Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes

Fanny Bonachéra,[†] Benjamin Parent,[†] Frédérique Barbosa,[‡] Nicolas Froloff,[‡] and Dragos Horvath*[,†]

Unite Mixte de Recherche 8576 Centre Nationale de la Recherche Scientifique − Unité de Glycobiologie Structurale & Fonctionnelle, Université des Sciences et Technologies de Lille, Bât. C9-59655 Villeneuve d'Ascq Cedex, France, and Cerep, Department of Molecular Modeling, 19 Avenue du Québec, 91951 Courtaboeuf Cedex, France

Received June 15, 2006

This paper introduces a novel molecular description—topological (2D) fuzzy pharmacophore triplets, 2D-FPT—using the number of interposed bonds as the measure of separation between the atoms representing pharmacophore types (hydrophobic, aromatic, hydrogen-bond donor and acceptor, cation, and anion). 2D-FPT features three key improvements with respect to the state-of-the-art pharmacophore fingerprints: (1) The first key novelty is fuzzy mapping of molecular triplets onto the basis set of pharmacophore triplets: unlike in the binary scheme where an atom triplet is set to highlight the bit of a single, best-matching basis triplet, the herein-defined fuzzy approach allows for gradual mapping of each atom triplet onto several related basis triplets, thus minimizing binary classification artifacts. (2) The second innovation is proteolytic equilibrium dependence, by explicitly considering all of the conjugated acids and bases (microspecies). 2D-FPTs are concentration-weighted (as predicted at pH = 7.4) averages of microspecies fingerprints. Therefore, small structural modifications, not affecting the overall pharmacophore pattern (in the sense of classical rule-based assignment), but nevertheless triggering a pK_a shift, will have a major impact on 2D-FPT. Pairs of almost identical compounds with significantly differing activities ("activity cliffs" in classical descriptor spaces) were in many cases predictable by 2D-FPT. (3) The third innovation is a new similarity scoring formula, acknowledging that the simultaneous absence of a triplet in two molecules is a less-constraining indicator of similarity than its simultaneous presence. It displays excellent neighborhood behavior, outperforming 2D or 3D two-point pharmacophore descriptors or chemical fingerprints. The 2D-FPT calculator was developed using the chemoinformatics toolkit of ChemAxon (www.chemaxon.com).

1. INTRODUCTION

Rational drug design[1,2] largely relies on the paradigm of site−ligand shape and functional group complementarity in order to explain the affinity of a ligand for its macromolecular receptor. While molecular modeling may offer a deeper insight into ligand recognition mechanisms—molecular dynamics simulations[3] or free energy perturbation calculations[4] might, in principle, also account for the entropic effects at binding—it did not succeed to displace the more straightforward concept of binding pharmacophores[5−7] from the minds of medicinal chemists.

The idea that ligand-site affinity can be broken down into pairwise contributions from interacting functional groups is, after all, not all that far-fetched. Ligand binding is entropically penalizing—a ligand would not restrict its freedom of translation, rotation, and conformational flexibility by binding to a receptor unless this cost is compensated by enthalpic gains. The existence of at least one ligand pose making favorable contacts with the active site is a necessary, albeit not sufficient condition—but even so, a virtual filtering procedure, discarding all molecules failing to show enough complementarity to the site, might well score significant enrichment in actives. Complementarity, in the pharmacophoric sense, must be understood as the ability to form stabilizing interactions—hydrophobic contacts, hydrogen bonds, and salt bridges—between a ligand and a site. The exact chemical nature of the interacting functional groups can be dropped in favor of their pharmacophore type[8] *T*—hydrophobic (Hp) or aromatic (Ar), hydrogen-bond acceptor (HA) or donor (HD), and positively charged (PC) or negatively charged (NC) ions. Pharmacophorically equivalent functional groups are considered replaceable, ignoring the specific ways in which their chemical environment may modulate their properties (the hydrogen-bonding strengths, for example). Formally, pharmacophore-type information can be represented under the form of a binary pharmacophore flag matrix $F(a,T)$, with $F(a,T) = 1$ if atom a is of type T and $F(a,T) = 0$ otherwise.

While the pharmacophore paradigm had been introduced as a purely qualitative framework to explain ligand affinity and specificity for a given site, it has been recently taken over and used as a fundament for various chemoinformatics approaches—empirical algorithmic approaches for rational in silico compound selection, on the basis of some numeric descriptors[9,10] of the distribution pattern of pharmacophoric groups in the molecule. This overall pattern, mathematically represented by a fingerprint (vector) in which every component refers to a specific combination of types at given separations, accounts for the nature and relative position (in terms of topology or geometry) of all of the groups that are

* Corresponding author tel.: +333-20-43-49-97; fax: +333-20-43-65-55; e-mail: dragos.horvath@univ-lille1.fr, d.horvath@wanadoo.fr.
[†] Université des Sciences et Technologies de Lille.
[‡] Cerep.

10.1021/ci6002416 CCC: $33.50
Published on Web 10/21/2006

potentially involved in site−ligand interactions (the actually involved ones are not necessarily known at this stage). Pharmacophore fingerprints may be exploited in both similarity searches[11] and predictive quantitative structure−activity relationships (QSARs).[12] Similarity searches assume that molecules described by covariant fingerprints have similar overall pharmacophore patterns and, hence, a higher chance to share a common binding pharmacophore (and to bind to a same target) than any pair of randomly chosen compounds. In QSAR, model fitting may select[13] several key fingerprint components as arguments to enter an empirical (linear on nonlinear) function estimating the expected activities.

Despite their simplicity and potential pitfalls,[14] pharmacophore-based empirical models have been shown to be successful chemoinformatics tools. A key factor to success is the proper definition of underlying pharmacophore descriptors, with a minimal loss of chemically relevant information. One widely used approach is to monitor the numbers of pharmacophore group pairs[9,15] as a function of the pharmacophore-type combination they represent and the distance separating them. Distribution density plots of such pairs with respect to geometric or topological distance have been shown to display excellent neighborhood behavior (NB),[16] in the sense of selectively attributing high pharmacophore similarity scores to compound pairs with similar experimental properties. The use of fuzzy logics[17] at the descriptor buildup and similarity scoring stages appeared to be paramount in order to smooth out conformational sampling or categorization artifacts. Higher-order descriptors[18−20] monitor the triplets or quadruplets of pharmacophore types and, therefore, furnish a much more detailed description of the overall pharmacophore pattern but become more costly to evaluate and, more important, much more prone to categorization artifacts. This is the case of the binary three-dimensional three- and four-point fingerprints, which were found to show deceivingly low NB compared to their fuzzy two-point counterparts.[16] The main reason for this is the uncertainty of the assignment of a pharmacophore-type triplet or quadruplet to one of the predefined basis triangles or tetrahedra corresponding each to one of the fingerprint elements. In the context of a binary three-point fingerprint (see Figure 1), a basis triangle i is fully specified by a list of three pharmacophore types $T_j(i)$—each type T_j being associated with a corner $j = 1-3$ of the triangle—plus a set of three tolerance ranges $[d_{kj}^{\min}(i), d_{kj}^{\max}(i)]$ specifying constraints for triangle edge lengths. Basis triangles should thus be understood as the meshes of a grid onto which a molecule is being mapped. Considering an atom triplet $\{a_1, a_2, a_3\}$ in a molecule, this triplet is said to match a basis triangle i if (1) each atom a_j is of pharmacophore type $T_j(i)$, in other terms, $F[a_j, T_j(i)] > 0$ for each corner j and (2) the calculated—geometric or other—interatomic distances $\mathrm{dist}(a_j,a_k)$ each fall within the respective tolerance ranges: $d_{kj}^{\min}(i) \le \mathrm{dist}(a_j,a_k) < d_{kj}^{\max}(i)$.

If in a molecule M an atom triplet simultaneously fulfilling the above-mentioned conditions can be found, then the fingerprint of M will highlight the bit i corresponding to this basis triangle. The risk taken here is that in a very similar compound M'—or, if $\mathrm{dist}(a_j,a_k)$ are taken as geometric interatomic distances, in a slightly different conformation of the same molecule M—the equivalent atom triplet $\{a'_1,a'_2,a'_3\}$ may fail to match the basis triangle i. It is sufficient to have one of the three distances $\mathrm{dist}(a'_j,a'_k)$ exceeding by little one of the boundaries in order to highlight a completely different basis triangle i' in the fingerprint of M'. Basis triangles i' and i are similar, but this is ignored by a binary similarity scoring scheme failing to find either bit i or bit i' set in both compounds. In two-point descriptors, where elements standing for successive distance ranges are assigned successive indices $i' = i \pm 1$, the fingerprint scoring function could be trained to account for the covariance of neighboring bins. Such a straightforward fuzzy logics correction is no longer applicable here. There are, for example, three "successive" triangles of i {with the same $[d_{kj}^{\min}(i), d_{kj}^{\max}(i)]$ ranges for two of the edges and using the successive tolerance range for the third} but only one slot at position $i + 1$ of the fingerprint. The direct consequence is that relatively small differences in interatomic distances may trigger apparently random jumps (symbolized by the arrow of Figure 1, upper part) of the highlighted bits from one location in the fingerprint to another.

This paper shows that fuzzy tricentric pharmacophore descriptors can be successfully constructed and used. The current work reports the buildup of the topological fuzzy pharmacophore triplets (2D-FPT) using shortest-path topological distances as an indicator of pharmacophore group separation. The descriptor reports basis triangle population levels in a molecule instead of a binary presence/absence indicator. An atom triplet in the molecule will contribute to the population levels of all of the related basis triangles by an increment which is directly related to their fuzzy matching degree (Figure 1, below). In the fuzzy approach, it is sufficient to characterize basis triangles i by a set of three nominal edge lengths $d_{jk}(i)$ instead of the above-mentioned tolerance ranges. The fuzzy degree by which an atom triplet is said to match a basis triangle will be 100% if interatomic distances perfectly equal nominal edge lengths, $\mathrm{dist}(a_j,a_k) = d_{jk}(i)$, and smoothly decrease—according to a law to be detailed further on—as discrepancies between real and nominal distances become important.

While 2D-FPTs are obviously not subject to conformational sampling artifacts, fuzzy-logics-based descriptors nevertheless present essential advantages:

- Their tolerance with respect to the limited variability of topological distances between pharmacophore groups mimics the natural fuzziness of ligand recognition by active sites, which may tolerate the insertion or deletion of linker bonds in a series of analogues.
- Their size may be significantly reduced by an appropriate choice of the basis triangle set. In the fuzzy approach, it is, for example, possible to keep only basis triangles with edge sizes being multiples of 2, 3, or 4. Within the strict buildup procedure, any atom triplet featuring two atoms separated by an odd number of bonds would fail to highlight any of the basis triangles of even edge lengths—it would, in other words, slip between the meshes of the grid. A fine grid enumerating all basis triplets with all possible combinations of nominal distances must then be used—but many more of these will be required in order to cover the same global span in terms of possible distances.

A second element of originality introduced here is the pharmacophore-type assignment scheme for ionizable compounds. Classical rule-based pharmacophore typing ignores the mutual long-range influence of multiple ionizing groups,

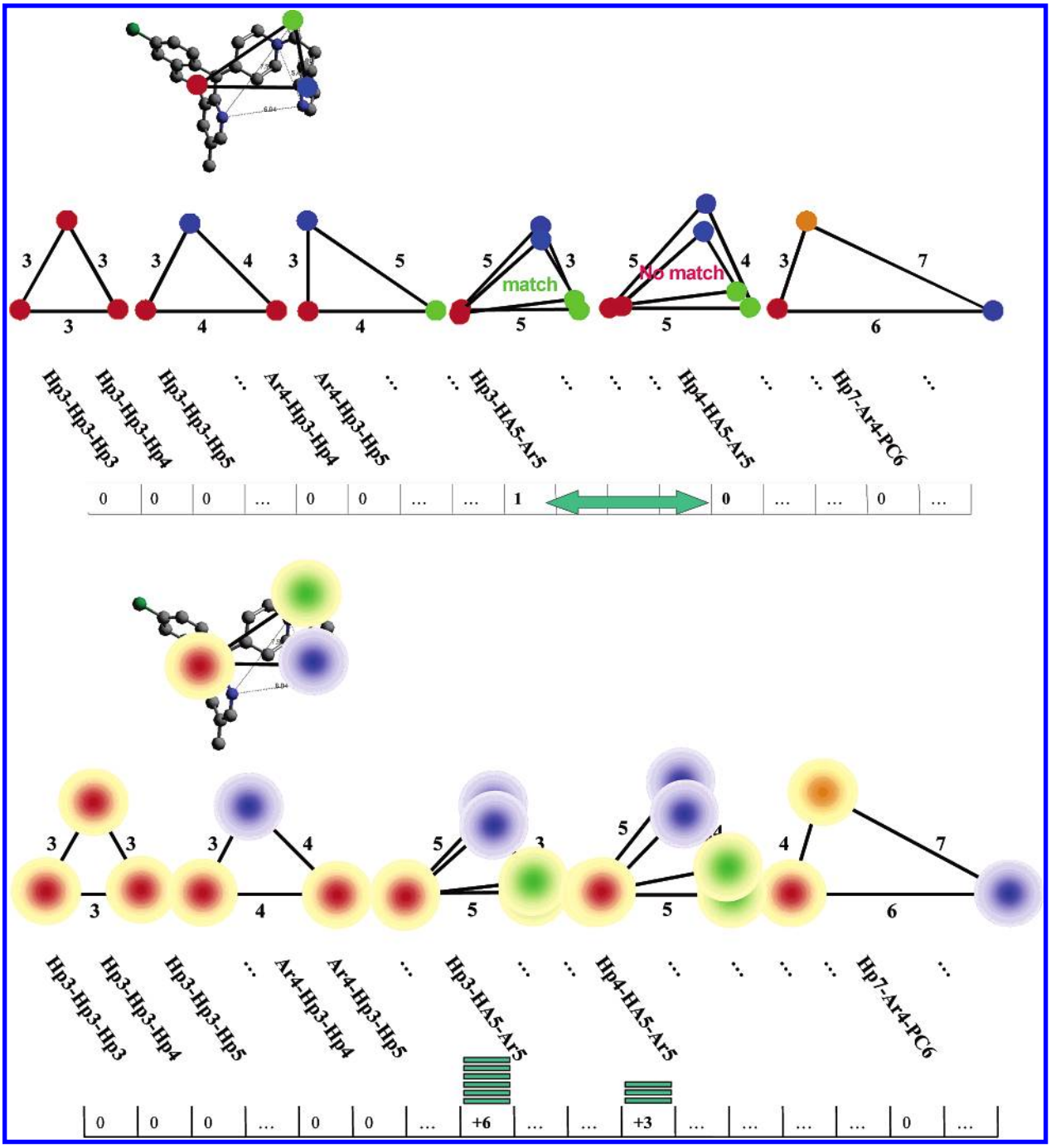

Figure 1. Buildup of a binary (above) and a fuzzy (below) pharmacophore triplet fingerprint, a vector in which every element stands for the presence (binary) or occurrence count (fuzzy) of given basis triplets. A triplet in a molecule (a) highlights a binary fingerprint component of the one best matching basis triangle or (b) increments the integer components of all of the matching basis triangles by amounts dependent on the match quality.

where each one of these is typed according to its protonation state of an isolated functional group at the considered pH. This leads to a typical overestimation of the occurrence of cation−cation or anion−anion pairs in polyamines and polyacids, respectively, and skews the molecular similarity measure upon the deletion of an ionizable group. Also, classical pharmacophore descriptors are not sensitive to electronic effects, being, for example, largely invariant upon the replacement of a methyl group (hydrophobe) by chlorine (another hydrophobe). This is acceptable unless, for example, the mentioned substitution prevents a neighboring amino group from accepting a proton in order to form a salt bridge at its binding site. To address these issues, 2D-FPT relies on the analysis of calculated[21] populations of all of the ionic or neutral forms involved in proton exchange equilibria—the "microspecies" μ, as they will be called throughout the paper—at a given pH. Each of these microspecies is mapped onto the basis triangle set, taking the actual anions and cations and donors and acceptors into account. The molecular fingerprint is rendered as the weighted average of microspecies fingerprints with respect to the predicted concentrations $c\%(\mu)$ of each microspecies μ at the considered pH of 7.4. In many cases, 2D-FPT-based analysis successfully proved that apparently near-identical compounds with puzzlingly different activities are not really as similar as they seem: the apparently minor (in the sense of classical rule-based

Table 1. Parameters Controlling 2D-FPT Buildup—Two Considered Setups

parameter	description	FPT-1	FPT-2
E_{min}	minimal edge length of basis triangles (number of bonds between two pharmacophore types)	2	4
E_{max}	maximal triangle edge length of basis triangles	12	15
E_{step}	edge length increment for enumeration of basis triangles	2	2
e	edge length excess parameter: in a molecule, triplets with edge length > $E_{max} + e$ are ignored	0	2
D	maximal edge length discrepancy tolerated when attempting to overlay a molecular triplet atop of a basis triangle	2	2
$\rho_{Hp} = \rho_{Ar}$	Gaussian fuzziness parameter for apolar (hydrophobic and aromatic) types	0.6	0.9
$\rho_{PC} = \rho_{NC}$	Gaussian fuzziness parameter for charged (positive and negative charge) types	0.6	0.8
$\rho_{HA} = \rho_{HD}$	Gaussian fuzziness parameter for polar (hydrogen bond donor and acceptor) types	0.6	0.7
l	aromatic−hydrophobic interchangeability level	0.6	0.5
	number of basis triplets at given setup	4494	7155

assignment) functional group substitutions actually had major impacts on ionization at the given pH. Many "activity cliffs" seen in classical descriptor spaces can be "leveled out" with pK_a-shift-sensitive 2D-FPT.

At last, the problem of appropriate similarity metrics to be used with 2D-FPT will be discussed, and an original scoring function, better adapted to such a high-dimensional descriptor, will be introduced. A plethora of various recipes have already been suggested[11] for comparing the descriptor sets (vectors) of two compounds m and M in order to determine a molecular dissimilarity score $\Sigma(m,M) = f[\vec{D}(M),\vec{D}(m)]$ (the distance in the structure space where each molecule is seen as a point localized by its vector of descriptors). 2D-FPT is, however, a large and potentially sparse fingerprint: out of the several thousands of basis triplets, only a few will be populated in simple molecules. Euclidean or Hamming distances may thus overemphasize the relative similarity of two simple molecules, while correlation coefficient-based metrics may be biased in favor of pairs of complex compounds. The original working hypothesis used here is to explicitly acknowledge that the simultaneous absence of a triplet in both molecules is a less-constraining indicator of similarity than its simultaneous presence, whereas its exclusive presence in only one of the compounds is a clear proof of dissimilarity. Specific partial distances are calculated with respect to the shared, exclusive, and null triplets in a fingerprint. A linear combination of these contributions leading to optimal neighborhood behavior was selected and used as the specific 2D-FPT similarity score.

For validation purposes, the NB of 2D-FPT was checked with respect to an activity profile featuring activity data (pIC_{50} values) of each molecule with respect to more that 150 targets, according to a previously outlined methodology.[22] Activity dissimilarity scores for $\sim 2.5 \times 10^6$ compound pairs were generated by Cerep, on the basis of the data in the BioPrint database[23,24] and according to a novel profile similarity scoring scheme. A second NB study has been carried out on publicly available data, by merging various QSAR data sets,[25−27] for different targets into an activity profile, assuming that each one of the molecules does not bind to any target except the one(s) for which pIC_{50} values above the micromolar threshold have been reported. Eventually, a validation study featuring virtual screening simulations will be presented. Virtual similarity screenings using 2D-FPT descriptors and metrics were performed by "seeding" a large commercially available compound collection (MayBridge) of 50 000 molecules with two sets of compounds (not used for 2D-FPT calibration) of known activities (featuring both actives and inactives) with respect to the dopamine receptor D2 and the tyrosine kinase c-Met, respectively. The ability of the 2D-FPT approach to retrieve the known actives and to avoid the selection of known inactives was benchmarked with respect to ChemAxon fuzzy pharmacophore fingerprints.[15]

2. METHODS

2.1. 2D-FPT Buildup: Fuzzy Mapping of Molecular Triplets onto Basis Triplets. Two prerequisite tasks must be completed prior to the actual construction of 2D-FPT.

Pharmacophore Flagging. This aspect will be detailed later on, because it is a central issue in ensuring the pK_a sensitivity of the fingerprints. At this time, the pharmacophore flag matrix $F_m(a,T)$, equaling 1 if atom a in the structure m is of type $T \in$ {"Hp", "Ar", "HA", "HD", "PC", "NC"} and zero otherwise, should be taken as granted. To account for the fact that aromatics and hydrophobes may, to some extent, interchangeably bind to the same binding pocket, in this work, aromatics are also flagged as lower-weight hydrophobes and vice versa. This requires the introduction of a fuzzy pharmacophore-type matrix $\Phi_m(a,T)$, identical to the binary flag matrix F for all of the polar types. For hydrophobes and aromatics, however, $\Phi_m(a,T) = \max[F_m(a,T), lF_m(a,T')]$ where T' stands for "aromatic" when T stands for "hydrophobic" and vice versa. $0 < l < 1$ is a tunable aromatic−hydrophobic compatibility parameter (Table 1). For example, an aromatic atom a has $F_m(a,\mathrm{Ar}) = \Phi_m(a,\mathrm{Ar}) = 1.0$, but $F_m(a,\mathrm{Hp}) = 0$ while $\Phi_m(a,\mathrm{Hp}) = l$.

Choice and Nonredundant Enumeration of the Basis Triplets Defining a Particular Version of 2D-FPT. The selection of a series of basis triplets to be monitored by the molecular fingerprint is essentially arbitrary and might be adapted to the specific problem for which 2D-FPTs are to be tailored. For the sake of concise specification, basis triplets are named $T_1d_{23}-T_2d_{13}-T_3d_{12}$, where T_i are the corner pharmacophore-type labels mentioned above and d_{ij} are the lengths of edges opposing each corner. For example, Ar4−Hp5−PC8 stands for a triangle in which the hydrophobe is four bonds away from the cation and eight bonds from the aromatic, while the aromatic and cation are five bonds apart. Basis triplets in this work were generated by systematic nonredundant enumeration, looping over each corner type, and respectively over each edge length from a user-defined minimal value E_{min} to a maximal E_{max}, with an integer step E_{step}. A pseudocode depiction of this procedure is given in Figure 2. Fingerprint element i hence monitors the population level of the basis triangle coded by the ith enumerated name in the list. The choice of E_{min}, E_{max}, and E_{step} (see Table 1) controls the coverage and graininess of the triplet basis set.

With these prerequisites, 2D-FPT buildup starts by the enumeration of all atom triplets $\{a_1, a_2, a_3\}$ in a molecule

```
for each T1 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {        #loop over type of corner1
   for each T2 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {     # ... corner 2
      for each T3 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {  #... and corner 3

        # Visit all the edge lengths from Emin to Emax with Estep
        for (d12=Emin,d12<=Emax,d12+=Estep) {

          #For 2nd edge, no need to loop over lengths below d12
          for (d13=d12,d13<=Emax,d13+=Estep) {

            # Only length combinations that may represent a triangle are enumerated
            # - third length may take only values verifying triangle inequalities
            dmin=max(Emin,|d12-d13|);
            dmax=min(Emax,d12+d13);
            for (d23=dmin,d23<dmax,d23+=Estep) {

              # Generate triangle corner labels Lk by concatenating types and
              # opposed edge lenght
              L1=T1d23;L2=T2d13;L3=T3d12;

              # Sort triangle corner label strings into a sorted list S.
              sort(L,S);


              # Final basis triplet name is obtained by concatenating corner labels in
              # their sorted alphabetical order
              NAME=S1'-'S2'-'S3;

              # Check whether this name had been generated previously;
              # if not add it to the list of basis triplets BLIST
              if !(BLIST.containsElement(NAME)) BLIST.add(NAME)

            } # end third edge length loop
          } # end second edge length loop
        } # end first edge length loop
      } # end third corner type loop
   } # end second corner type loop
} # end first corner type loop
```

Figure 2. Pseudocode rendering of the basis triplet enumeration procedure.

m, such that (1) the shortest topological distance between any two atoms equals or exceeds the minimal edge length E_{min} in basis triplets and (2) the longest one does not exceed the maximal edge length E_{max} by more than a tunable excess parameter e (Table 1).

To avoid confusion, in the following, the notation $t(a_k,a_j)$ to denote the (shortest-path) topological distance between two atoms will replace the generic interatomic distance dist(a_k,a_j) used in the introductory discussion on pharmacophore triplets. An atom triplet [note that the atoms of a triplet must be ordered such as to conveniently assign atoms to triangle corners; {a_1, a_2, a_3} should not be understood as a list of three atoms taken according to their sequential ordering in the structure but the permuted list with the aromatic atom in position 1 if $T_1(i)$ = Ar etc.] is said to "potentially match" a basis triplet i if (1) each atom a_j features the pharmacophore type $T_j(i)$, in other terms, $\Phi_m[a_j,T_j(i)] > 0$ for each corner j, and (2) the topological distances $t(a_j,a_k)$ are close to the corresponding nominal edge lengths $d_{kj}(i)$, in the sense that $|t(a_j,a_k) - d_{kj}(i)| \leq \Delta$, the latter being a user-defined tolerance parameter (Table 1).

If a basis triangle is found to be a potential matcher of the triplet, their actual degree of similarity is calculated according to a simplified triangle overlay procedure related to the ComPharm[28] algorithm. Both the basis triplet i and the molecular triplet are represented as triangles of given (integer) edge lengths in the Euclidean plane. Each atom a_j in corner j is a source of a "pharmacophore field" $\psi_j(T,P)$ of type T. The intensity of such a pharmacophore field at any point P of space located at a distance d_{jP} from corner j is postulated to decrease according to a Gaussian function $\Phi(a_j,T)\exp(-\rho_{Tj}d_{jP}^2)$ of this distance, scaled by the extent $\Phi(a_j,T)$ to which atom a_j represents the pharmacophore type T. A 2D-superposition procedure translating and rotating the basis triangle with respect to the molecular triplet in order to achieve a relative alignment maximizing the covariance of these pharmacophore fields is launched after an initial triangle prealignment placing equivalent corners as closely together as possible. The fuzziness parameters ρ_T are treated as independent user-defined parameters of the method (Table 1).

Triplet-to-basis triangle overlay calculates a pharmacophore field covariance score ranging (in principle) between 0 (no match at all) and 1 (congruence). This score $O(i,\{a_k\})$ is an implicit function of the present pharmacophore types (and their intrinsic fuzziness parameters ρ_T), the nominal edge lengths of the basis triangle, and the actual topological distances within the atom triplet. In reality, covariance scores of 0 are never obtained, because the overlaid objects are filtered potential matchers. Actually, triangles sharing a common edge are guaranteed to score at least 0.67 (two conserved features out of three), no matter how far their third corners fall apart. Therefore, only covariance scores above the 2/3 threshold are considered:

$$O^*(i,\{a_k\}) = \max[0.0, O(i,\{a_k\}) - 2/3] \qquad (1)$$

The increment of the basis triplet population level due to the presence of a given atom triplet in m is proportional to $O^*(i,\{a_k\})$. Given the potentially large 2D-FPT fingerprint size, it is more practical to operate with integer rather than real population-level values. A scale-up factor of O^* has been introduced such that a basis triplet represented in a molecule by a single, perfectly congruent triplet reaches an arbitrary population level of 50. The ith 2D-FPT element $D_i(m)$, representing the total population level of a basis triplet i in species m, becomes

$$D_i(m) = \mathrm{int}[150 \times \sum_{\mathrm{atomtriplets}\{a_k\}\mathrm{in}m} O^*(i,\{a_k\})] \qquad (2)$$

2.2. Proteolytic Equilibrium-Dependent Fingerprint Buildup. The 2D-FPT generator uses ChemAxon's molecular reader classes[29] to input compounds in various formats and to standardize[30] connectivity and bond-order tables of compounds admitting several equivalent representations. Standardization rules were formally defined as chemical reactions in an XML configuration file read by the ChemAxon standardizer object (setup file in the Supporting Information).

On the basis of the standardized internal representations, the pharmacophore-type assignment procedure begins by submitting the current molecule to the ChemAxon pK_a plug-in.[31] This plug-in first predicts pK_a values for the ionizable groups of the molecule, then generates all of the possible conjugated acids and bases—the microspecies μ—together with their expected concentration $c\%(\mu)$, in percent, at the given pH (equal to 7.4 throughout this work). Next, the ChemAxon pharmacophore mapper tool (PMapper[15]) is used to flag the pharmacophore types within every microspecies. Specific pharmacophore flag matrices $F_\mu(a,T)$ and $\Phi_\mu(a,T)$ will be generated for each microspecies μ. PMapper is controlled by an XML file specifying flagging rules. A set of relevant substructures is specified as SMARTS[32] with labeled key atoms. Functional groups matching such sub-

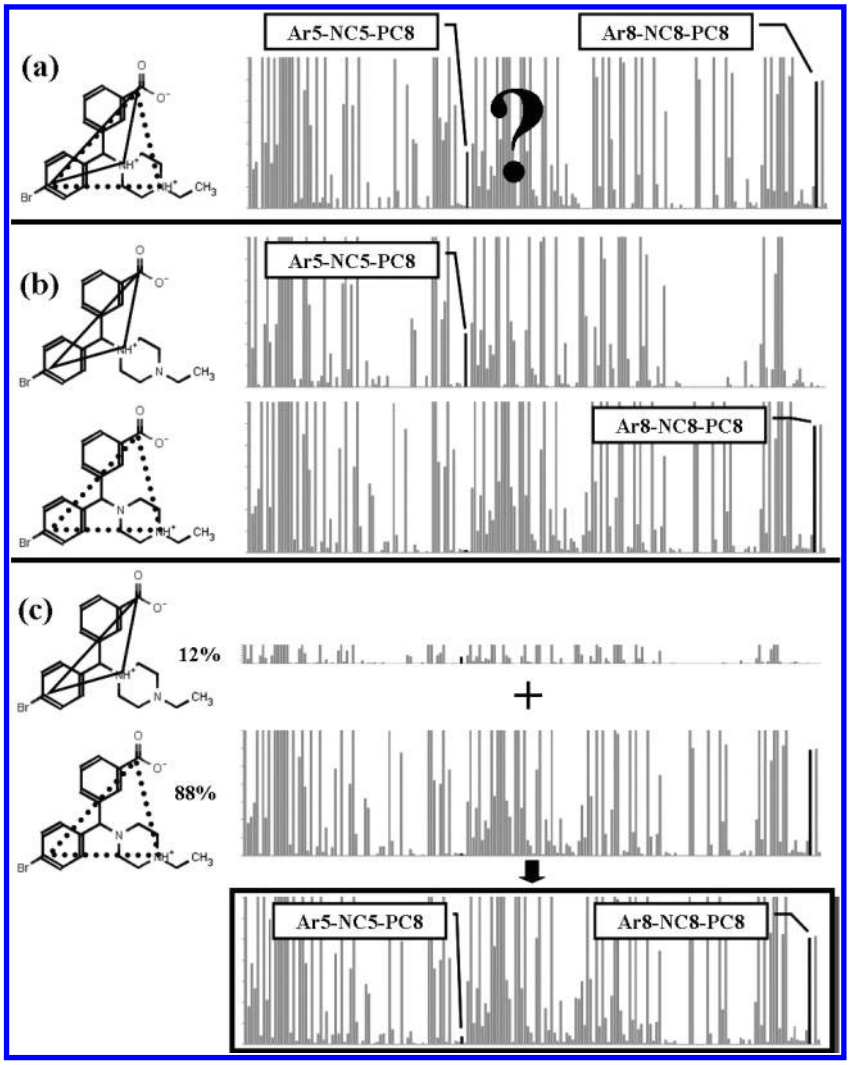

Figure 3. Graphical example of the principle of the construction of pK_a-sensitive 2D-FPT fingerprints: (a) rule-based pharmacophore flagging would assume three charged types in the molecule. Two triplets, both populated according to rule-based flagging, are localized in the sample fingerprint shown (bar sizes display population levels D_i, while the *x* axis enumerates the basis triplet counter *i*). Atom triplets that respectively contributed to each of the highlighted D_i's are marked in the structure. (b) The molecule actually appears at pH = 7 under the form of these two zwitterions. Each form carries only one of the triplets exemplified above. (c) The actual molecular fingerprint is obtained by weighed averaging of the microspecies fingerprints and, therefore, will resemble more the one of the zwitterionic forms predicted to occur at a concentration of 88% at equilibrium.

structures and the corresponding key atoms are detected in the molecule. An atom is assigned a given pharmacophore flag if it matches a certain substructure but not others. However, because formal charges are rigorously set in each microspecies, the assignment of PC and NC flags directly relies thereon. Any atom *a* carrying a positive formal charge (matching SMARTS "[*+]")—except for the nitrogen in nitro groups or nitrogen oxides—in the current microspecies μ will be assigned a flag $F_\mu(a,\mathrm{PC}) = 1$. By contrast, a classical flagging scheme would rely on the recognition of protonable group SMARTS and detect a potential cation even if it was not represented as such in the input molecule. Hydrogen-bond donor and acceptor flags are also set on the basis of specific rules pertaining to the microspecies. For example, a formally protonable N with a free electron pair, but not actually protonated in the current microspecies, will not be assigned an acceptor flag unless its pK_a value exceeds 5. Therefore, amide nitrogens will never be labeled as acceptors, but aniline nitrogens will unless they are strongly deactivated by electron-withdrawing groups. Oxygens always count as acceptors and −OH groups as donors. The recognition of aromatics is directly provided by ChemAxon's tools, while hydrophobes are defined as any carbon or halogen that is not aromatic and not charged.

The molecular fingerprint is thus obtained as a weighed average of microspecies fingerprints:

$$D_i(M) = \mathrm{int}\left[\sum_{\mathrm{microspecies}\,\mu\,\mathrm{of}\,M} \frac{c\%(\mu)}{100} D_i(\mu)\right] \qquad (3)$$

where $D_i(\mu)$'s are obtained for each microspecies μ, according to eq 2 using the specific pharmacophore flag matrix of the current microspecies for the estimation of the overlay score. The principle of proteolytic equilibrium-sensitive 2D-FPT buildup is illustrated in Figure 3. In the following, the notation D_i will, unless otherwise noted, implicitly refer to molecular average 2D-FPTs calculated according to eq 3.

2.3. FPT Similarity Scores. The appropriate choice of the similarity score $\Sigma(m,M) = f[\vec{D}(M),\vec{D}(m)]$ comparing the 2D-FPT vectors of two molecules *m* and *M* is critical in order to ensure good NB. With classical metrics, such as the

Euclidean or Dice formulas, a first question is whether descriptors should be used as defined in eq 3 or after average/variance rescaling, leading to the set of normalized $\mathscr{D}_k(M)$: where $\alpha(D_k) = \langle D_k(m)\rangle_{\text{all } m}$ stands for the average of the

$$\mathscr{D}_k(M) = \frac{D_k(M) - \langle D_k(m)\rangle_{\text{all}m}}{\sqrt{\langle D_k{}^2(m)\rangle_{\text{all}m} - \langle D_k(m)\rangle_{\text{all}m}{}^2}} = \frac{D_k(M) - \alpha(D_k)}{\sigma(D_k)} \tag{4}$$

population level of triplet k over the BioPrint drugs and reference compounds[24] and $\sigma(D_k)$ stands for the corresponding variance. A further choice consisted in introducing a weighting scheme to specific triplets that are significantly populated in relatively few classes of compounds and absent from all of the others. These may be subject to an up to 10-fold increase of their relative importance with respect to ubiquitously present ones:

$$W_k = \min\left[10.0, \frac{\langle D_k(m)\rangle_{m\text{with}D_k(m)>0}}{\alpha(D_k)}\right] \tag{5}$$

Throughout this paper, structural dissimilarity metrics used with 2D-FPT will be denoted by the symbol Σ superscripted by the type of the metric, with an index informing on the use of normalized descriptors (N) as given in eq 4 or the weighting scheme (W) defined in eq 5. For example, the weighed Dice dissimilarity score using normalized descriptors is defined below, with N_T being the total number of basis triplets of the given 2D-FPT setup:

$$\Sigma^{\text{Dice}}_{N,W}(m,M) = 1 - \frac{2\sum_{k=1}^{N_T} W_{kk}(m)\mathscr{D}_k(M)}{\sum_{k=1}^{N_T} W_k\mathscr{D}_k{}^2(m) + \sum_{k=1}^{N_T} W_k\mathscr{D}_k{}^2(M)} \tag{6}$$

Indices N and W are omitted unless the metric explicitly relies on normalization and weighting and in cases of specific metrics (see below) or metrics from third-party software, whenever normalization and weighting options are no longer available.

The third, main, original contribution of this paper is the introduction of Σ^{FPT}, a specific metric of the dissimilarity of fuzzy pharmacophore triplets. Classical similarity scores, however, are generic metrics, applicable in arbitrary vector spaces, for example, independent of the actual nature of molecular descriptors associated with the degrees of freedom of the structure space. As this work will show, the specific design of a similarity scoring scheme based on an actual interpretation of the information in the fingerprint may significantly improve NB.

Concretely, the knowledge that $D_i(M)$ represents population levels of basis triplets, and that the simultaneous absence of a triplet in two molecules is a less-constraining indicator of similarity than its simultaneous presence, will be actively exploited. A first prerequisite in this sense is the introduction of a measure of the significance $S_k(M)$ of a triplet k for a molecule M, with respect to the observed averages and variances of each triplet population level:

$$S_k(M) = \begin{cases} 0 & \text{if } D_k(M) < 0.7\alpha(D_k) \\ 1 & \text{if } D_k(M) > 0.7\alpha(D_k) + \sigma(D_k) \\ \dfrac{D_k(M) - 0.7\alpha(D_k)}{\sigma(D_k)} & \text{otherwise} \end{cases} \tag{7}$$

A triplet k in a pair of molecules (m,M) may fall into one of the following categories: shared (++), for example, significant—in the above-mentioned sense—for both m and M, null (− −), for example, not significant for either, and exclusive (+ −), for example, significant for either m or M but not for both.

Rather than assigning it to one and only one of these, its fuzzy levels τ of association to each of the categories are defined in order to always sum up to 1:

$$\tau_k^{++}(m,M) = \frac{S_k(M)\,S_k(m)}{\text{norm}}$$

$$\tau_k^{-\,-}(m,M) = \frac{[1 - S_k(M)][1 - S_k(m)]}{\text{norm}}$$

$$\tau_k^{+\,-}(m,M) = \frac{|S_k(m) - S_k(M)|}{\text{norm}}$$

$$\text{norm} = S_k(M)\,S_k(m) + [1 - S_k(M)][1 - S_k(m)] + |S_k(m) - S_k(M)| \tag{8}$$

The fraction of triplets in a category c therefore becomes

$$f^c(M,m) = \frac{1}{N_T}\sum_{k=1}^{N_T}\tau_k^c(M,m) \tag{9}$$

Classical distance functions are typically calculated on the basis of the differences observed for each component k of the molecular descriptors $\delta_k(m,M) = |\mathscr{D}_k(m) - \mathscr{D}_k(M)|$. The herein introduced originality consists of a separate monitoring of these contributions for the shared, exclusive, and null triplets. Rather than simply summing up all $\delta_k(m,M)$ contributions (leading to a Hamming-type dissimilarity score), weighed partial distances $\Pi^c(m,M)$ are estimated in order to monitor how much of the difference stems from triplets in each category:

$$\Pi^c_{W,N}(m,M) = \frac{\sum_{k=1}^{N_T} W_k\tau_k^c(m,M)\,\delta_k(m,M)}{\sum_{k=1}^{N_T} W_k} \tag{10}$$

The working hypothesis adopted here was that a meaningful dissimilarity score can be expressed as some linear combination involving certain of the three fractions defined in eq 9 as well as the three partial distances (eq 10). Successive trials monitoring the NB of the resulting metric with respect to a subset of the entire learning set (see the following section) led to the following expression:

$$\Sigma^{\text{FPT}}(m,M) = 0.1323\Pi_{W,N}{}^{+-}(m,M) + 0.6357\Pi_{W,N}{}^{++}(m,M) + 0.2795[1 - f^{++}(m,M)] \tag{11}$$

The NB of the herein proposed scoring scheme was benchmarked with respect to classical dissimilarity metrics in various validation studies.

2.4. Experimental Data and Validation Studies. The performance of 2D-FPT in similarity searches has been assessed and compared to that of other 2D and 3D pharmacophore descriptors, following the previously published methodology[16] for monitoring the NB of in silico similarity scores. In the current work, activity profiles of 2275 nonproprietary (commercial drugs and drug precursors) molecules from the BioPrint database of Cerep were used to calculate the activity dissimilarity scores $\Lambda(m,M) = f[\vec{p}(M),\vec{p}(m)]$ expressing the amount of difference between the response patterns of the two molecules with respect to the considered battery of targets. Profiles $p_t(m)$ report measured $pIC_{50} = -\log IC_{50}$ (mol/l) values of every molecule m against each of $N_{targets} = 154$ different biological targets t (enzymes, receptors). $p_t(m) = 9/6/3$ means that molecule m is a nano-/micro-/millimolar binder of t, respectively. The actual algorithm used for estimating the activity profile dissimilarity score $\Lambda(M,m)$ is outlined in Appendix A.

An alternative NB study has been conducted on the basis of an activity profile compiled from publicly available data sets[25−27] (see the Supporting Information). Unlike the highly diverse BioPrint data, this study features a compilation of 112 compounds tested on the angiotensin converting enzyme (ACE), 111 on acetylcholine esterase (AchE), 163 on the benzodiazepine receptor (BzR), 321 on cyclooxygenase-II (Cox2), 641 on dihydrofolate reductase (DHFR), 66 on glycogen phosphorylase B, 67 on thermolysin, and 88 on thrombin (THR)—a total of 1569 molecules from eight activity classes. Each activity class is represented by a typical QSAR set, featuring variations of one or a few central scaffolds and including both actives ($pIC_{50} > 6$) and inactives in roughly equal proportions. The actual compilation of 1569 compounds has been realized by standardizing[30] the structures of molecules from the cited sources, then merging the sets and discarding duplicate compounds with conflicting activity data (associated activity values for a same target differing by more than one pIC_{50} log). In the absence of experimental data about the affinity of a compound m with respect to a target t, inactivity was assumed and $pIC_{50}(m,t)$ set to 3.5 in order to fill up the structure−activity profile matrix. Under this assumption, activity dissimilarity scores $\Lambda(M,m)$ were calculated according to Appendix A, with the conversion function ψ in equation A6 modified so as to return 1.0 only if its argument exceeds 12.5% of the number of targets in the profile (that is, one difference with respect to eight targets—the 5% threshold used with the much larger BioPrint profile makes no sense when $N_{targets} = 8$). With these specifications, an active compound M appears as equally distanced—at $\Lambda(M,m) = 1$—from any confirmed inactive of its own class, as well as from all of the molecules belonging to different classes. $\Lambda(M,m) = 0$ only if m and M are both actives within the same class. An inactive is set at $\Lambda(M,m) = 0.1$ from any other inactive, within its own series or not, but such pairs were consistently discarded, like in the BioPrint study case.

In the comparative NB studies, the experimental activity dissimilarity $\Lambda(M,m)$ is confronted to various calculated molecular dissimilarity scores $\Sigma(M,m)$. The purpose of such a benchmark is assessing in how far molecules (m,M) that are predicted to be neighbors in a given "structure space"—low $\Sigma(M,m)$—are systematically found to also be neighbors in "activity space"—low $\Lambda(M,m)$. The statistical formalism used to quantitatively evaluate NB is briefly revisited in Appendix B. NB can be graphically assessed by plotting the optimality criterion Ω against the consistency χ at various structural similarity thresholds s. For simplicity, the plots were truncated at $\chi = 0.4$—displaying only the high-consistency range. Therefore, the characteristic U shape of $\Omega-\chi$ plots[16] may not always be apparent, but this is of little relevance for the discussion: the rule of thumb for the interpretation of the obtained graphs is that low Ω at high χ signals good neighborhood behavior.

2.4.1. Benchmarked Descriptors and Metrics. The NB of the 2D-FPT has been compared to the ones of different two-point pharmacophore descriptors, including fuzzy bipolar pharmacophore autocorrellograms (FBPA),[9] a 3D descriptor, and ChemAxon's topological fuzzy pharmacophore fingerprints.[15] The latter were calculated using both the recommended standard configuration (PF) and employing the "-R/--ignore-rotamers" (PFR) option of the ChemAxon descriptor generation tool. This option suppresses the default hypothesis according to which more fuzziness is applied when generating descriptor elements corresponding to more distanced atom pairs, as these have more options to experience important relative movements in the real molecule subjected to thermal agitation. ChemAxon's Chemical Fingerprints[33] (CF) were also used for benchmarking, as a representative of fragment-based fingerprints. To explicitly monitor the benefit of the novel-type flagging technique used with 2D-FPT, an alternative FPT relying on the same rule-based procedures used with PF/PFR has been generated. Molecular dissimilarity scores based on third-party descriptors were calculated according to the metrics best adapted for each—the Tanimoto score with ChemAxon's PF and CF and the fuzzy FBPA metric, respectively. XML setup files used for PF and CF descriptor and dissimilarity score calculations (PF.xml and CF.xml respectively) are included in the Supporting Information.

2.4.2. Virtual Screening of Seeded Compound Collections. A set of 50 000 random compounds—excluding organometallic derivatives and compounds of molecular mass above 1000 g/mol—from the MayBridge[34] vendor catalog were used as a reference chemical space to which molecules of known activities were added: (1) 194 compounds with reported c-Met tyrosine kinase activities from the literature,[35−37] including 72 actives with $IC_{50} \leq 10^{-7}$ M and (2) 460 molecules that were tested against the dopamine D2 receptor[38] (219 with $IC_{50} \leq 10^{-7}$ M). Both sets covered activity ranges from nanomolar to low millimolar values of IC_{50}. For each, the pharmacophorically most diverse three representatives were picked out of the respective subsets of very potent inhibitors ($IC_{50} < 10^{-8}$ M) and used as lead compounds for virtual screening according to both the 2D-FPT (FPT-2) and the PF-based Tanimoto metrics. The numbers of both confirmed actives ($IC_{50} \leq 10^{-7}$ M) *and* confirmed inactives ($IC_{50} > 10^{-7}$ M) were monitored within the sets of 200 nearest neighbors from the seeded chemical space found by each metric around each of these six leads.

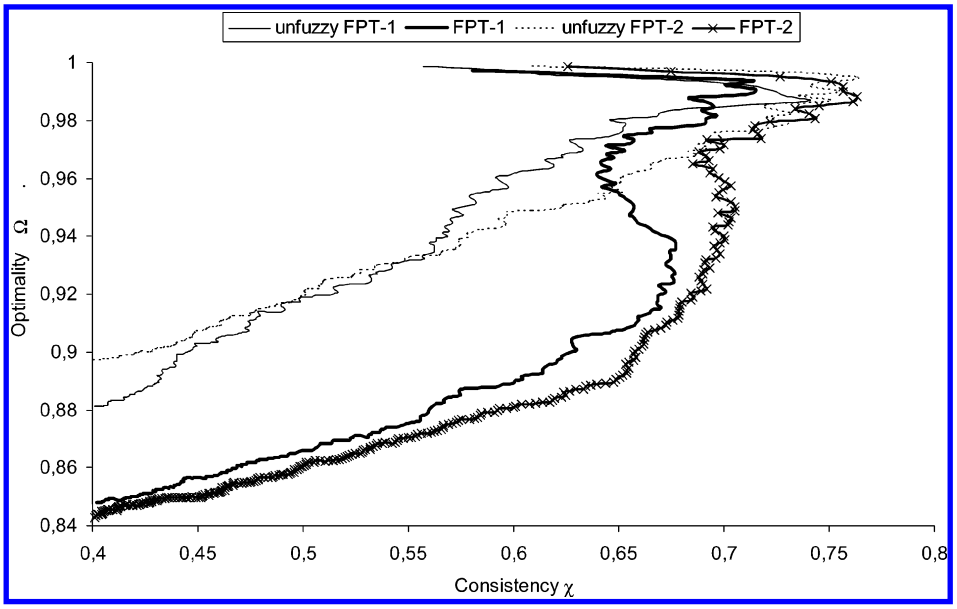

Figure 4. Comparative $\Omega-\chi$ plots illustrating the improvement of NB upon enabling the fuzzy mapping of atom triplets onto basis triplets, for both fingerprint versions FPT-1 and FPT-2, using the 2D-FPT specific similarity score Σ^{FPT} (BioPrint data set).

3. RESULTS AND DISCUSSIONS

3.1. The Importance of Fuzzy Mapping. To explicitly quantify the importance of fuzzy atom triplet mapping onto the basis triangles, the fuzziness factors ρ of considered FPT versions from Table 1 were temporarily set to 5.0 in order to generate comparative $\Omega-\chi$ plots for the corresponding unfuzzy fingerprints (the specific Σ^{FPT} score was used in all cases). At such high values of ρ, atom triplets will strictly highlight basis triplets of identical edge lengths. They will fail to highlight any basis triplet if the given combination of interatomic separations is not represented in the basis set. The corresponding curves in Figure 4 differ very little at their origins, where the selected pairs mostly include analogues with the same molecular scaffold and therefore are made of almost exactly the same atom triplets. However, the use of fuzzy logics is essential for extending the selection beyond these very first close analogues, to encompass pairs of compounds for which the underlying pharmacophore pattern similarity is not necessarily backed by a skeleton similarity. With fuzzy logics, many more activity-related compound pairs can be successfully picked without allowing pairs of different activities to enter the selection. Ω is observing a significant decrease without a loss of consistency, which is not seen when fuzzy mapping is turned off.

3.2. Importance of the pK_a-Dependent Fingerprint Buildup Strategy. The introduction of pK_a-dependent pharmacophore-type weights is expected to significantly contribute to the chemical meaningfulness of FPT. For example, a rule-based "educated guess" typically used to recognize potentially ionized groups in organic compounds would rely on the axiom that aliphatic amines are protonated, for example, must be flagged as cations and donors. Accordingly, N-alkylpiperazine-containing organic compounds will be assumed to harbor a cation−cation pair (see example in Figure 3). However, at pH = 7, only one of the two nitrogens is likely to carry a proton, its charge preventing the second one to do so. The cation−cation pair hence only appears in a minority of molecules, and its weight in the overall pharmacophore pattern should be adjusted accordingly. Piperazine may in reality be closer related to cyclohexylamine or morpholine than the rule-based pharmacophore pattern matching would suggest. Of course, rules can be tentatively optimized to avoid these kind of pitfalls: for example, the ChemAxon default pharmacophore mapping rules do not include tertiary amines into the cation category. This makes sense in medicinal chemistry, where the majority of amino groups in drugs are tertiary. The undue hypothesis of polycation patterns in the pharmacophore motif may hence be avoided, though at the cost of failing to perceive the similarity between secondary and tertiary amines.

An accurate prediction of the ionization status of protonable groups is a prerequisite for the success of the herein advocated flagging strategy. The NB of the fingerprints relying on ChemAxon's pK_a prediction plug-in outperforms the strategy of rule-based protonation state setup (Figure 5). This is thus an indirect proof of the accuracy of the pK_a prediction tool, offering an accurate estimation of expected protonation states. The rules used to build the alternative 2D-FPT (all other setup parameters being equal to FPT-1 values) were ChemAxon's default rules, the same used to construct the PF two-point pharmacophore fingerprints. A total of 59 pairs of compounds with identical activity profiles, ranking among the top 1000 most similar according to the pK_a-based approach, would lose their top-ranking positions and regress by more that 10000 ranks in the ordered pair list according to the rule-based method. Conversely, 50 activity-related pairs are perceived as similar by the rule-based metric, but not by the pK_a-based scoring scheme. The significant differences appear with respect to the distribution of activity-unrelated compound pairs. A total of 14 "violators" of the pK_a-based scheme (pairs with $\Lambda = 1$ but nevertheless ranked among the top 1000) are correctly reranked among the structurally dissimilar by the rule-based procedure. By contrast, 100 of the rule-based violators are successfully eliminated by the pK_a-based approach. Four typical examples of these latter ones are given in Figure 6. The similarity of compound pair a is clearly overstated by

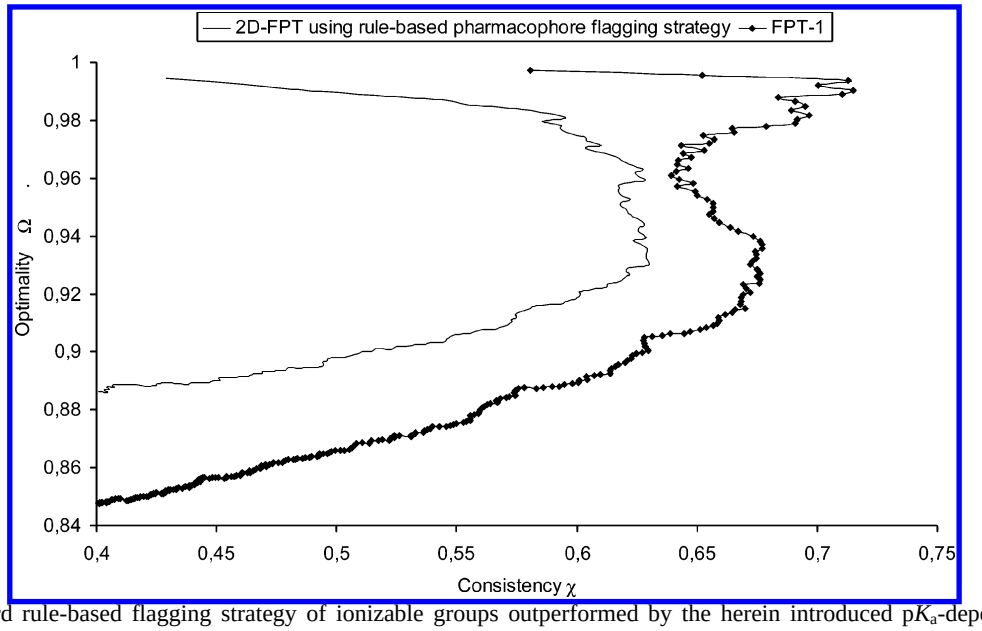

Figure 5. Standard rule-based flagging strategy of ionizable groups outperformed by the herein introduced pK_a-dependent fuzzy-type assignment procedure.

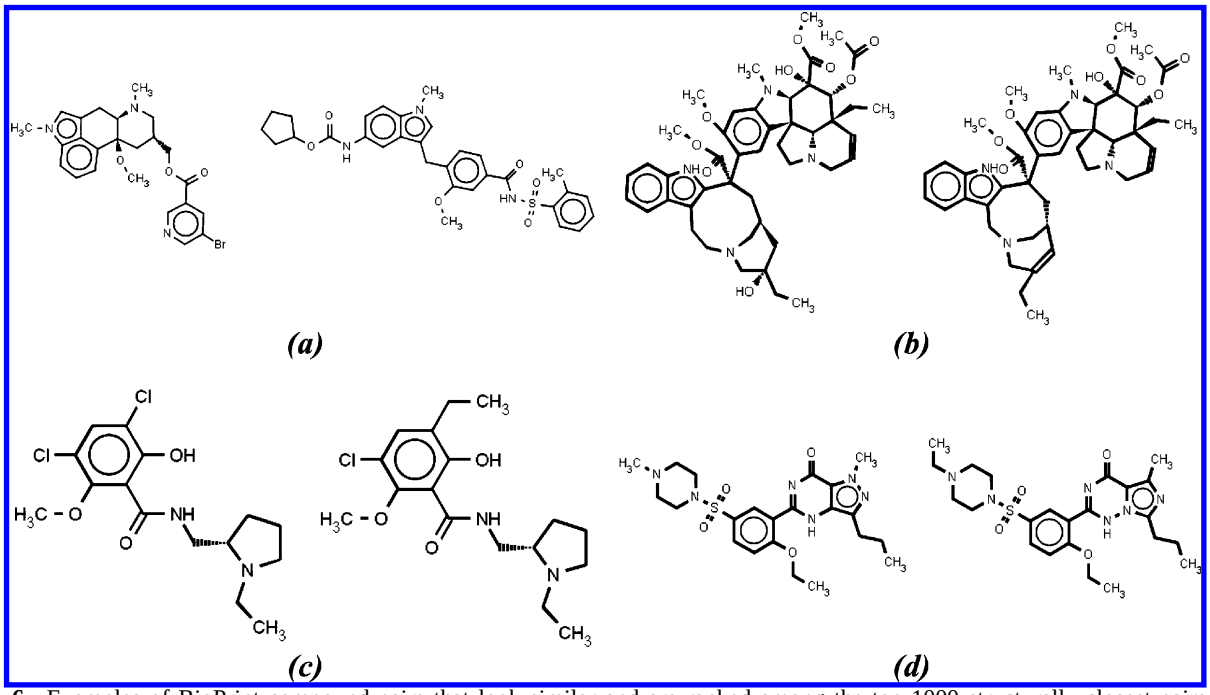

Figure 6. Examples of BioPrint compound pairs that look similar and are ranked among the top 1000 structurally closest pairs by the rule-based pharmacophore flagging scheme but, in reality, display radically different activity profiles and are correctly perceived as structurally different by the pK_a-based pharmacophore flagging scheme.

the rule-based scoring scheme, which regards both molecules as neutral species—acylsulfonamides are not declared as potential anions, and tertiary amines are not declared as cations in the ChemAxon default setup file pharma-frag.xml. Pair a stands thus for the numerous examples of activity-unrelated violator pairs that might have been avoided by redefining some of the flagging rules. In cases b, c, and d, however, pharmacophore dissimilarity cannot be accounted whatsoever by detailed flagging rule definitions: subtle substitution effects are seen to trigger relatively small pK_a shifts, but with dramatic impacts on the overall populations at proteolytic equilibrium. In compound pair c, the dissimilarity stems from the much more important ionization of the dichlorophenol compared to the monochlorophenol. While the left-hand compound mainly appears (according to the ChemAxon pK_a tool) under its zwitterionic form at pH = 7.4, the right-hand counterpart is predominantly positively charged. Even more dramatically, in example d, the addition of a simple methyl group enhances the protonation of the tertiary amine (70% cation at pH = 7.4 compared to 40% only in the left-hand molecule). Unless this effect is explicitly accounted for, a pharmacophore dissimilarity metric might never be able to explain the important activity differences observed upon the addition or deletion of a single hydrophobic center. Of course, the success of the approach relies on the precise pK_a estimation, or else the overestimated equilibrium population shifts that fortuitously explain observed activity differences might as well prevent the metric from recognizing the real pharmacophore similarity of activity-related pairs. As many com-

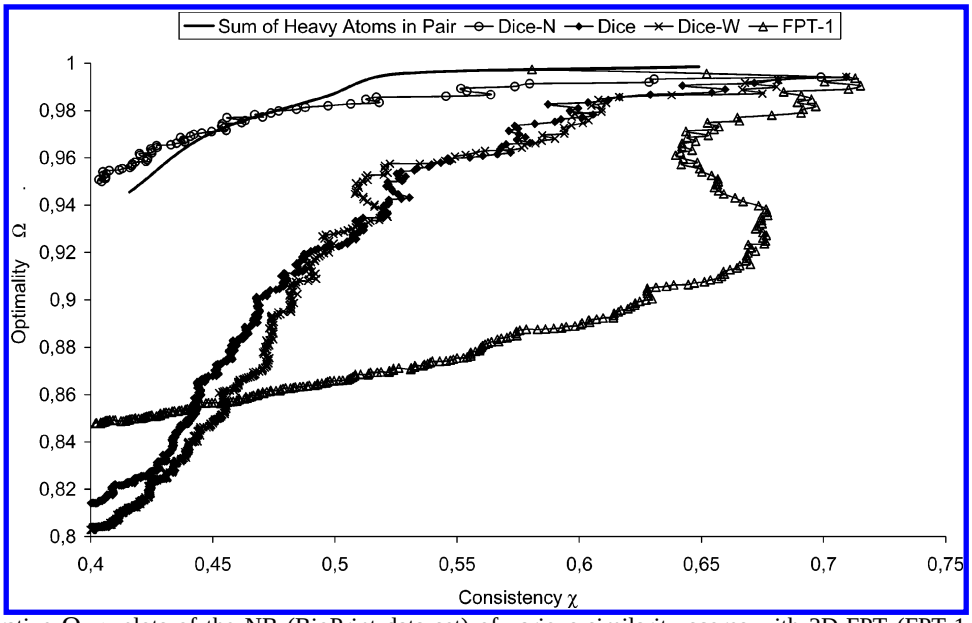

Figure 7. Comparative $\Omega-\chi$ plots of the NB (BioPrint data set) of various similarity scores with 2D-FPT (FPT-1 setup). Considered metrics are variants of the Dice formula: Σ^{Dice} ("Dice" in Figure legend), Σ_N^{Dice} ("Dice-N" in legend), and Σ_W^{Dice} ("Dice-W" in legend), as well as the 2D-FPT specific similarity score Σ^{FPT} ("FPT" in legend, eq 11).

pounds in this study are well-known drugs and reference molecules that are likely to have served for the pK_a tool calibration, further validation on the basis of original compound collections might be welcome. This notwithstanding, it can be concluded that one of the notorious limitations of pharmacophore-based similarity, the inability to explain activity shifts accompanying slight substitution pattern changes—a thorny issue raising fundamental questions about the validity of the neighborhood principle—might be successfully overcome in quite numerous cases of pK_a shift-related activity differences.

3.3. The Relative Performance of the Specific FPT Similarity Score. The NB of the various similarity scoring schemes using 2D-FPT (built according to setup 1 in Table 1) has been assessed, the results being shown in Figure 7.

The uppermost, solid curve represents the behavior of a fake dissimilarity score equaling the sum of heavy atoms in the molecule pair (*m*,*M*). It is nevertheless a well-shaped $\Omega-\chi$ plot, proving that activity-relatedness is statistically more likely to occur within subsets of small molecule pairs. This size effect is due to the fact that the smaller (~10 heavy atoms) of the employed molecules are unlikely to be strong binders to targets in the activity panel. Activity profiles of such compounds will be mostly empty, and their comparison returns low Λ scores (of about 0.1). Significant accumulation of such compound pairs at the top of the by-size sorted pair list ensures a significant consistency level of more than 60% within the top 20 lightest pairs (right-most point on the curve). Compound pairs with Λ scores of 0 (hitting common targets) are not contributing to these initial high consistency scores. The artifactual NB of size would have been even more marked if a bonus for binding to a same target would not have been included in Λ (results not shown).

Any rational pair selection strategy must therefore do better than (e.g., lay below) the size-driven NB curve. This is, unsurprisingly, not the case for the Dice metric based on normalized descriptors, which is quite sensitive to the complexity of the pharmacophore patterns of molecules, and implicitly to molecular size (see Figure 8). Small molecules with few populated triplets run an artificially high chance to be ranked as very similar: at $D_k(m) = 0$, $\mathscr{D}_k(m)$ simply relates to $-\alpha_k(m)$. The lesser the number of populated triplets is, the closer to the vector of average triplet populations—and the more correlated—the vectors $\mathscr{D}_k(m)$ and $\mathscr{D}_k(M)$ will be.

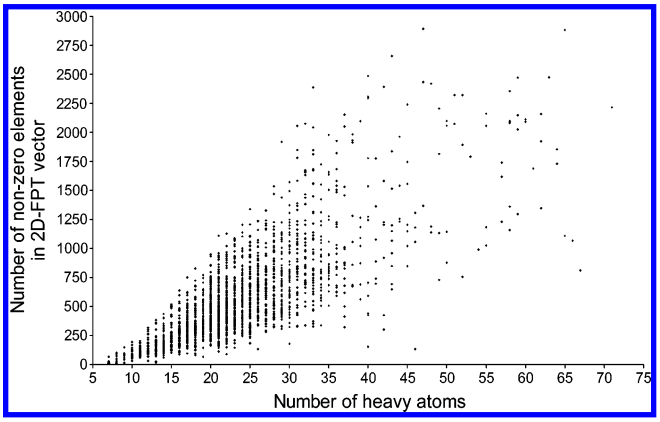

Figure 8. Dependence of the number of populated triplets on molecule size.

The same effect can be noticed with Euclidean scores (not shown). When $D_k(m) > 0$ and $D_k(M) > 0$, the chances that $D_k(m) = D_k(M)$ are quite small. Molecule pairs with a significant common set of populated basis triplets will, because of the summation of small but numerous residuals $\delta_k(m,M)$, typically end up at a higher Euclidean dissimilarity than pairs of small molecules with $D_k(m) = D_k(M) = 0$ for an overwhelming majority of triplets *k*. For example, the introduction of a methyl group in a large molecule *M* would trigger changes in the population levels of many more triplets *k* than the introduction of the same $-CH_3$ in a small compound *m*. Therefore, the calculated Euclidean distance score for a methyl/normethyl compound pair would counterintuitively increase with molecule size.

The Dice scores with or without the weighting of rare pharmacophore triplets can be successfully used to compare brute 2D-FPT, although they are clearly outperformed by the spe-

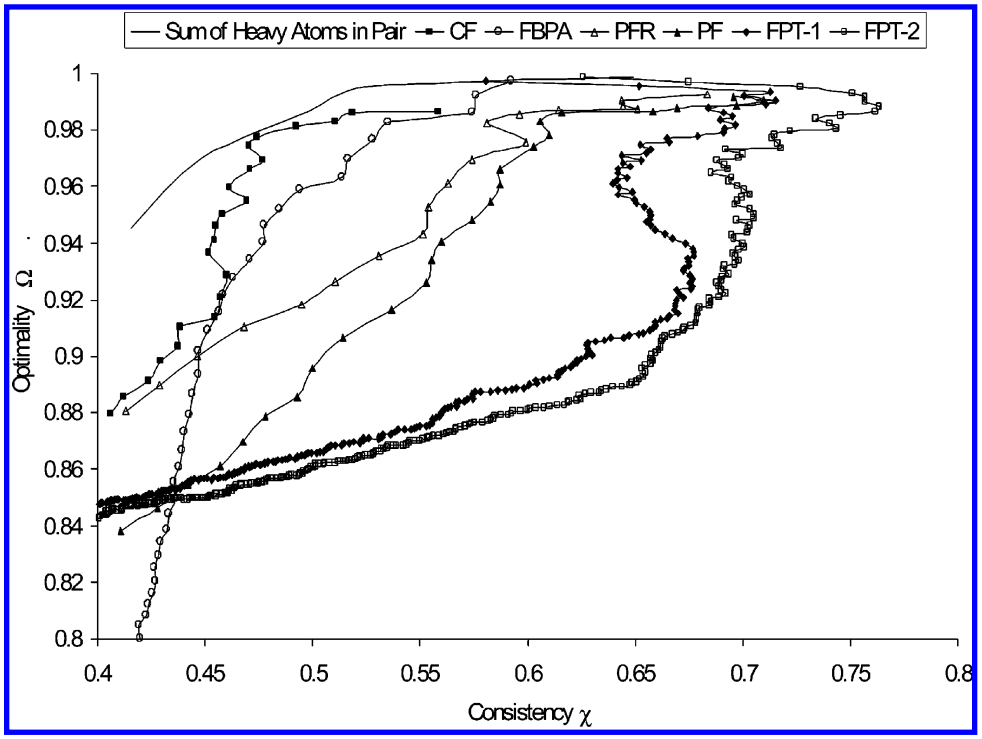

Figure 9. Comparative $\Omega-\chi$ plots illustrating the NB of 2D-FPT (both setups, using the specific Σ^{FPT}) with respect to other descriptors and associated metrics (BioPrint data set).

cific FPT metric. In the Dice formula using 2D-FPT without any further norming or rescaling, the main criterion controlling dissimilarity is the number of common nonzero descriptor elements, as these are the only contributing to the sum of $D_k(m)D_k(M)$. Any molecules having no nonzero D_k values in common will be considered 100% dissimilar. However, two large molecules with less-sparse 2D-FPT vectors are much more likely to achieve some fortuitous overlap of their fingerprints than two small molecules. Even if an overwhelming number of exclusively populated D_k's exist, having $D_k(m)D_k(M) > 0$ for at least one k automatically ensures that such a molecule pair will nevertheless be ranked as more similar than any pair of small molecules with no shared triplets at all.

A general problem in molecular similarity scoring—be it molecular descriptor comparison or activity profile matching—appears to be the appropriate handling of the uncertain "null" situations describing the absence of an item (pharmacophore triplet, affinity with respect to a target) from both molecules. On one hand, it may be argued that the two compounds share the absence of an item, which makes them more similar. On the other, sharing the presence is clearly a stronger proof of similarity than sharing the absence, and the question is, how much stronger? Also, how can shared presence and shared absence be counterbalanced against the number of differences observed in the fingerprint, to achieve a meaningful final score?

The excellent NB of the dedicated dissimilarity score defined in eq 11 suggests an appropriate balancing of the contributions for the specific case of 2D-FPT. The dissimilarity score Σ^{FPT} is seen to increase in response to (a) observed differences between population levels of exclusively populated basis triplets and (b) observed differences between population levels of shared triplets. The coefficient of the latter is more important—however, it is the former that statistically contributes the most to the dissimilarity scores because situation a occurs more often.

Furthermore, Σ^{FPT} decreases as the total fraction of shared triplets increases—with the effect that $\Sigma^{FPT}(M,M)$ will decrease with molecule size: larger molecules (with richer pharmacophore patterns, strictly speaking) are "more similar to themselves" than smaller ones. This is not paradoxical if we give up considering Σ^{FPT} as a similarity metric, but consider it as a substitution score not unlike the ones used for sequence matching in bioinformatics:[39] the conservation of the rarer, larger, and functionally specific tryptophane in two sequences is seen as more significant and given a larger bonus than the conservation of a ubiquitous alanine.

3.4. Neighborhood Behavior of 2D-FPT, Compared to the Other Descriptors. Figure 9 compares the NB of 2D-FPT using Σ^{FPT} to that of other descriptor spaces and metrics. In can be seen that CF chemical fingerprints, which are tailored for (sub)structure recognition, do not fare better than size-driven artifacts. All of the pharmacophore descriptors, however, perform better than cumulated size. At low selection sizes (large Ω), PF outperform the fuzzy three-dimensional FBPA. However, although the latter metric tends to be too permissive (allowing compound pairs with different activities among its top-scoring pairs), it is nevertheless able to retrieve a maximum of existing activity-related pairs while maintaining a reasonable consistency of the selection (deep Ω minimum). Interestingly, applying higher fuzziness levels for more distant pharmacophore point pairs (default behavior in ChemAxon's pharmacophore fingerprint calculator) seems counterproductive in this benchmarking test: better results (PFR) are obtained when this approach is switched off.

It is remarkable that the 2D-FPT curves and notably the one obtained with the smaller triangle basis set (FPT-1) originate at relatively low consistency levels. As the selection is extended, the fraction of activity-related among the co-

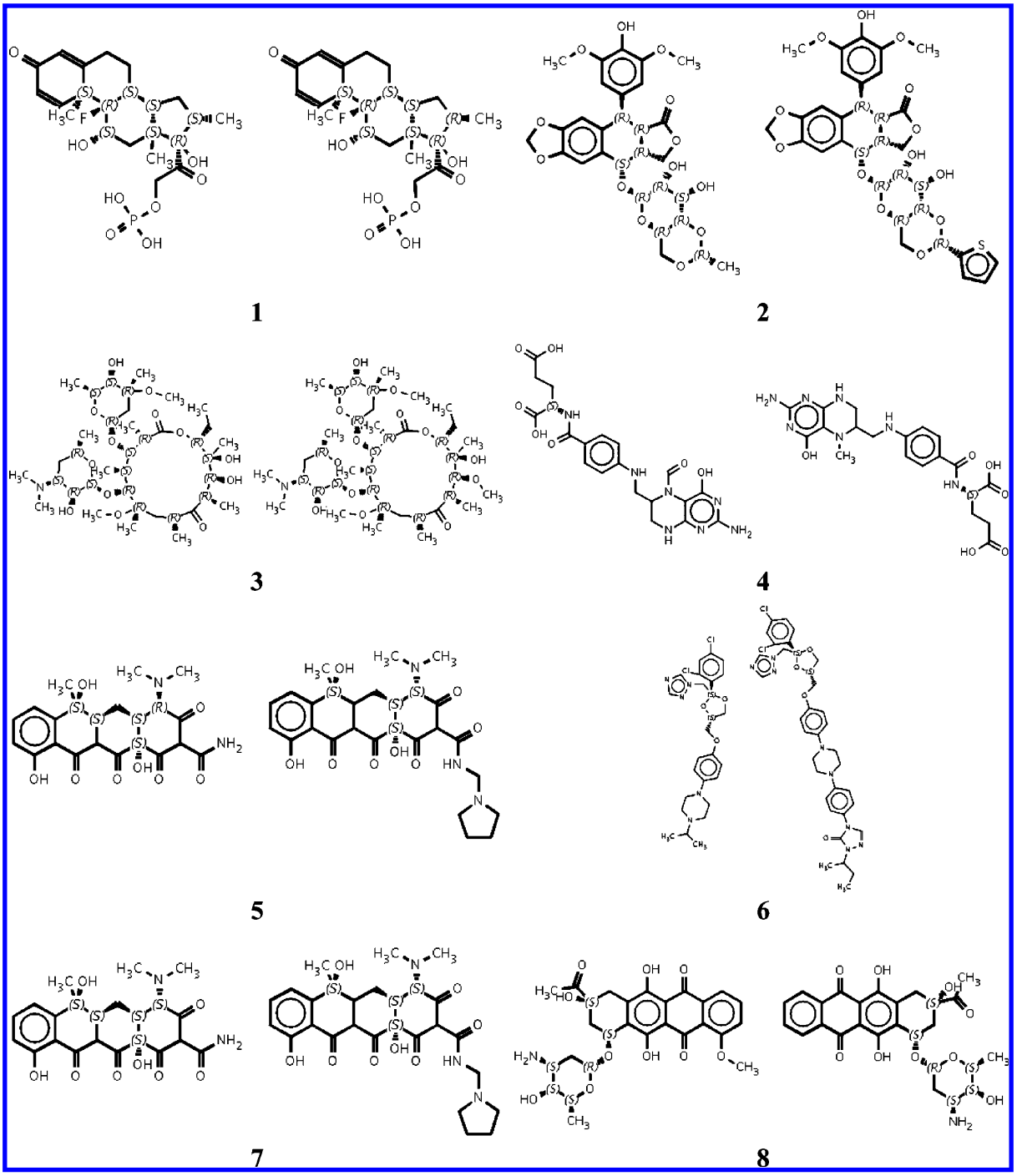

Figure 10. The eight pairs with highly dissimilar activity profiles found among the 50 most similar pairs according to 2D-FPT similarity scoring (FPT-1 setup).

opted pairs becomes much larger than that seen within the first top scorers. At high consistency values (0.5−0.7), significantly more activity-related compound pairs are retrieved by 2D-FPT than by any of the other scoring schemes.

Such behavior might be expected with topological descriptors such as 2D-FPT, because pairs of diastereomers (*M*,*M**) score as much as a compound scores with respect to itself: $\Sigma^{FPT}(M,M^*) = \Sigma^{FPT}(M,M)$. The hypothesis that the initial inconsistency is due to the accumulation of activity-unrelated diastereomer and enantiomer pairs at the top of the similarity-sorted pair list must however be discarded. PFs, for example, are also topological distance-based and use a classical Tanimoto-based scoring scheme, so that $\Sigma^{PF}(M,M^*) = \Sigma^{PF}(M,M) = 0$ and diastereomers are always top scorers. However, the very high consistency of the right-most data point of the PFR curve proves that the 105 compound pairs with $0.00 \leq \Sigma^{PFR} < 0.01$, the herein included pairs of diastereomers, are not overwhelmingly activity-unrelated.

Actually, Σ^{FPT} no longer guarantees diastereomer pairs to rank among top scorers. $\Sigma^{FPT}(M,M) > 0$ decreases with the complexity of *M*, and pairs of slightly differently substituted analogues (*M*,*M*′) sharing a highly complex pharmacophore pattern may score better than pairs of less complex molecules (*m*,*m**) with identical fingerprints. Although $\Pi^{+-}(m,m^*) = \Pi^{++}(m,m^*) = 0$, having $f^{++}(M,M') > f^{++}(m,m^*)$ may eventually let the pair of close analogues score lower Σ^{FPT} values than the pair of diastereomers. The consistency inversion observed with 2D-FPT is, unexpectedly, not a

consequence of ignoring stereochemical information but actually stems from pairs of closely related analogues of very high molecular complexity. Among the best-ranked 100 pairs of compounds according to the FPT-1 setup of 2D-FPT scoring scheme, 66 have $\Lambda > 0.2$, 30 have $\Lambda > 0.5$, and 15 have $\Lambda > 0.8$. By contrast, in the pair subset ranked from 100 to 200, there are only 21 at $\Lambda > 0.2$, 13 at $\Lambda > 0.5$, and 6 at $\Lambda > 0.8$, for example, less than half as many NB violators than in the first 100 pairs. Violator pairs are, beyond doubt, chemically similar (to the point that finding the difference when looking at the structures is not always easy; Figure 10, except for examples 6 and 7, where substitution differences involve the introduction of a heterocycle and a cationic group, respectively). It is difficult to "blame" the 2D-FPT metric for having selected them. However, such "me-too" close analogue pairs are always among the top scorers of all of the similarity metrics, including PF and FBPA, but they are not seen to distort either of the herein-obtained NB curves. It can be safely assumed that, statistically speaking, closely related analogues differing in terms of either the stereochemistry or minor substituent changes tend to have similar biological activities, the exceptions to this rule being relatively rare (but widely publicized[40]). The previous section showed that 2D-FPTs are able to successfully explain some of these "activity cliffs" on the basis of predicted pK_a shifts. It appears however that they also tend to specifically pinpoint another subset of activity cliffs, pertaining to a specific series of close analogues that tend to score better than the ubiquitous activity-related "me-too" pairs. The 2D-FPT score-driven ranking of the BioPrint compound pairs evidenced a top-ranking subset of highly complex and very similar compound pairs with an increased propensity to form activity cliffs versus that of "typical me-too" pairs. At this point, it is however unclear whether this finding may be generalized to suggest that more-complex molecules are more likely to have their biological properties strongly affected by small chemical alterations. This is certainly not true with respect to overall physicochemical properties: methylation of a macrocycle like the third example in Figure 10 would hardly affect properties such as the octanol−water partition coefficient; by contrast, the methylation of methanol leads to the physicochemically different dimethyl ether. It is however important to remark that most of the compound pairs in Figure 10 are natural compounds or derivatives of natural compounds, optimized by Darwinian evolution to be perfect binders to a given target. From this viewpoint, it seems understandable that any small chemical alteration on the natural ligands may have dramatic changes in affinity. Synthetic drug molecules appear to be much less well-adapted to their targets and therefore, statistically spoken, much more tolerant to structural variations. 2D-FPT might provide a very useful metric for molecular complexity and implicit lead-likeness or drug-likeness—issues[41] that will be explored elsewhere.

The second parametrization attempt FPT-2 turned out to be more successful, but although the subsets of top scorers are significantly less marked by the accumulation of activity-unrelated pairs, the previously discussed consistency inversion does not vanish. Its better performance can be mainly ascribed to the shift of the minimal and maximal topological edge lengths from 2 to 4 and from 12 to 15, respectively. Monitoring triplets including directly bound, geminal or vicinal atoms does not enhance NB. This makes sense: binding pharmacophores typically include anchoring points from different parts of the ligand. Triplets involving, for example, both the carbonyl =O and the hydroxyl −OH in a hydroxamic acid RC(=O)−NH−OH are not accounted for in any of the versions—a specific fitting for metal enzyme inhibitors might prove necessary under these circumstances. The coverage of long-range molecular triplets seems to be very important: it also seems a good idea to extend the size of actually considered molecular triplets by $e = 2$ more bonds beyond E_{max}.

The initial choice of a grid of basis triplets having a mesh size (edge increment E_{step}) of 2 appears to be the good compromise. An E_{step} of 3 would have reduced the basis set size dramatically—however, molecular triangles with edge size values not appearing in the basis triplets would have been at risk to fall through the grid meshes, in failing to match any one of the basis triplets. Successful 2D-FPT setups with $E_{step} = 3$ may exist but must be actively searched for in the setup parameter space. $E_{step} = 1$ would, on the contrary, engender much larger grid sizes, thus causing significantly more practical problems with the handling of the resulting descriptors. Given the excellent behavior at $E_{step} = 2$, potential benefits of denser basis sets are unlikely to outweigh the descriptor size-related inconveniences.

A first key observation in Figure 11, monitoring the NB of various metrics with respect to the public data set obtained by merging eight independent QSAR series, is the much lower Ω values compared to what had been seen within the BioPrint set. Unsurprisingly, detecting structurally similar pairs of related activities is a much harder problem within the diverse set of drugs than within an artificially constructed set of series of analogues around a limited number of scaffolds. In this latter case, a simple discrimination between structural families—telling benzodiazepine-like chemotypes apart from acetycholine-like ligands and so forth—is sufficient to ensure significant NB. There are, for example, 65 active and 47 inactive ACE binders in the set; for example, $65/1569 = 4.14\%$ of ACE actives in the entire set. Any metric that would consistently score lower dissimilarity between any two ACE set members than between an ACE and a non-ACE compound pair effectively discriminates between the ACE set and the rest of compounds. Within the ACE set, the rate of actives is however $65/112 = 58\%$, which represents a $58/4.14 = 14$-fold enrichment in actives. Under these circumstances, dissimilarity scoring based on chemical fingerprints does display a significant NB, in sharp contrast to the observations made on the BioPrint set. The discrimination between the various chemical families that make up the public data set is readily achievable by all three metrics monitored in Figure 11: all of them avoided ranking any of the pairs of compounds from different series within the top 550 pairs corresponding to the checkpoints highlighted on the plots. All NB violators—in the sense of $\Lambda(m,M) > 0.5$—encountered at these checkpoints are intraseries activity cliffs regrouping an active and a structurally very close inactive. Within the top 550 pairs selected by the CF metric, the 128 observed NB violation instances break down into 15 ACE, 27 AchE, 5 BzR, 20 Cox2, 43 DHFR, and 18 THR compound pairs. Pharmacophore-based metrics should go beyond activity class recognition and successfully tell apart actives and inactives on the basis of a common scaffold. This

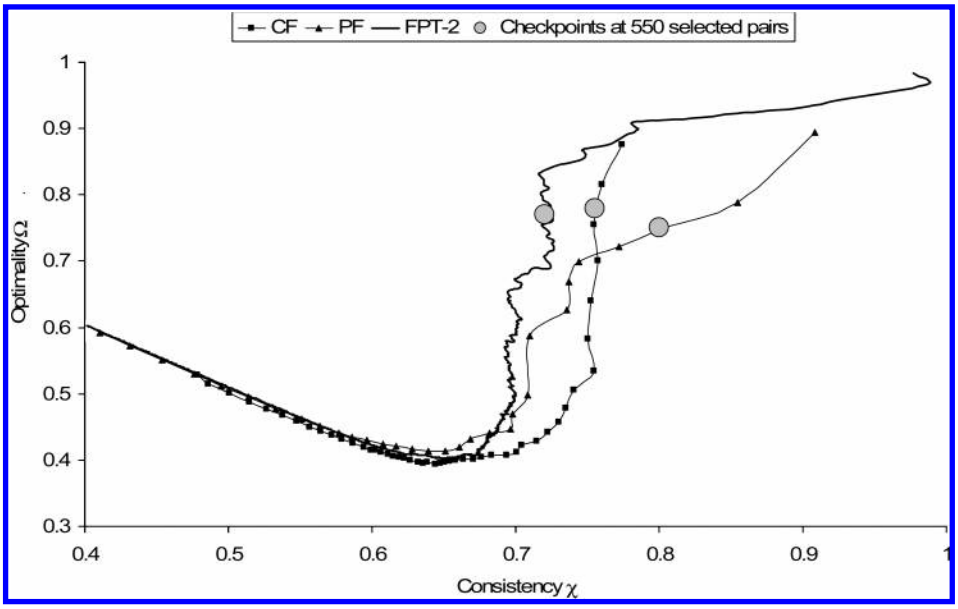

Figure 11. Comparative $\Omega-\chi$ plots illustrating the NB of 2D-FPT (setup FPT-2, using Σ^{FPT}) with respect to ChemAxon chemical and pharmacophore descriptors and associated metrics (public data set regrouping 1569 compounds from eight QSAR series).

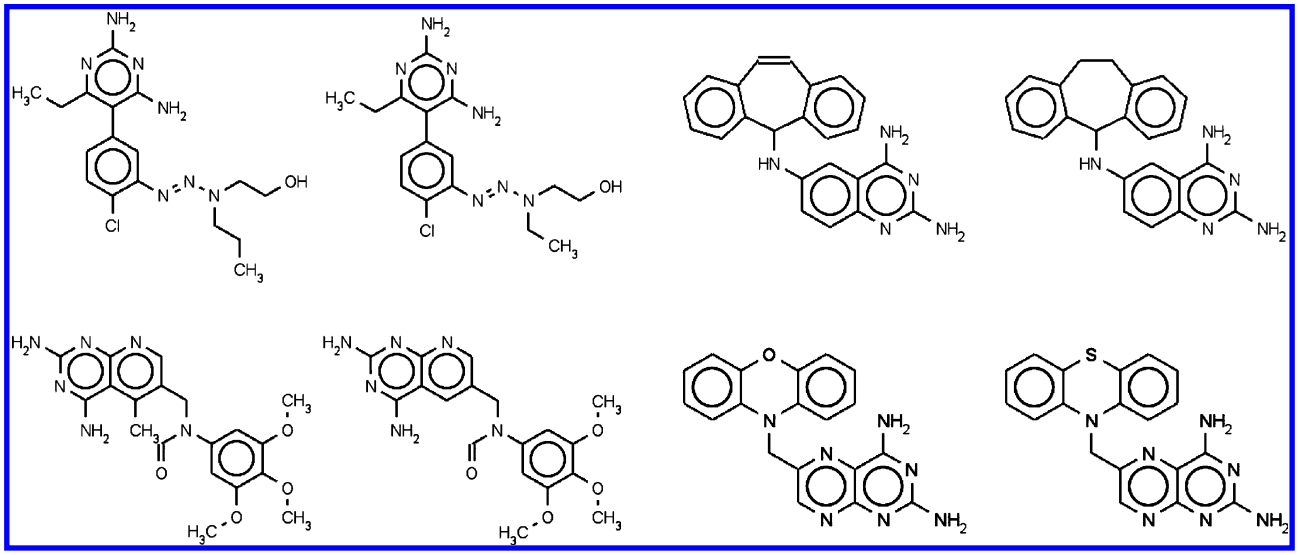

Figure 12. Typical "activity cliffs" of dihydrofolate reductase—very similar compound pairs with significantly differing DHFR activities ($\Lambda > 0.5$). Such compound pairs are consistently perceived as similar by all metrics—however, only the Σ^{FPT} formalism ranks these relatively complex compound pairs among the top 550.

is indeed observed with both PF and FPT metrics: both of these and particularly the latter reach out into higher consistency domains, not accessible to the CF approach. Unlike in the BioPrint study case, PF-driven NB reaches relatively better optimality scores at a same consistency or relatively higher consistencies at the same selection size (0.8 instead of 0.7 for the top 550 selected pairs, see checkpoints). An analysis of NB violators reveals that PF retrieved 92 such pairs within the top 550: 7ACE, 4 AchE, 3 BzR, 59 Cox2, and 19 DHFR, whereas FPT retrieved 138: 5 ACE, 48 Cox2, 83 DHFR, and 2 THR. The FPT approach thus experiences a sharp decrease of its NB criteria because of a local accumulation of DHFR activity cliffs, some typical examples of which are depicted in Figure 12. These are clearly structurally highly related compounds scoring very low dissimilarity values within both FPT and PF formalisms. However, only the former score includes a bonus for pharmacophore complexity, or it can be seen that DHFR ligands are among the most complex compounds in this set. DHFR pairs are therefore relatively better ranked than other intraset pairs when using FPT. Unfortunately, DHFR appears to display a rugged structure−activity landscape ridden by activity cliffs that cannot be conveniently explained by any of the herein explored metrics. This may be an illustration—but still no definite proof—of the possible correlation between ligand complexity and the propensity for activity cliffs, previously cited as an envisageable explanation for the observed consistency inversion of the FPT metric within the BioPrint set.

3.5. Virtual Screening Results of Seeded Compound Collections. Such simulations directly address the ability of the metrics to discover actives from databases but are less well-suited for rigorous benchmarking than the general NB analysis reported previously, insofar as the following are concerned:

- While a retrieval of a maximum of hidden actives among the top neighbors of each lead compound is desirable, it is not clear how many of the hidden actives are genuinely

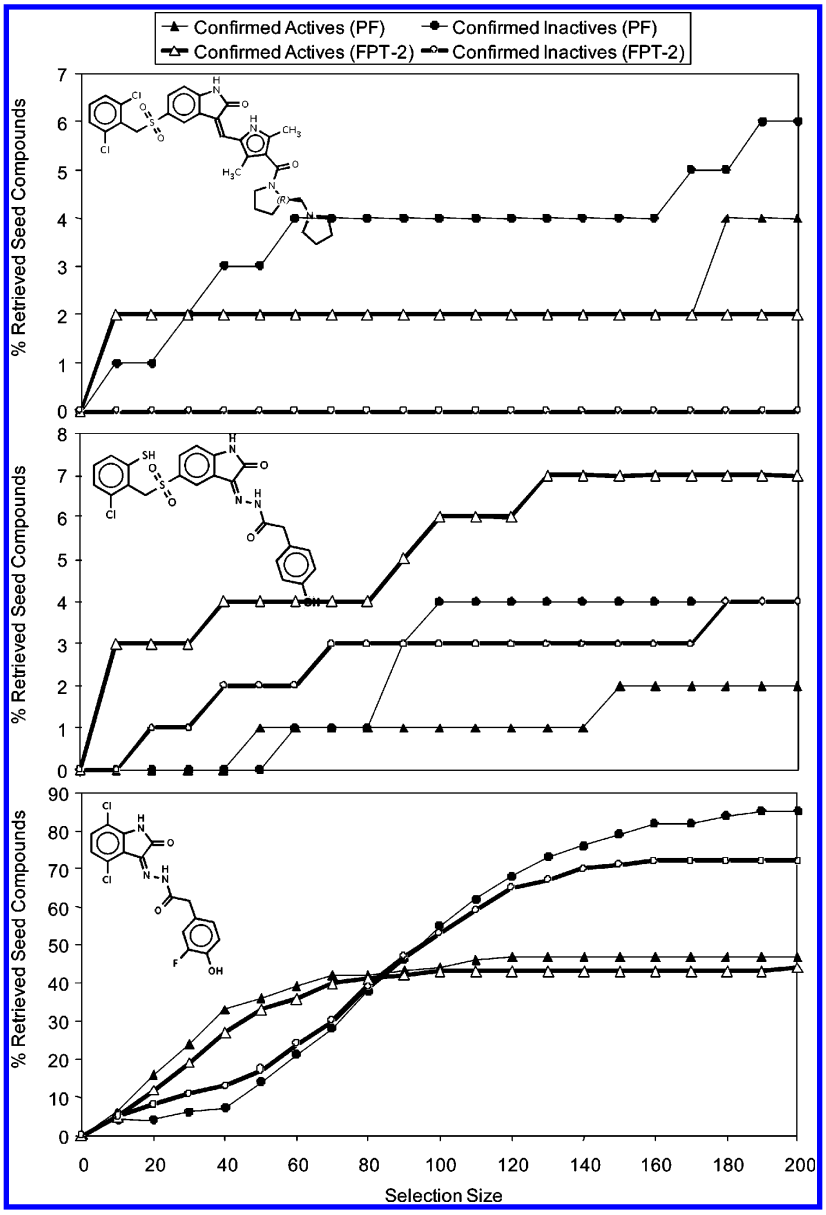

Figure 13. Results of virtual screening, probing each of the shown references against the MayBridge collection, seeded with compounds of known c-Met affinity (including actives with $pIC_{50} \geq 7$). Plots report the number of known actives and known inactives within subsets of nearest neighbors (subset size on the *x* axis) retrieved by the 2D-FPT (FPT-2 setup) and PF metrics, respectively.

similar to the lead and therefore eligible to be a virtual hit. Similarity to an active lead may be a sufficient but is clearly not a necessary condition. Unlike in virtual screening approaches based on QSAR or docking scores, successful similarity scoring is not expected to systematically score all of the actual active "ligands" better than the inactive "decoys"—if the set to be screened includes actives that are genuinely dissimilar to the reference, this subset of ligands might actually systematically score worse than decoys. The distributions of active ligands with respect to their similarity scores might actually be bi- or multimodal, complicating even more the statistical assessment of its robustness.[42] The selection criterion being the match of overall pharmacophore patterns—including those parts in which variability is not detrimental to binding—a search around a single lead may be too narrow.[43] In the present work, searches around single leads were performed with two different metrics (FPT and PF) and will be discussed in terms of relative retrieval rates.

- The key uncertainty in exploiting these results is the unknown activity status of the compounds from the bulk collection. The total number of actives present within the top neighbors is unknown, unless those compounds are ordered and tested against the target under study. Therefore, this study used both known actives and inactives for seeding. Selective enrichment in known actives, all while keeping the known inactives (often closely related analogues from the same series) out of the top neighbor set, is a strong indication of an increased probability to discover real actives among the hits from the bulk collection.

In the c-Met tyrosine kinase study case, the first two out of three lead compounds appear to be located at the rims of the cluster of the literature compounds of known activities. Both the PF and 2D-FPT-based metrics agree on the fact that the first lead (top plot in Figure 13) appears to have only two other known actives in its immediate neighborhood, with PF finding two more within the (arbitrary) limit of 200

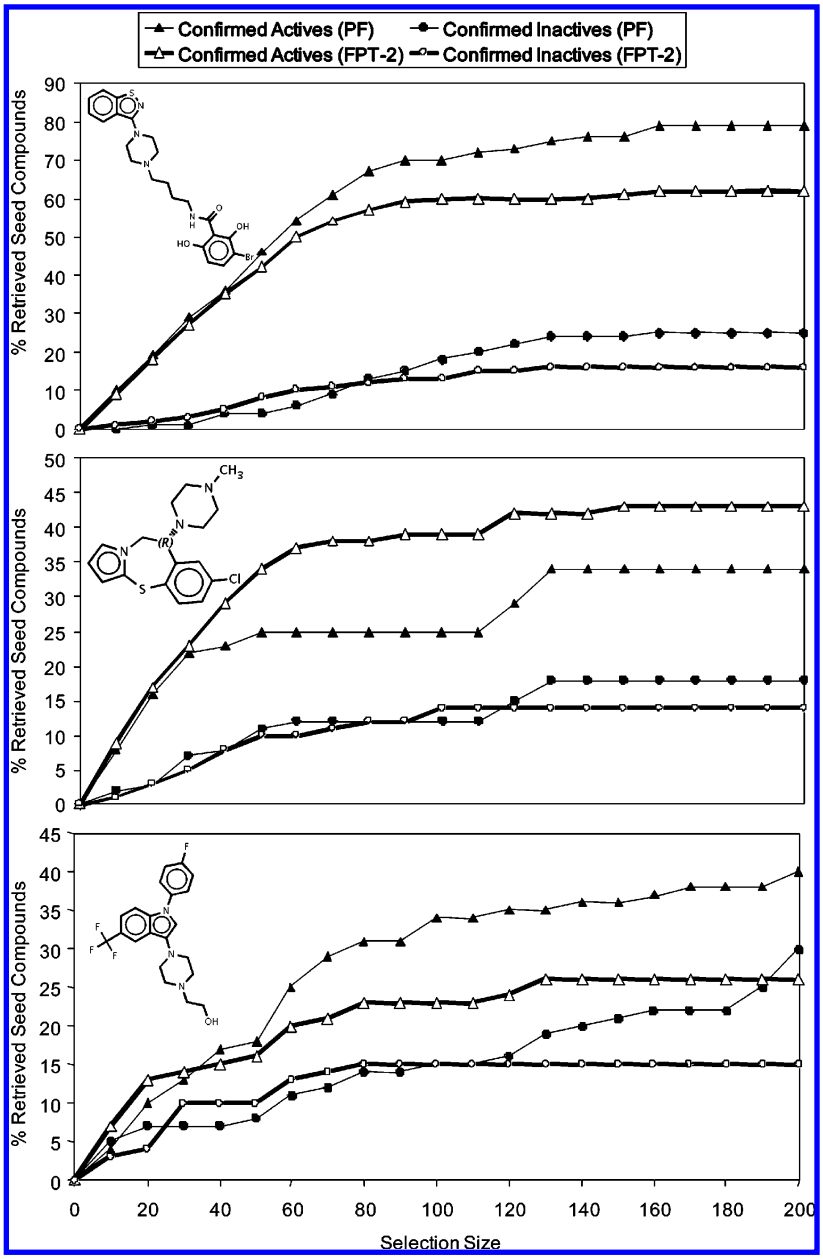

Figure 14. Virtual screening results for the D2 ligand study case (see legend of Figure 10 for details).

selected neighbors. However, the PF approach also co-opts four to six known inactives, which 2D-FPT successfully avoids. The results around the second lead compound are also clearly better with 2D-FPT, which recognizes roughly three times more known actives at basically equal numbers of co-opted inactives. The third c-Met lead appears, according to both metrics, to lay at the center of the c-Met compound cluster. Within the top 120 neighbors, retrieval levels closely match each other—with a slight advantage in favor of the PF approach, while at bigger selection sizes, the number of inactives co-opted by the PF significantly increases.

The study cases involving dopaminergic D2 compounds (Figure 14) showed that in all three situations lead molecules were well-surrounded by neighbors within the series. The first experiment may be considered a success of the PF approach—although it is still co-opting more inactives, it does better in known active retrieval by a clear margin. 2D-FPT clearly wins the second screening round, by simultaneously maximizing actives and minimizing co-opted inactives. The third experiment, eventually, is less clear-cut as the PF approach manages to retrieve more actives but only at the price of co-opting many more inactives than 2D-FPT.

Overall, the 2D-FPT-driven virtual screening appears to be more consistent—with respect to known actives and inactives—in the sense that higher active retrieval rates by PF are always accompanied by higher inactive retrieval rates as well. 2D-FPT systematically keeps the inactive retrieval rate equal or lower while nevertheless managing to improve the active retrieval rate in certain examples.

4. CONCLUSIONS

The insofar proven success of 2D-FPT-based similarity scoring compared to other fuzzy 2D and 3D pharmacophore descriptors is not surprising, as the three key innovations introduced here with respect to classical state-of-the-art descriptors and metrics are straightforward, chemically meaningful, and therefore expected to trigger improvements:

(1) The fuzzy mapping of molecular triplets on basis triplets is beneficial even in the context of topological distances (and assumed essential in a 3D context prone to conformational artifacts). It allows to accommodate the natural tolerance of receptors with respect to the number of bonds separating two binding groups and, from a practical point of view, allows a significant reduction of the descriptor dimension to a few thousands compared to >50 000 in binary fingerprints.

(2) The pK_a-dependent pharmacophore-type weighting scheme is able to correct many of the unavoidable inconsistencies that are introduced by rule-based flagging. Furthermore, local substituent swaps that, per se, would not translate to any significant pharmacophore pattern change as far as rule-based flagging is concerned may cause pK_a values to drift across the pH threshold and therefore trigger dramatic changes in the equilibrium population (and compound activity). Some of the "activity cliffs" in the structure−activity landscape of classical descriptor spaces are thus proven to be artifacts due to the failure of the latter to account for proteolytic equilibrium shifts. In the 2D-FPT space—for the first time, to our knowledge—this particular cause of landscape ruggedness has been successfully dealt with (insofar as the pK_a prediction tool is accurate, which appears to well be the case of the ChemAxon pK_a calculator employed in this work).

(3) The original similarity scoring scheme developed here recalls the simple truism that similarity due to the fact that a type is absent from both molecules is weaker than similarity due to the fact that both molecules contain the same type. As, in our hands, none of the classical scoring schemes managed to find the appropriate balance between contributions from shared, null, or exclusive triplets, such an optimal balance has been actively searched for—and found.

FPT as well as other pharmacophore-based descriptors have shown significant NB with respect to both diverse compound sets (BioPrint) and sets composed of several series of analogues. It is generally speaking much easier to demonstrate NB with respect to the latter situation, where simple discrimination between the main chemotypes at the basis of the various analogue series may suffice. The conclusions drawn on the basis of such studies may however be subject to different sources of bias due to relative size, chemical complexity, and other peculiarities of the considered analogue series. Mining for the underlying pharmacophore similarity in series with few representatives for each represented scaffold is much more challenging but successfully achieved by the FPT methodology. An interesting and recurring observation made in this work, requiring further investigation, is the possible correlation between the average pharmacophore complexity of the ligands of a target and its propensity for activity cliffs.

ACKNOWLEDGMENT

Special thanks to the ChemAxon (www.chemaxon.com) team, for allowing academics to freely use their software and for quick and effective hotline help. Sunset Molecular Inc. (http://sunsetmolecular.com/) and Tudor Oprea are acknowledged for providing the dopamine D2 data set. Nicole Dupont and Alexandre Barras (Institut de Biologie de Lille) are acknowledged for gathering the c-Met activity data from the literature. Thanks to Dr. Guy Lippens (University of Lille 1) for careful reading and important suggestions. ACCAMBA project members (http://accamba.imag.fr/) are acknowledged for encouraging this work.

APPENDIX A: THE ACTIVITY DISSIMILARITY SCORE

Similarity is an empirical concept, and there are no fundamental laws determining whether the activity profiles of two bioactive organic molecules are intrinsically similar or not. Like in the case of structural similarity, activity dissimilarity awaits for empirical definitions to be tried, validated, or discarded with respect to their usefulness in quantitative NB studies. Neighborhood behavior is necessarily a boot-strapping problem: its key assessment—that neighbors in a first (calculated) property space are likely to also be neighbors in a second (activity) property space—relies on two independent definitions of what "neighborhood" is supposed to mean in each one of the spaces.

For the above-mentioned reasons, this work postulates an activity dissimilarity score on the basis of plain medicinal chemistry common sense. Examples in which classical metrics (Euclidean, vector dot product, etc.) return counterintuitive dissimilarity measures will be discussed in order to highlight the need for a novel scoring scheme. Its implicit validation however comes from the fact that this definition of closeness in activity space respects the NB principle with respect to various molecular similarity metrics in structure space. In the following, the working hypotheses and parameters adopted in order to estimate the similarity of two activity profiles will be briefly outlined.

Profile similarity is determined by the behavior of a molecule pair (M,m) with respect to each target t. The target-specific response difference $\Delta_t(M,m)$ is defined as

$$\Delta_t(M,m) = \begin{cases} 0 & \text{if } |p_t(M) - p_t(m)| \le 0.5 \\ 1 & \text{if } |p_t(M) - p_t(m)| \ge 2.0 \\ \dfrac{|p_t(M) - p_t(m)| - 0.5}{1.5} & \text{otherwise} \end{cases} \quad \text{(A1)}$$

$\Delta_t(M,m)$ expresses a typical medicinal chemist's approach to activity comparison: two compounds with pIC_{50} values within 0.5 log units are said to have roughly the same activity; if however the pIC_{50} difference exceeds two log units, the molecules are beyond any doubt of different activity. In many situations, two log units is used as a landmark for selectivity: more than 2 orders of magnitude of affinity difference may not make any practical difference.

The activity index $\alpha_t(m)$ of a molecule m with respect to a target t is defined as a step function of the actual pIC_{50} value, such that compounds with affinities better than or equal to 1 μM count as active. A micromolar landmark for activity is widely used, especially in early stages of lead discovery.

$$\alpha_t(m) = \begin{cases} 0 \text{ if } p_t(m) < 6.0 \\ 1 \text{ otherwise} \end{cases} \quad \text{(A2)}$$

On the basis of definitions A1 and A2, $N_{diff}(m,M)$ and $f_{diff}(m,M)$—the index and respectivel fraction of significant differences in the profiles of molecules M and m are defined

as

$$N_{\text{diff}}(m,M) = \sum_{t=1}^{N_{\text{targets}}} [\alpha_t(m) + \alpha_t(M) - 2\alpha_t(m)\,\alpha_t(M)]\,\Delta_t(m,M)$$

$$f_{\text{diff}}(m,M) = \frac{N_{\text{diff}}(m,M)}{N_{\text{targets}}} \quad \text{(A3)}$$

In the N_{diff} index, the first factor plays the role of logical exclusive or it equals 1 if and only if either $\alpha_t(m) = 1$ or $\alpha_t(M) = 1$. If so, N_{diff} is incremented by the amount of the target-specific response difference $\Delta_t(M,m)$: a pair (M,m) of approximately micromolar affinities on opposite sides of the 1 μM threshold will not contribute. Intuitively, N_{diff} is a fuzzy counter of the obvious activity differences in the profile.

The index and respective fraction of similarities $N_{\text{sim}}(m,M)$ and $f_{\text{sim}}(m,M)$ observed in the activity profiles of the two molecules are defined as

$$N_{\text{sim}}(m,M) = \sum_{t=1}^{N_{\text{targets}}} \alpha_t(m)\,\alpha_t(M) \times [1 - \Delta_t(m,M)]$$

$$f_{\text{sim}}(m,M) = \frac{N_{\text{sim}}(m,M)}{N_{\text{targets}}} \quad \text{(A4)}$$

N_{sim} is the fuzzy counter of targets with respect to the two compounds having both strong $[\alpha_t(m) = \alpha_t(M) = 1]$ and similar $[\Delta_t(M,m) < 1]$ activities. Positive N_{sim} signals that the two compounds both interact with the same active site(s) and are therefore likely to include some common pharmacophore elements—insofar as most receptors tend to display a set of key interaction points that are always used in ligand binding, next to less important specific anchoring groups that form specific interactions with specific ligands. It is important to note that N_{diff} and N_{sim} do however not sum up to the total number N_{targets}. With respect to a pair of molecules, the set of targets making up the activity profile can be split into three domains: similarity, difference, and uncertainty, of sizes N_{sim}, N_{diff}, and $N_{\text{targets}} - N_{\text{diff}} - N_{\text{sim}}$, respectively. The uncertainty domain regroups targets for which molecules m and M display neither clear-cut different nor obviously similar behaviors. These include the (few) cases when compounds display significant potency differences despite both being active and the (ubiquitous) targets with respect to which m and M similarly fail to bind. A mutual lack of activity brings little information: molecules may be both inactive because of their similarity, or they may be each inactive in their own way.

The final activity dissimilarity score $\Lambda(m,M)$ associated with the activity profiles of molecules m and M is defined according to the following equation:

$$\Lambda(m,M) = \psi[f_{\text{diff}}(m,M) - \lambda \times f_{\text{sim}}(m,M)] \quad \text{(A5)}$$

with the conversion function $\psi(x)$ defined below:

$$\psi(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \le 0.05 \\ 0.1 + 18x & \text{if } 0 \le x < 0.05 \end{cases} \quad \text{(A6)}$$

In our opinion, this piecewise context-depending similarity scoring scheme returns a calculated profile activity score in agreement with medicinal chemistry and pharmaceutical know-how. Λ is a compromise between the sizes of the difference and similarity domains, with an empirical $\lambda = 5$ empirically chosen to emphasize the importance of observing actual similarities. The role of the conversion function $\psi(x)$ is to ensure the following:

- Only compound pairs sharing at least one significant (better than 1 μM) common hit in the profile may qualify to score top profile similarity (e.g., minimal $\Lambda = 0$), provided that the number of observed differences is low enough.
- If difference compensates for similarity, or if neither differences nor similarities could be evidenced (fully "uncertain" profiles, in the above-mentioned sense), a compromise score of 0.1 is returned. This value was chosen such as to signal that such profiles are clearly not different but should nevertheless not be allowed to compete in ranking with doubtlessly similar profiles at $\Lambda = 0$.
- Clearly different profiles, with $N_{\text{diff}} > \lambda N_{\text{sim}}$ score Λ values above 0.1, reach an upper limit of 1.0 if the excess differences make up more than 5% of the total number of targets in the profile.

It must be noted that Λ is not, strictly speaking, a metric: $\Lambda(M,M) = 0$ only if M binds at least to one target, with more than 1 μM of affinity. It is important to note that the conception of the Λ score ensures, unlike Euclidean or block distance metrics, a context-dependent activity difference interpretation. For example, the situation $p(m,t) = 5.0$ and $p(M,t) = 7.0$ marks an important difference between m and M, in the sense that selecting m from a database by means of a similarity screening experiment with respect to M might count as a failure. However, if $p(m,t) = 7.0$ and $p(M,t) = 9.0$, the discovery of m starting from M typically goes as a success, although the same 2 orders of magnitude of activity were lost. In the former case, target t contributes +1 to $N_{\text{diff}}(m,M)$, while in the latter, t contributes zero to both N_{diff} and N_{sim}. Eventually, if $p(m,t) > 7.0$ and $p(M,t) = 9.0$, target t becomes a contributor to N_{sim}. The Λ score therefore ranks a compound pair of activities (8,9) as more similar than a pair of activities (7,9) with respect to the target in question—like any Euclidean or Hamming score. Unlike these latter, however, Λ also meaningfully prioritizes the (7,9) pair over the (5,7) pair.

The failure of classical similarity metrics to respond differently to compound pairs that are both active and respectively both inactive often leads to an inappropriate, counterintuitive estimation of activity dissimilarity, as exemplified in Figure 15. The two bar plots represent comparative activity profiles—biological targets are aligned along the x axis, while the empty and filled bars respectively represent the pIC_{50} values of the compared molecules with respect to each target. Practically, IC_{50} values are only measurable starting from a certain activity threshold of the ligand—for compounds that are not active enough, a baseline pIC_{50} value of 3.0 is assumed (this also applies to BioPrint data). The left-hand graph displays a pair of molecules which have measurable pIC_{50} values with respect to a single target in the profile, and only one of them binds strongly enough to qualify as a potential hit or lead. A significant activity difference of three log units can be observed—obviously,

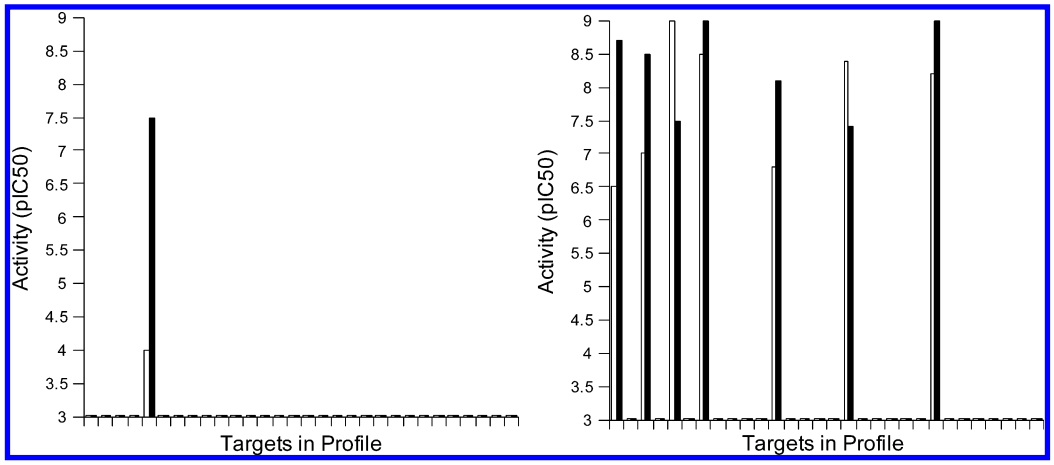

Figure 15. Two bar plots representing comparative activity profiles.

these molecules have different activity profiles. No other targets contribute to the Euclidean activity dissimilarity score, which therefore equals 3. The right-hand plot displays, by contrast, a pair of molecules with almost ideally covariant activities: they bind to the same targets, with comparable and significant—although not identical—affinities. However, every such target, rather than counting as a bonus in the profile similarity scoring, actually contributes some increment to the Euclidean profile dissimilarity score, which exceeds the dissimilarity level of the left-hand "different" compound pair and reaches 3.68. It is highly unlikely to expect identical activity values from binders to a same target, but it is guaranteed to get identical entries in the profile vector if none of the compounds have measurable pIC_{50} values—therefore, compound pairs with low hit rates in the profile will be spuriously favored by Euclidean scoring. A vector dot-product-based scoring metric would hardly perform better—as, in the left-hand plot, the only signals above the basis level stem from the same target; scores close to 1.0 (maximum similarity) are expected no matter what precise formula is used to calculate the profile correlation coefficient.

APPENDIX B: NEIGHBORHOOD BEHAVIOR CRITERIA.

NB analysis relies on monitoring activity dissimilarity within the subset $P(s)$ of molecule pairs (m,M) having calculated structural dissimilarity scores $\Sigma(M,m)$ below a variable dissimilarity threshold s. Let $N(s)$ represent the number of pairs retrieved by the selection $P(s)$ and which represent a fraction $f(s) = N(s)/N_{all}$ out of the total number of molecule pairs in the study. The consistency score $\chi(s)$ is defined in eq B1 by situating the average activity dissimilarity $\langle\Lambda(m,M)\rangle_{P(s)}$ of the $N(s)$ pairs in the actual selection at threshold s, in the context of (1) its upper baseline, the global average $\langle\Lambda(m,M)\rangle_{all}$ of all of the pairs in the study, which $\langle\Lambda(m,M)\rangle_{P(s)}$ approaches if selection at threshold s leads to a subset $P(s)$ as poor in activity-related pairs as a randomly picked one, and (2) its lower, ideal baseline, representing $\langle\Lambda(m,M)\rangle_{N(s)}^{MIN}$, the average Λ of the $N(s)$ compound pairs with the lowest Λ among the given N_{all} pairs.

$$\chi(s) = \frac{\langle\Lambda(m,M)\rangle_{all} - \langle\Lambda(m,M)\rangle_{P(s)}}{\langle\Lambda(m,M)\rangle_{all} - \langle\Lambda(m,M)\rangle_{N(s)}^{MIN}} \qquad \text{(B1)}$$

The overall optimality criterion $\Omega(s)$ renders a weighted account of two molecule pair counts in the actual selection of pairs $P(s)$ and randomly picked pairs:

- The first is the number of false similar pairs N_{FS} [structurally similar pairs with dissimilar activity profiles: $\Sigma(M,m) \le s$ and $\Lambda(M,m) > \kappa$]. A scaling factor $K > 1$ is applied to N_{FS} in order to take into account that, in virtual screening applied to drug discovery, the selection of pairs with diverging activity profiles is more penalizing than a failure to select all of the activity-related pairs (see below). In this work, $K = 100$.
- The second is the number of potentially false dissimilar pairs N_{PFD} [activity-related molecule pairs, apparently not structurally similar enough to be selected: $\Sigma(M,m) > s$ and $\Lambda(M,m) \le \kappa$].

The determination of N_{FS} and N_{PFD} requires in principle[16] a choice of the tolerated activity dissimilarity threshold κ—in the current context, however, every selected molecule pair (M,m) in $P(s)$ is fuzzily contributing an increment of $\Lambda(m,M)$ to N_{FS} and $1 - \Lambda(M,m)$ to N_{PFD}. In a random selection process, a set of size $N(s)$ would include activity-related and activity-unrelated pairs in a proportion equal to their overall occurrence in the total pair set and therefore

$$\Omega(s) = \frac{KN_{FS} + N_{PFD}}{KN_{FS}^{rand} + N_{PFD}^{rand}} = \frac{K\sum_{P(s)}\Lambda(M,m) + \sum_{All-P(s)}[1 - \Lambda(m,M)]}{K\frac{N(s)}{N_{all}}\sum_{all}\Lambda(m,M) + \left[1 - \frac{N(s)}{N_{all}}\right]\sum_{all}[1 - \Lambda(M,m)]} \qquad \text{(B2)}$$

NB can be graphically assessed by plotting the optimality criterion Ω against the consistency χ at various structural similarity thresholds s. Low Ω at high χ signals good neighborhood behavior.

Supporting Information Available: The public data set complied from eight QSAR series, including calculated FPT descriptors (FPT-2) and the .xml setup files controlling compound standardization and generation of ChemAxon PF and CF descriptors. This material is available free of charge via the Internet at http://pubs.acs.org. Activity dissimilarity $\Lambda(M,m)$ and FPT dissimilarity scores $\Sigma^{FTP}(M,m)$—not shared via

pubs.acs.org for technical reasons (files too large)—are available upon request (dragos.horvath@univ-lille1.fr).

REFERENCES AND NOTES

(1) Adam, M. Integrating Research and Development: The Emergence of Rational Drug Design in the Pharmaceutical Industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.

(2) Geney, R.; Sun, L.; Pera, P.; Bernacki, R. J.; Xia, S.; Horwitz, S. B.; Simmerling, C. L.; Ojima, I. Use of the Tubulin Bound Paclitaxel Conformation for Structure-Based Rational Drug Design. *Chem. Biol.* **2005**, *12*, 339−48.

(3) Ivanov, A. A.; Baskin, I. I.; Palyulin, V. A.; Piccagli, L.; Baraldi, P. G.; Zefirov, N. S. Molecular Modeling and Molecular Dynamics Simulation of the Human A2B Adenosine Receptor. The Study of the Possible Binding Modes of the A2B Receptor Antagonists. *J. Med. Chem.* **2005**, *48*, 6813−20.

(4) Bernacki, K.; Kalyanaraman, C.; Jacobson, M. P. Virtual Ligand Screening against Escherichia coli Dihydrofolate Reductase: Improving Docking Enrichment Using Physics-Based Methods. *J. Biomol. Screening* **2005**, *10*, 675−81.

(5) Barreca, M. L.; Ferro, S.; Rao, A.; De Luca, L.; Zappala, M.; Monforte, A. M.; Debyser, Z.; Witvrouw, M.; Chimirri, A. Pharmacophore-Based Design of HIV-1 Integrase Strand-Transfer Inhibitors. *J. Med. Chem.* **2005**, *48*, 7084−8.

(6) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **2005**, *48*, 6997−7004.

(7) Low, C. M.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. Scaffold Hopping with Molecular Field Points: Identification of a Cholecystokinin-2 (CCK2) Receptor Pharmacophore and Its Use in the Design of a Prototypical Series of Pyrrole- and Imidazole-Based CCK2 Antagonists. *J. Med. Chem.* **2005**, *48*, 6790−802.

(8) Güner, O. F. *Pharmacophore Perception, Use and Development in Drug Design*; International University Line: La Jolla, CA, 2000.

(9) Horvath, D. High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and its Role in the Drug Discovery Laboratory. In *Combinatorial Library Design and Evaluation. Principles, Software Tools, and Applications in Drug Discovery*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 429−472.

(10) Makara, M. G. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.

(11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(12) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of Validated QSAR Models of d(1) Dopaminergic Antagonists for Database Mining. *J. Med. Chem.* **2005**, *48*, 7322−32.

(13) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-Protein-Coupled Receptor Affinity Prediction Based on the Use of a Profiling Dataset: QSAR Design, Synthesis, and Experimental Validation. *J. Med. Chem.* **2005**, *48*, 6563−74.

(14) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004.

(15) For details on the two-point topological pharmacophore descriptors developed by ChemAxon, see http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Sept 2006).

(16) 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.

(17) 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; special issue "Machine Learning Methods in QSAR Modeling".

(18) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−23.

(19) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1995**, *38*, 144−150.

(20) Menard, J. P.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−13.

(21) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86*, 865−71.

(22) 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.

(23) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon J. C. Predicting ADME Properties and Side Effects: The BioPrint Approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470−80.

(24) http://www.cerep.fr/cerep/users/pages/Collaborations/Bioprint.asp (accessed Sept 2006).

(25) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.

(26) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(27) The above-mentioned data sets are also available via http://www.cheminformatics.org/ (accessed Sept 2006).

(28) Horvath, D. ComPharm − Automated Comparative Analysis of Pharmacophoric Patterns and Derived QSAR Approaches, Novel Tools in High Throughput Drug Discovery. A Proof of Concept Study Applied to Farnesyl Protein Transferase Inhibitor Design. In *QSPR/QSAR Studies by Molecular Descriptors*; Diudea, M., Ed.; Nova Science Publishers: New York, 2001; pp 395−439.

(29) http://www.chemaxon.com/jchem/doc/api/ (accessed Sept 2006).

(30) http://www.chemaxon.com/jchem/index.html?content=doc/user/Standardizer.html (accessed Sept 2006).

(31) http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html#pka (accessed Sept 2006).

(32) http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Sept 2006).

(33) http://www.chemaxon.com/jchem/doc/user/fingerprint.html (accessed Sept 2006).

(34) http://www.maybridge.com/ (accessed Sept 2006).

(35) Christensena, J. G.; Burrows, J.; Salgiab R. c-Met as a Target for Human Cancer and Characterization of Inhibitors for Therapeutic Intervention. *Cancer Lett.* **2005**, *225*, 1−26.

(36) Vojkovsky, T.; Koenig, M.; Zhang, F.-J.; Cui, J. Tetracyclic Compounds as c-Met inhibitors. Patent WO2005004808, 2005.

(37) Koenig, M. Indolinonehydrazides as c-Met Inhibitors. Patent WO200500-5378, 2005.

(38) Compounds and activity data taken from the WOMBAT database of Sunset Molecular, Inc. (http://sunsetmolecular.com/products/?id=4) courtesy of Tudor I. Oprea, 2005.

(39) Altschul, S. F. Amino Acid Substitution Matrices from an Information Theoretic Perspective. *J. Mol. Biol.* **1991**, *219*, 555−65.

(40) Kubiny, H. Structure-Based Design of Enzyme Inhibitors and Receptor Ligands. Second European Workshop in Drug Design, Certosa di Pontignano, May 17−24, 1998; oral presentation.

(41) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−63.

(42) Seifert, M. H. J. Assessing the Discriminatory Power of Scoring Functions for Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456−1465.

(43) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049−54.

CI6002416