. Pascal, Strasbourg Cedex, 67008 France

<sup>‡</sup>Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Leninskiy prospect, 31a, 119991 Moscow, Russian Federation

**ABSTRACT:** This work addresses the problem of similarity search and classification of chemical reactions using Neighborhood Behavior (NB) and Condensed Graphs of Reaction (CGR) approaches. The CGR formalism represents chemical reactions as a classical molecular graph with dynamic bonds, enabling descriptor calculations on this graph. Different types of the ISIDA fragment descriptors generated for CGRs in combination with two metrics – Tanimoto and Euclidean – were considered as chemical spaces, to serve for reaction dissimilarity scoring. The NB method has been used to select an optimal combination of descriptors which distinguish different types of chemical reactions in a database containing 8544 reactions of 9 classes. Relevance of NB analysis has been validated in generic (multiclass) similarity search and in clustering with Self-Organizing Maps (SOM). NB-compliant sets of descriptors were shown to display enhanced mapping propensities, allowing the construction of better Self-Organizing Maps and similarity searches (NB and classical similarity search criteria – AUC ROC – correlate at a level of 0.7). The analysis of the SOM clusters proved chemically meaningful CGR substructures representing specific reaction signatures.



## 1. INTRODUCTION

Most of chemoinformatics 2D approaches such as structure–property modeling, similarity search, and clustering have been developed for individual molecules, which can be encoded by bit strings (fingerprints) or descriptors vectors. Chemical reactions involving several molecular species of two classes (reactants and products) are difficult objects because *a priori* it is not clear how to combine the information about all species involved. Several efforts in this direction have been reported. They concern using either physicochemical<sup>1–6</sup> or fragment descriptors.<sup>7–11</sup> Thus, Gasteiger and collaborators used an ensemble of some physicochemical descriptors (total charge,  $\sigma$ -electronegativity,  $\pi$ -electronegativity, polarizability, and aromaticity indices) calculated only for selected atom(s) of the product(s)<sup>1,2</sup> or substrates.<sup>3</sup> Self-Organized Maps (SOMs) calculated on these descriptors were further used for reactions classification. Aires de Sousa et al.<sup>4</sup> invented MOLMAP descriptors extracted from individual SOMs for each product and reactant. For the whole reaction, the descriptors were calculated as a difference between MOLMAPs of products and reactants. This approach has been successfully used to classify enzymatic reactions.<sup>5</sup> To classify enzymatic reactions Ridder et al.<sup>6</sup> generated fingerprints based on Sybyl atom types for reactant and product molecules separately, followed by calculation of the difference fingerprints (defined by the differences in occurrence of each atom type in the reactant and product fingerprints).

In most of the studies exploiting fragment descriptors, the descriptor vector for a chemical reaction is calculated as a difference between occurrences of selected substructural fragments, e.g. "reaction signatures" by Faulon et al.<sup>7</sup> and RMNA descriptors by Borodina et al.<sup>8</sup> Daylight fingerprints of

reactions are computed as the difference between the fingerprint of the reactant molecules and the fingerprint of the product molecules. These reflect the bond changes which occur during the reaction.<sup>9</sup>

An original approach reported by Varnek et al.<sup>10</sup> suggests to generate descriptor vector of reaction from Condensed Graph of Reaction (CGR) introduced earlier by Fujita as an "imaginary transition state".<sup>11</sup> A CGR represents a chemical reaction involving several molecular species as single pseudomolecule described by special dynamic bonds (see the Method section). This opens an opportunity to extend on reactions classical chemoinformatics approaches designed for individual molecules. However, one cannot calculate any physicochemical parameter for a pseudomolecule; only some topological or fragment descriptors can be generated from a CGR. In particular, in ref 10 ISIDA fragments descriptors (atom/bond sequences or atom-centered fragments)<sup>12</sup> have been used.

In this paper, ISIDA descriptors generated on CGRs will be used for generic (multiclass similarity) search in reaction databases (DB) and reaction DB classification. The question arises: which particular descriptor types are most suitable for this purpose?

For chemical reactions, we will adopt the following working hypothesis to be equivalent to the similarity principle: "Similar reactions are likely to belong to the same reaction class", i.e. likely occur according to the same chemical transformation. Reactions will be mapped, like molecules, to points in some "Chemical Space" (CS).<sup>12</sup> Then, the hypothesis as to whether

Received: March 17, 2012

“neighboring” points in this CS (according to the employed CS metric<sup>13</sup>) systematically correspond to reactions of the same class will be tested for statistical relevance.

Until now, in spite of the existence of many meaningful reaction representations, no previous study was based on benchmarks of various description hypotheses. This is done according to the Neighborhood Behavior (NB) criterion,<sup>14,15</sup> used to select an optimal combination of fragment descriptors and metric in order to distinguish, at best, several types of reaction, from calculated distances between organic reactions. Following Patterson et al.<sup>16</sup> and Dixon et al.,<sup>17</sup> Horvath et al.<sup>14</sup> proposed some criteria in order to quantify the NB.

This paper advocates a *rigorous NB benchmarking* as a guidance criterion for the mapping and analysis of chemical reaction spaces from various points of view, in highlighting both the common trends and the specificity of various approaches. It proposes an answer to these three questions:

1. Similarity-driven searches for “analogue” reactions are performed, starting from multiple queries of several reaction types, and assessed in terms of their specificity to retrieve processes of the same class. The success of similarity search associated with the optimal descriptor/metric is an intrinsic expression of NB quality – but how far are classical similarity search success scores, such as the Area under the ROC curve, correlated to NB scores?

2. Is the NB property a relevant indicator of the propensity of a CS to allow for meaningful nonlinear CS maps to be built? Self-Organizing Maps (SOMs) were constructed and assessed in terms of mapping quality, in order to check whether NB-compliant CS are also the optimally described by SOMs.

3. Last but not least, can we use NB analysis to facilitate the specific substructural reaction keys determination?

## 2. METHODS

This section first describes the data used, then the construction of considered CS based on fragment descriptors, going next to a presentation of quantitative NB, alternative virtual screening criteria (Area under ROC curve), and Kohonen mapping tools and quality criteria, used to probe the relevance of these CS. Readers already familiar with this methodology may directly proceed to §2.7, the last, but essential chapter of this subsection which describes the herein performed numeric simulations and benchmark tests, using the previously introduced concepts and algorithms.

**2.1. Data Sets and Simulation Design.** The reaction DB used here represents a sample of 8544 chemical reactions of 9 classes (Table 1) retrieved by substructural searches from the ChemInform and Replib databases. ChemInform is composed of some 900 000 historical reactions filed by FIZ CHEMIE Berlin since 1900 to date. Data contain information on structures and reaction conditions. Replib is a MDL database, last updated in 1991. It is composed of 209 800 organic and organometallic reactions (with structures and links to publications).

This data set has been split into the training set (TS) used to carry out the NB scoring and the validation set (VS) used to check how well these CS do in actual reaction mapping and virtual screening studies. The relative sizes of TS and VS have been set by means of a NB robustness study. Unlike customary chemoinformatics studies, the VS here represents the larger part of DB (size ratio TS/VS = 1/4 – see the Results section). Please note that further on all NB (global and local) scores are implicitly obtained on TS molecules. By contrast, Kohonen

**Table 1. Classes and Numbers of Reactions in the Database, Size of the Validation, and the Training Sets**

reaction classes	index	database	test set	training set
nucleophilic substitutions on carbon 2	C2	274	219	55
nucleophilic substitutions on carbon 4	C4	76	60	16
Sonogashira reactions	SO	1938	1550	388
Diels–Alder reactions	DA	3048	2438	610
Diels–Alder reactions on heteroatom(s)	DA-HET	935	748	187
Domino–Heck Diels–Alder reactions	DH-DA	28	22	6
dihydroxylations	HYD	1041	832	209
epoxydations	EPO	496	396	100
metathesis	MET	708	566	142

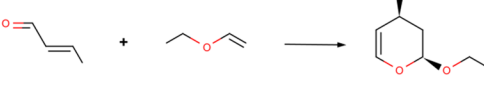
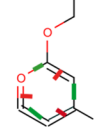
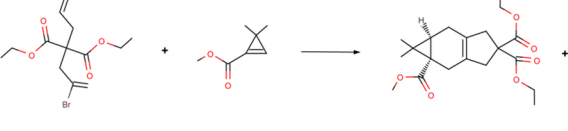
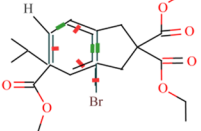
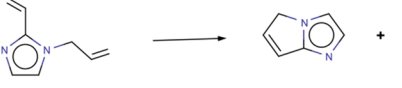
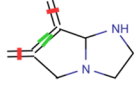
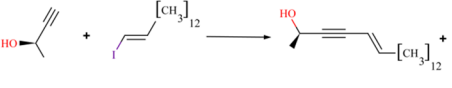
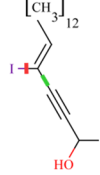
maps were used to map the VS reactions. “Training” here does not mean that the TS served to build a model to be challenged against VS reactions but refers to the “learning” (choosing) of a *Chemical Space*, to be validated in terms of its propensity to generate a valuable Kohonen map of the VS. Final Kohonen quality criteria and statistics are always reported with respect to VS reactions. Area-under-Curve values of ROC curves are also reported for searches, within the VS, of similar reactions for given query processes.

**2.2. Condensed Graphs of Reaction.** The “Condensed Graph of Reaction” (CGR) is a pseudomolecule which results from the superposition of reactant and product molecular graphs into one single graph.<sup>12</sup> The chemical transformations are tacitly taken into account through dynamic bonds. Dynamic bonds represent covalent bonds being broken, formed, or transformed during the reaction (Figure 1). Atom-to-atom mapping should be performed in order to identify superposed atoms in reactants and products. Here, we used ChemAxon tools for atom-to-atom mapping and the ISIDA-CGR Designer program in order to convert all studied reactions into CGRs.

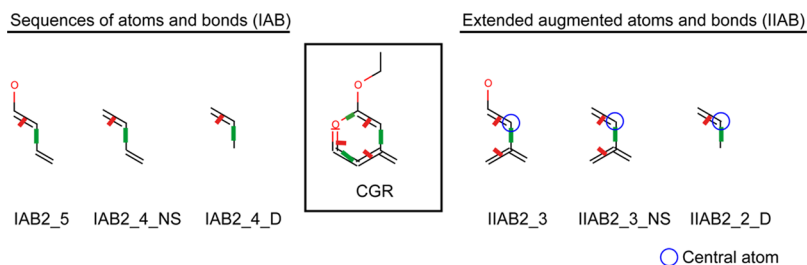
The data manager ISIDA EdiSDF of Structure-Data Files (SDF)<sup>12,18,19</sup> and the chemical editor ISIDA EdChemS<sup>12,18,20</sup> were extended for CGR handling and visualization. Presentation of various dynamic bonds was designed and built in the programs (Figures 1 and 2) and (sub)structure search was realized for CGRs.

**2.3. Reaction Descriptors: Building the Chemical Spaces.** CGRs were fragmented using the ISIDA Fragmentor,<sup>21</sup> which is the original program appreciably extending the corresponding module of the ISIDA/QSPR program.<sup>19,22,23</sup> ISIDA Fragmentor generates substructural molecular fragments as subgraphs of molecular graphs. A fragment occurrence is a descriptor value. CGRs were represented with implicit hydrogen atoms. Two types of fragment descriptors were tested: sequences (I) and extended augmented fragments (II) (Figure 2 and Table 2).

Fragments of type I correspond to the shortest topological paths along bonds between every atom pair, and the path length is defined by the number of atoms in the sequence. Fragments of type II are fragments centered on one atom, and its environment is included in the  $x^{\text{th}}$  sphere(s) around the centered atom. Two subclasses of sequence fragments are produced: sequences of atoms and bonds (IAB) and bonds sequences (IB – ignoring the nature of involved atoms). Two subclasses of type II are tested as well: extended augmented

CLASSES	REACTION	CGR
DA-HET		
DH-DA		
MET		
SO		

**Figure 1.** Examples of similarity search queries (at left) and the associated Condensed Graph of Reactions (CGRs) (at right): transformed, broken, and created bonds are called dynamic bonds. Dynamic bonds are represented by green bonds if bonds are created and by red bonds if they are broken.



**Figure 2.** Illustration of different types of fragmental descriptor used in this study: only fragments containing at least one dynamic bond (red if broken, green if created) are considered. With the NS option, fragmentation does not include any single bonded moieties. With D, only substructures based exclusively on dynamic bonds are analyzed. In the bond (IB and IIB) fragments, the nature of atoms is ignored.

atoms and bonds (IIAB) and extended augmented bonds (IIB), respectively. Moreover, three particular options are used to deal with CGR. Effectively, what is important in a chemical reaction is the reaction center (RC). Therefore, descriptors are limited to counting the fragments that contain either (a) at least one dynamic bond, or (b) at least one dynamic bond and no single bond around the RC (option NS), or (c) only dynamic bonds (option D).

For the second key parameter of the CS, the employed metric, two classical alternatives have been explored: Euclidean and Tanimoto<sup>13</sup> (for the latter, the distance is being defined as one minus the similarity coefficient).

**2.4. Neighborhood Behavior Criteria. 2.4.1. Concept of Neighborhood Behavior.** The fragment descriptors, used in this study, contain at least one dynamic bond. The similarity principle applied to chemical reactions can be tentatively defined as a hypothesis that similar reactions have similar reaction centers, and, therefore, similar chemical reactions belong most probably to the same reaction class.

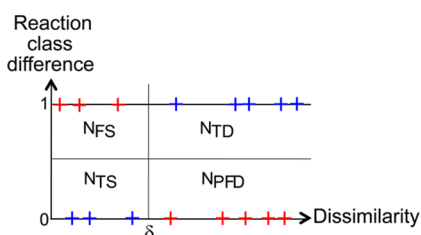
Neighborhood Behavior<sup>24</sup> quantifies the extent to which this similarity principle holds for the objects in the CS based on the above-mentioned fragment counts, in conjunction with the employed similarity metric. Optimal NB is observed when similar objects (in terms of the reaction class defined) are *selectively* grouped together in the CS.<sup>25–27</sup>

Assuming that reaction pairs of high calculated similarity should belong to the same class, the NB optimality score at given (dis)similarity threshold  $\delta$  represents (Figure 3) a weighed sum of both (a) misclassified pairs of reactions: “false” similars, not belonging to the same class, and (b) missed pairs of reactions: “potentially false” dissimilars which do belong to the same class, albeit their calculated similarity was not high enough to pass the selection threshold  $\delta$ . The balance is biased in favor of false similar pairs (a), which are a more serious problem in similarity-based virtual screening. The impact of the choice of the NB bias factor  $\kappa > 1$  on the relative ranking of chemical spaces was explicitly addressed in this work (§3.2). The above-mentioned sum is compared to a random

**Table 2. Types of Tested Fragment Descriptors: Sequences (Type I) and Extended Augmented Fragments (Type II) Appear under Two Possible Embodiments: Including Both Atom and Bond (AB) Information or Bond Information Only (B) Are Considered<sup>a</sup>**

index	subclass of fragment types	Min	Max	definition
1	IAB	2	[2;7]	atoms and bonds sequences of length $\text{Min} < n < \text{Max}$
2	IB	2	[2;7]	bonds sequences of length $\text{Min} < n < \text{Max}$
3	IIAB	2	[2;5]	atoms and bonds spheres centered on one atom of size $\text{Min} < n < \text{Max}$
4	IIB	2	[2;6]	bonds spheres centered on one atom of size $\text{Min} < n < \text{Max}$
5	IAB_NS	2	[2;5]	atoms and bonds sequences without single bonds near the reaction center (RC) of length $\text{Min} < n < \text{Max}$
6	IB_NS	2	[2;8]	bonds sequences without single bonds near the RC of length $\text{Min} < n < \text{Max}$
7	IIAB_NS	2	[2;3]	atoms and bonds spheres centered on one atom of size $\text{Min} < n < \text{Max}$ without single bonds near the RC
8	IIB_NS	2	[2;4]	bonds spheres centered on one atom of size $\text{Min} < n < \text{Max}$ without single bonds near the RC
9	IAB_D	2	[2;8]	atoms and bonds sequences with only dynamic bonds of length $\text{Min} < n < \text{Max}$
10	IB_D	2	[2;8]	bonds sequences with only dynamic bonds of length $\text{Min} < n < \text{Max}$
11	IIAB_D	2	[2;8]	atoms and bonds spheres centered on one atom of size $\text{Min} < n < \text{Max}$ with only dynamic bonds
12	IIB_D	2	[2;8]	bonds spheres centered on one atom of size $\text{Min} < n < \text{Max}$ with only dynamic bonds

<sup>a</sup>Consequently, the nomenclature rule is the following: *Type\_InformationContent\_Size\_Option* where type is for I or II, content for AB or B, size outlines minimal and maximal atom numbers: Min, Max and option (NS or D) is the filter used to pick relevant fragments, involved in the actual transformation. All fragments must contain, at least, one dynamic bond.



**Figure 3.** Neighborhood Behavior scoring is based on the class difference of selected reactions pairs (0 if the two reactions are of same classes and 1 on the contrary) at a given dissimilarity threshold ( $\delta$ ). Each cross represents a reaction pair: blue crosses correspond to pairs of NB-compliant reactions (TS - true similars, reactions of the same class, furthermore structurally related according to the calculated dissimilarity score on X, and respectively TD - true dissimilars, which are of different classes, in agreement with their high structural dissimilarity exceeding the threshold  $\delta$ ). Red crosses represent pairs of NB-violating reactions (FS - false similars, of different classes albeit ranked as structurally similar, and PFD - potentially false dissimilars, which are of the same class although structurally unrelated).

selection of a subset of pairs of the same size as the pair subset selected at threshold  $\delta$  (denominator of eq 1).

Therefore, the optimality score  $\Omega(\delta)$  is defined as

$$\Omega(\delta) = \frac{kN_{FS}(\delta) + N_{PFD}(\delta)}{kN_{SEL}(\delta) \frac{N_p^1}{N_p} + (N_p - N_{SEL}(\delta)) \frac{N_p^0}{N_p}} \quad (1)$$

Here,  $N_{FS}(\delta)$  is the number of reactions pairs of different classes which are similar,  $N_{PFD}(\delta)$  is the number of pairs of same type which are not similar enough to be selected,  $N_{SEL}(\delta)$  is the number of selected couples for a given  $\delta$ ,  $N_p^1$  is the total count of pairs of reactions in different classes,  $N_p^0$  is the total of pairs from the same class, and  $N_p$  is the total number of reactions couples. If the CS guiding the selection of similar pairs makes sense, an optimal dissimilarity threshold  $\delta^*$  will emerge, at which  $\Omega$  reaches a minimum  $\Omega^*$  of false similars and potentially false dissimilars with respect to random expectations. Note that the denominator contains the expectation values for an ideally randomized selection of pairs, whereas specific random picks may lead to slightly varying counts and give birth to fake minima of  $\Omega$ . In order to eliminate noise, scrambled  $\Omega$  values are estimated by randomly permuting the calculated similarity scores associated with the pairs. Each of the 20 attempted scrambling operations produces a randomized curve  $\Omega(\delta)_{random}$  as a function of the threshold. From this new twenty optimality scores, a random mean  $\langle \Omega(\delta)_{random} \rangle$  - expectedly converging to 1.0, ( $\forall \delta$ ) - is computed along with the standard deviation  $\sigma(\Omega(\delta)_{random})$ . Therefore, the ascertained (noise-corrected) optimality function  $\Xi^{-1}(\delta)$  is defined as

$$\Xi(\delta) = \langle \Omega(\delta)_{random} \rangle - \sigma(\Omega(\delta)_{random}) - \Omega(\delta) \quad (2)$$

By contrast to  $\Omega$ ,  $\Xi(\delta)$  will reach a *maximum*  $\Xi(\delta)^*$  at its optimal dissimilarity cutoff  $\delta^*$ . The choice of descriptors and metric of the CS may therefore be driven by the maximization of  $\Xi(\delta)^*$  with respect to the given set of reaction pairs. For practical purposes, and specifically in order to compare the response curves  $\Xi(\delta)$  for metrics that have diverging value ranges (bounded Tanimoto - unbounded Euclidean), it is recommended to plot  $\Xi(\delta)$  on Y against the fraction of selected pairs  $f_{sel} = N_{SEL}(\delta)/N_p$  at  $\delta$  on X. In this way, the dissimilarity cutoff is hidden as an implicit variable of the plots, which now have a common reference frame  $[\Xi(\delta); f_{sel}]$  and can be directly superimposed and compared. The optimal combination corresponds to the maximal  $\Xi(\delta)^*$ .

The enrichment  $E$  of a particular class  $i$  can be computed according to eq 3

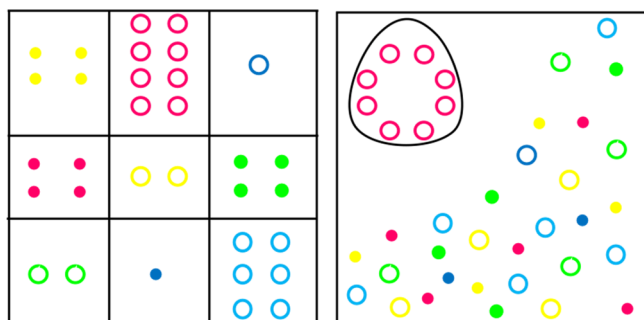
$$E = \frac{N_{SELi}(\delta^*)}{N_{SEL}(\delta^*)} \times \frac{N_p}{N_{pi}} \quad (3)$$

with  $N_{SELi}(\delta^*)$  representing the number of selected reactions of a particular class  $i$ , and  $N_{pi}$  representing the total number of pairs of a particular class  $i$ .

**2.4.2. Global Neighborhood Behavior.** Equation 2, applied to a representative subset of the entire DB contents, measures the Global Neighborhood Behavior (GNB) index. In this case, the particular notation  $\Gamma$  will be used instead of the generic  $\Xi$ :  $\Gamma$  will reach a maximum  $\Gamma^*$  at its optimal dissimilarity cutoff  $\delta^*$ . Note that  $\Gamma$  will be most impacted by the densest clusters of reactions in CS, for these will contribute the foremost ranked pairs of reactions, in terms of calculated similarity (see Figure 4, but also discussion in ref 28, where the distinction between global and local NB has been first introduced).

The generically best descriptors/metric combinations - recommended for generic use for similarity-driven querying of a database, are the ones maximizing  $\Gamma^*$ .





**Figure 4.** Left: Reactions of various classes positioned in an ideal CS, maximizing both GNB and LNB with respect to each reaction class. Right: A less than ideal CS in which LNB with respect to one reaction class (magenta circles) is nevertheless optimal.

**2.4.3. Local Neighborhood Behavior.** Contrary to the GNB, Local Neighborhood Behavior (LNB) specifically focuses on a given reaction (the query process  $Q$ ) and checks whether the points that are closest to it in CS actually correspond to “analogue” reactions of the same class (Figure 4). The associated local optimality criterion  $\Lambda^Q(\delta)$  is therefore computed, like  $\Gamma$ , according to (2), except that the associated local optimality criterion in (1) is specifically taken only over the pairs in which one of the reactions is the query process  $Q$ . The average of the LNB criterion  $\Lambda^*$  (taken as the average of local maxima  $\Lambda^{*Q}$  for each  $Q$ ) over many relevant queries  $Q$  is an alternative index of the overall compliance of a CS with the NB principle. Furthermore, the variance of LNB scores is a measure of robustness of similarity-based virtual screening.<sup>28</sup>

**2.5. AUC Monitoring – the Classical Alternative to NB.** The Area Under the ROC curve ( $AUC$ )<sup>29,30</sup> is classically used to quantify virtual screening success. The ROC is the curve of the purity (sensitivity) in function of the specificity

$$\text{sensitivity} = \frac{N_{TS}}{N_{TS} + N_{FS}} \quad (4)$$

$$\text{specificity} = \frac{N_{FS}}{N_{TOT} - N_{TS} - N_{FS}} \quad (5)$$

where  $N_{TS}(\delta)$  and  $N_{FS}(\delta)$  are the numbers of reactions pairs of the same class, respectively different classes, with computed dissimilarity below  $\delta$ , and  $N_{TOT}$  is the total number of reaction pairs.

By analogy to local NB,  $AUC$  calculations were performed around specific query reactions, representative of each class. Consequently, for one tested combination descriptors/metrics, one query by reaction class  $j$  is submitted and one  $AUC_j$  is computed with the package *pROC* of the R program. An  $AUC_{mean}$  is calculated as follows

$$AUC_{mean} = \frac{\sum_j AUC_j}{\text{numberOfClasses}} \quad (6)$$

$AUC_j$  may take values between 0.5 (random screening) and 1 (finding all the other members of the class  $j$  at the top of the reaction list ranked by similarity with respect to the query).

**2.6. Clustering and Self-Organizing Maps.** CS that are well compliant with the NB principle are more likely to generate clusters regrouping reactions of the same class. Among unsupervised clustering algorithms, Kohonen Self-Organizing Maps (SOM) are neural network-based models mapping the initial points of a CS onto one of the predefined “neurons”

(nodes) in a topographic 2D-lattice.<sup>4,5,31</sup> Albeit the mapping of the CS onto the 2D lattice of nodes is nonlinear, it is neighborhood-compliant: similar objects represented by close points in CS will be placed in the same or in neighboring lattice nodes. The Self-Organizing Map program package, version 3.1,<sup>32</sup> was used in this work. The R software was used for graphical rendering of obtained SOMs. Balanced accuracy, number of empty nodes, topographic error ( $TopEr$ ),<sup>33</sup> quantization error ( $QE$ , serving as objective function to be minimized when building the maps),<sup>34</sup> and the topographic product ( $ToPr$ )<sup>35</sup> are the quality criteria chosen to discuss SOM quality.<sup>35–45</sup> For the definition of the other SOM quality criteria, see Appendix A.

Balanced accuracy ( $BA$ )<sup>46</sup> is the main map criterion used in this work and is defined as

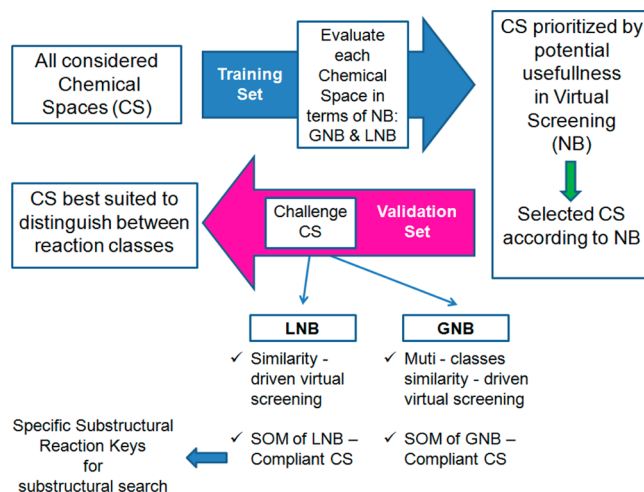
$$BA = \frac{\sum_i \frac{WC_i}{WC_i + BC_i}}{\text{numberOfClasses}} = \frac{\sum_i \frac{WC_i}{\text{numberOfInstancesOfClass}(i)}}{\text{numberOfClasses}} \quad (7)$$

Here,  $i$  denotes the current class,  $WC_i$  denotes the number of “well classified” instances of class  $i$ , while  $BC_i$  is the number of instances of class  $i$  in neurons which are not populated as a majority by the class  $i$ . Consequently,  $WC_i$  represents the number of true positives of class  $i$ , and  $(WC_i + BC_i)$  represents the total number of instances of class  $i$ .  $BA$  becomes the mean of true positives rates (or recall) by class. To illustrate the notions of “well classified” and “bad classified”, take an example of map with 2 neurons A and B. If the neurons A contains 10 reactions of class a and 2 reactions of class b and that the neuron B contains 2 and 8 reactions of classes a and b respectively, the node A is considered to be of class “a”, and B to be of class “b”. So, the 10 reactions of class a in the neurons A are  $WC_a$ , the 2 reactions of class a in the neurons B are  $BC_a$ , whereas the 8 reactions of class b in the neurons B are  $WC_b$  and the 2 reactions of class b in the neurons A are  $BC_b$ . When  $BA$  equals 1, all classes are correctly separated.

## 2.7. Overview of the Performed Numeric Simulations.

Figure 5 depicts the herein reported workflow. Its key steps and the associated calculations are briefly outlined in the following.

The database of 8544 reactions is divided into two subsets: the training set (TS) and the validation set (VS). TS are submitted to GNB and LNB analysis, to rank the considered CS (all mentioned fragmentation type/similarity metric



**Figure 5.** Workflow used in this study.

combinations) with respect to their hypothesized, potential usefulness in virtual screening (GNB and/or LNB). Allegedly useful vs allegedly poor CS are selected according to decreasing  $\Gamma^*$  or  $\Lambda^*$ . Next, the actual virtual screening propensities of selected CS will be assessed, this time with respect to the validation molecules. CS selected with respect to the GNB criterion will be checked with respect to their propensities to serve in generic (multiclass) similarity-driven virtual screening and to host relevant Self-Organizing Maps of reactions, aimed at separating the various processes. The quality of multiclass similarity searches (i.e. starting from an arbitrary query, of arbitrary class) and of Kohonen maps (used as classification tools) are expected to decrease with  $\Gamma^*$ . Since LNB focuses, by definition, on a particular class, LNB results are applied to class-specific similarity searches (optimized searching for analogous processes when the class of the query is known beforehand – i.e. tackling questions like ‘What is the best CS to search for analogues of a Diels–Alder process?’, in contrast to ‘What default CS to use, when the nature of the query is not predefined?’). Also, LNB-optimal CS are challenged with Kohonen map building but for a different purpose. These SOMs serve to extract specific substructural reaction keys, to be used in a substructural search in the whole DB.

**2.7.1. Robustness of GNB Criterion.** For a given CS, the global NB of the entire reaction DB can be defined by a single  $\Gamma^*$  value. In order to ensure that this is indeed a characteristic of the CS and not an artifact due to the particular choice of reactions in the DB, its robustness with respect to the size of the set of reactions has been systematically studied. For each of the considered CS, progressive global  $\Gamma^*$  values were estimated for DB subsets of increasing sizes. The scan began with one tenth of the DB, up to one-fourth of it, considering all DB subsets of size  $1/n$ , with  $n = 10 \dots 4$ . Each subset was composed by randomly picking one out of  $n$  representatives of each reaction class, thus ensuring a constant relative population of different reactions. With increasing subset size, global  $\Gamma^*_n$  should progressively converge toward the actual global  $\Gamma^*$  over the entire DB. For benchmarking purposes of CS, the absolute global  $\Gamma^*$  values are not relevant. In this context, it is enough to prove that the relative ranking of NB quality of the competing CS remains fixed beyond a given subset size.

**2.7.2. Benchmarking of Chemical Spaces in Terms of GNB: Global Neighborhood Behavior Analysis.** The calculated  $\Gamma^*$  over TS compounds in various CS have been compared, these CS were rank ordered in terms of decreasing  $\Gamma^*$ , and tentative explanations for the varying NB with respect to the nature of the descriptors have been proposed. Furthermore, the impact of the choice of the NB bias factor  $\kappa$  on the relative ranking has been addressed.

**2.7.3. GNB and Chemical Space “Mapability”.** Best, average, and worst performing descriptors according to the  $\Gamma^*$  value were selected to build Kohonen maps, the quality of which has been assessed and compared to their NB performances, in order to verify that NB criteria are a general measure of the easiness (“mapability”) to generate any kind of rational nonlinear map of CS. For these best, average, and worst descriptor sets, the same series of SOMs of 36 (6 by 6) up to 520 (20 by 26) neurons (32 different X-Y configurations) with either rectangular topology or hexagonal topology and of either neighborhood types (Gaussian or Bubble) were computed, using validation set reactions. In the same spirit, these CS were used to conduct similarity-based virtual screening studies based on classical success scores, such as

the AUC (Area Under ROC) criterion, in order to check as to how far these criteria match the observed  $\Gamma^*$ .

**2.7.4. Local vs Global Neighborhood Behavior.** Average local optimality criteria  $\langle \Lambda^* \rangle$  were calculated for each reaction class, by using each representative thereof from the TS as a query and then taking the average  $\Lambda^*$  over all queries of a class. CS benchmarking was then performed for each class, in order to highlight the CS optimally separating a given class from all the others – and which might differ from the “globally” best CS as highlighted in the GNB-based study. The “winning” CS with respect to every reaction class will be presented and discussed.

**2.7.5. Extraction of Specific Substructural Reaction Keys for Diels–Alder in Their Local NB-Preferred Chemical Space.** LNB analysis has been performed for Diels–Alder (DA) reactions. Winning descriptors have been used to build 128 different SOMs – at different size, topology, and neighboring types as previously reported with the GNB. The relevance of these maps has been now assessed from a novel point of view: for each of the neurons specifically populated by DA reactions, the common substructure of their residents has been extracted. These motives, called specific substructural reaction keys, served as keys for a substructural search within the entire DB, using the in order to assess the retrieval rate of DA processes. Both motif extraction from resident reactions of each neuron as well as the substructure search were performed with the ISIDA EdiSDF tool.<sup>19</sup> A high retrieval rate implies that the herein generated substructure keys are chemically significant signatures of the DA reaction.

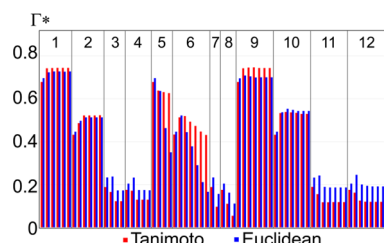
### 3. RESULTS AND DISCUSSION

**3.1. Robustness of NB Calculations.** For all the sizes of the DB samples used here (from 1/10 to 1/4 of DB), the relative ranking with respect  $\Gamma^*$  scores leads to the same top 15 combinations descriptors and metrics. While the relative ranking of  $\Gamma^*$  scores seems thus robust, their absolute numerical values are nevertheless fluctuating with the subset size but remain stable for sizes larger than 1/5 of the DB. Therefore, for each type of reaction, one representative out of five was assigned to the TS, and the remaining four to the validation set. Consequently, the 1:4 splitting ratio used to define the TS and VS in the Methods section (1713 reactions for training and 6831 in the validation set - Table 1) has been determined in agreement with this robustness study.

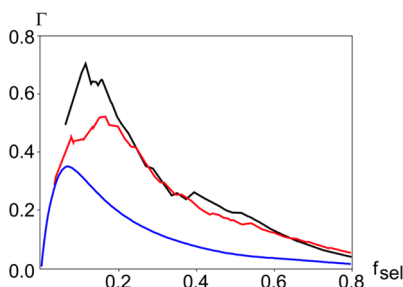
**3.2. Global Neighborhood Behavior Analysis.** The GNB analysis was conducted for 130 CS setups = 65 fragment descriptors (from 12 fragment type classes, see Table 2)  $\times$  2 metrics. In Figure 6, for Tanimoto and Euclidean metrics, the GNB scores  $\Gamma^*$  (at default  $\kappa = 3$ ) associated with the descriptor sets outlined along the X axis (and regrouped by their fragment type class) are shown as red and respectively blue bars. Figure 7 shows three  $\Gamma - f_{sel}$  (see 2.4.1) plots obtained in three different Euclidean CS, in order to illustrate the shapes of curves corresponding to the different  $\Gamma^*$  values. Curves superseding the others signal better global NB: fragmentation scheme IAB2\_4\_NS (black) outperforms IB2\_4\_D (red) and IAB2\_5 (blue). It is not necessary to represent the entire curves: their maxima  $\Gamma^*$  are in most cases a good indicator of NB.

From Figure 6, several conclusions emerge:

1. Both Tanimoto and Euclidean metrics are similarly successful in separating reaction classes: their  $\Gamma^*$  scores are comparable. Observed differences are of no practical concern. This is not astonishing either, since discrepancies between Euclidean and correlation coefficient-based metric (Tanimoto)



**Figure 6.** Global optimality criteria ( $\Gamma$ ) of 130 tested descriptor/metric combinations: 65 fragment descriptor types of 12 classes (see Table 2), combined with Tanimoto (red) or Euclidean distance (blue). Each bar corresponds to  $\Gamma$  of one combination descriptor/metric according to the fragment maximum size. The sizes of the 12 areas are not equal because the number of the generated fragments varies with the descriptors type.



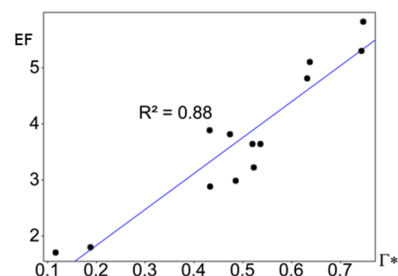
**Figure 7.** Example of NB analysis involving Euclidean metrics and different types of fragment descriptors. Curves,  $\Gamma = f(f_{sel})$  with  $f_{sel}$  the fraction of selected pairs, correspond to “good” (black curve), “fair” (red curve), and “bad” (blue curve) descriptors.

are known to mainly depend on the relative complexity of the compared structures or reactions. Indeed, two low complexity structures – both featuring few populated fragments, thus both having a sparse “light” descriptor vector – will likely have their similarity overstated by Euclidean and understated by correlation coefficients. The opposite applies to complex entities with very “dark” descriptor vectors. However, most of the herein monitored descriptor sets specifically monitor sequences of dynamic bonds, which are rather conserved within each class, leaving no room for large fluctuations of descriptors “darkness” throughout the reaction classes. The most general descriptors, IAB2\_Max and IB2\_Max, not restricted to dynamic bonds (only one of the bonds in the sequence being required to be dynamic) are the ones most sensitive to the chemical environment of the RC (beyond actual formed/broken bonds). They are thus more able to capture the chemical diversity of the structures and are spanning a broader range of “darkness”. Unsurprisingly, it is these (types 5 and 6 in Figure 6) that display the only significant NB shifts when switching to the Euclidean metric. Types 7 and 8 – augmented atom counts that are also sensitive to the RC environment – do not behave differently with respect to the metric, since they fail no matter what metric is used (see below).

2. Augmented atom counts fail in NB tests. The reason for this behavior is, as will be shown, largely due to the nature of the major reaction class (Diels–Alder, DA) in the considered DB. The DA “transition state” captured in the CGRs (Figure 1) consists of a six-membered ring of dynamic bonds and is a very distinct signature with respect to the other reaction classes (except hetero-Diels processes, from which they cannot be distinguished by B-type descriptors, in which coloring by atom

symbol is toggled off). This six-membered ring does indeed translate into a class-characteristic combination of atom/bond sequences. Additionally, each DA reaction will also match specific sequences, combining atoms within the ring with some exocyclic substituents, but it is certain that in the descriptor vector of any two DA processes, a subset of commonly populated elements will be found – the ones associated with the common, ring sequences. In other words, there is a certain guarantee that DA processes will be recognized as related, *because sequences that are fully contained in this ring do exist*. This is not true with Augmented Atoms, which monitor coordination spheres inspired of organometallic compounds, since Augmented Atoms correspond to all atoms and bonds or bonds if  $n$  sphere(s) are considered around a particular atom of the structure. Consider, for example, the addition of the same diene to bromo- and chloroethene, respectively. The characteristic ring will be broken up between the Augmented Atoms centered on the ring atoms – but some of these fragments (depending on the user-defined size) will include the halogen substituent. The information about the common six-membered ring will be partly rerouted into *different* fingerprint elements (those corresponding to the Br- vs Cl-containing fragments) – and there is no bonus for matching a brominated to a chlorinated augmented atom: those are distinct descriptor elements. Therefore, the inter-DA reaction dissimilarity scores are systematically higher with IAB-type fragments, thus there are relatively less DA/DA pairs entering the list of selected “true similar” at optimal dissimilarity cutoff. Out of 662 IAB2\_4\_NS sequences, 50 are present in at least one of the DA processes, whereas four are ubiquitously present in virtually each DA process. There are 1363 IAB2\_4\_NS augmented atoms found in DB, out of which 124 are encountered at least once within the DA subset, and five are ubiquitous. Clearly, the descriptor subset encoding for DA-representative fragments represents only 5/124 out of the entire vector needed to encode DA processes – whereas, with sequences, a much more significant ratio of 4/50 is observed. Since DA is the dominant reaction class, having the DA signature concentrated in a common subset of descriptor elements allows for selection of large pair subsets with few false dissimilar.

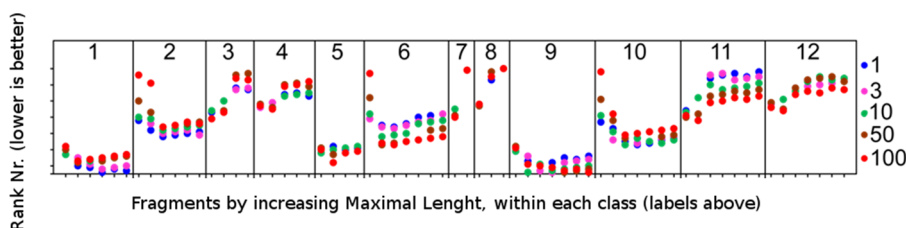
3. Figure 8 shows that the enrichment rate in DA/DA pairs within the subset of selected pairs at optimal cutoff is strongly



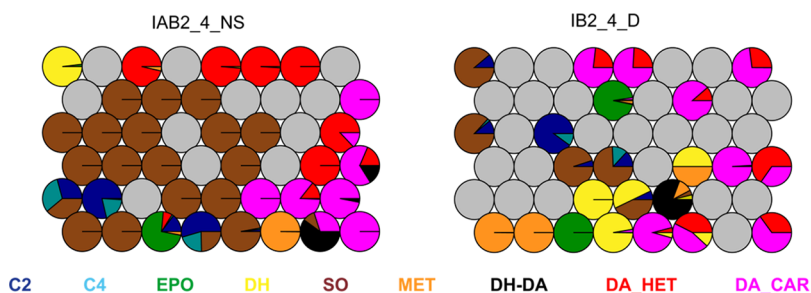
**Figure 8.** Correlation between the Enrichment Factor (EF) in DA/DA pairs within the subset of selected pairs at optimal cutoff and the overall global optimality criteria ( $\Gamma$ ).

correlated with the overall optimality scores, thus proving the above-proposed explanation of the relative lack of performance of augmented atom descriptors. It should be kept in mind that this relative lack of performance is specific for the reaction classification problem, where the focus is on the recognition of common patterns. The fundamental limitation of this study is





**Figure 9.** Relative ranks of the various CS as a function of the NB bias factor  $\kappa$  (see eq 1), descriptors' types and the maximal fragment length (see Table 2). The best rank (rank #1) corresponds to highest  $\Gamma^*$  score. In terms of  $\Gamma^*$  scores (best  $\Gamma^*$  meaning "Gold Medal", rank #1) as a function of the NB bias factor  $\kappa$  ( $\kappa = 1; 3; 10; 50; 100$ , see a color code on the right). Fragments length is between 2 and Max atoms.



**Figure 10.** SOMs built with "good" (left) and "less good" (right) descriptors selected according to NB calculations with the Euclidean distance. Statistical parameters of the maps are given in Table 3.

that it does not cover any unfeasible processes: the data set features no *hypothetical* DA processes which seem correct as far as "paper chemistry" may tell but practically do not occur. So far, in our study, every item that "looks like" a DA process is a DA process.

4. Sequences of atoms and bonds separate better chemical reaction classes than sequences of bonds alone. Effectively, as a CGR represents an imaginary transition state, the RC corresponds, in most cases, to a sequence of 2 (or maximal 3) dynamic bonds in which heteroatoms are important for the reactive mechanism. Unsurprisingly, most of the benefits of using AB-type fragments rather than B come from enabling the separation between homo- and hetero-DA processes, for which the CGR description in terms of bonds only is identical. By contrast, using atom labels will implicitly highlight the halogen-related subclasses of the Sonogashira and nucleophilic substitution processes, which will now highlight different descriptor elements corresponding to  $-\text{Cl}$ ,  $-\text{Br}$  or  $-\text{I}$  containing sequences. The same "signature dilution" effect described in connection with augmented atom-based encoding of DA processes is here effective. Yet, this has less impact on general NB, as the cited reaction subsets are not dominant in the database.

5. Sequences of atoms and bonds with the D or NS options (types 1 and 9 in Figure 6, respectively) return roughly equal GNB optimality criteria, irrespective of the maximal length of the monitored fragments. They are much stronger performers than the counts of all sequences including at least one dynamic bond (type 5) – most likely another consequence of "reaction key dilution", with a generalized impact over the DB.

If the Euclidean distance is used as a dissimilarity measure, the optimal descriptors are IAB2\_4\_NS i.e. substructural fragments of atoms and bonds, of length between 2 and 4 atoms, resulting in CGRs fragmentation after elimination of nondynamic single bonds. With the Tanimoto score, IAB2\_5\_NS are optimal but not significantly better than the slightly shorter IAB2\_4\_NS. The good performance of these

fragmentation schemes is not surprising, since IAB2\_4\_NS fragments cover at most three bonds, and 6246 of 8544 chemical transformations involve 3 dynamic bonds or less (dihydroxylation, metathesis, and epoxydation). However, DA, DA-HET, and DH-DA have a signature of 6 dynamic bonds and 6 atoms, which is easily captured by a series of multiple overlapping sequences.

As can be seen from Figure 9, the relative ranks of the various CS in terms of  $\Gamma^*$  scores (best  $\Gamma^*$  meaning "Gold Medal", rank #1) would not be significantly altered by a different choice of the NB bias factor  $\kappa$ , allowed to scan the range from 1 (no additional penalty for "false similar" pairs) to 100 (strong bias against similar pairs). Overlapping dots in the figure simply mean that ranking of the performance of descriptors is robust with respect to  $\kappa$ . Strong rank fluctuations, if any, are observed for "uninteresting" CS, which never make it into the top performers. One noteworthy but unsurprising exception can nevertheless be observed: at low  $\kappa$ , NS sequences (column #9 on the plot) partly lose their dominant ranking in favor of the less constraining D sequences (column #1). The latter being less focused on the dynamic bonds, and therefore more sensitive to the outer chemical environment, are more likely to let "false similars" enter the selection of closest pairs – hence their better performance at higher level of tolerance of false similars.

**3.3. GNB and Chemical Space "Mapability" – Kohonen Maps.** Eight different descriptor sets were selected to confront their GNB proficiency to the quality of Kohonen maps built on their basis. GNB optimality criteria of these descriptors are within [0.3,0.8].

An optimal SOM is minimizing the number of empty clusters (found to range between 1 and 491), the quantization error [0.17], the topographic error [0.1], while maximizing the balanced accuracy [0.4..0.9] and having a topographic product [-194..3742] closest to 0. Out of the 1024 generated SOMs, all maps with either (a) more than 20 empty neurons, (b) absolute topographic product above 10, or (c) topographic error above



0.3 were discarded. Out of the remaining, the map with the best balanced accuracy was selected (Figure 10).

The winner is a hexagonal bubble Kohonen map of size 48 ( $6 \times 8$ ), which achieved its best performance in the IAB2\_4\_NS descriptor space (incidentally, this was the GNB challenge winner). The hexagonal-bubble setup is actually the configuration preferred by Kohonen.<sup>34</sup> Table 3 declines the

**Table 3. For These Descriptor Sets (Column #1), Global Optimality Scores (with Euclidean Metric, Column #2), and SOM Quality Criteria Are Shown<sup>a</sup>**

descriptors	$\Gamma^*$	#empty	ToPr	BA	QE	ToEr
IAB2_4_NS	0.70	12	−1.00	0.74	1.65	0.26
IAB2_4_D	0.73	25	18.41	0.69	0.77	0.51
IB2_4_NS	0.55	26	26.37	0.68	1.01	0.18
IB2_3	0.52	12	11.47	0.68	1.24	0.55
IB2_4_D	0.51	25	40.99	0.68	0.48	0.43
IB2_4	0.44	12	−0.58	0.63	3.55	0.49
IB2_2_D	0.45	32	78.71	0.54	0.05	0.50
IAB2_5	0.35	5	−2.58	0.6	6.91	0.34

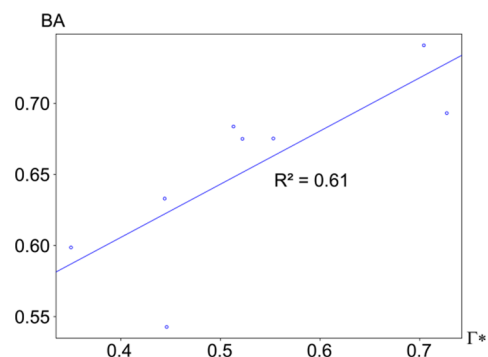
<sup>a</sup>These correspond to a SOM of hexagonal topology, a bubble neighborhood function, and 48 ( $6 \times 8$ ) output neurons, and are as follows: Nr. of empty neurons, Topographic Product, Balanced Accuracy, Quantization Error, and Topographic Error. The two highlighted lines pertain to the maps discussed in more detail (Figure 10).

performance indices of the eight  $6 \times 8$  hexagonal bubble maps built using the eight herein considered descriptor sets. Two of these, based on descriptor sets IAB2\_4\_NS (a) and respectively IB2\_4D (b), are depicted in Figure 10.

The former is based on a GNB-winning descriptor set, while the second is built with less NB-compliant descriptors. The former is also a better map, by all standard statistical criteria shown in Table 3.

Inspection of Figure 10 shows that cluster purity decreases and the number of empty clusters increases with the less NB-compliant fragments. Moreover, similar reaction classes are not grouped in the same area. For example, the Metathesis are grouped in only one cluster with IAB2\_4\_NS but in four unrelated nodes clusters with the IB2\_4\_D fragments. Actually, some clusters of the best map are not pure. This is not surprising, given the nature of CGRs and fragment descriptors. Some DH-DA and Sonogashira reactions are grouped into the same cluster because IAB2\_4\_NS fragments are similar. In these Sonogashira reactions, a cycle is formed, as in DH-DA. For the same reason, some DH-DA are grouped with some DA and/or DA-HET and/or metathesis clusters. Moreover, DH-DA are similar to DA (DA or DA-HET) and metathesis classes<sup>47</sup> – the frontier separating these classes is, objectively, less sharp than the difference between say metathesis and dihydroxylation.

For the eight instances of the winning map configuration applied to the CS shown in Table 3, the correlation coefficient between balanced accuracy of the maps and  $\Gamma^*$  is  $R^2 = 0.61$  (Figure 11). This is a moderate correlation level, showing nevertheless that the overall NB propensity of a CS has a significant impact on its ability to generate useful Kohonen maps. However, if the point corresponding to IB2\_2\_D, at  $\Gamma^* = 0.44$  and  $BA = 0.54$  is considered an outlier and removed, and the correlation coefficient  $R^2$  becomes 0.80. This point effectively represents an atypical GNB scenario, as this  $\Gamma^*$  is



**Figure 11.** Balanced accuracy (BA) of SOM-based models as a fraction of optimality criteria ( $\Gamma$ ). The Euclidean distance has been used both in NB and in SOM calculations.

associated with a high fraction of selected pairs (28%) and a medium level of enrichment within selection. This is an understandable consequence of the rather low-detail fragmentation scheme, considering only dynamic bonds – which is not enough to discriminate between nucleophilic substitutions and Sonogashira reactions or, respectively, homo- and hetero-Diels–Alder processes. Since co-optation of false similar pairs in the selection is unavoidable,  $\Gamma^*$  is optimized by enlarging the selection.

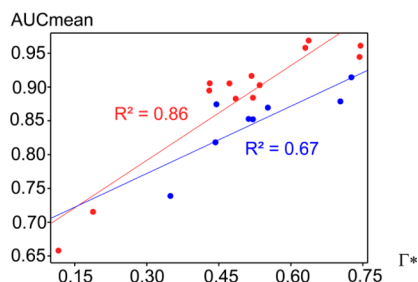
Nevertheless, balanced accuracy is not the absolute SOM quality criterion, either. A respectable BA level may be achieved either by a map providing a fair separation of every reaction class from all the others or by one perfectly segregating some classes, while others are not being discriminated at all. Moreover, as hinted above, some classes are chemically close to each other, while others describe processes with radically different reaction mechanisms. It is thus difficult to quantitatively define a specific penalty function to precisely account for failure to separate a given pair of reaction classes. For example, in the IB\_2\_4 descriptor-based map (Figure 10 b;  $BA = 0.68$  and  $\Gamma^* = 0.5$ ), BA is relatively high because most nodes have purity values above 60%, but very few tend to an absolute purity level of 100%. Therefore, it can be concluded that NB and Kohonen net proficiencies are fairly well correlated – good NB seems to be a good prerequisite of descriptors envisaged to serve for Kohonen map build-up. Nevertheless, there is no ideal, unambiguous definition of Kohonen map quality: the consensus choice based on the herein outlined criteria appeared to have allowed picking a chemically meaningful winner (Figure 10 a).

Herein, DA processes are neatly separated from substitution reactions (including Sonogashira processes), themselves distinct from metathesis and dihydroxylation processes. Furthermore, many neurons are dedicated to refine the mapping of large families (Sonogashira and DA). Reaction classes seen to mingle with the same neuron represent chemically related transformations (nucleophilic substitutions with Sonogashira, hetero- and homo- DA, etc.).

**3.4. GNB and AUC-Based Similarity-Driven Screening for Analogue Reactions.** Next, GNB results were confronted to classical success scores of multiclass similarity searches performed on the VS: one query for each class (9 classes) is used to rank remaining test set compounds with respect to their similarity scores. AUC values are calculated from each of the nine resulting ROC curves, and the mean thereof,  $AUC_{mean}$  is

taken as a “classical” neighborhood criterion of the current descriptor space. Queries are exemplified in Figure 1.

Figure 12 represents the  $\Gamma^*$  as a function of the  $AUC_{mean}$  and the correlation curve associated with descriptors/Euclidean

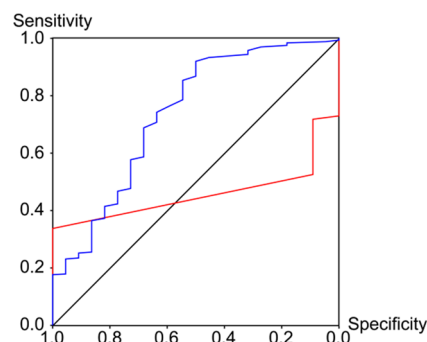


**Figure 12.** Multiclass similarity search:  $AUC_{mean}$  vs  $\Gamma^*$  for Tanimoto metrics (red points) and the Euclidean distance (blue points).  $AUC_{mean}$  correlates better with  $\Gamma^*$  calculated with Tanimoto metrics rather than with  $\Gamma$  calculated with the Euclidean distance.

distance and descriptors/Tanimoto coefficient. Note that the correlation coefficient  $R^2$  is about 0.7 for the Euclidean distance, 0.9 for the Tanimoto coefficient. The overall  $\Gamma^*$ - $AUC_{mean}$  correlation coefficient corresponding to all the tested combinations is about 0.8.  $AUC$  is nowadays the state-of-the-art criterion in similarity search quality evaluation, so the good correlation to  $\Gamma^*$  speaks in favor of the meaningfulness of the latter criterion as well. However, the indices are not identical, and when discrepancies are observed,  $\Gamma^*$  seems to be the more trustworthy index: some descriptors associated with high  $AUC_{mean}$  and medium  $\Gamma^*$  are not effective in discriminating all the reaction classes.

While the mean of  $AUC$  values is, expectedly, correlating with global NB indices, the family specific  $AUC$  scores incorporated in  $AUC_{mean}$  may be regarded as an alternative illustration of local NB. However, NB analysis has been specifically developed because sometimes a good  $AUC$  is associated with a ROC curve in which the enrichment in items of wanted properties (“hits”) does not occur among the very first nearest neighbors but among “medium” near neighbors (i.e. no hits at all are found within say 10% of the top closest neighbors, then all the hits appear next, among the following 5% of “medium” close neighbors). This situation yields a better  $AUC$  score than a scenario in which all the top 2% of neighbors are hits, whereas the remaining hits are ranked at the very end of the virtual screening order. Yet, the latter situation correctly describes a *better* NB: similarity is not a prerequisite for similar properties, and there is nothing bad with having hits being out of the scope of similarity scoring. All that matters in NB is the fact that examples of items of different properties tend to be rare among the top nearest neighbors. A real-case example of the above-mentioned scenario is shown in Figure 13. Consequently, we can say NB scores are highly indicative of the expected success rate to retrieve reactions of the same class as the query process, based on ISIDA descriptor similarity.

**3.5. Local Neighborhood Behavior and Specific Substructural Key Extraction.** Unlike GNB, local NB focuses on the specific neighborhoods of reactions of a given class. Therefore, in an ideal CS maximizing the GNB criterion – the left-hand box of Figure 4 – local NB is also guaranteed, for each reaction class. The reciprocal is not true – see the right-hand box of a putative CS in which LNB with respect to the magenta circle class is excellent, but globally the NB



**Figure 13.** Neighborhood Behavior versus the alternative  $AUC$  criterion: the red ROC Curve (IAB2\_4\_NS) corresponds to low  $AUC$  and high LNB score ( $\Lambda = 0.836$ ,  $AUC = 0.458$ ), the (IAB2\_6\_D) has high  $AUC$  but low LNB ( $\Lambda = 0.508$ ,  $AUC = 0.736$ ). Reason: NB criteria are very sensitive to the hit rate among the very first, top ranked pairs. In the first CS, top-ranked neighbors of the DH-DA query are all DH-DA, whereas the remaining DH-DA reactions are not considered similar. In the second, the similarity is tweaked toward retrieving a larger DH-DA subset among the similars of the considered queries – but this happens, unfortunately, at the cost of co-opting reactions of other classes into the selection.

principle is not well respected. A further reason for GNB/LNB discrepancies are putative density artifacts, as discussed in a previous work<sup>28</sup> (refer to Figure 4 therein). After calculating, for each reaction class and every CS, the average  $\langle \Lambda^* \rangle$  around every class member of the training set, the various CS were ranked with respect to their local optimality with respect to the given class. The winners for each class – which, unsurprisingly, are recurrent winners for several, related classes are shown in Table 4. The specific pattern of nucleophilic substitutions and

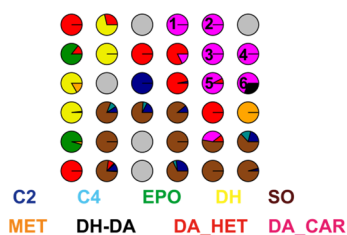
**Table 4. Specific Descriptor Sets to Particular Reaction Transformations According to Local Neighborhood Behavior Analysis**

descriptors	reaction mechanism
IB2_4_NS	C2, C4, SO
IAB2_4_NS	DA-HET, DH-DA, HYD, EPO
IAB2_6_D	DA, MET

related Sonogashira processes is best captured by IB2\_4\_NS descriptors. Indeed, the exact nature of the leaving group (in this study: various halogen atoms) is irrelevant for these classes: therefore, capturing the nature of atoms would artificially subdivide these processes into subclasses, depending on the departing halogen. This decreases overall NB.

IAB2\_4\_NS – the GNB benchmark winner – also guarantees the optimal discrimination of dihydroxylation, epoxidation, and “special” DA processes: Domino-Heck (DH-DA) and Heteroatomic (DA-HET) variants (Figure 10). For classical DA and metathesis reactions, the pattern of dynamic bonds is very specific: fragment counts based solely on these latter (IAB2\_6\_D) are sufficient to distinguish them from other reactions.

As local NB suggests IAB2\_6\_D as specific descriptors set to discriminate between DA processes and the rest of reactions, these were employed to generate SOMs, in order to check how these would render DA processes (compared to SOMs based on GNB-favored CS). Out of the 128 explored architectures, the SOM regrouping a maximal number of DA processes into pure nodes (Figure 14) corresponds to a rectangular topology,



**Figure 14.** SOM built on the “best” descriptors selected in LNB calculations with Euclidean distances for the Diels–Alder reaction. Nodes 1–6 mostly populated by DA provide with representative subset of reactions from which DA specific CGR signatures are extracted (see Figure 15).

with a bubble neighborhood function, of 36 neurons ( $6 \times 6$ ). The map features four neurons 100% occupied by DA reactions, one neuron also hosting related DH-DA processes, and one node also hosting DA-HET processes. The quality criteria are not outstanding:  $ToPr = 5.72$ ,  $QE = 0.96$ ,  $BA = 0.65$ ,  $TopEr = 0.68$ . Notice that this Kohonen map based on local NB-favored descriptors grouped all DA in a particular area, while other neurons are ‘dirty’, i.e. indiscriminately host

processes of several types. The neurons containing DA reactions were analyzed, and 6 specific substructural reaction keys were extracted (Figure 1). The 6 specific substructural reaction keys were tested by using them as substructural queries of the whole database: 3046 DA reactions out of the 3048 present in the database are returned. The two missed examples are actually not DA reactions, but photochemical processes, according to an a posteriori bibliographic search.<sup>48,49</sup> Therefore, it can be concluded that regrouping of the DA processes on the SOM neurons was meaningful: each node produced its own specific common substructures, and the reunion of all the substructure hits found by either of these covers the totality of present DA processes (including those that did not serve to train the SOM).

#### 4. CONCLUSION

This work showed how, by means of CGR technology, chemical processes can be treated like classical compounds and thus subjected to the whole battery of similarity-based virtual screening approaches and benchmarking studies (NB, AUC monitoring, SOM-driven mapping).

Cluster	CGR signature	Example of CGR	Reaction corresponding to the CGR
1			
2			
2			
3			
4			
4			
6			

**Figure 15.** Substructural keys specific to Diels–Alder (DA) reactions are able to retrieve the ensemble of DA reactions in a multiple substructural search in the reaction database (see Table 1).



The relationship between the specifics of various descriptor spaces and their NB is coherent and chemically interpretable: AUC monitoring is highly correlated with optimality criteria.

Moreover, global NB is a good indicator of CS mapability, using SOMs. However, SOMs are complex unsupervised models, and their quality may be assessed from various different points of view (quality indices). The one most important in this study – *balanced accuracy*, monitoring the propensity of the map to segregate reactions by their class – is also the one found to best correlate global GNB scores. In a classification or similarity problem, the data representation (description) is crucial. NB efficiently pinpointed to descriptor spaces of good proficiency in classification and similarity searching, and these descriptors confirmed their usefulness in alternative similarity-driven (AUC monitoring) or SOM-based mapping approaches. NB is a convenient way to learn how to represent the data in a chemical reactions DB.

Furthermore, when NB is monitored in a local manner, with respect to a specific reaction class, the hereby selected optimal descriptors tend, in a map selected for the purity of neurons harboring the targeted class (DA), to be less proficient in segregating the other classes and create more ‘dirty’ neurons. This is coherent with the definition of local NB, as a criterion focusing on the surrounds of a specific query, or – under its average form, as used here – on the segregation between a single reaction type and the ‘rest of world’. Clearly, local, class-specific and global optimality constraints may differ, as each class displays its own specific signature – whereas global NB should focus on compromise descriptors which characterize all reaction classes conveniently well.

For chemical reactions, like for simple molecules, similarity-based and substructure-based searches are complementary database querying methods. The clustering of DA processes with LNB-compliant ISIDA fragment descriptors allowed to quickly detect the various substructural patterns representing alternative definitions of DA processes, discovered among the reactions residing on different neurons. The analysis of the various successes and failures of different fragment descriptors in various NB and virtual screening experiments could also be explained in the light of specific reaction signatures. In this work, dedicated to chemical process classification, similarity-based and substructure-based searches are closely related, especially since similarity-based virtual screening and SOM-driven mapping rely on substructure counts as descriptors.

ISIDA fragments of CGRs are both information-rich and generic enough to allow a convenient mapping of reaction classes. They do not need to exhaustively cover the entire set of structural patterns corresponding to chemical functions – focusing on fragments containing dynamic bonds is enough, at least in order to discriminate between the diverse reaction classes. However, if – at the next stage – the modeling of actual chemical reactivity is targeted, then recognition of formed/broken bonds may not be enough to decide whether a putative process of a class will actually happen, under the assumed reaction condition. The inclusion of physicochemical properties in ‘colored’ fragments<sup>50</sup> might account for remote electronic and steric effects.

## ■ APPENDIX A

Topographic error (*TopEr*) is defined as

$$TopEr = \frac{1}{N} \sum_{i=1}^N \mu(\mathbf{x}_i) \quad (\text{A.1})$$

This measure represents the proportion of the mapped reactions (represented by the descriptor vectors  $\mathbf{x}_i$ , out of the total on  $N = 6831$  members of the VS) for which first and second best-matching neurons (BMN) are not adjacent in the map. The function  $\mu(\mathbf{x}_i)$  is 1 if the first and second BMNs of  $\mathbf{x}_i$  are adjacent and, 0 otherwise.

The quantization error (QE; A.2)

$$QE = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{w}_{bmn(i)}\| \quad (\text{A.2})$$

sums the distances between the input vectors of entries  $i$  and the code vectors  $\mathbf{w}_{bmn(i)}$  of the best matching neuron  $bmn(i)$  is computed by determining the average distance of  $N$  sample vectors to the cluster centroids by which they are represented.

In order to introduce the topographic product (*ToPr*), let  $\mathbf{W}$  represent the two-dimensional position vector (of integers) of a neuron in a map, and  $D(m,n)$  the  $\mathbf{W}$ -based “output” Euclidean distances between neurons  $m$  and  $n$ . By contrast, let  $d(m,n)$  stand for the  $\mathbf{w}$ -code vector-based “input” Euclidean distances in the descriptor space. Let  $K^d(m,k)$  represent the  $k$ -th nearest neighbor of neuron  $m$  in the input space: in an ascending sorted list of  $d(m,n)$ , the term  $d[m, K^d(m,k)]$  would be ranked at the  $k$ -th position. Analogously, let  $K^D(m,k)$  return the  $k$ -th nearest neighboring neuron of  $m$  on the map (in output space).

$$ToPr = \sum_{m=1}^M \sum_{k \neq m} \frac{1}{2k} \ln \left\{ \frac{d[m, K^D(m,k)]}{d[m, K^d(m,k)]} \times \frac{D[m, K^D(m,k)]}{D[m, K^d(m,k)]} \right\} \quad (\text{A.3})$$

Both neuron indices  $k$  and  $m$  run over the entire set of  $M$  neurons of the map.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: varnek@unistra.fr.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

A.d.L. thanks to the Région Alsace for providing a Ph.D. fellowship. Dr. G. Niel is acknowledged for the help with data preparation.

## ■ ABBREVIATIONS:

DB, database; CGR, condensed graph of reaction; RC, reaction center; QSPR, quantitative structure properties relationship; SOM, self-organizing map; CS, chemical space; VS, validation set; TS, training set; ROC, receiver operating characteristic; AUC ROC, area under the curve; *AUCmean*, mean of AUC ROC; NB, Neighborhood Behavior; GNB, Global Neighborhood Behavior; LNB, Local Neighborhood Behavior

## ■ REFERENCES

(1) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033–4042.

- (2) Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. Classification of Organic Reactions: Similarity of Reactions Based on Changes in the Electronic Features of Oxygen Atoms at the Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 210–219.
- (3) Sacher, O.; Reitz, M.; Gasteiger, J. Investigations of Enzyme-Catalyzed Reactions Based on Physicochemical Descriptors Applied to Hydrolases. *J. Chem. Inf. Model.* **2009**, *49*, 1525–1534.
- (4) Zhang, Q.-Y.; Aires-de-Sousa, J. Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers. *J. Chem. Inf. Model.* **2005**, *45*, 1775–1783.
- (5) Latino, D. A.; Aires-de-Sousa, J. Genome-Scale Classification of Metabolic Reactions: A Chemoinformatics Approach. *Angew. Chem., Int. Ed.* **2006**, *45*, 2066–2069.
- (6) Ridder, L.; Wagener, M. SyGMA: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008**, *3*, 821–832.
- (7) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.
- (8) Borodina, Y.; Rudik, A.; Filimonov, D.; Kharchevnikova, N.; Dmitriev, A.; Blinova, V.; Poroikov, V. A New Statistical Approach to Predicting Aromatic Hydroxylation Sites. Comparison with Model-Based Approaches. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1998–2009.
- (9) Daylight Chemical Information Systems. Fingerprints - Screening and Similarity. <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html> (accessed July 24, 2012).
- (10) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction As One Single Pseudo Molecule. <http://dtai.cs.kuleuven.be/ilp-mlg-srl/papers/ILP09-5.pdf> (accessed July 24, 2012).
- (11) Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205–212.
- (12) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.
- (13) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (14) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in Silico Structural Spaces with Respect to in Vitro Activity Spaces—a Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680–690.
- (15) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in Silico Structural Spaces with Respect to in Vitro Activity Spaces—a Benchmark for Neighborhood Behavior Assessment of Different in Silico Similarity Metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691–698.
- (16) 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.
- (17) Dixon, S. L.; Merz, K. M. One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation. *J. Med. Chem.* **2001**, *44*, 3795–3809.
- (18) Solov'ev, V. P.; Kireeva, N. V.; Tsivadze, A. Y.; Varnek, A. A. Structure-Property Modelling of Complex Formation of Strontium with Organic Ligands in Water. *J. Struct. Chem.* **2006**, *47*, 298–311.
- (19) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuck, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191–198.
- (20) Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. Quantitative Structure-Property Relationship Modeling of beta-Cyclodextrin Complexation Free Energies. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529–541.
- (21) InfoChimie. ISIDA Fragmentor. <http://infochim.u-strasbg.fr/spip.php?rubrique49> (accessed July 24, 2012).
- (22) InfoChimie. ISIDA/QSPR program. <http://infochim.u-strasbg.fr/spip.php?rubrique53> (accessed July 24, 2012).
- (23) Varnek, A.; Solov'ev, V. P. "In Silico" Design of Potential Anti-HIV Actives Using Fragment Descriptors. *Comb. Chem. High Throughput Screening* **2005**, *8*, 403–416.
- (24) Papadatos, G.; Cooper, A. W. J.; Kadiramanathan, V.; Macdonald, S. J. F.; McLay, I. M.; Pickett, S. D.; Pritchard, J. M.; Willett, P.; Gillet, V. J. Analysis of Neighborhood Behavior in Lead Optimization and Array Design. *J. Chem. Inf. Model.* **2009**, *49*, 195–208.
- (25) Horvath, D.; Mao, B. Neighborhood Behavior – Fuzzy Molecular Descriptors and their Influence on the Relationship between Structural Similarity and Property Similarity. *QSAR Comb. Sci.* **2003**, *22*, 498–509.
- (26) Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46*, 2457–2477.
- (27) Bonachera, F.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2008**, *48*, 409–425.
- (28) Horvath, D.; Koch, C.; Schneider, G.; Marcou, G.; Varnek, A. Local Neighborhood Behavior in a Combinatorial Library Context. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 237–252.
- (29) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874.
- (30) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinf.* **2011**, *12*, 77–84.
- (31) Latino, D. A.; Zhang, Q. Y.; Aires-de-Sousa, J. Genome-Scale Classification of Metabolic Reactions and Assignment of EC Numbers with Self-Organizing Maps. *Bioinformatics* **2008**, *24*, 2236–2244.
- (32) Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J. SOM\_PAK. The Self-Organizing Map Program Package. [http://www.cis.hut.fi/research/som\\_pak/som\\_doc.txt](http://www.cis.hut.fi/research/som_pak/som_doc.txt) (accessed July 24, 2012).
- (33) Uriarte, E. A.; Martin, F. D. Topology Preservation in SOM. *Int. J. Math. Comput. Sci.* **2005**, *1*, 19–22.
- (34) Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J. CiteULike. SOM PAK: The Self-Organizing Map program package. <http://www.citeulike.org/user/Jaykul/article/771854> (accessed July 24, 2012).
- (35) Bauer, H.-U.; Pawelzik, K. R. Quantifying the Neighborhood Preservation of Self-Organizing Feature Maps. *IEEE Trans. Neural Netw.* **1992**, *3*, 570–579.
- (36) Kaski, S.; Lagus, K. Comparing Self-Organizing Maps. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN '96)*, Bochum, Germany, July 16–19, 1996; Malsburg, C.; Seelen, W.; Vorbrügger, J. C.; Sendhoff, B., Eds. Springer: Berlin, 1996; pp 809–814.
- (37) Villmann, T.; Der, R.; Herrmann, M.; Martinetz, T. M. Topology Preservation in Self-Organizing Feature Maps: Exact Definition and Measurement. *IEEE Trans. Neural Netw.* **1997**, *8*, 256–266.
- (38) Bauer, H. U.; Herrmann, M.; Villmann, T. Neural Maps and Topographic Vector Quantization. *Neural Networks* **1999**, *12*, 659–676.
- (39) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343–357.
- (40) Venna, J.; Kaski, S. Neighborhood Preservation in Nonlinear Projection Methods. An Experimental Study. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN '01)*, Vienna, Austria, August 21–25, 2001; Dorffner, G.; Bischof, H.; Hornik, K., Eds.; Springer: Berlin, 2001; pp 485–491.
- (41) Polani, D. Measures for the Organization of Self-Organizing Maps. In *Self-Organizing Neural Networks. Recent Advances and*

*Applications*; Jain, L., Seiffert, U., Eds.; Springer: New York, 2001; pp 13–44.

(42) Pözlbauer, G. Survey and Comparison of Quality Measures for Self-Organizing Maps. In *Proc. 5th Workshop on Data Analysis (WDA 2004)*, Vysoké Tatry, Slovakia, June 24–27, 2004; Paralič, J., Pözlbauer, G., Rauber, A., Eds.; Elfa Academic Press: Vysoké Tatry, Slovakia, 2004; pp 67–82.

(43) Fyfe, C. The Topographic Product of Experts. In *Proceedings of the 15th international conference on Artificial Neural Networks: biological Inspirations. Part I*, Warsaw, Poland, September 11–15, 2005; Duch, W., Oja, E., Zadrozny, S., Eds.; Springer-Verlag: Berlin, 2005; pp 397–402.

(44) Steil, J. J.; Sperduti, A. *Indices to Evaluate Self-Organizing Maps for Structures*. [http://biocollub.uni-bielefeld.de/volltexte/2007/139/index\\_en.html](http://biocollub.uni-bielefeld.de/volltexte/2007/139/index_en.html) (accessed July 24, 2012).

(45) Kirt, T.; Vainik, E.; Vöhandu, L. A Method for Comparing Self-organizing Maps: Case Studies of Banking and Linguistic Data. In *Proceedings of eleventh East-European conference on advances in databases and information systems*, Varna, Bulgaria, September 29–October 3, 2007; Ioannidis, Y., Novikov, B., Rachev, B., Eds.; Springer: Berlin, 2007; pp 107–115.

(46) Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence*, Hobart, Australia, December 4–8, 2006; Sattar, A., Kang, B.-H., Eds.; Springer: Berlin, 2006; pp 1015–1021.

(47) Varnek, A. Fragment Descriptors in Structure–Property Modeling and Virtual Screening. In *Cheminformatics and Computational Chemical Biology. Methods in Molecular Biology*; Bajorath, J., Ed.; Springer: New York, 2011; Vol. 672, pp 213–243.

(48) Theis, R. J.; Dessy, R. E. The Photochemical Reaction of 1,3,4,6-Tetraphenylhexatriene. *J. Org. Chem.* **1966**, *31*, 4248–4249.

(49) Barton, J. W.; Shepherd, M. K. Benzocyclo-octenes. Part 5. Thermal Rearrangements of Benzocyclo-octenes to benzo[a]-cyclopropa[c,d]pentalenes. *J. Chem. Soc., Perkin Trans. 1* **1986**, 961–966.

(50) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855–868.