J Comput Aided Mol Des (2011) 25:237–252
DOI 10.1007/s10822-011-9416-2

# Local neighborhood behavior in a combinatorial library context

**Dragos Horvath · Christian Koch · Gisbert Schneider · Gilles Marcou · Alexandre Varnek**

Received: 25 November 2010 / Accepted: 31 January 2011 / Published online: 12 February 2011


**Abstract** This article revisits a particular aspect of the molecular similarity principle—the Neighborhood Behavior (NB) concept. Earlier, the NB optimality criterion was introduced to select descriptor spaces, combining a given descriptor set and a similarity metric, which optimally comply with the similarity principle. Here, we focus on a "local" analysis based on the neighborhood of individual bioactive compounds. The defined NB-score measures similarity-based virtual screening success when using individual actives as queries. Systematic studies of local NB have been performed on a large combinatorial library of compounds with reported $IC_{50}$ values for five proteases, involving more than 140 descriptor/metric combinations of various fragment- and pharmacophore-based descriptors and different similarity metrics. Although, for each descriptor/metrics combination, the NB-score heavily depends on the query compound, on the average, 2D pharmacophore-based descriptors outperformed their 3D counterparts.

**Keywords** Neighborhood behavior · Pharmacophore descriptor · Fuzzy pharmacophore · Similarity · Virtual screening · Protease inhibition

D. Horvath (✉) · G. Marcou · A. Varnek
Laboratoire d'InfoChime, UMR 7177 Université de Strasbourg-CNRS, Institut de Chimie, 4, rue Blaise Pascal, 67000 Strasbourg, France
e-mail: horvath@chimie.u-strasbg.fr; d.horvath@wanadoo.fr

C. Koch · G. Schneider
Swiss Federal Institute of Technology (ETH), Institute of Pharmaceutical Sciences, Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland

## Introduction

Neighborhood Behavior (NB) [1–4], the relationship between structural similarity of molecules and their experimental properties, is a quantitative expression of the classical similarity principle [5, 6] stating that "similar molecules have similar properties". As a key to rational analysis of Structure–Activity Relationships (SAR), it is a central research topic in chemoinformatics and medicinal chemistry. NB is a property of the chosen descriptor space (DS), in which each molecule $M$ is represented by a point defined by the $N_D$-dimensional vector of molecular descriptors $D_i(M)$, $i = 1 \ldots N_D$, where the distance (metric) between two such points represents the calculated dissimilarity score [7] $\Sigma(m,M) = f[D_i(m), D_i(M)]$ between the associated molecules $m$ and $M$.

The aim of this article is to revisit the NB issue from a "local" point of view in the context of a combinatorial library that was tested on five different proteases. The size and quality (experimental homogeneity) of the data qualify it as an opportunity to investigate the behavior of similarity-based virtual screening tools in the context of combinatorial library design. The specifically addressed points are:

*Definition of the local NB criterion*: Often NB criteria are based on the simultaneous analysis of all possible molecule pairs within a data set, assuming that compound density is roughly constant throughout the entire chemical space. In order to avoid potential artifacts arising due to inhomogeneous compound density in descriptor space, the mean of "local" NB criteria that specifically address the vicinity of the most active hits is introduced, by contrast to more "global" NB indices.

*Noise-suppressed "ascertained" NB criteria*: The original NB optimality score [3] has been amended by

subtraction of expected noise levels, obtained from data scrambling experiments, which monitor the risk of artifactual NB resulting from oversampling of molecule pairs with similar properties.

*Definition of Component Merits (CM):* Asserting the relative merit of the three different strategic choices, or "components" making up a descriptor space (descriptors, normalization strategy, metric) to the NB in a particular descriptor space.

*NB optimization by descriptor selection and creation of a privileged protease-specific chemical space from the most information-rich available target:* Supervised quantitative structure–activity relationship (QSAR) models involving (a) selection of important descriptors, and (b) weighing the latter in order to fit quantitative equations predicting the targeted affinity value, may only be derived in the presence of a body of known examples including both active and inactive compounds. However, such models support only limited extrapolation to related biological targets (unless their ligand binding sites are virtually identical). In the QSAR build-up process, target specificity is most often learned only during step (b)—descriptor weighing—whereas step (a), selection, provides a list of descriptors that may represent a meaningful, privileged DS with respect to entire families of related targets. The parsimonious use of QSAR-extracted information from step (a) alone may provide a chemically valuable balance between the accuracy of specific activity predictions and cross-target extrapolation.

This study will enquire whether such "unachieved" learning on hand of the Tryptase binding data, which allows for the construction of a protease-oriented chemical space with potentially improved NB with respect to three other proteases (Factor $X_a$, Trypsin and UPA). In this context, an extensive benchmarking of various descriptor spaces has been performed. This study features ISIDA [8, 9] fragment descriptors and various versions of more or less "fuzzy" two-point and three-point pharmacophore descriptors (CATS [10–14], LIQUID [15], FPT [16, 17], ChemAxon PF [18]). Various similarity scoring schemes were applied in the considered descriptor spaces.

## Methods

To facilitate reading, further on used symbols and abbreviations are listed in Table 1.

**Table 1** Explained key symbols and abbreviations

| Symbol | Meaning |
|---|---|
| CATS | *C*hemically *A*dvanced *T*emplate *S*earch descriptors (150-dimensional "D") |
| DPRED | QSAR-calculated Tryptase affinity used as a 1D descriptor |
| F3-6 | ISIDA fragment count descriptors (561 D) |
| FPT | Fuzzy pharmacophore triplet descriptors (3,508 D) |
| LIQUID | *LI*gand-based *QU*antification of *I*nteraction *D*istributions descriptors (120 D) |
| PF | ChemAxon pharmacophore fingerprints (210 D) |
| SEL | Descriptor space defined by terms entering the Tryptase QSAR model (22 D) |
| $\Sigma(m,M)$ | Calculated dissimilarity score between molecules m and M: neighborhood is defined by a dissimilarity cutoff $\Sigma(m,M) < d$, the optimal cutoff being denoted d* |
| FS, TS, TD, PFD | False similar, true similar, true dissimilar and potentially false dissimilar molecules pairs at given dissimilarity cutoff—the basis of NB criteria |
| NB | Neighborhood behavior |
| $\Xi$ | Ascertained optimality criterion |
| $\Omega$ | "Classical" optimality criterion |
| $\langle x \rangle$, $\sigma(x)$ | Mean & standard deviation of variable $x$ |
| CM | Component merits (synthetic performance score over various contexts) |

### Compounds, targets, core subset

A combinatorial library of 15,840 compounds, synthesized according to Ugi-type [19, 20] three-component reactions (Fig. 1), constitutes the base of the data set. The library was created using 15 amines, 44 aldehydes and 24 isonitriles in a one-pot-reaction by Dr. Lutz Weber and coworkers (personal communication) [12, 21, 22]. Inhibitory Concentration at 50% ($IC_{50}$) of enzyme activity (relative to uninhibited enzyme activity) had been determined for all compounds against five serine proteases [data courtesy of Morphochem AG, Munich]: Chymotrypsin, Factor $X_a$ (FXA), Trypsin, Tryptase, Urokinase-type Plasminogen Activator (UPA) [12, 23]. Upper measuring bounds were set to 100 μM.

Due to expectedly high computational cost of the upcoming analysis, a core subset of 2,500 compounds was selected for use throughout the study (except for Tryptase affinity machine learning simulations, see below). This

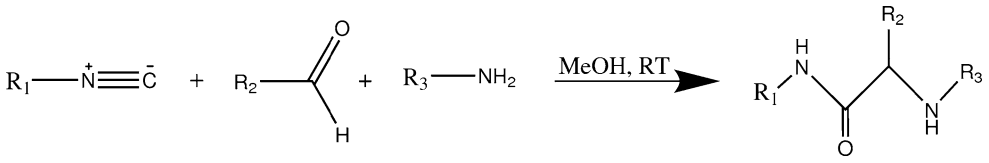


**Fig. 1** Scheme of the Ugi-type three-component reaction. Isonitriles, aldehydes and amines groups are coupled, yielding substituted amino acids

**Table 2** Number of actives ($pIC_{50} \geq 4.9$) and $IC_{50}$ statistics for the UGIsel core subset

| Target | Chymotrypsin | FXA | Trypsin | Tryptase | UPA |
|---|---|---|---|---|---|
| Actives | 12 | 81 | 3 | 216 | 11 |
| Median-$IC_{50}$ [μM] | 2.9 | 8.5 | 7.1 | 6.6 | 8.7 |
| Min.-Max. $IC_{50}$ [μM] | 2.0–12.0 | 0.16–12.6 | 3.80–8.1 | 0.09–12.6 | 0.04–11.98 |

selection contains all compounds that are active against at least one of the five targets—272 molecules with at least one $pIC_{50} \geq 4.9$ ($\approx 12.5$ μM) against any given target—and randomly selected inactives (Table 2). This rather low activity threshold was defined in accord with a lead discovery perspective. A pair of compounds was counted as having "similar property" with respect to a target if their difference $\Delta pIC_{50} < 0.5$.

## Descriptor space definition

### *Descriptors*

Several classes of descriptors were used in this work, and, within each such class, one or more different specific versions of the respective fingerprints—corresponding to different descriptor build-up parameterization schemes—were generated, as follows.

*Fragment descriptors*, representing counts of all the present atoms-and-bonds sequences of at least three and at most six atoms, were generated with the ISIDA [24, 25] software. They will be further on labeled as '**F3-6**'.

*ChemAxon Pharmacophore Fingerprints* '**PF**', were obtained with the *generatemd* [18] tool operating in default mode.

*CATS2D* [14] (*C*hemically *A*dvanced *T*emplate *S*earch) is based on topological cross-correlation of generalized atom types resulting in a pharmacophore fingerprint. Atoms in a molecule were assigned to one of five different potential pharmacophore point types T; namely hydrogen-bond acceptor (A), hydrogen-bond donor (D), positively charged (P), negatively charged (N) and lipophilic (L). Ten distance bins, with distances of length 0–9, respectively, were considered. This histogram count of 15 pair types times ten distance bins leads to a $N_{cats} = 150$-dimensional descriptor. Pair counts in CATS2D can be either reported as such, or scaled with respect to (a) total number of non-hydrogen atoms or, respectively (b) the total number of populated PPP-type pairs (the population level in a bin *T1–T2@n* to be divided by the total count of *T1–T2* pairs). Further on, these three CATS scaling schemes will be labeled **R**—raw pair counts, **A**—atom number-scaled CATS and **P**—pair count-scaled CATS, respectively. Furthermore, fuzzy logics-driven binning was considered: a type pair at a relative separation of $n$ bonds not only contributes an increment of $+1$ to the default bin associated to $n$, but also some user-defined increment $f < 1$ to neighboring bins $n-1$ and $n + 1$. Four different levels of fuzziness, where $f = (0, 0.25, 0.5, 0.75)$ for raw as well as scaled versions (R,S,P) led to a set of 12 CATS descriptor sets: from R1—raw, non fuzzy counts ($f = 0$) to P4—pair count-scaled, fuzzy ($f = 0.75$).

*LIQUID* [15] (*LI*gand-based *QU*antification of *I*nteraction *D*istributions) is a fuzzy 3D-pharmacophore descriptor based on correlation vectors. Atoms are assigned to three different ellipsoid-shaped pharmacophore types (lipophilic, hydrogen-bond acceptor, hydrogen-bond donor) modeled by trivariate Gaussians. The six associated type pairs are binned with consecutive intervals (width: 1 Å), for distances ranging from 1 to 20 Å, leading to $N_{LIQUID} = 6$ pairs $\cdot$ 20 bins = 120-dimensional descriptor. After atom typing (assignment of upto two matching types), "local feature densities" are calculated for every typed interaction point, which captures the occurrences of same types of interaction points within user defined cluster radii. Then interaction points are clustered according to local feature density. The cluster radii define how far from an interaction point to search for points with higher feature densities. For every cluster the geometrical center is calculated around which an ellipsoid PPP is mathematically modeled by trivariate Gaussians, thus adding tolerance ("fuzziness"). The size and orientation of the PPPs were identified by principal component analysis. Respective bins in the vector contain the mean probabilities of occurrences of a specific PPP-type pairs at a certain distance. Two cluster radii settings ([2,4]) were considered—let the former be further on labeled "localized" (**L**) and the latter "fuzzy" (**F**). In addition to the raw point pair counts (let this be designed by the label "**R**"), two scaling schemes were considered:

- Maximum-to-One, labeled "**M**": setting the highest valued bin of all 120 bins to 1 and accordingly adjusting all the others.
- Surface-to-One, labeled "**S**": the sum of the population levels of the 20 bins of every PPP-type pair is scaled to 1.

In total, six LIQUID-based DS—**RL**, **RF**, **ML**, **MF**, **SL**, **SF**—were employed in this work. Single 3D conformations of molecules used for LIQUID descriptor calculations were computed using CORINA [26] (version 2007).

*Fuzzy Pharmacophore Triplets (FPT)* [16, 17] represent fuzzy counts of monitored pharmacophore feature triplets, at given topological inter-feature distances, i.e. "edge lengths" of the considered triangles. Out of the considered FPT setups discussed in the original publication [17], used in this work were both the default FPT1 and the rule-based pharmacophore typing-based FPT (not relying on predicted group p$K_a$ values—therefore labeled FPT-nopK). In order to evidence the impact of the actual pH value on the NB of the pH-sensitive FPT1 (which correspond to a default pH = 7.4), two additional FPT-pH1 and FPT-pH5 (corresponding to pH values of 1.0 and 5.0, respectively) were also considered in this study. This lead to a number of four descriptor sets of the FPT family.

*Normalization strategies*

The above-mentioned descriptors may either be used as such in dissimilarity computing (no normalization) or they can be mean–variance normalized prior to their use. Mean–variance normalization (sometimes referred as Z-transformation [27]) replaces initial descriptor values $D$ by normalized $D^{norm} = (D - \langle D \rangle / \sigma(D)$, where the mean $\langle D \rangle$ and standard deviation $\sigma(D)$ are taken over all molecules of the UGIsel set. Labels **N** (no normalization) and **Z** (mean/variance normalization) were associated to the two strategies.

*Dissimilarity scores and metrics*

Three different distance metrics (generically denoted as $\Sigma$) have been employed as structural neighborhood criteria, in association with above-mentioned descriptors. The first two—Euclidean distance (labeled **EUCLID**) and the Dice coefficient-based distance **DICE**, are well known and need no further introduction [7]. The third is the Fraction of Differences (**FDIFF**)—a count of the fraction of features that are differently populated in a pair of molecules ($m$,$M$). In absence of normalization (**N**), the FDIFF count is incremented by 1 for each descriptor element $i$ if the fragment or pharmacophore element (pair, triplet) coded for by the position $i$ of the descriptor is populated in either $m$ or $M$, i.e. if $D_i(m) = 0$ and $D_i(M) > 0$ or $D_i(M) = 0$ and $D_i(m) > 0$. FDIFF equals the total count of elements fulfilling this condition, related to the dimension $N_D$ of the vector space of $D$. Its lower bound is 0, for molecules with near-identical fingerprints, in the sense that all the monitored patterns are either simultaneously present—albeit not necessarily in a same quantity—or simultaneously absent from $m$ and $M$. A pair of compounds in which any feature present in $M$ is absent from $m$ and vice versa would score a maximal FDIFF of 1. In the **Z** normalization scenario, a feature counts as "present" if its population level $D_i$ exceeds the mean population level $\langle D \rangle$, instead of $D_i > 0$.

*Descriptor spaces*

A descriptor space is a combination of a set of descriptors $D$, a normalization strategy and an associated metric. The one set of fragment descriptors (F3-6), plus the ChemAxon PF, the 12 CATS, the 6 LIQUIDS and the 4 FPT variants sum up to 24 different descriptor types. Multiplied by the two normalization strategies and the three distance scoring schemes, 144 distinct Descriptor Spaces were analyzed in this work. In addition to these "unsupervised" DS, customized descriptor spaces were generated by machine-learning-driven descriptor selection (vide infra).

QSAR model of tryptase affinity

The Stochastic QSAR Sampler (SQS) [16, 24, 28] was employed to select sets of descriptors that optimally correlate with experimental Tryptase affinity. Tryptase, the hit-richest of the five proteases against which the library was profiled, was the obvious choice for QSAR model training (the low occurrence rate of even moderately actives against other targets is a major obstacle for QSAR model training). 25% (3,960 compounds) of the initial library were randomly picked for training (2,640) and internal validation (1,320 molecules, respectively) of QSAR models expressing the observed p$IC_{50}$ values as linear combination of descriptors selected, by the evolutionary tool, out of the initial candidate pool regrouping FPT, ISIDA fragments and ChemAxon PF. Tryptase activity distribution in the training set was similar to the one in the entire library, i.e. showed a strong imbalance in favor of inactives and included notably few strong (micromolar or better) binders. Therefore, this might not be an appropriate set for building robust, predictive Tryptase affinity models to serve as such in lead optimization. Since weak actives are, after inactives, the second largest subset, the resulting QSARs will, at best, learn to discriminate weak from non-binders, but not nano- from micromolar ligands. However, the current goal is not Tryptase affinity QSAR modeling, but exploring whether the descriptors that are selected during QSAR build-up define any "privileged" descriptor space, with enhanced NB, and whether such descriptor space displays any generic "protease-friendly" features—in the sense of optimized NB not only with respect to the target used for fitting, but also with respect to other proteases.

SQS may, in principle, enumerate thousands of linear and non-linear models based on various descriptor selection schemes. Only linear models were explored here, and a "privileged spaces"—further on labeled SEL, for "selected"—were defined on hand of the 22 descriptors entering the model with the top internal validation correlation coefficient. The SEL descriptor set, in conjunction with the

two normalization strategies and the three distance scores, contributed six new DS to the study.

Alternatively, the predicted Tryptase affinity according to this model (labeled DPRED) was considered as a mono-dimensional descriptor space, and used as such in NB analysis—with respect to Tryptase and the other proteases alike. Although at first sight a surprising idea, the use of a descriptor space of $N = 1$ dimensions is both technically and conceptually straightforward: do compound pairs predicted to have similar Tryptase affinities according to the model show similar experimental affinities with respect to a (Tryptase or another) protease? Technically, however, only the Euclidean metric makes sense in the DPRED space, and normalization is irrelevant—DPRED therefore contributes a single new descriptor space to the study.

### The neighborhood behavior criterion Ω

Previously [3], the NB optimality criterion $\Omega(d)$, a function of the dissimilarity threshold $d$ was defined as given in Eq. (1).

$$\Omega(d) = \frac{kN_{FS}(d) + N_{PFD}(d)}{kN_{FS}^{(null)} + N_{PFD}^{(null)}} = \frac{kN_{FS}(d) + N_{PFD}(d)}{ks(d)N_{\neq} + [1 - s(d)]N_{=}} \tag{1}$$

Above, $N_{FS}$ stands for the count of "false similar" compound pairs with different properties, but nevertheless selected due to their calculated dissimilarity level $\Sigma(M,m) \leq d$. These $N_{FS}$ compound pairs represent a subset of all the pairs of compounds with diverging experimental properties, $N_{\neq}$. Conversely, $N_{PFD}$ represents the count of "potentially false dissimilar" compound pairs, which display similar properties, but were not selected because $\Sigma(M,m) > d$. The $N_{PFD}$ pairs belong to the $N_{=}$ pairs of compounds with similar experimental properties, and $N_{\neq}+N_{=}$ obviously equals $N_P$, the total number of considered compound pairs. Optimal NB corresponds to a choice of metric $\delta$ and optimal cutoff $d^*$ that minimize the weighted sum $kN_{FS} + N_{PFD}$. Here $k > 1$ ($k = 5$ in the present work) outlines the higher importance of keeping the number of false similar as low as possible, while failure to minimize $N_{PFD}$ translates in poor retrieval rates of actives in similarity-based virtual screening (less critical, for many actives are not *supposed* to be retrievable by these methods, since they may be genuinely dissimilar with respect to the query molecule, and nevertheless active). The sum to be minimized is related to its expectation value—the denominator of the fraction—corresponding to the null hypothesis that Σ does not exhibit any NB at all. Let $s(d)$ be the fraction of selected compound pairs with $\Sigma(m,M) \leq d$. If they are randomly distributed, then compound pairs of equivalent and diverging properties will be represented, within the selection of $s(d) \times N_P$ pairs, proportionally to their overall occurrence rates $N_{=}/N_P$ *and* $N_{\neq}/N_P$, respectively. Therefore, the expectation value for $N_{FS}$ is simply $s(d) \times N_{\neq}$, while for $N_{PFD}$ this amounts to the *not* selected fraction of $N_{=}$.

### Local optimality criterion Ω

The global optimality criterion [1, 3] was estimated on the basis of all the compound pairs *m,M* in the data set (excluding the ones in which both *m* and *M* were inactives), and plotted with respect to the dissimilarity cutoff $d$ in order to determine the optimal dissimilarity radius $d^*$ minimizing $\Omega(d^*)$. Alternatively, the "local" NB monitoring scheme advocated here consists in estimating, for each active molecule *M*, its local optimality score $\Omega^M(d^*)$ specifically over the $N-1$ pairs excluding the monitored active *M*—*later* on referred to as the virtual screening "query". Then, the mean local NB score over all actives $\langle \Omega^M(d^*) \rangle_M$ can be defined as the mean local optimality index. In this work, the global $\Omega(d^*)$ score has been adapted to locally monitor NB around each of the actives ($pIC_{50} > 4.9$) of a target (except for Tryptase, the hit-richest enzyme with 216 actives according to this definition, for which only the top 100 inhibitors were considered). For each active molecule *M*, a local neighborhood index $\Omega^M(d^*)$ was specifically calculated with respect to the $N-1 = 2{,}499$ compound pairs featuring *M*.

### Ascertained optimality criterion Ξ

The definition of the NB criterion—be it global or local—is based on the expectation values for false similar (FS) and potentially false dissimilar (PFD) compound pairs corresponding to a random selection of a pair subset of a given size *s*. In practice, however, if a series of random numbers is fed into Eq. (1) instead of the actual dissimilarity scores $\Sigma(m,M)$, observed $N_{FS}$ and $N_{PFD}$ may significantly differ from expectation values in the denominator. In particular, if the data set is unbalanced, with $N_{\neq} \gg N_{=}$ for example, the chance to have the $N_{=}$ few occurrences of property-wise similar compound pairs fortuitously ranked at the top of the Σ-ordered pair list—artifactually minimizing Ω—may not be negligible. Also, noise may be negligible for global NB calculations involving $O(N^2)$ compound pairs for a set of $N$ compounds, but become important in local NB monitoring based on $N-1$ pairs.

In order to quantitatively assess this risk, and ensure that monitored optimality systematically surpasses the "twilight zone" of Ω values ascribable to random noise, each calculation of the Ω profile as a function of $d$ according to a given metric $\Sigma$ in the current DS was systematically

accompanied by $n_{rand} = 20$ data scrambling simulations. Each scrambling run returned a calculated profile of $\Omega^{rand}$ as a function of $d$, after having randomly swapped the actual $\Sigma(m,M)$ values amongst the compound pairs. For each considered $d$ within the metric value range, it is thus possible to calculate the mean $\langle\Omega^{rand}\rangle_d$ of all the observed $\Omega^{rand}(d)$ values—expected (and observed) to approach the $d$-independent expectation value of 1.0. Furthermore, the standard deviation $\sigma(\Omega^{rand})_d$ of the $\Omega^{rand}(d)$ values defines the $d$-dependent thickness of the fluctuation-prone "twilight zone" of the $\Omega(d)$ profile (see Fig. 2). Actual neighborhood behavior at a cutoff $d$ may be ascertained only if the $d$,$\Omega(d)$ point of the profile lays below the "twilight zone" delimiter $\langle\Omega^{rand}\rangle_d - \sigma(\Omega^{rand})_d$. Therefore, this work introduces the Ascertained Excess Optimality Criterion $\Xi(d)$ as an updated, fluctuation artifact-free NB descriptor—see Eq. (2).

$$\Xi(d) = \langle\Omega^{rand}\rangle_d - \sigma(\Omega^{rand})_d - \Omega(d) \quad (2)$$

In order to avoid the confusion caused by the fact that a minimal $\Omega$ implied maximal NB, the new criterion $\Xi(d)$ was defined such that maximal, positive values are markers of optimal NB. The peak $\Xi(d^*)$ at the optimal cutoff $d^*$ is considered to be the measure of NB for a given data set located in a specified descriptor space.

## From ascertained optimality to rank-based quality scores

In principle, the mean $\langle\Xi^M(d^*)_M\rangle$ and its variance, representing the distribution of optimality criteria in a descriptor space DS, may be compared to the equivalent distribution in some other descriptor space DS′, in order to check whether these distribution significantly differ and, if so, to design the one biased towards higher $\langle\Xi^M(d^*)_M\rangle$. However, the distribution of $\Xi^M(d^*)$ values may not follow any simple density distribution rules, because it cannot be precluded (see further discussions) that some query molecule $M$ may show intrinsically high, and others intrinsically low $\Xi^M(d^*)$ values, irrespectively of the descriptor space. Note that such bias might invalidate the use of t-test significance [29]. As a precaution measure we recommend that, for purposes of comparison and unless explicitly stated otherwise, $\Xi^M(d^*)$ values should be converted to equivalent rank indices, based on a simple soccer-inspired scoring scheme, as follows:

Each active molecule $M$ on every target may be considered a "match", in which DS challenges DS′. An optimality relevance threshold $\Delta\Xi$, set to ten percent of the absolutely best $\Xi^M(d^*)$ over all actives of the considered target, and all the considered descriptor spaces, was empirically chosen to represent the tolerable fluctuation limit. DS wins $M$ and obtains three points if $\Xi^M(d^*)_{DS} - \Xi^M(d^*)_{DS'} > \Delta\Xi$. A draw, yielding one point to both DS and DS′ is called if $|\Xi^M(d^*)_{DS} - \Xi^M(d^*)_{DS'}| \leq \Delta\Xi$; otherwise, DS′ is declared winner and earns three points. The final performance criterion associated to a DS with respect to a given target will then be the *mean* point number scored per confrontation—a value between 0, for the DS systematically beaten by all the others, and 3, for a DS systematically outperforming all others for all $M$.

## Component merits

Descriptors, normalization strategy and choice of the dissimilarity metric jointly characterize a descriptor space and determine its NB. A relative benchmarking of the individual merit of each set of descriptors, of every metric and of each normalization strategy may therefore be difficult,

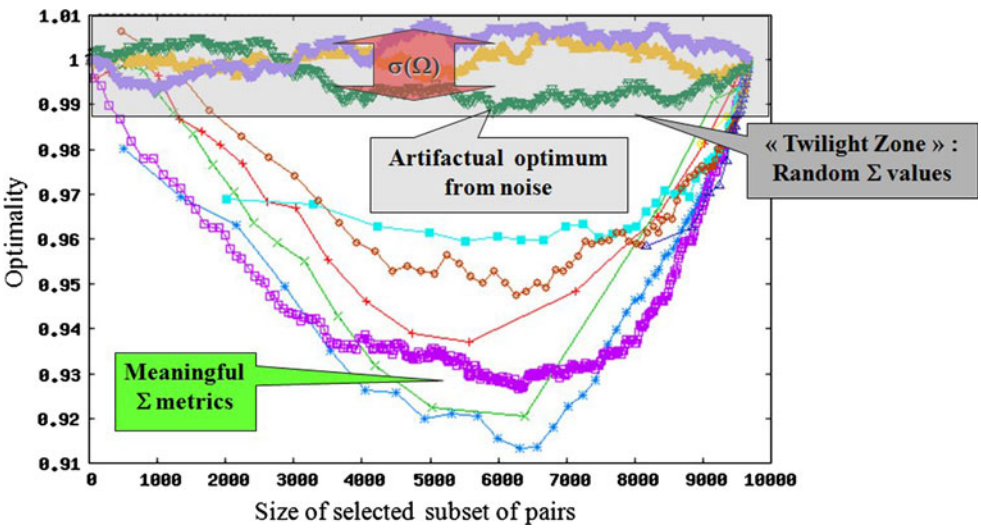


**Fig. 2** Typical U-shaped Optimality plot, highlighting the "twilight zone" covered by curves obtained with random numbers instead of actual dissimilarity metric values $\Sigma(m,M)$

for there are no guarantees that, for example, a descriptor space based on—say—fragment descriptors systematically shows better NB than any other descriptor space based on FPT. The former may, for example, do better with a Euclidean metric and z-normalization, while the latter may reach comparable NB performances only when used in conjunction with **FDIFF** without normalization. Which is then the better descriptor, to be recommended as default choice for similarity-based virtual screening? Similar questions may be asked with respect to metrics and normalization schemes.

The previous paragraph has outlined the definition of synthetic rank scores for the different descriptor space. The question addressed here is how to highlight the relative merit of the different components of a descriptor space (that is, descriptors, metric and normalization scheme) to the obtained rank score point. Continuing the sports-inspired analogy, a DS is considered a "team" of three components—descriptors, normalization strategy and metric. Here, component merits (CM) will be empirically taken as means over the three best performances in which the component acted and is therefore asymmetric. We argue that if a component is intrinsically flawed, it should never be able to display significant NB. The occurrence of a limited number of successes is enough to maintain the interest in that component, as the failures, even if they outnumber successes, might be due to a bad choice of the other descriptor space parameters. CM may be further averaged over the five considered targets/tournaments and, together with its standard deviation, shown as error bars will then represent the intrinsic quality of the component. Two components A and B (of the same category—descriptors, metrics, normalization schemes) may then be assessed for statistically significant differences, according to some simple mean/variance comparison test of the distributions.

Note that all the above considerations refer to the classical scenario, having a given set of descriptors used in the context of two normalization strategies and three different metrics. The mono-dimensional DPRED-based descriptor space, used only in conjunction with the Euclidean metric (its NB is furthermore invariant to normalization) is an exceptional case and was therefore not used in the outlined tournament scheme.

## Results and discussion

### Interpretation and significance of NB optimality scores

The NB optimality score was developed as an unbiased answer to the classical question 'How similar is "similar"?', symbolic of the difficulty of interpretation of similarity-based virtual screening results. Given a neighbor of an active reference compound at some dissimilarity score $d$, how can this abstract, descriptor and metric-dependent number $d$ be translated into the optimal binary decision—to synthesize and test the active analogue, or to ignore the inactive analogue, respectively? Like belief theory [30, 31] (addressing the same problem, but from a different perspective), NB optimality analysis has been designed to be dissimilarity-metric independent, applicable to any monitored, continuous or binary studied (bio)chemical property. As outlined in the initial articles [3, 32], a plot of $\Omega(d)$ as a function of the cutoff $d$ should be U–shaped, with a deeper minimum at $d^*$ denoting better Neighborhood Behavior. Note that the optimality criterion is herein *postulated* to represent a quality index of the NB of a descriptor space, and, implicitly, of the degree of success of thereon based similarity-driven virtual screening. Preference for this formalism, by contrast to other reported NB and virtual screening success criteria, can hardly be "proven" experimentally. For example, in linear regression analysis, the "best" linear model is the one minimizing the sum of *squares* of errors—not the sum of absolute errors, nor the one of their fourth powers (valid, but intellectually less appealing choices). The main empirical arguments in favor of an $\Omega(d^*)$-based NB monitoring are the following:

- It is based on straightforward counts of *False Similar* and *Potentially False Dissimilar* pairs, and behaves like an intuitive, tunable balanced accuracy criterion, in which the relative importance of FS and PFD pairs may be fine tuned to fit the specific needs of an experimental domain. It has recently [1] been pointed out that these indices may yield results that are more intuitive to a chemist's expectation of virtual screening quality than other, more artificial NB scoring schemes, and that a related index (the Cohen $\kappa$ statistics [33], unknown to us when designing $\Omega$) had already been used in cognitive sciences.
- It allows for an unambiguous definition of the optimal dissimilarity radius, reflecting the relative importance of FS versus PFD.

An in-depth analysis of the relationship between $\Omega(d^*)$—more precisely, its noise-free "ascertained" equivalent $\Xi(d^*)$ introduced here—and other NB monitoring scores is currently under preparation. Obviously, the relationship between $\Omega(d^*)$ and the concrete success of a similarity-based virtual screening campaign is anything but obvious, for concrete success is a multiobjective issue, depending on many aspects beyond the actual virtual screening: (1) availability of the virtual hits and their ability to undergo the experimental test (solubility, permeability, etc.), (2) originality ('novelty') of virtual hits

(discovery of scaffold-hopping: novel series may—but need not—compensate for lower hit rates). Achieving high performance in terms of benchmarking criteria is a necessary and desirable, but cannot guarantee actual success in prospective studies.

The reason for introducing the mean scored for local neighborhood behavior as a complementary criterion to the global index was to address two potential caveats of global scoring:

*Lack of validation*

In the global scheme, a single value $\Omega(d^*)$ based on the optimal dissimilarity cutoff $d^*$ is used to characterize the behavior of a descriptor space. NB monitoring serves to pinpoint the optimal choice of descriptor space for similarity-based virtual screening and in global monitoring of neighborhood behavior the entire data set represented a single global virtual screening experiment. Fitting an optimal virtual screening scheme means, in this context, picking the best descriptor space out of the possible options, based on their relative $\Omega(d^*)$ values: virtual screening success, defined by $\Omega(d^*)$, is a function of the employed descriptor space. However, global $\Omega(d^*)$ represents a single "training" quality score value, just like the training $R^2$ in linear regression. By contrast, local scores may, within the same descriptor space, fluctuate with respect to the various queries. Whereas $\langle \Omega^M(d^*) \rangle_M$ characterizes the entire data set, like $\Omega(d^*)$, its variance with respect to the various queries $M$ offers an implicit measure of robustness, in the way a cross-validation experiment helps pinpointing overfitting artifacts. Low variance means that any "fitted virtual screening model", i.e. a descriptor space chosen for its best performance on hand of a ("training") query, qualifies to maximize virtual screening performance for other ("validation") queries of a same target. High variance means that the virtual screening setup scheme learned on hand of any given query might not be safely extrapolated to other experiments. Note that, in principle, the optimal dissimilarity radius $d^*$ might also be considered an additional adjustable parameter of the "virtual screening model", next to the main degree of freedom—the choice of the descriptor space. A given $d^*$ determined with respect to a "training" query should then be used, unmodified, for all other virtual screening campaigns based on the current descriptor space. In the present work, when estimating the local scores $\Omega^M(d^*)$, $d^*$ is free to readjust for each query $M$—the more accurate notation of the above score would herewith be $\Omega^M(d^{*M})$, but will not be employed for reasons of conciseness. In as far, $\langle \Omega^M(d^*) \rangle_M$ thus represents the mean of successes of *ideal* virtual screening experiments lead around each active $M$.

*Density artifacts*

If the compounds were homogeneously distributed in the considered descriptor spaces, then Eq. (1) is expected to return a global estimate of the degree of therein observed neighborhood behavior. If, however, the molecule sample turns out to form several clusters of different densities according to the local metrics, then $\Xi(d)$ will be biased by the neighborhood behavior of the densest cluster(s), as these contain most of the short-distance pairs $m,M$. In Fig. 3a, the neighborhood behavior within the upper left panel, denser, smaller cluster of compounds is low, but excellent within the less dense lower right agglomeration. The opposite is shown in Fig. 3b. In both cases, compound pairs within the dense cluster control the $\Xi(d)$ profile, and it cannot be excluded that all of the pairs within the other may be relegated amongst the potentially false dissimilar if all the involved $\Sigma(m,M) > d^*$. Packing actives tightly together in descriptor space is arguably good neighborhood behavior: a high $\Xi(d^*)$ score is thus expected for the scenario in Fig. 3b, by sharp contrast to its counterpart (Fig. 3a). However, one would fail when picking the scenario of Fig. 3b for a similarity-based virtual screening campaign, while using the active molecule depicted by the central, thick red dot in the lower right cluster as a reference. When centered on $M$, the dissimilarity value scale will implicitly adapt to local density conditions—it now becomes irrelevant whether the nearest neighbors of $M$ happen to be "far" according to a distance scale tuned on denser neighborhoods. The two scenarios presented in Fig. 3 feature both actives specifically surrounded by actives—contributing high $\Xi^M(d^*)$ peaks—and actives mingling with inactives. On average, $\langle \Xi^M(d^*) \rangle_M$ of the two depicted descriptor spaces might not show large discrepancies. Although case A displays a better neighborhood behavior (for the denser cluster successfully regroups actives together), the expectancy of a successful virtual screening experiment starting from some arbitrary lead—which is not necessarily part of the densest cluster—is roughly equivalent.

## Benchmarking results

*Descriptor benchmarking*

Figure 4 represents the mean component merits (CM) and standard deviations of a given descriptor set, throughout all targets. Plotting was performed in order of decreasing CM means. Notably, the observed CM decrease upon a change of the employed descriptor set turned out to be quite smooth. While the differences observed between the leftmost top performers and the rightmost poor performers are

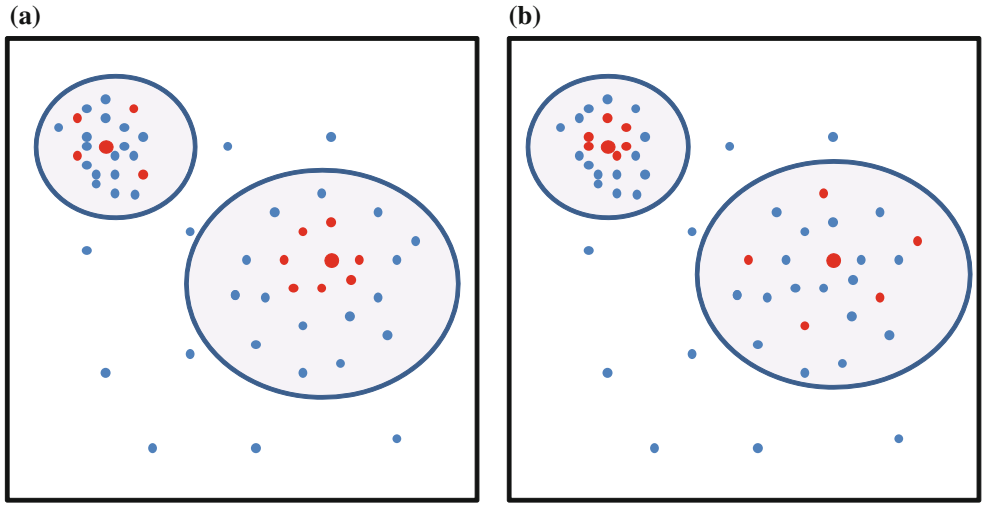


**Fig. 3** Inhomogeneous distributions of molecules (*blue* inactives, *red* actives) in the Descriptor Space may bias the globally estimated degree of neighborhood behavior

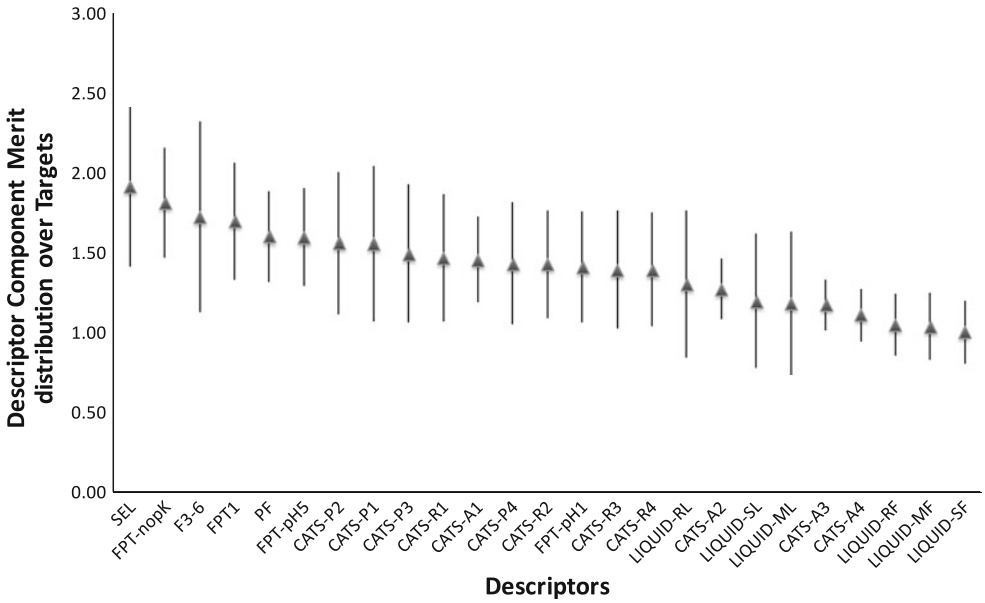


**Fig. 4** Synthetic rank scores of descriptor performance over all the targets: means (*triangles*) and standard deviations (*bars*) of Component Merits (Y)—success scores associated to the descriptor sets on X

statistically significant at Student's $t \geq 5$, the small differences between neighboring descriptors on the *x*-axis are not. The top-performing descriptor set, SEL, displays a statistically different behavior at $t \geq 5$ only with respect to CATS-A3, CATS-A4 and the rightmost three LIQUID variants. Its better performance with respect to FPTs, fragments F3-6, PF and CATS-A may not be entirely cleared from the suspicion of being a chance artifact.

Thus, the only stark observation that can be statistically supported is that the top performers, all of which are topological descriptors, systematically beat the 3D terms (LIQUID). This may yet be another illustration of the classical [34] conundrum of 3D descriptors: no conformational sampling at all (reliance on topological distances) may outperform insufficient conformational sampling.

The observed CM variances are also worth analyzing: among the top-performing descriptors, SEL and F3-6 display, in spite of good overall means, high variances: they are top rankers with respect to some queries, but fail with respect to others. Pharmacophore triplets (FPT) display, by contrast, a much more homogeneous behavior—they do not win any of the query-specific tournaments (see Table 3), but consistently appear among the top performers. Fragment descriptors, are potent NB tools with respect

**Table 3** Component Merits (synthetic descriptor rank scores) of best-ranked descriptors for each target

| Chymo-trypsin | CM | FXA | CM | Trypsin | CM | Tryptase | CM | UPA | CM |
|---|---|---|---|---|---|---|---|---|---|
| F3-6 | 2.67 | SEL | 2.32 | SEL | 2.32 | F3-6 | 1.93 | SEL | 2.00 |
| FPT-pH1 | 1.99 | FPT-nopK | 2.13 | CATS-P1 | 2.12 | SEL | 1.85 | CATS-P2 | 1.88 |
| FPT-pH5 | 1.95 | FPT1 | 2.07 | CATS-R1 | 2.09 | LIQUID-RL | 1.72 | FPT-nopK | 1.81 |
| FPT-nopK | 1.90 | CATS-P1 | 1.89 | CATS-P2 | 2.06 | PF | 1.69 | CATS-P3 | 1.76 |
| LIQUID-RL | 1.89 | PF | 1.82 | CATS-P3 | 2.05 | FPT1 | 1.65 | CATS-P1 | 1.75 |
| FPT1 | 1.89 | FPT-pH5 | 1.75 | CATS-P4 | 1.86 | LIQUID-ML | 1.63 | CATS-P4 | 1.69 |
| LIQUID-SL | 1.78 | CATS-P2 | 1.69 | CATS-A1 | 1.80 | FPT-nopK | 1.56 | CATS-R2 | 1.63 |
| LIQUID-ML | 1.71 | CATS-R1 | 1.63 | CATS-R3 | 1.78 | LIQUID-SL | 1.51 | CATS-A1 | 1.55 |
| PF | 1.65 | CATS-A1 | 1.55 | CATS-R4 | 1.77 | FPT-pH5 | 1.50 | CATS-R3 | 1.49 |
| CATS-A1 | 1.25 | CATS-P3 | 1.46 | CATS-R2 | 1.77 | CATS-P4 | 1.47 | CATS-R4 | 1.48 |
| SEL | 1.05 | CATS-R2 | 1.42 | F3-6 | 1.65 | CATS-R4 | 1.46 | CATS-A2 | 1.48 |

to chymotrypsin and tryptase, but perform less well with respect to the other three protease targets.

*Impact of the pH-dependence of fuzzy pharmacophore triplets on neighborhood behavior*

Fuzzy Pharmacophore Triplets FPT1 are occurrence-based means of the fingerprints of populated ionization states at a given pH (default 7.4), unlike their "rule-based" variant FPT-nopK, where default ionization states at pH = 7.4 are assumed. In a global neighborhood behavior study [17] against a reference set of drugs and drug-related molecules, the former were shown to dramatically improve neighborhood behavior and smoothen several apparent activity cliffs—compounds of diverging activities, but falsely assumed to be structurally very close unless p$K_a$ prediction is used to show that subtle p$K_a$ changes triggers significant changes of the predominant ionization states. This result could not be reproduced here—FPT1 and FPT-nopK somehow expectedly appear as equally valid descriptors in this study. The rule-based flagging strategy is very often equivalent to the p$K_a$ -based one for all the molecules with single—or multiple, but remote—ionizable groups, which represent the majority of the Ugi library. Cases where specific p$K_a$ effects come into play appear to be rare. However, p$K_a$-based (at pH = 7.4) and rule-based pharmacophore flagging both outperform pharmacophore flagging schemes at pH = 5, and the even worse setup at pH = 1 ($t$ scores of hypothesis FPT-nopK being better that FPT-pH5 and respectively FPT-pH1 are of 1.7 and 3.1, respectively).

Chemically relevant pharmacophoric feature flagging is thus important. The protonation state assignment does not however have to be absolutely correct, but relatively consistent—in the sense that equivalent functional groups will be equivalently treated by the protonation tool. Two carboxylic acids perceived as similar under their carboxylate forms will continue to be similar in their protonated state—although their similarity scores with respect to other categories of chemicals may significantly change.

*Impact of the employed metrics and normalization strategies on neighborhood behavior*

All three considered dissimilarity scores—Euclidean, Dice, Fraction of Differences—are compatible with good neighborhood behavior: over all the explored descriptor sets and normalization strategies, none of the three approaches was shown to perform significantly worse. The same is true for normalization–mean/variance normalization does not seem to play any important role. This is not unexpected since descriptor spaces were defined on hand of homogeneous descriptor sets, having feature counts as elements, all supposedly covering the same orders of magnitude. It should thus not be concluded that normalization is unimportant, but merely that normalization was implicitly achieved in most of the considered descriptor space, without need for further mean/variance rescaling. The impact of the normalization strategy is strongest with the SEL subset, a composite collection of pharmacophore triplet, pair and fragment counts, which no longer cover the same range.

*Privileged descriptor spaces*

The SEL and DPRED descriptor spaces, issued from SQS-driven model building with respect to the tryptase activity, may be regarded as "learned" DS, i.e. artificial constructs based on selected descriptors deemed important for explaining tryptase affinity. DPRED represents the predicted tryptase affinity, and is the DS that greedily exploits existing affinity information—selected descriptors are being accounted for and weighed with respect to their impact on tryptase activity.

**Table 4** SEL descriptors entering the tryptase affinity QSAR model, together with their QSAR coefficients

| Descriptor | Coeff. | Remarks |
|---|---|---|
| Ar2-HD4-PC4 | −0.0031 | Fuzzy pharmacophore triplets: |
| Ar4-HA12-PC12 | 0.0194 | Every feature is suffixed by the length (topological distance) of the opposing edge. |
| Ar4-Hp8-PC8 | −0.0015 | |
| Ar4-PC2-PC4 | 0.0136 | |
| Ar6-HD6-PC2 | −0.0034 | |
| HA4-HD2-NC4 | −0.0137 | Ex: Ar2-HD4-PC4 is a triplet having the donor HD at two bonds from the positive charge, while Ar-HD and Ar-PC are, respectively, 4 bonds apart |
| HA8-HA12-PC6 | 0.0024 | |
| HA8-NC4-PC8 | −0.089 | |
| HA10-PC4-PC12 | 0.5137 | |
| HD2-HD8-HD8 | 0.0088 | |
| HD4-HD8-Hp8 | −0.0020 | |
| HD4-Hp6-PC4 | −0.0202 | Hp: hydrophobe, Ar: aromatic, HBA: hydrogen bond acceptor, HBD: donor, PC: positive charge, NC: negative Charge |
| HD8-HD8-PC4 | 0.0077 | |
| Hp6-Hp10-Hp12 | −0.0002 | |
| Hp10-PC6-PC12 | 0.4932 | |
| Hp12-Hp12-NC2 | −0.0073 | |
| C*C–C = N | 0.1284 | F3-6 atoms & bond sequence counts (* stands for "any bond") |
| N–C–C*C*C–N | 0.3448 | |
| C–C*S*C | −0.0582 | |
| C = C–C*C | 0.0400 | |
| PF10 | 0.0424 | Elements #10 and 66 of default ChemAxon pharmacophore counts [18] |
| PF66 | 0.0602 | |

An intercept of 3.98 should be added to the coefficient-weighted sum of descriptor contributions, in order to obtain DPRED the actual predicted tryptase p$IC_{50}$

SEL, by contrast, is a descriptor space resulting from an "unachieved" learning process: it relies on the information regarding the relevance of descriptors for tryptase affinity, but ignores how exactly these are related to that experimental endpoint. Descriptors entering SEL, and their coefficients of the linear combination returning DPRED, are given in Table 4.

Table 5 reports the three top positions occupied by SEL-based descriptor space (all metrics and normalization strategies confounded), and, respectively, the DPRED space, in the local optimality-based "tournaments" with respect to each target. Recall that the rank scores reported in parentheses, span a range from 3 (if a descriptor space were systematically significantly better than all the others) to 0 (a descriptor space which is systematically and significantly outperformed by all the others). A value of 2 corresponds to a scenario where the given descriptor space is significantly better than 50% of its "competitors", and equally potent (slightly, but not significantly better or worse) than all the others. Alternatively, 2 may also be achieved in being significantly better than two-thirds of the competitors, and significantly worse with respect to the remaining third.

**Table 5** Absolute ranks of the three top performing SEL-based DS and of the DPRED-based DS, respectively, with respect to each of the considered targets (point numbers achieved by the respective DS in the "tournaments" relative to each targets are given in parentheses)

| Target | SEL | DPRED |
|---|---|---|
| TRYPTASE | #2 (2.06), #3 (1.95), #26 (1.52) | #1 (2.28) |
| FXA | #1 (2.37), #3 (2.33), #9 (2.17) | #4 (2.32) |
| UPA | #1 (2.15), #2 (2.00), #9 (1.83) | #106 (1.12) |
| TRYPSIN | #1 (2.74), #12 (2.12), #13 (2.11) | #127 (0.70) |
| CHYMOTRYP | #55 (1.06), #56 (1.06), #61 (1.02) | #131 (0.75) |

Unsurprisingly, DPRED is the uncontested winner with respect to tryptase, for which it was calibrated (with, furthermore, a significant degree of overlap between training set compounds and the molecules used in this study). It seems to be a poor descriptor space for the remaining proteases (although it maintains a top status for FXA, the target closest to tryptase in term of ligand binding behavior).

By contrast, SEL maintains excellent neighborhood behavior with respect to three more targets (chymotrypsin excluded). Hence, unachieved learning can escape overfitting artifacts, and allows the selection of descriptors that are generally relevant for protease inhibition. It may thus be used to construct "target-class-friendly" descriptor spaces, successfully supporting inhibitor design for novel targets, endowed by little experimental information, on the basis of data from related, well-explored proteins.

The deceiving behavior of SEL with respect to chymotrypsin appears to be an artifact due to the peculiar set of weak actives that serve here as reference queries for

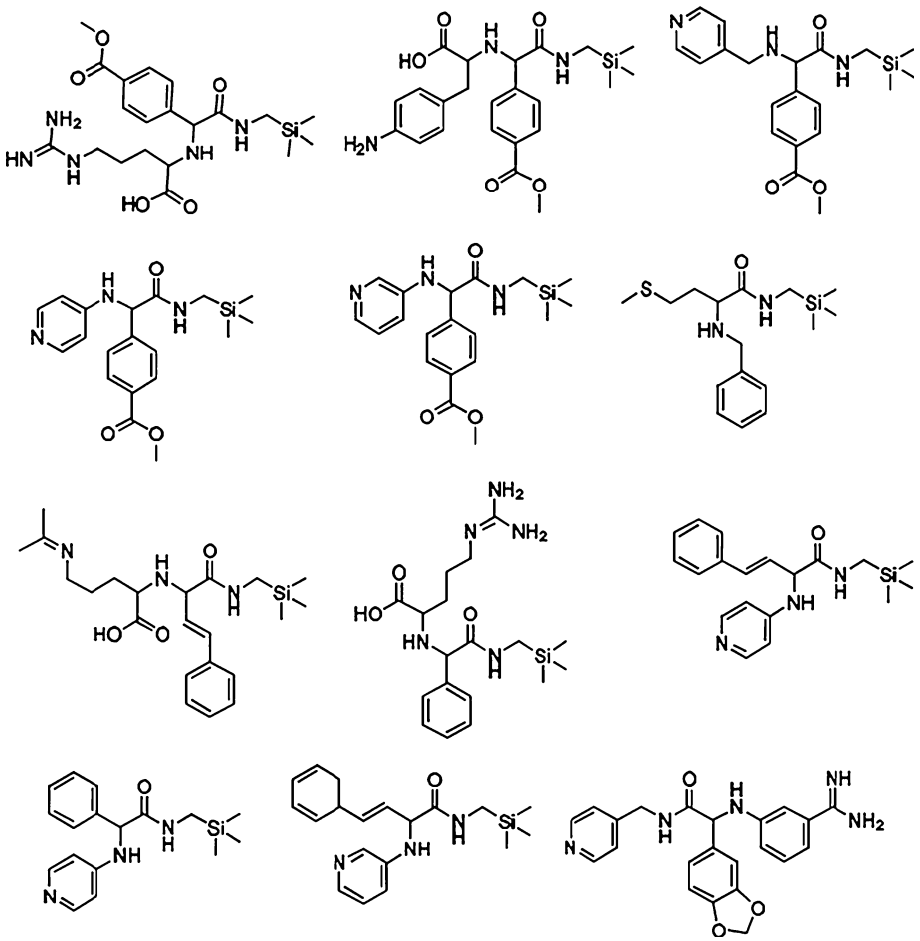

**Fig. 5** Structures of the 12 chymotrypsin "actives" ($pIC_{50} \geq 4.9$), most of which contain a key $-SiMe_3$ moiety

local optimality assessment. As can be seen from Fig. 5, this set is clearly dominated by moieties containing the trimethylsilyl fragment. This group might be responsible for much of the (rather weak) chymotrypsin activity of these compounds. The library features very many similar structures with different hydrophobic substituents instead of $-SiMe_3$, but none of these display any relevant chymotrypsin affinity. By contrast to F3-6 fragments, 2D pharmacophore-based DS cannot make the difference between $-SiMe_3$ and, say, t-Bu. F3-6 capture the exact nature of atoms: they do not convey the information of $-SiMe_3$ bulkiness, but merely record the fact that $-SiMe_3$ and $-CMe_3$ are not identical. Conversely, the difference is of no relevance when it comes to tryptase inhibition—thus, the relevant Si-containing fragments do not enter SEL—which therefore fails for chymotrypsin. 3D LIQUID descriptors, which account for the C–C and Si–C bonds length difference, all while considering both groups as hydrophobes, achieve their overall best benchmarking results with respect to the chymotrypsin data. This observation also supports the hypothesis of the key role of bulky $-SiMe_3$.

## Local neighborhood behavior: success of similarity-based virtual screening depends on the query

Generally speaking, similarity-based virtual screening may fail because of inappropriate choices of descriptors and metrics, and/or because there are no similar actives to be retrieved in the vicinity of the given query compound. Figure 6 presents a representative view of the distributions of local optimality scores that are obtained when using each individual active $M$ as "query" for a similarity-based virtual screening experiment. Distributions for the two most hit-rich targets, only in descriptor space based on the Dice metric and local normalization strategy, are. This plot is not meant for descriptor benchmarking, but to illustrate the fact that none of the employed descriptor space was able to systematically guarantee the success of similarity-based VS, irrespective of the chosen query. In each descriptor space, molecules $M$ with $\Xi^M(d^*) \approx 0$ can be found.

In addition to the mean local optimality, the query-related spread of optimality scores is an alternative quality criterion. Using a descriptor space scoring slightly lower

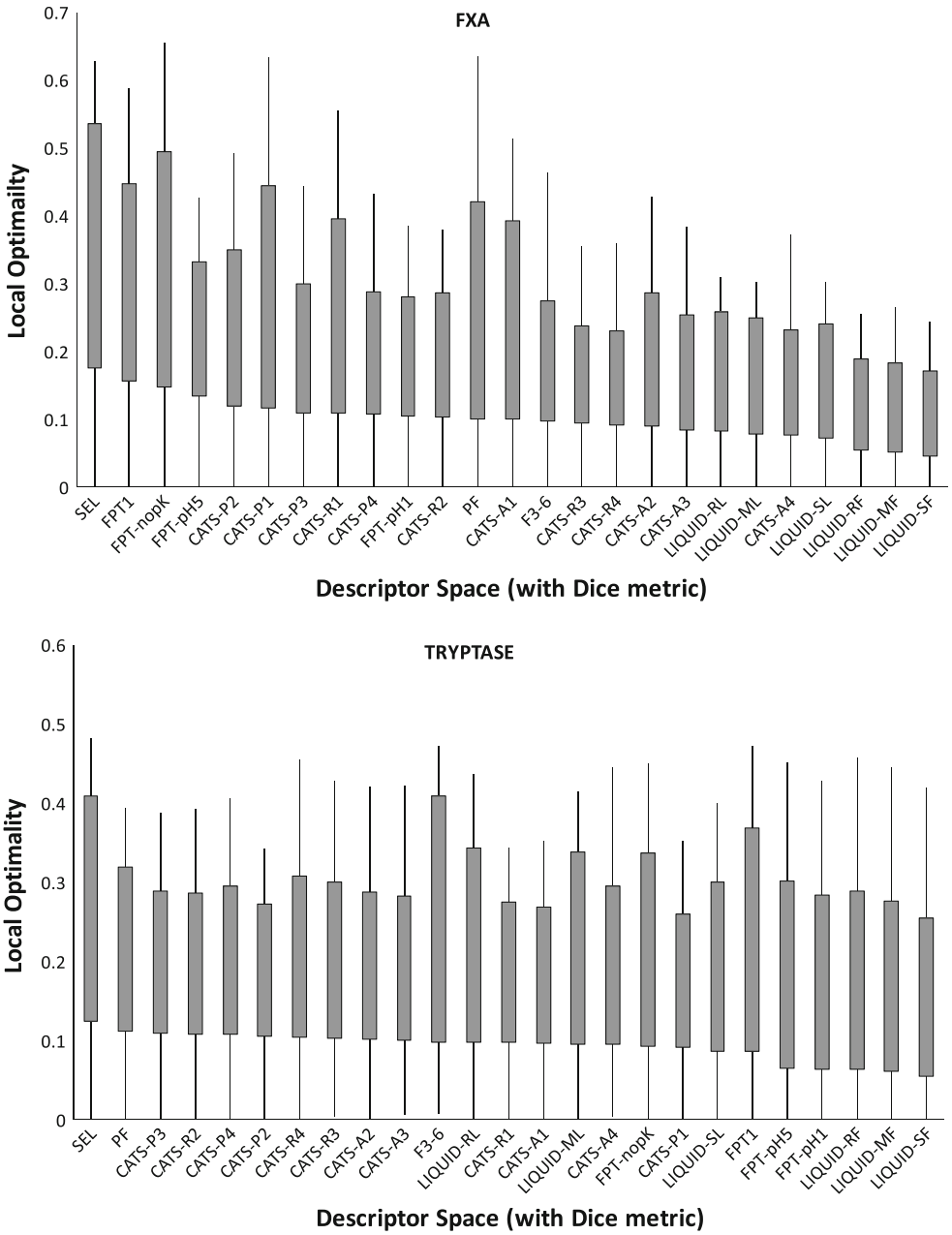


**Fig. 6** Distributions of local optimality indices $\Xi^M(d^*)$ over the set of individual active molecules **M**, for fXa and tryptase, in descriptor spaces based on the Dice metric and local rescaling strategy (associated descriptors listed on *X*). *Y* represents optimality. For each descriptor space, the associated vertical bar covers the observed optimality value range $min[\Xi^M(d^*)]_M$ to $max[\Xi^M(d^*)]_M$, while the rectangle covering $\langle \Xi^M d^* \rangle_M - \sigma[\Xi^M(d^*)]_M$ to $\langle \Xi^M(d^*) \rangle_M + \sigma[\Xi^M(d^*)]_M$ is centered on the mean optimality score and spans over two optimality standard deviation units. Enumeration order on *X* is based on $\langle \Xi^M(d^*) \rangle_M - \sigma[\Xi^M(d^*)]_M$ (rectangle lower bound)

mean optimality may be a fair compromise if the latter displays less sensitivity with respect to the input query. However, no ideal scenarios—high mean local optimality, but low optimality variance—could be evidenced here: Low variance was observed only in cases when, irrespective of the employed query molecules, the obtained local optimality scores were systematically low. In other words, similarity-based virtual screening either fails completely or succeeds partially for some reference compounds, but never seems to succeed irrespectively of the query.

Additional insight into the possible reasons of similarity-based virtual screening failure can be gained by monitoring the distribution of local optimality scores of an active molecule within various descriptor spaces. Actives scoring very low optimality values irrespective of the employed descriptor space may be "genuine singletons" (i.e. not conveniently surrounded by near analogues, with unique structural patterns: no metric can be expected to detect analogies to other actives). Alternatively, actives with a significant proportion of very close and predominantly inactive analogues (so that they would naturally be eligible as top-ranking neighbors, irrespective of the chosen descriptor space i.e. "genuine activity cliffs") also share the same signature of consistently low optimality scores, irrespective of the employed descriptor space.

By contrast, a molecule *M* displaying high local optimality scores irrespective of the employed descriptor space must be member of a cluster of active "me too" near neighbors—in the sense that their structural relatedness is obvious, and successfully captured by any reasonable similarity scoring scheme (including the medicinal chemist's perception).

Large variances of local optimality upon a descriptor space change imply a profound reshuffling of the neighborhood of *M*. This seems to be, by and large, the main type of behavior observed in this study (Fig. 7). For tryptase, 69% of the queries witness a variance of their local neighborhood optimality by more than 0.1. For Factor $X_a$, 50% of the active molecules display optimality variances above 0.15 and 80% are above 0.1. This comes even more as a surprise when considering that all the descriptor spaces (24 out of 25), except for the ISIDA fragment space F3-6, rely on pharmacophore patterns. The shift of similarity criteria from topology to spatial pharmacophore features does not account for much of the observed local optimality variance.

The strong dependence of virtual screening success on both the used methods and descriptor space has been previously evidenced [35, 36]. It is no surprise to witness the

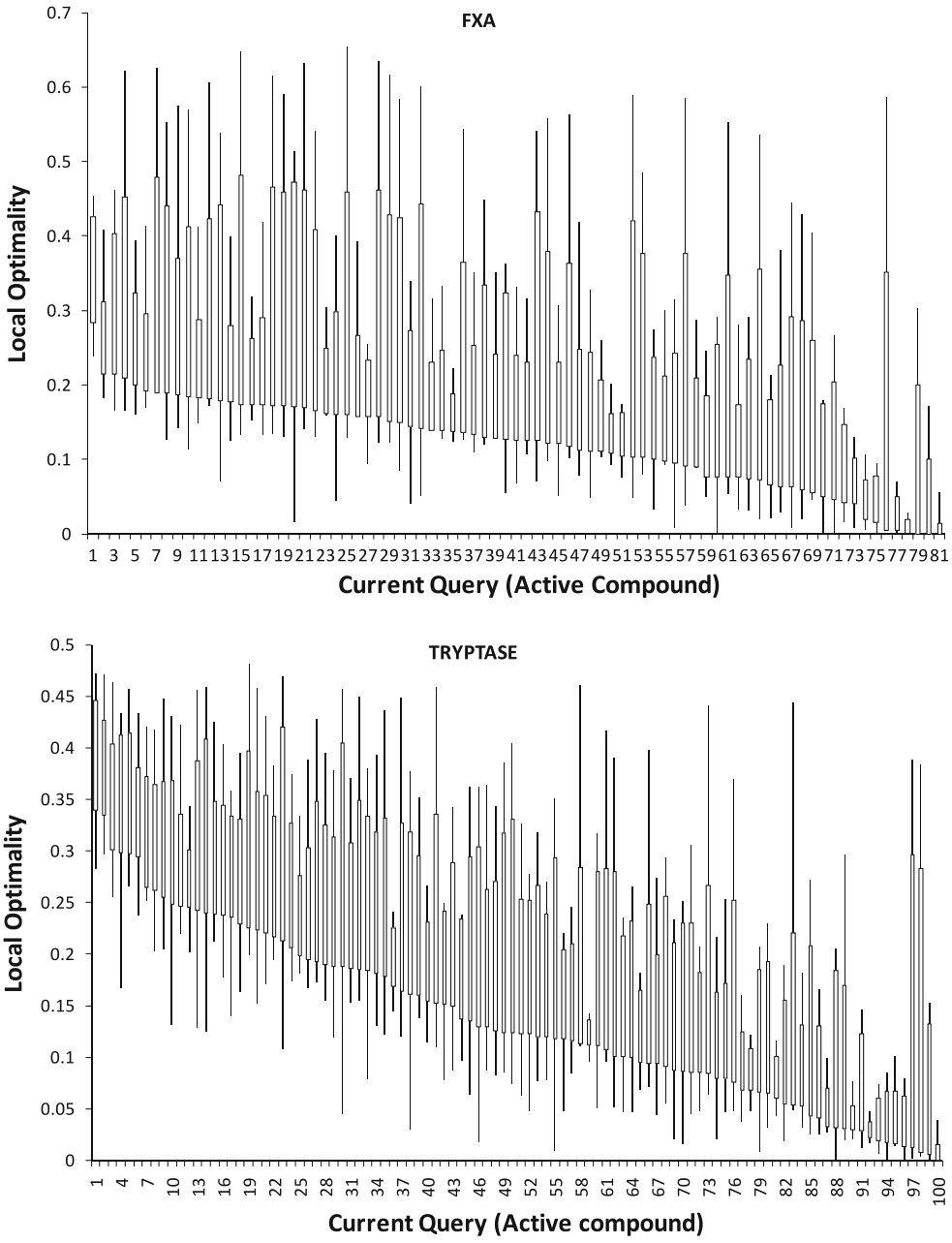


**Fig. 7** Distributions (on *Y*) of local optimality indices of each of the actives **M** (enumerated on *X*), for fXa and tryptase, over the Dice-based descriptor spaces with local normalization strategy. Y represents optimality. For each *M*, the associated vertical bar covers the observed optimality value range $min[\Xi^M(d^*)]_{DS}$ to $max[\Xi^M(d^*)]_{DS}$, while the rectangle covers $\langle \Xi^M(d^*) \rangle_{DS} - \mathrm{var}[\Xi^M(d^*)]_{DS}$ to $\langle \Xi^M(d^*) \rangle_{DS} + \mathrm{var}[\Xi^M(d^*)]_{DS}$

strong variance of the relative performance of a method when applied to virtually screen for ligands of different targets, or that a given similarity-based query may return dramatically different sets of virtual hits when expressed in conceptually different descriptor spaces (say hashed fingerprints vs. pharmacophores). However, the present work outlined that such "chaotic" behavior of similarity-based virtual screening is by no means conditioned by radically different chemical information being captured by the considered descriptor spaces, as it could be well evidenced within a single descriptor family—pharmacophore fingerprints. They only differ in terms of the empirical choices made at encoding the pharmacophore pattern under the form of a vector of numbers, and yet—these empirical choices are often sufficient to lead to a complete "reshuffling" of compound neighborhoods in the herein studied, densely packed combinatorial context. This is an important lesson learned from this study.

## Conclusions

To conclude, this study shows that, irrespective of the employed descriptor space, failure of similarity-based virtual screening experiments strongly depends on the employed query, and is unavoidable, in the sense that, even within the best scenarios that were tested here, some starting points will inevitably lead to deceiving results. However, the starting points bound to fail also are descriptor space-dependent, i.e. failure of the neighborhood search for active analogues around a query active *M* will not be systematic throughout all the descriptor spaces considered. The almost chaotic behavior of the virtual screening success with respect to the descriptor space and query compound might be a specific consequence of the combinatorial nature of the data set. This space contains many related compounds, and it takes little in the DS setup to change the relative ranking in terms of calculated similarity scores (if there is an excess population of molecule pairs with, say, Dice dissimilarity scores between 0.10 and 0.11, the third decimal position of calculated distances—one highly likely to change from one descriptor space to another—will be rank-determinant, and control neighborhood behavior).

Intuitively, one would expect a dependence of virtual screening success with respect to the query, but also believe that "inappropriate" starting points (singletons and molecules on the edge of activity cliffs) would be intrinsic to the studied target, largely independent on the used descriptor space and identifiable as such. If so [5, 37, 38], then it might be envisaged to classify biological targets with respect to the relative importance of "rough" regions of the structure–activity landscape, aiming towards definition of a generic druggability index of a target accounting for the easiness to obtain analogues by virtual screening. Unfortunately, these results do not support such expectations: a successful query in a given descriptor space DS may fail in some other DS′ albeit, on the average, the latter is not a less pertinent choice than DS.

It had been pointed out [39] that the Tanimoto cutoff at 0.85 with Daylight fingerprints tends to return rather poor results (less than 30% of success). Earlier, we have advocated the principle of optimizing descriptors and metrics in order to enhance similarity-based virtual screening performance [3, 32]. The current results however tend to reinforce and generalize the cited Tanimoto/Daylight-derived observations, irrespectively of employed descriptor spaces. Our study failed to support the intuitive idea that, among state-of-the-art molecular descriptors, some are generally better suited for virtual screening than others. The best and worst performing descriptor spaces happen to be largely different in each particular case, which makes that on the average a "winning" descriptor space is hardly emerging at all—with one clear-cut exception: 3D LIQUIDs were, statistically significantly, outperformed by their topological counterparts. Also, the SEL descriptor set was shown to represent a privileged descriptor space.

Any further attempts advocating rational design of "improved" similarity-based virtual screening tools should explicitly address the challenge of proving that the method is more likely to succeed on the average over multiple queries [40]. This goal should yet be pursued, in the light of encouraging results from the "unachieved learning" approach, and also considering a reinforcement of the capture of topological information. A fusion of topological information and pharmacophore flagging may be the key for improvement in terms of strong and homogeneous neighborhood behavior, as ongoing work of our team seems to suggest.

**Acknowledgments** We are grateful to Dr. L. Weber for providing the Ugi data set. Dr. Y. Tanrikulu is thanked for programming the LIQUID software.

## References

1. Papadatos G, Cooper AWJ, Kadirkamanathan V, Macdonald SJF, McLay IM, Pickett SD, Pritchard JM, Willett P, Gillet VJ (2009) Analysis of neighborhood behavior in lead optimization and array design. J Chem Inf Model 49(2):195–208. doi:10.1021/ci800302g
2. Horvath D, Barbosa F (2004) Neighborhood behavior—the relation between chemical similarity and property similarity. Curr Trends Med Chem 4:589–600
3. Horvath D, Jeandenans C (2003) 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 43:680–690

4. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. J Med Chem 39(16): 3049–3059
5. Maggiora GM (2006) On outliers and activity cliffs—why QSAR often disappoints. J Chem Inf Model 46:1535
6. Johnson M, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York
7. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Model 38:983–996
8. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J Comput Aided Mol Des 19(9–10):693–703
9. Solov'ev VP, Varnek AA (2004) Structure-property modeling of metal binders using molecular fragments. Russ Chem Bull 53(7):1434–1445
10. Schneider G, Schneider P, Renner S (2006) Scaffold-hopping: how far can you jump? QSAR Comb Sci 25:1162–1171
11. Renner S, Schneider G (2006) Scaffold-hopping potential of ligand-based similarity concepts. Chem Med Chem 1:181
12. Schuller A, Fechner U, Renner S, Franke L, Weber L, Schneider G (2006) A pseudo-ligand approach to virtual screening. Comb Chem High Throughput Screen 9(5):359–364
13. Schneider G, Lee M-L, Stahl M, Schneider P (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. J Comput Aided Mol Des 14:487–494
14. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed Engl 38:2894–2896
15. Tanrikulu Y, Nietert M, Scheffer U, Proschak E, Grabowski K, Schneider P, Weidlich M, Karas M, Goebel M, Schneider G (2007) Scaffold hopping by "fuzzy" pharmacophores and its application to RNA targets. Chem Bio Chem 8:1932–1936
16. Bonachera F, Horvath D (2008) Fuzzy tricentric pharmacophore fingerprints. 2. Application of topological fuzzy pharmacophore triplets in quantitative structure-activity relationships. J Chem Inf Model 48(2):409–425
17. Bonachera F, Parent B, Barbosa F, Froloff N, Horvath D (2006) Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. J Chem Inf Model 46:2457–2477
18. ChemAxon (2007) Screen user guide. http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html. Accessed Feb. 2009 2010
19. Ugi I, Steinbrückner C (1960) Über ein neues Kondensationsprinzip. Angew Chem 72:267–268
20. Ugi I, Meyr R, Fetzer U, Steinbrückner C (1959) Versuche mit Isonitrilen. Angew Chem 71:368
21. Weber L (2002) Multi-component reactions and evolutionary chemistry. Drug Discovery Today 7:143–147
22. Weber L, Wallbaum S, Gubernator K, Broger C (1995) Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. Angew Chem Int Ed Engl 34: 2280–2282
23. Weber L (2002) The application of multi-component reactions in drug discovery. Curr Med Chem 9:2085–2093
24. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko IV, Marcou G (2008) ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. Curr Comput Aided Drug Des 4(3): 191–198
25. Varnek A, Fourches D, Solov'ev V, Klimchuk O, Ouadi A, Billard I (2007) Successful "in silico" design of new efficient uranyl binders. Solvent Extr Ion Exch 25(4):433–462. doi: 10.1080/07366290701415820
26. CORINA (2005) Molecular Networks. 3.2 edn. GmbH, Erlangen
27. Kornhuber J, Terfloth L, Bleich S, Wiltfang J, Rupprecht R (2009) Molecular properties of psychopharmacological drugs determining non-competitive inhibition of 5-HT3A receptors. Eur J Med Chem 44(6):2667–2672
28. Horvath D, Bonachera F, Solov'ev V, Gaudin C, Varnek A (2007) Stochastic versus stepwise strategies for quantitative structure-activity relationship generation. How much effort may the mining for successful QSAR models take? J Chem Inf Model 47:927–939
29. Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. Biometrika 34:28–35
30. Martin YC, Muchmore S (2009) Beyond QSAR: lead hopping to different structures. Qsar Comb Sci 28(8):797–801. doi:10.1002/qsar.200810176
31. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Model 48(5):941–948. doi:10.1021/ci7004498
32. Horvath D, Jeandenans C (2003) 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 43:691–698
33. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20:37–46
34. Brown RD, Martin YC (1996) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J Chem Inf Comput Sci 36:572–584
35. Sheridan RP, Kearsley SK (2002) Why do we need so many chemical similarity search methods? Drug Discov Today 7(17): 903–911
36. Sheridan RP (2007) Chemical similarity searches: when is complexity justified? Expert Opin Drug Discov 2(4):423–430. doi:10.1517/17460441.2.4.423
37. Guha R, VanDrie JH (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48(3): 646–658
38. Petalson L, Bajorath J (2007) SAR index: quantifying the nature of structure-activity relationships. J Med Chem 50:5571–5578
39. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? J Med Chem 45(19):4350–4358. doi:10.1021/jm020155c
40. Chen BN, Mueller C, Willett P (2010) Combination rules for group fusion in similarity-based virtual screening. Mol Inf 29 (6–7):533–541. doi:10.1002/minf.201000050