JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

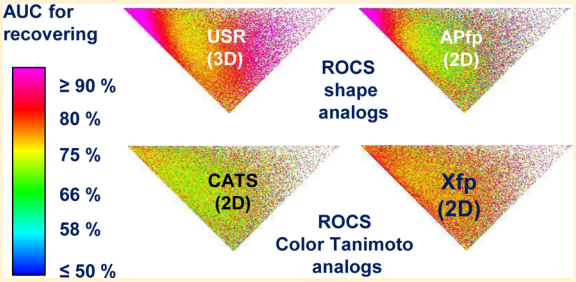
Atom Pair 2D-Fingerprints Perceive 3D-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of ZINC and GDB-17

Mahendra Awale and Jean-Louis Reymond*

Department of Chemistry and Biochemistry, University of Berne, Freiestrasse 3, 3012 Berne Switzerland

Supporting Information

ABSTRACT: Three-dimensional (3D) molecular shape and pharmacophores are important determinants of the biological activity of organic molecules; however, a precise computation of 3D-shape is generally too slow for virtual screening of very large databases. A reinvestigation of the concept of atom pairs initially reported by Carhart et al. and extended by Schneider et al. showed that a simple atom pair fingerprint (APfp) counting atom pairs at increasing topological distances in 2D-structures without atom property assignment correlates with various representations of molecular shape extracted from the 3D-structures. A related 55-dimensional atom pair fingerprint extended with atom properties (Xfp) provided an efficient pharmacophore fingerprint with good performance for ligand-based virtual screening such as the recovery of active compounds from decoys in DUD, and overlap with the ROCS 3D-pharmacophore scoring function. The APfp and Xfp data were organized for web-based extremely fast nearest-neighbor searching in ZINC (13.5 M compounds) and GDB-17 (50 M random subset) freely accessible at www.gdb.unibe.ch.



INTRODUCTION

The high attrition rates encountered in drug development projects have led to questioning whether the initial choices of molecules selected for activity screening should be reconsidered from first principles.[1−3] In particular the notion of molecular shape,[4−12] and its analysis in the context of successes and failures shows that molecules with a substantial three-dimensional (3D) shape have a better chance of becoming drugs.[13] On the contrary, medicinal chemistry has mostly focused on planar and even rodlike molecules, in part due to the availability of aromatic coupling reactions.[14] To help expand the repertoire of possible drugs, we recently carried out a systematic enumeration of all possible organic molecules up to a defined size following simple rules of synthetic feasibility and chemical stability, and produced the Chemical Universe Databases GDB-11 (generated database of all molecules up to 11 atoms of C, N, O, F, 26.4 million structures), GDB-13 (up to 13 atoms of C, N, O, S, Cl, 977 million structures) and GDB-17 (up to 17 atoms of C, N, O, S, halogens, 166.4 billion structures).[15−18] Analysis of molecular shape in small subsets of the GDB showed that these databases are much richer in 3D molecular shapes compared to databases of known compounds such as ChEMBL,[19] PubChem,[20] or ZINC.[21]

While the GDBs probably contains many innovative bioactive molecules in interesting regions of 3D-molecular shape space, these very large databases pose an almost unsurmountable problem in terms of virtual screening,[22−26] i.e. the ability to predict which compounds have interesting potential bioactivities.[27−30] Indeed the size of the GDBs implies that they can only be virtually screened using simple scoring functions based on the 2D-structure of molecules such as molecular fingerprints.[31] Our uses of GDB in drug discovery projects so far have therefore relied either on 2D-fingerprints alone[32,33] or on scoring of selected small subsets by docking or 3D-shape similarity analysis.[34−37] This difficulty stems in part from the necessity to address the stereochemical and conformational diversity of each molecule by generating up to several tens of stereoisomers and hundreds of 3D-conformers from each 2D-structure.

Molecular fingerprints rapidly calculable from 2D-structures but capable of encoding 3D-shape information would be highly desirable to enable the analysis and exploitation of very large databases such as the GDB or simply the ZINC database, collecting millions of commercially available compounds. Numerous recent studies have carried out comparisons of 2D-fingerprints vs 3D-fingerprints for their performance in recovering actives from decoys in the directory of useful decoys (DUD), a broadly accepted method to benchmark virtual screening methods.[38−42] However, these studies did not analyze whether the 2D-fingerprints actually reproduced 3D-shapes of the molecules. We therefore set out to identify a shape-encoding 2D-fingerprint considering not only ligand-based virtual screening (LBVS) performance but also its ability to actually reproduce 3D-shape features.

In our search for a shape-encoding 2D-fingerprint, we were intrigued by the notion of atom pairs originally proposed by Carhart et al.[43] These authors defined an atom pair as a pair of

Received: April 15, 2014

   dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

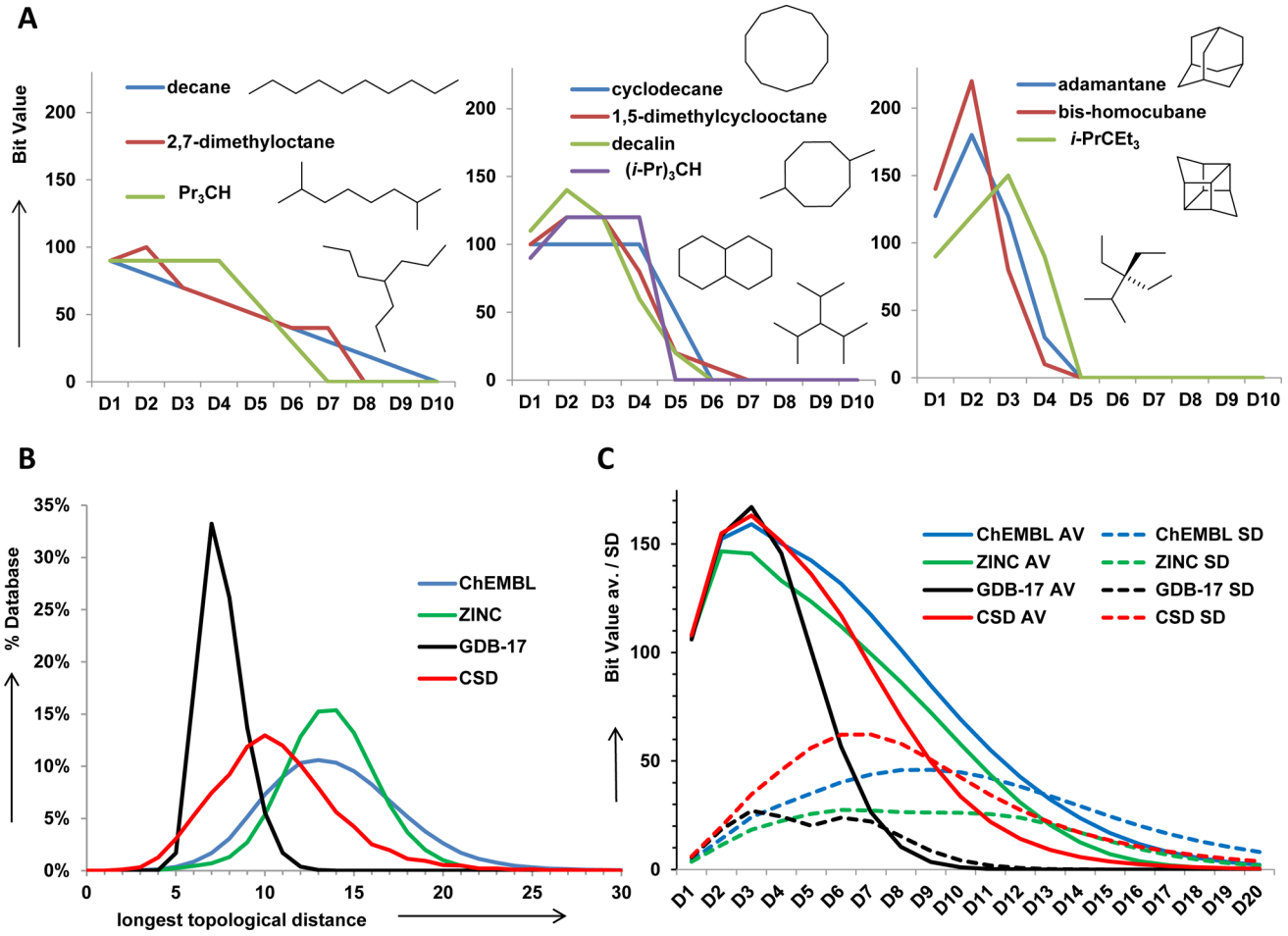


Figure 1. APfp bit value statistics. (A) APfp Bit value for model hydrocarbons. The APfp bit value is Di/HAC, where Di is the number of atom pairs separated by i-bonds and HAC is the heavy atom count. The bit values are stored in rounded percentage counts. (B) Percentage of the database molecule as a function of increasing longest topological distance. (C) Average and standard deviation of the bit value in different databases.

atoms with properties and separated by a distance measured as the shortest topological path between the two atoms counted in bonds. The approach was later revisited by Schneider et al.,[44] who pointed out that topological distances are independent of conformational equilibria and thus simplify the analysis of pharmacophores. Both authors used atom pairs as a pharmacophore fingerprint by assigning properties to each atom and counting the number of same-property atom pairs and cross-property atom pairs at increasing topological distances. Carhart et al. defined a different property for each element, the number of π-electrons on the atom, and its number of non-hydrogen neighbors, and used all topological distances occurring in the molecule, resulting in a variable but large number of same pairs and cross pairs. Schneider et al. focused on only five properties deemed relevant for pharmacophores, namely lipophilic, H-bond acceptor, H-bond donor, positive charge, and negative charge, and computed all possible same pairs and cross pairs at topological distances of 1−10 bonds (with the D10 bits accumulating all pairs with topological distances longer than 10), creating a 150-dimensional fingerprint called CATS. Both authors demonstrated the efficiency of their fingerprint for similarity searching by identifying scaffold-hopping analogs; however, they did not report whether their atom pair 2D-fingerprints actually represented 3D-shapes and pharmacophores.

To test if such 2D-fingerprints did indeed encode 3D-molecular shape, we analyzed the properties of the parent atom pair fingerprint (APfp, without assignment of atomic properties) in terms of the correlation between topological distances and through-space distances. The correlation between distances separating molecules in different fingerprint spaces measured in city-block distance (CBD),[45] a similarity measure performing similarly to the Tanimoto coefficient for enrichment studies but which allows to preorganize large databases for fast searching,[46] was also analyzed, as well as the ability of APfp to enrich shape analogs defined by 3D-shape fingerprint similarity. APfp indeed correlated well with shape descriptors computed from the 3D-structures of the molecules,[47−52] including the principal moments of inertia scaled to molecular weight collected in a scalar fingerprint (PMIfp),[4] the Ultrafast Shape Recognition (USR) fingerprint,[53,54] and the Rapid Overlay of Chemical Structures (ROCS) shape similarity function.[5,55] The correlation of APfp with these measures of 3D-shape was better than for other molecular 2D-fingerprints such as the binary daylight-type substructure fingerprint (Sfp) and an Extended Connectivity fingerprint (ECfp4),[56,57] our recently reported Molecular Quantum Numbers (MQN, 42 counts for atom, bonds, polarity, and topology),[58,59] and the SMILES fingerprint (SMIfp, counts for 34 characters in SMILES).[60]

The use of APfp as a pharmacophore fingerprint was also re-examined with the aim of reducing fingerprint complexity by

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

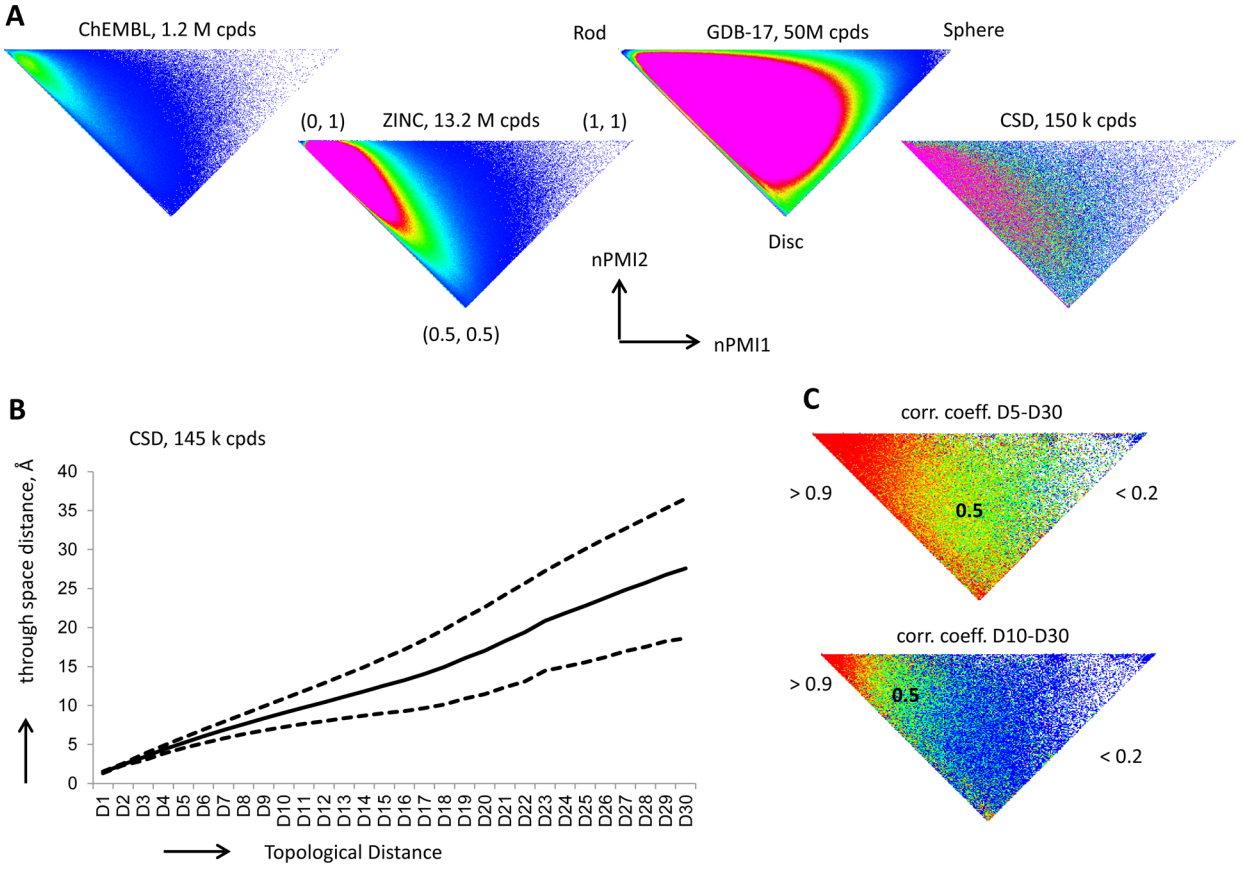


Figure 2. Shape analysis and distance correlations. (A) Occupancy of the PMI triangle by database compounds. The heat map is colored from blue (minimum value) to magenta (maximum value: > 400 cpds/pixel for ChEMBL, ZINC, GDB-17, > 20 cpds/pixel for CSD). (B) Through-space distance as a function of topological distance between atom pairs in CSD molecules. The average ± standard deviation is shown. (C) Correlation coefficient between through-space and topological distances between atom pairs in CSD molecules as a function of molecular shape, analyzed considering pairs at topological distance D5 and longer (upper plot) or D10 and longer (lower plot).

focusing on only four atomic properties relevant for pharmacophores. We considered the hydrophobic, H-bond donor, H-bond acceptor, and planarity, and selected only the four same-property pairs and the H-bond donor−H-bond acceptor as the only cross-pair. This produced an extended 55-dimensional pharmacophore fingerprint Xfp. Xfp recovered actives from the DUD and from ZINC as well as the more complex 150-dimensional CATS. Furthermore, APfp and Xfp showed interesting similarities to the ROCS shape and ROCS pharmacophore scoring functions, respectively. To facilitate the use of APfp and Xfp as scoring functions, interactive web-browsers were enabled for searching ZINC and GDB-17. These browsers are freely available at www.gdb.unibe.ch, need only a few seconds per search, and should provide a useful support for drug discovery projects as a rapid substitute for 3D-pharmacophore similarity searching.

RESULTS AND DISCUSSION

Atom Pair Fingerprint, APfp. APfp was set to count atom pairs up to 10−20 bonds. Similarly to the approaches by Carhart and Schneider, the counts were scaled to the heavy atom count (HAC) to obtain a simple proportionality to molecular size since the number of possible atom pairs increases with the square of HAC. Bit values were expressed in percent and rounded to the integer value.

It seems quite intuitive that summing up atom pairs at increasing topological distances might reflect molecular shape. Elongated molecules contain atom pairs that are far apart, while in globular molecules most atom pairs are at relatively short topological distances. This can be exemplified by considering the scaled atom pair counts up to 10 bonds for C10-molecules with different shapes. For extended rodlike molecules such as *n*-decane or 2,7-dimethyl-octane the bit values decrease linearly with increasing topological distance. The disclike molecules cyclodecane, 1,5-dimethyl-cyclooctane, and decalin have a flat bit value profile at lower distances and zero-values at high distances. The propeller-topology molecule tripropylmethane features an intermediate between rodlike and disclike fingerprints, while the corresponding tri-isopropyl-methane is clearly disclike. Finally the spherical molecules adamantane, bishomocubane and triethyl-isopropyl-methane have an even sharper bit value profile at low distances (Figure 1A).

To help in fingerprint design and performance analysis we analyzed molecules up to HAC = 50 in the databases ChEMBL (1.12 M cpds), ZINC (13.5 M cpds), CSD (145 k cpds), and the Chemical Universe Database GDB-17 (50 M random subset). In these databases most molecules had their longest topological distance between atom pairs in the range of 6−18 bonds (Figure 1B). The profile suggested considering topological distances between 1 and 10−20 bonds to cover most of the occurring distances to form the corresponding atom pair fingerprints APfp10 to APfp20. The average fingerprint bit value peaked at D3, reflecting the high frequency of short fragments; however, the standard deviation of the bit

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

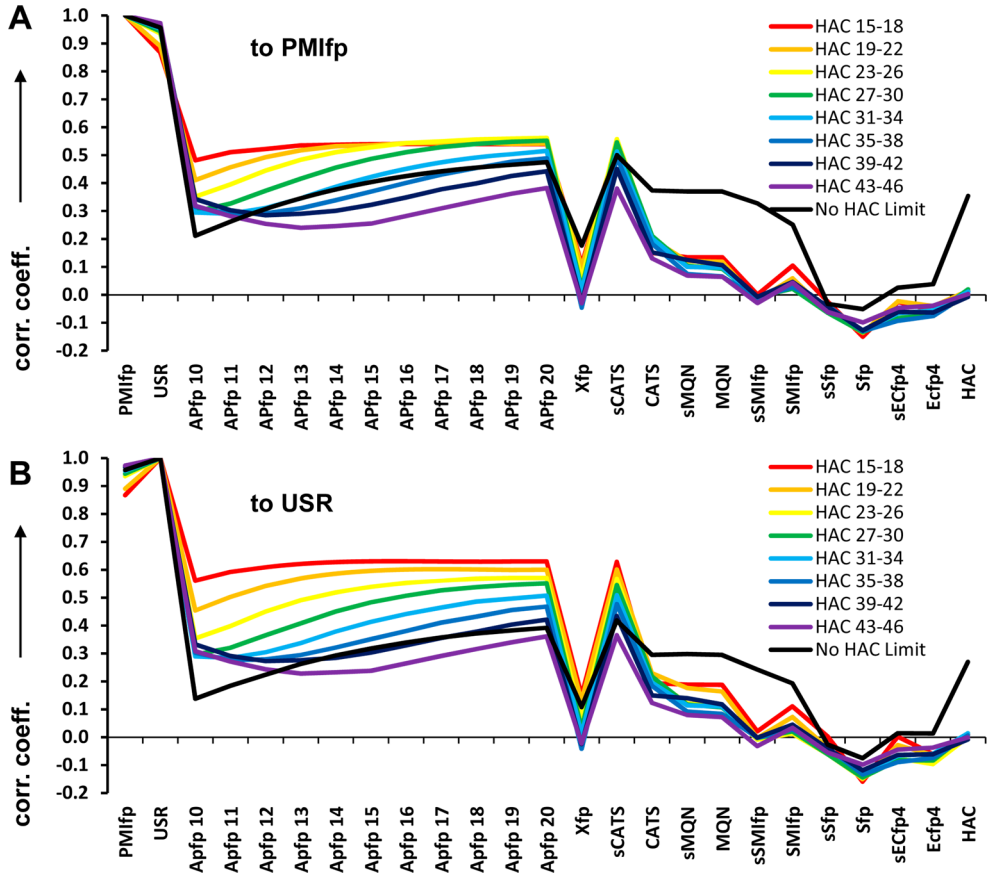


Figure 3. Pearson Correlation Coefficient for distances between 100,000 random pairs of compounds from CSD in various fingerprint spaces. Random pairs were drawn either from size-focused sets or from all sizes. (A) Correlation to distances in the PMIfp fingerprint space. (B) Correlation to distances in the USR fingerprint space. See Methods for details.

values in all four databases was quite significant across the entire range considered, suggesting that all bits might play a role in the fingerprint performance (Figure 1C).

Correlation between 3D-Shape and APfp. To test if APfp could perceive the 3D-shape of molecules, fingerprint comparisons were performed using CSD as a reference collection of molecules. CSD reports experimentally determined 3D coordinates, and its contents cover a broad range of molecular shapes as measured by the principal moment of inertia (PMI) triangle,[4] with significant coverage of disklike and spherical shapes (Figure 2A).

Topological distances and actual through-space distances between atom pairs were compared first. While the two measures were highly correlated below four bonds, there was an increasing variance in the through-space distance at long topological distances (Figure 2B). The average through-space distance at a topological distance of 20 bonds was 15.3 ± 7.4 Å, which is significantly shorter than the 25.0 Å expected for an alkane chain in extended conformation and reflects the occurrence of ring systems in most large molecules. The correlation between topological distances and through-space distances was influenced by molecular shape, as shown by the average value of the correlation coefficient as a function of the molecule position in the PMI triangle (Figure 2C). The correlation was quite good for rodlike molecules close to the (0,1) corner, but was very low for disclike (0.5,0.5) and spherelike molecules (1,1), in particular when analyzing only distances above D10 where the relationship between topological distance and through-space distance is lower.

In a second analysis, APfp was compared with shape fingerprints derived from the 3D-structure of the molecules. Three-dimensional-shape was encoded as a PMI fingerprint (PMIfp) consisting of the three values of the molecular principal moments of inertia scaled to the molecular weight (which results in simple proportionality to size while the normalized PMIs used by Sauer et al. are size independent),[4] and as the Ultrafast Shape Recognition fingerprint (USR), a recently reported efficient shape fingerprint.[53,54] APfp performance was also compared with other 2D-fingerprints including our scalar fingerprints Molecular Quantum Numbers (MQN)[58] and SMILES fingerprint (SMIfp),[60] and the 1024 bit binary substructure fingerprint (Sfp)[31] and extended connectivity fingerprint (ECfp4).[56] Simplified fingerprints, sMQN, sSMIfp, sSfp, and sECfp4, considering only a single atom and bond type were also included to measure the pure shape perception of MQN, SMIfp, Sfp, and ECfp4, and the heavy atom count (HAC) as a reference 1D descriptor for size.[61,62] Schneider's CATS fingerprint and a simplified version with only a single-atom type (sCATS) as well as the extended atom pair fingerprint Xfp discussed below were added for comparison.

To probe the similarity of the various 2D-fingerprints to 3D-shape fingerprints, the correlation between the distance separating 100,000 random pairs of molecules in the various fingerprint spaces (measured using the city-block distance CBD) was determined. The molecule pairs were drawn at random either from the entire CSD, or from several size-focused sets covering the range 15 ≤ HAC ≤ 46 to remove the effect of molecular size. The highest correlation to distances in

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

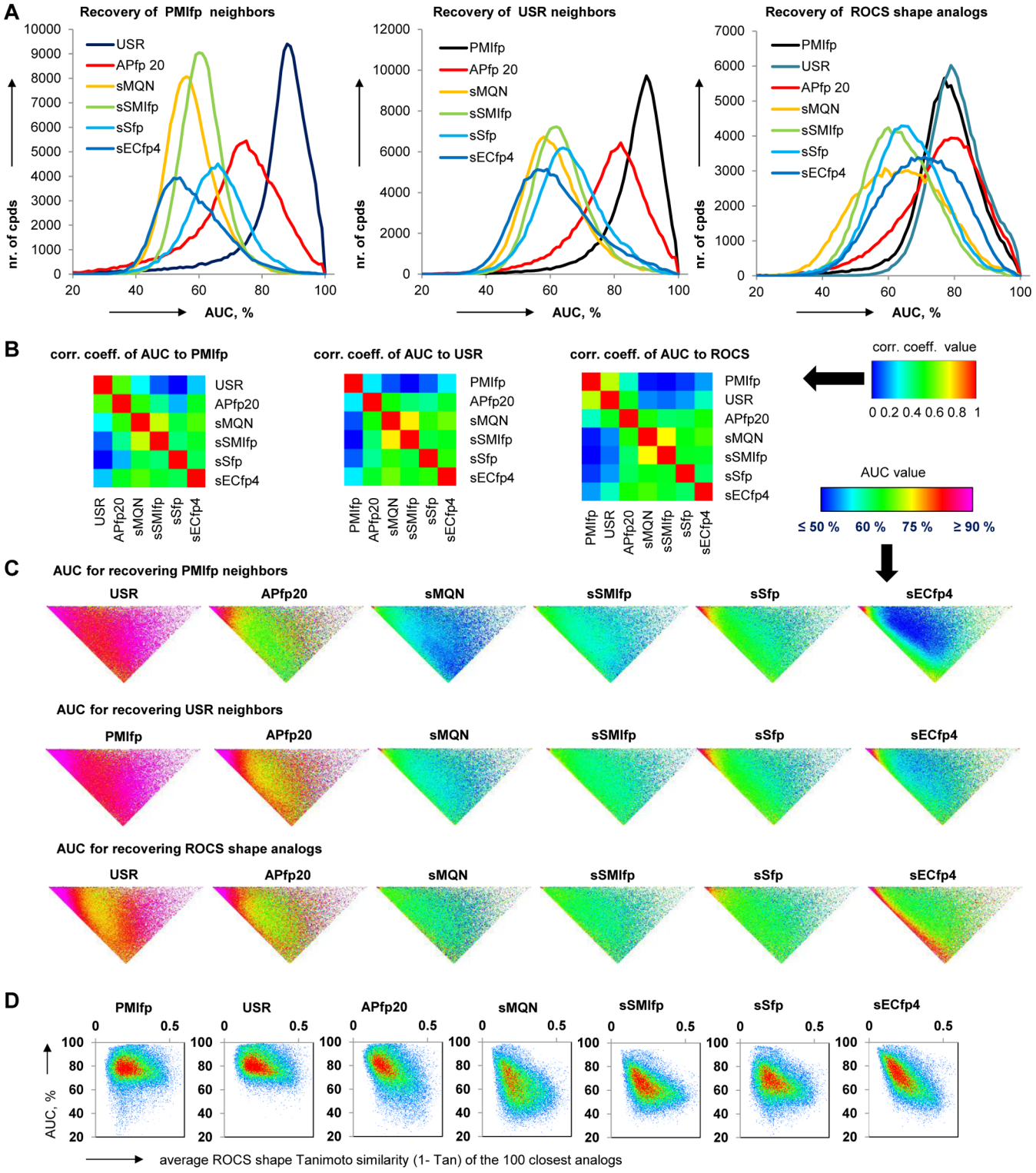


Figure 4. Recovery statistics of 100 nearest neighbors according to PMIfp, USR, and ROCS shape Tanimoto using various fingerprints, for each of the 145,000 molecules in CSD from their size-constrained subsets (all CSD molecules within HAC = query ± 2). (A) Frequency histogram of AUC values. Curves for APfp10-APfp19 and for sCATS are almost exactly superimposed with the curve of APfp20 and are not shown. (B) Correlation coefficients of AUC values from different fingerprints across all 145,000 molecules. (C) Average AUC value as a function of position in the shape triangle, continuous color scale: AUC ≤ 50%: blue, 58%: cyan, 66%: green, 75%: yellow, 80%: red, ≥ 90%: magenta. (D) Occupancy heat map of scatter plot of AUC value as a function of the average ROCS shape similarity of the 100 closest analogs with their size-constrained subset of CSD (HAC = query ± 2). See also Figure S1 in the SI for data with full fingerprints CATS, Xfp, MQN, SMIfp, Sfp, and ECfp4.

PMIfp- or USR-space was obtained with one another, which is not surprising since both fingerprints use the same set of 3D-coordinates as input (Figure 3). This correlation was visible both in the entire set containing various molecule sizes ("HAC no limit", black line) and within the different size-constrained set (colored lines). Distances in APfp-spaces were also

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

significantly correlated with PMIfp and USR distances across the various sets, with the general trend that the correlation increased by including counts at longer topological distances in APfp. Interestingly the performance of APfp20 was closely matched by sCATS, the simplified version of Schneiders's CATS fingerprint without atom categories. This fingerprint only counts atom pairs between 1 and 10 bonds, but sums up all pairs 10 or more bonds apart in the D10 bit, which enables a better perception of 3D shape with only 10 bits compared to APfp10. Distances in all other fingerprint spaces showed no significant correlation apart from a size-dependent correlation (HAC no limit, black line) for both PMIfp and USR occurring with CATS, MQN/sMQN and SMIfp/sSMIfp, a consequence of the fact that these fingerprints are size dependent, while the binary fingerprints Sfp/sSfp and ECfp4/sECfp4 are not.

Shape Similarity Enrichment Studies with APfp. The above distance correlation study revealed the overall similarity of the various fingerprint spaces across all possible fingerprint-space distances. To probe the similarity of APfp to 3D-shape fingerprints at close range, which is more relevant to LBVS applications, we tested the ability of APfp-similarity to enrich for 3D-shape analogs, defined as nearest neighbors in the PMIfp and USR-spaces, as well as the closest analogs retrieved by the ROCS (Rapid Overlay of Chemical Structures) shape scoring function, a well-established 3D-shape comparison algorithm.[5] The AUC (area under the curve) was determined for the ROC (receiver operator characteristics) curve in the recovery of the 100 nearest neighbors in PMIfp- or USR-space or the 100 most ROCS similar compounds of each of the 145,000 CSD molecules among all similar-size molecules (HAC ± 2) in CSD. The analysis was performed for the reference 3D-fingerprints PMIfp and USR and for the 2D-fingerprints APfp10 to APfp20, sCATS/CATS, sMQN/MQN, sSfp/Sfp, and sECfp4/ECfp4.

The frequency histogram as a function of AUC showed that the AUC values were always highest when recovery was attempted with another 3D-fingerprint, i.e. recovering PMIfp neighbors with USR, recovering USR neighbors with PMIfp, and recovering ROCS shape analogs with PMIfp or USR (Figure 4A and Figure S1A in the SI). However, the atom pair fingerprints (APfp10−APfp20 and sCATS) also performed remarkably well in all three cases, with average AUC values of 72.5 ± 13.6% for recovering PMIfp neighbors, 78.7 ± 11.4% for recovering USR neighbors, and 76.2 ± 12.2% for recovering ROCS shape analogs. All other fingerprints gave lower AUC values, although sECfp4 showed a surprising spread of AUC values for recovering ROCS shape analogs.

The AUC values obtained with the different fingerprints were not strongly correlated, although the pairs sMQN/sSMIfp and Xfp10/CATS showed a relatively high correlation of AUC values in the recovery of PMIfp, USR and ROCS shape analogs (Figure 4B and Figure S1B in the SI). APfp20, and to a lesser extent sECfp4, stood out by its moderate yet significant correlation of AUC values with both the 3D-fingerprints and the 2D-fingerprints, suggesting that APfp20 encoded both 3D-shape and 2D-substructure information.

The relationship between the AUC value and the position of each molecule in the PMI triangle showed that pure rodlike molecules (left corner of the triangle) gave very high AUC values with all fingerprints (Figure 4C and Figure S1C in the SI). While the performance of sMQN/MQN, sSMIfp/SMI, and sSfp/Sfp was confined to the rodlike molecules, sECfp4 and to a lesser extent ECfp4 performed well along the lower left edge of the triangle representing planar, rodlike-to-disclike molecules. APfp20 stood out together with the 3D-fingerprints USR, PMIfp by its high AUC values across the entire shape triangle in each of the three enrichment studies.

The ease of recovery of shape analogs in the case of pure rodlike molecules can be understood in terms of the high correlation between through-space and topological distances in that region of the shape triangle (Figure 2C). For PMIfp, USR and APfp20 giving good AUC values spread across the entire PMI triangle, an additional contributing effect could be the occurrence of many close analogs of the query molecule, which would be highly similar not only by shape but by other structural features perceived by different fingerprints, thereby facilitating the enrichment. In fact a correlation was observed between AUC values for recovery of the 100 closest ROCS shape analogs and the average ROCS shape similarity score of these 100 analogs (Figure 4D and Figure S1D in the SI). However, this effect occurred for sMQN/MQN, sSMIfp/SMIfp, sSfp, and sECfp4 but only to a lesser extent for APfp20, showing that this correlation cannot explain the better performance of APf20 in recovering ROCS shape analogs compared to the other 2D-fingerprints.

Taken together, the analysis above showed that APfp, although only based on topological distances computed from the 2D-structures, recognizes the molecular shape of molecules as encoded from the 3D-coordinates by PMIfp, USR, and ROCS shape similarity quite efficiently and significantly better than other 2D-fingerprints.

Property Extended Atom Pair Fingerprint Xfp. The extension of the APfp to a pharmacophore fingerprint by adding atomic properties was investigated next. Carhart et al. and Schneider et al. considered a relatively large number of properties and counted all same-pairs and cross-pairs systematically (see Introduction). However, any new property pair and cross-pair create an additional discrimination between molecules, which might not necessarily help in a pharmacophore similarity search where a too high level of detail should be avoided to allow relatively dissimilar molecules to be identified as analogs. We hypothesized that a relatively small set of properties and cross-pairs might be sufficient to obtain a good pharmacophore fingerprint.

To identify a small but efficient set of properties, the enrichment performance of property extended APfp in recovering the actives in the directory of useful decoys (DUD),[38] either from the decoys or from the entire ZINC database, was investigated using various same-property pairs and cross-property pair combinations, scaling each count either to the HAC or to the number of property atoms, with bit values expressed in percent rounded to the integer value. Topological distances were counted up to 10−20 bonds, including also the zero distance value to perceive the possible presence of a single atom with a given property and of cross properties on the same atom (e.g., HBA−HBD on hydroxyl groups).

After extensive testing the following selection was chosen to form a 55-dimensional category extended atom pair fingerprint Xfp: Same-property atom pairs were retained for hydrophobic (Hyb), H-bond acceptor (HBA), and H-bond donor (HBD) as key pharmacophoric features, and planar atoms (sp^2-hybridized) perceiving an element of molecular shape. HBA−HBD was included as the only cross-property pair contributing positively to DUD enrichment performance. The best performance was obtained when same-property counts were scaled to the total number of atoms with that property and the HBA−

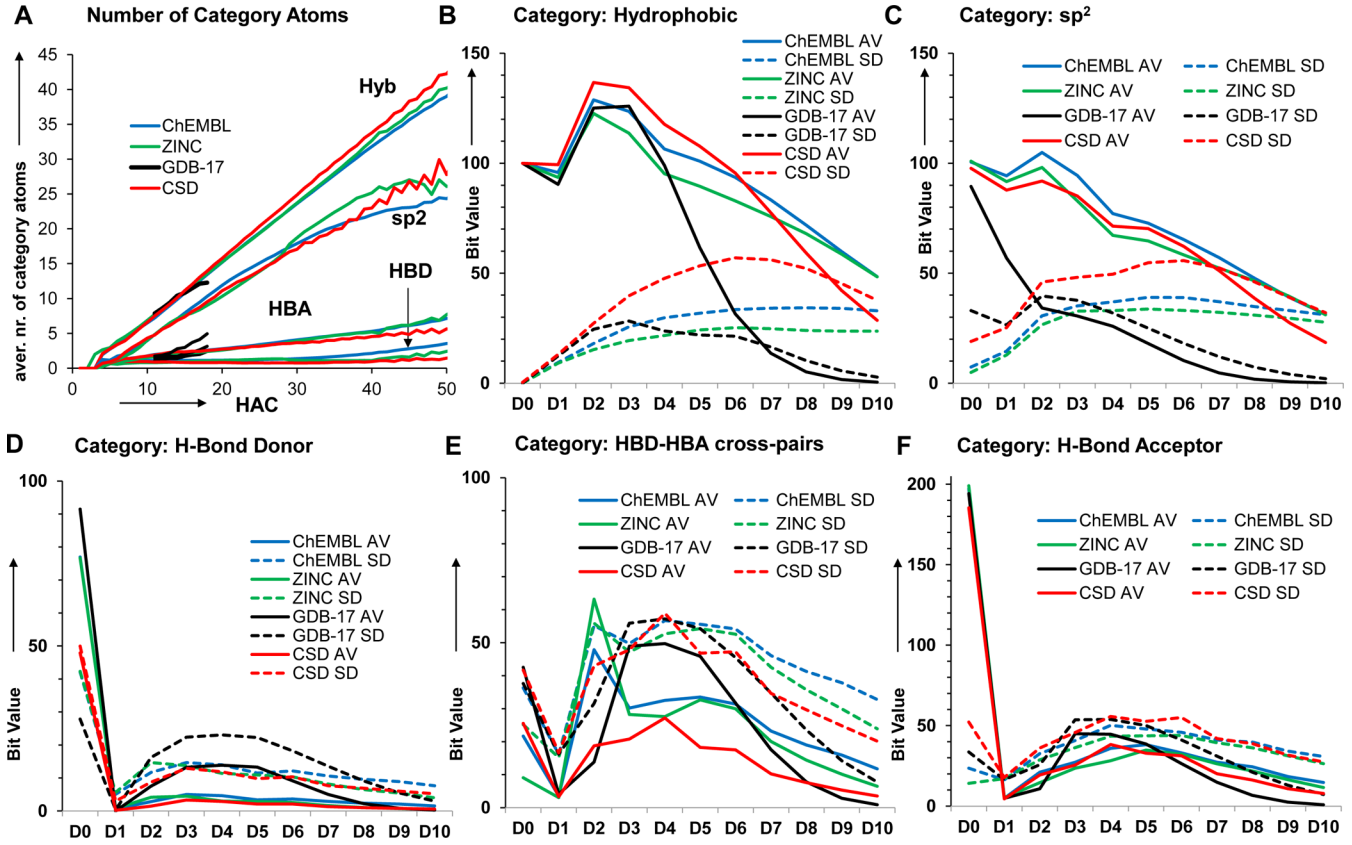


Figure 5. Atom category frequency (A) and Xfp bit values (B–F) in different databases.

HBD cross-pairs were scaled to HBA. Furthermore, the bit values in HBA same-property pairs were doubled to increase their weight. Atom pairs were only counted up to 10 bonds since the inclusion of additional pairs at longer topological distances did not improve DUD recovery performance, in line with the shape enrichment studies above (see SI Table S9 and Figure S7).

Analyzing the occurrence of properties selected for Xfp in various databases showed that molecules in ChEMBL, ZINC, and CSD contained approximately 80% hydrophobic atoms, 40–60% planar atoms, 15% HBA and 5–10% HBD (Figure 5A). The proportions were similar for GDB-17 except that only 28% of all atoms are planar in GDB-17 because this database contains a much lower fraction of aromatic compounds, and 10% are HBD resulting from the abundance of amines in GDB-17. In terms of the bit values themselves, the contribution of each atom property was more even since the counts were scaled to the category count (Figure 5B–E). Most bits contributed to fingerprint diversity with the exception of the hydrophobic self-pair (D0) due to the fact that all molecules analyzed contained at least one carbon atom. The low D1 value with HBD, HBA, and HBD–HBA cross-pairs resulted from the rarity of heteroatom–heteroatoms bonds, while the large D2 value for HBA–HBD cross-pairs reflected the relative abundance of the HX–C=X (X = N, O) group, in particular amides, amidines, and guanidines, in organic molecules, which are key pharmacophores in many bioactivity classes.

Ligand-Based Virtual Screening Performance of Xfp. Since Xfp had been optimized for DUD enrichment performance, its AUC and enrichment factors were evaluated in comparison to other fingerprints (Figure 6). Despite containing only 55 bits, Xfp performed slightly better than the 150-dimensional CATS in recovering actives from decoys both in terms of AUC values and in terms of the enrichment factor at 1% screening ($EF_{1\%}$). The results were similar for recovery from the entire ZINC database in terms of AUC and $EF_{0.1\%}$. Both fingerprints outperformed MQN and SMIfp, which themselves were better than the pure shape fingerprints PMIfp, USR, and APfp20, highlighting the necessity to include atom and functional groups to achieve LBVS performance. Interestingly, the absence of positive or negative charge information in Xfp did not measurably influence performance. For example Xfp was among the best fingerprints for recovering the mostly anionic GART (glycinamide ribonucleotide transformylase) ligands (e.g., from decoys: CBD_{Xfp}: AUC = 96.7%, EF1% = 31.6) and of the mostly cationic thrombin ligands (e.g., from decoys: CBD_{Xfp}: AUC = 85.0%, EF1% = 36.3, see SI Tables S1–S8).

These results were obtained using the city-block distance (CBD), and were indistinguishable from those obtained using the Tanimoto coefficient as similarity measure. It should also be noted that CATS performed better with CBD than with the Euclidean distance similarity originally used by Schneider et al.[44] The best performance in DUD enrichment occurred with Sfp and ECfp4 due to the fact that most active families in DUD contain many variations on common substructures, and also because the decoys were chosen to have low substructure similarity to that of the actives, as also noted previously by other authors.[38,40,63] However, Xfp featured the "scaffold-hopping" capability previously documented for Carhart's atom pairs fingerprint and for CATS, i.e. the ability to identify bioactive analogs with substantially different structural types (Figure 6EF and Figure 7).[44]

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

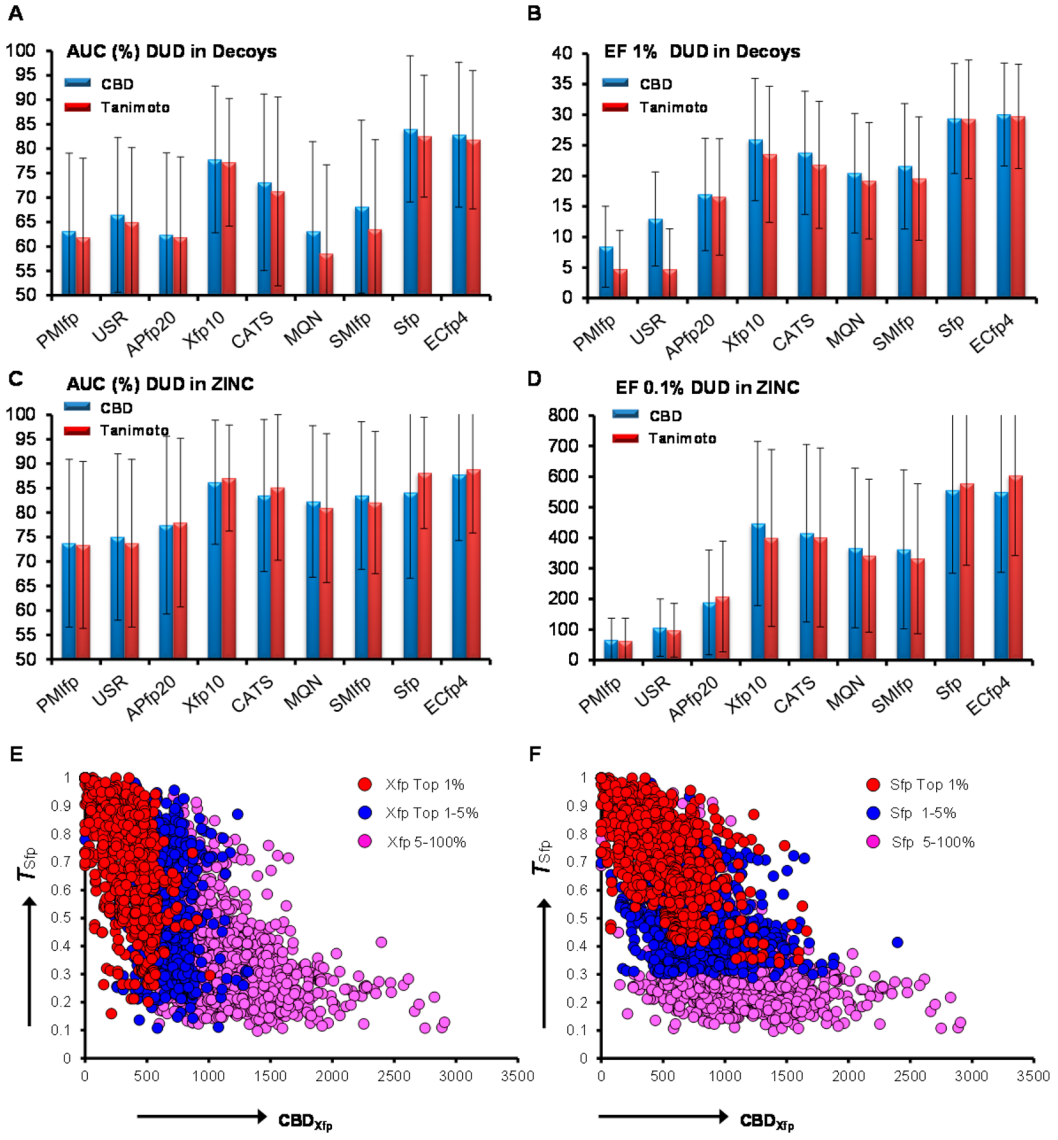


Figure 6. Recovery of DUD actives using various fingerprints. (A) Average AUC values for recovery of 40 sets of actives in directory useful decoys (DUD) from the corresponding decoys set by various fingerprints, using $CBD_{fingerprint}$ (blue bars) and $T_{fingerprint}$ (red bars) as scoring functions. (B) Enrichment factors at 1% of screen database. (C) AUC values for recovery from ZINC. (D) $EF_{0.1\%}$ values for recovery from ZINC. (E,F) Scatter plots of substructure similarity (T_{Sfp}) versus Xfp-distance (CBD_{Xfp}) for the DUD actives. The points are color-coded according to their position in the ROC curves for sorting of DUD database by CBD_{Xfp} (E) or T_{Sfp} (F). AUC values, Enrichment factors and ROC curves are provided in the SI in Tables S1−S10 and Figures S2−S7.

Xfp Similarity Enrichment of 3D-Pharmacophore Analogs. Considering the good performance of Xfp in the recovery of DUD actives, we asked the question whether this performance might be related to its ability to encode 3D-pharmacophores. Xfp was therefore compared with other fingerprints in virtual screening of ROCS Color Tanimoto analogs. The Color Tanimoto is a 3D-pharmacophore model comparable to the ROCS shape used above to evaluate APfp, but considers atomic property matching in the comparison.[5] Indeed the Xfp fingerprint performed remarkably well, with an average AUC value of 78.4 ± 8.9% (red curve in Figure 8A). As for the DUD and shape enrichment studies above, the performance of Xfp (Xfp10) was not significantly increased by extending the atom pair count up to 20 bonds (Xfp20), confirming the choice of a shorter fingerprint. Among the other 2D-fingerprints, the next best performers were CATS and MQN, which gave similar values of 74.1 ± 10%. The pure 3D-shape fingerprints USR and PMIfp and the 2D-shape fingerprint APfp20 performed comparably, followed by ECfp4, while SMIfp and Sfp showed the lowest AUC values.

The AUC values obtained with the different fingerprints were significantly correlated in the case of the shape fingerprints PMIfp, USR, and APfp20, as well as for the pairs MQN-SMIfp, Sfp-ECfp4, and Xfp-CATS, which reflected the similarity between fingerprint types (Figure 8B). The performance of Xfp and the other fingerprints in recovering pharmacophore analogs was only marginally affected by molecular shape as measured by the PMI triangle, as expected from the fact that the ROCS Color Tanimoto focuses on pharmacophore features rather than pure shape recognition (Figure 8C). Interestingly there was a slight correlation between recovery efficiency and the average ROCS Color Tanimoto similarity of the first 100 neighbors for all the better performing fingerprints (MQN, ECfp4, Xfp, CATS), which can be explained by the fact that

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

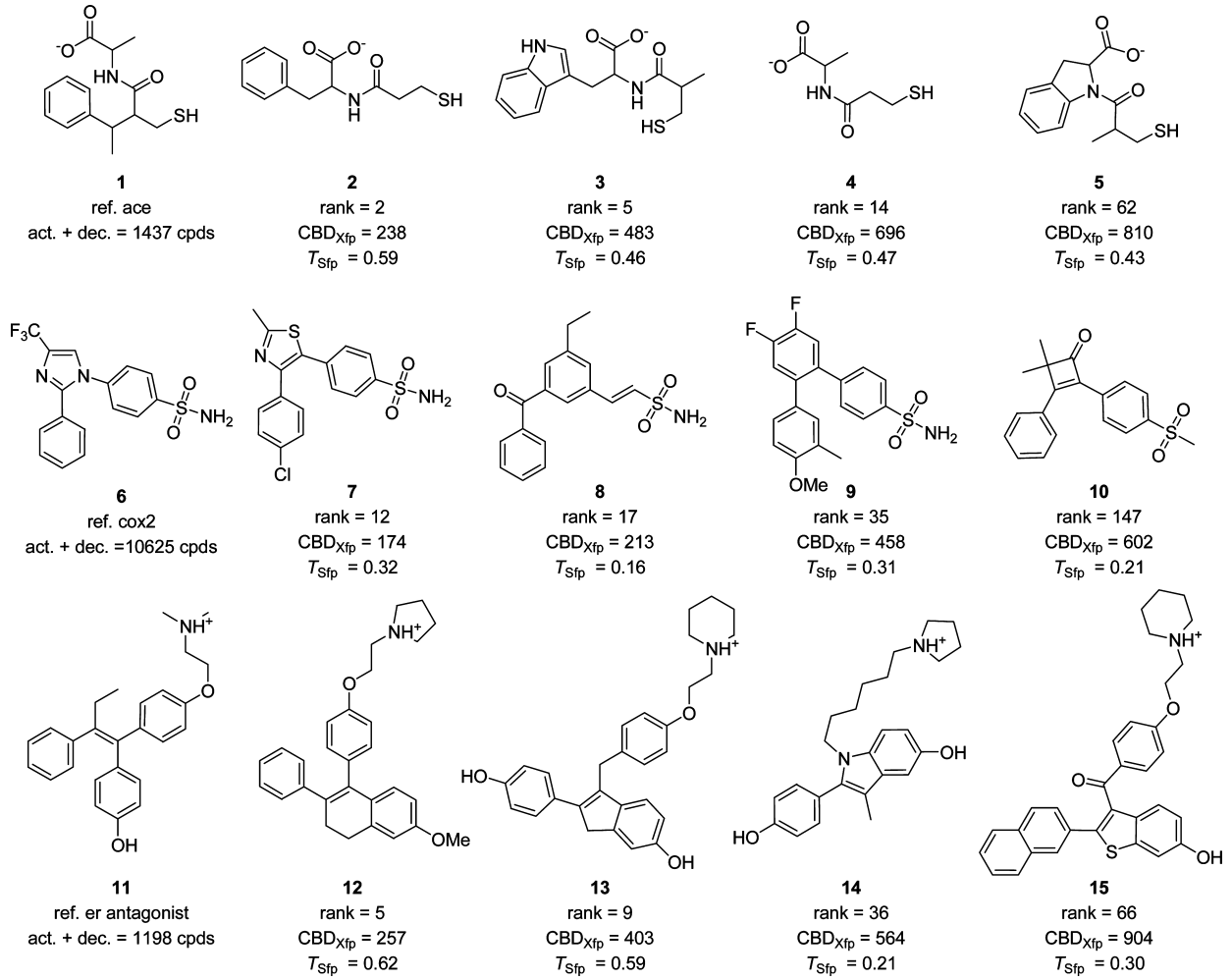


Figure 7. Structures of selected DUD actives of angiotensin converting enzyme (ace) inhibitors, cyclooxygenase-2 (cox2) inhibitors, and estrogen receptor antagonists (er antagonist), with their rank in the CBD_{Xfp} sorted list, CBD_{Xfp} and T_{Sfp} values relative to reference structure (**1**, **6**, and **11**) respectively. act. + dec is the number of unique molecules in the corresponding DUD set (actives + decoys) after removing stereochemical information.

high pharmacophore similarity is on average higher when molecules are otherwise structurally similar (Figure 8D). However, the effect was less pronounced than that observed with the pure shape recognition study above (Figure 4). Taken together, the recovery of ROCS Color Tanimoto analogs showed that Xfp was particularly well suited as a 2D-pharmacophore fingerprint.

Comparison of APfp and Xfp Similarity in ZINC with Other Fingerprints. To additionally test whether APfp (as APfp20) and Xfp (as Xfp10) offered interesting virtual screening possibilities in comparison to other fingerprints, the 10,000 nearest neighbors (0.073% of ZINC) of 15 drugs of different shapes and sizes were retrieved from the entire ZINC database using APfp, Xfp, the 3D-fingerprints PMIfp, USR, and the 2D-fingerprints CATS, MQN, SMIfp, Sfp, and ECfp4. The nine series of 10,000 nearest neighbors and 10,000 compounds randomly selected from ZINC were evaluated for their similarity to their parent drug using the ROCS Tanimoto Comboscore, which combines the ROCS shape and ROCS pharmacophore measures discussed above and is well validated for 3D-comparisons of ligands in terms of similar bioactivity potential.[5,55] The compounds were classified as Combo hits (ROCS Tanimoto Comboscore ≥1.4), scaffold-hopping (SH) virtual hits (ROCS Tanimoto Comboscore ≥1.4 AND Tan_{Sfp} < 0.5), and nonscaffold-hopping (non-SH) virtual hits (ROCS Tanimoto Comboscore ≥1.4 AND Tan_{Sfp} ≥ 0.5).

The number of Combo and SH hits from each series of 10,000 nearest neighbors was strongly influenced by molecule size, with the smallest drugs such as nicotine giving the largest number of hits, indicating that ZINC contained many close analogs for these smaller molecules, but was more sparsely populated at larger molecule size (Table 1). Across the 15 drugs APfp gave remarkably high hit rates (Combo: 9.2%, SH: 8.9%) probably reflecting its ability to sense the molecular shape encoded in ROCS. The performance of Xfp was lower (Combo: 6.8%, SH: 5.5%) and comparable to that of the substructure fingerprints Sfp and ECfp4, although the SH hit rate of Xfp (5.5%) was higher than that of Sfp (3.0%) and ECfp4 (4.9%). The hit rates were not correlated with fingerprint sophistication and type, with remarkably high hit rates for both MQN (Combo: 13.5%, SH: 12.3%) and SMIfp (Combo: 9.1%, SH: 7.6%) despite the fact that these fingerprints do not consider substructures and only count isolated features (atoms, bonds, cycles, SMILES characters), while the lowest hit rates occurred with the most sophisticated 3D-shape fingerprints PMIfp (Combo: 4.2%, SH: 3.9%) and USR (Combo: 4.3%, SH: 4.1%).

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

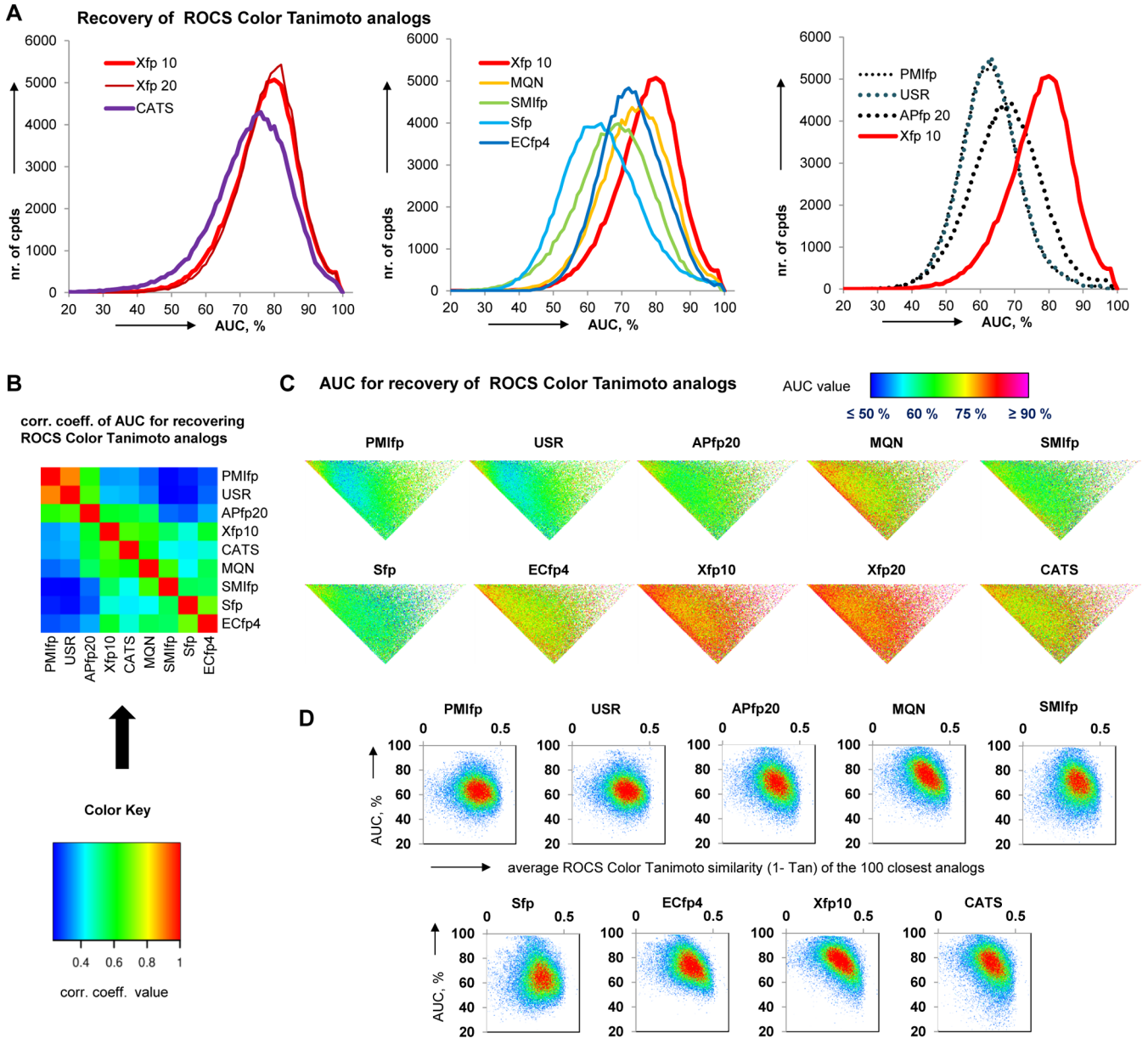


Figure 8. Recovery statistics of 100 closest analogs according to ROCS Color Tanimoto (3D-pharmacophore) using various fingerprints, for each of 145,000 molecules in CSD from their size-constrained subsets (all CSD molecules within HAC = query ± 2). (A) Frequency histogram of AUC values. (B) Correlation coefficients of AUC values from different fingerprints across all 145,000 molecules. (C) Average AUC value as a function of position in the shape triangle, continuous color scale: AUC ≤ 50%: blue; 58%: cyan; 66%: green; 75%: yellow; 80%: red; ≥90%: magenta. (D) Occupancy heat map of scatter plot of AUC value as a function of the average ROCS Color Tanimoto similarity of the 100 nearest neighbors with their size-constrained subset of CSD (HAC = query ± 2).

Despite their comparable performance in terms of hit rates, the different fingerprints all retrieved substantially different lists of analogs. Thus, the pairwise overlap between the 150,000 nearest neighbors retrieved by each of the nine fingerprints was on average less than 25%, and each fingerprint featured approximately 50% of unique nearest neighbors not found by any of the other fingerprints (Figure 9A). Each fingerprint also featured 25−40% of unique SH hits, with usually less than 25% pairwise overlap with the other fingerprints (Figure 9B). On the other hand the vast majority (>80%) of non-SH hits identified by the various fingerprints were expectedly also found among the Sfp-nearest neighbors (Figure 9C). In this focused non-SH category the hits from Xfp, CATS, MQN, and SMIfp also contained the majority (50−80%) of the hits identified by the other fingerprints except Sfp and ECfp4. By contrast the non-SH hits from the pure shape fingerprints PMIfp, USR, and APfp only contained a significant fraction of each others' hits (up to 60%), reflecting the fact that these fingerprints do not include atom-type information.

The average recovery of all ROCS Comboscore hits identified by the various fingerprints as a function of CBD_{APfp} showed that APfp had a preference for SH hits, with recovery of 34% of the SH hits and 16% of the non-SH hits up to CBD_{APfp} = 6, corresponding to 0.2% coverage of ZINC. Extending to CBD_{APfp} = 10 recovered 55% of the SH-hits and 35% of the non-SH hits with 1.1% coverage of ZINC (Figure 9D). Xfp showed significantly higher recoveries than APfp at much lower coverage of ZINC, with recovery of 36% of the non-SH hits

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. LBVS for ROCS Tanimoto Comboscore Analogs in ZINC[a]

drug	MW	PMIfp	USR	APfp	Xfp	CATS
nicotine	162.1	1431/1466	1680/1722	3713/3763	2187/2402	2506/2688
levetiracetam	170.2	597/641	500/521	1043/1085	908/1137	873/1013
gabapentin	171.2	613/705	1207/1270	1112/1147	821/1232	2135/2427
memantine	179.3	517/644	294/350	601/742	494/806	273/518
rimantadine	179.3	1007/1156	763/849	732/867	1410/1815	1741/2018
dexmedetomidine	200.3	240/244	194/197	307/309	185/190	239/243
varenicline	211.3	164/165	52/55	454/457	208/213	266/272
fencamfamine	215.3	1088/1115	1215/1243	4968/5025	1471/1787	2670/3018
acyclovir	225.2	5/8	14/18	21/28	111/154	85/117
aminoglutethimide	232.3	118/119	104/106	161/167	238/284	250/280
oseltamivir	312.4	0/1	0/1	0/3	0/4	0/5
oxycodone	315.4	0/8	1/20	1/43	1/40	1/46
esomeprazole	345.4	27/29	62/71	159/163	123/149	266/298
quetiapine	383.5	1/4	2/7	17/24	31/40	22/32
imatinib	493.6	0/3	0/6	0/8	1/15	0/10
av % ROCS Combo >1.4		4.2%	4.3%	9.2%	6.8%	8.7%
av % SH (Tan_{Sfp} < 0.5)		3.9%	4.1%	8.9%	5.5%	7.6%
drug	MW	MQN	SMIfp	Sfp	ECfp4	Rand.
nicotine	162.1	7398/7582	4535/4976	727/2300	2032/2740	123/124
levetiracetam	170.2	1170/1346	653/842	0/412	619/951	24/25
gabapentin	171.2	1369/1687	1183/1630	388/1140	1110/1560	10/10
memantine	179.3	1322/1699	580/887	419/880	576/859	4/4
rimantadine	179.3	2459/2951	690/1079	670/1340	1403/1835	14/15
dexmedetomidine	200.3	636/647	857/872	124/144	81/90	13/13
varenicline	211.3	460/466	224/230	47/57	175/184	6/6
fencamfamine	215.3	2942/3105	2106/2384	246/1381	743/896	140/141
acyclovir	225.2	150/193	123/193	174/311	59/123	1/1
aminoglutethimide	232.3	261/286	40/59	0/103	193/368	0/0
oseltamivir	312.4	0/5	0/5	1/7	1/7	0/0
oxycodone	315.4	1/45	1/39	1/63	0/48	0/0
esomeprazole	345.4	219/240	436/481	1612/1836	318/418	16/16
quetiapine	383.5	7/15	4/13	119/140	67/85	1/1
imatinib	493.6	0/8	0/10	1/17	1/16	0/0
av % ROCS Combo >1.4		13.5%	9.1%	6.8%	6.8%	0.2%
av % SH (Tan_{Sfp} < 0.5)		12.3%	7.6%	3.0%	4.9%	0.2%

[a]The 10,000 nearest neighbor of each indicated drug using $CBD_{fingerprint}$ as similarity measure were retrieved. The number of SH hits (ROCS Tanimoto Comboscore >1.4 AND Tan_{Sfp} < 0.5)/Combo hits (ROCS Tanimoto Comboscore >1.4) found among these 10,000 nearest neighbors is indicated. The average percentage of Combo hits and SH hits across all 15 drugs is indicated in the last two lines.

and 15% of the SH hits up to CBD_{Xfp} = 33, corresponding to only 0.02% coverage of ZINC. Extending to CBD_{Xfp} = 59 (only 0.35% coverage of ZINC) recovered 50% of the SH-hits and 70% of the non-SH hits (Figure 9E). This performance was slightly better than for the recovery using CBD_{MQN}, which retrieved 20% of non-SH hits and 14% of SH hits up to CBD_{MQN} = 13 corresponding to a database coverage comparable to Xfp (0.02% of ZINC), while extending to CBD_{MQN} = 23 recovered 49% of SH hits and 50% of non-SH hits at slightly higher 0.56% coverage of ZINC (Figure 9F).

The fact that Xfp recovered the high-substructure similarity, non-SH hits more efficiently than APfp or MQN reflects the fact that Xfp perceives substructural elements better than APfp or MQN. This better recognition of substructures by Xfp is probably caused by the association of atom pair topological distances, which only perceive shape as in APfp, with atom types, which only perceive overall features but not their relative position, as in MQN. In particular atom-pair topological distances in association with atom types allow distinguishing regioisomeric relationships not recorded by APfp or MQN. Overall this analysis of fingerprint nearest neighbors indicated that APfp and Xfp offered unique collections of hits with good recovery of ROCS Tanimoto Comboscore analogs from ZINC.

Rapid Virtual Screening of Large Databases Using the APfp and Xfp Browsers. Considering that both APfp and Xfp showed very good performance in the shape and pharmacophore similarity studies and, for Xfp, in separating ligands from decoys in the DUD recovery, in particular outperforming the MQN-similarity search, we set out to format ZINC and GDB-17 for fast APfp- and Xfp-similarity searching in form of web-based APfp- and Xfp-browsers similar to the MQN-browser previously reported for various databases.[46,60,64,65] Indeed, while 2D-fingerprint similarity scoring is usually much faster than scoring with 3D methods such as ROCS due to the fact that multiple conformers must not be precomputed, we have shown recently, at the example of MQN and SMIfp similarity sorting, that searching of very large databases can be further accelerated if the fingerprints are precomputed and CBD is used as similarity function because the databases can be preorganized for fast searching using the sum of all bit values as hash function.[60,64,65]

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

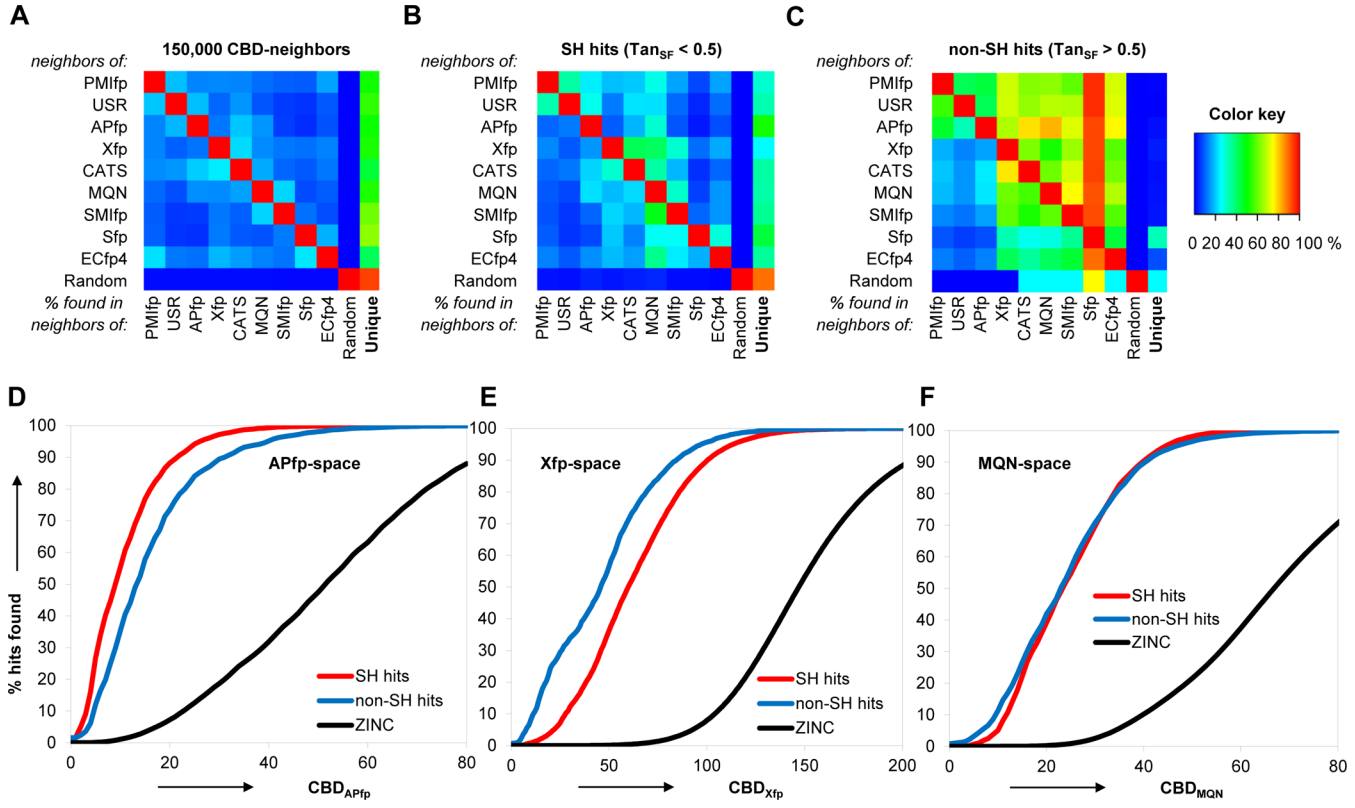


Figure 9. ROCS Comboscore properties of fp-nearest neighbors in ZINC. The nearest neighbors (NN) of each of the 15 drugs in Table 2 (15 × 10,000 in each fingerprint space) were classified as SH hits (ROCS Tanimoto Comboscore >1.4 AND Tan$_{Sfp}$ < 0.5) and non-SH hits (ROCS Tanimoto Comboscore >1.4 AND Tan$_{Sfp}$ > 0.5). (A) Percentage NN of one fp found by another fp and percentage NN unique to this fp. (B) Percentage SH-hits of one fp found by another fp and percentage SH-hits unique to this fp. (C) Percentage non-SH hits of one fp found by another fp and percentage non-SH hits unique to this fp. (D−F) Cumulated recovery of ZINC, SH-hits, and non-SH hits from the reference drugs as a function of CBD$_{APfp}$, CBD$_{Xfp}$, and CBD$_{MQN}$. CBD$_{APfp}$ and CBD$_{Xfp}$ values are computed from the fingerprints containing percentage values divided by 10 and rounded to unity as detailed in the "browser " section below.

Table 2. Bin Occupancy Statistics for the Different Fingerprints[a]

database (size)	CSD (145 k)		ChEMBL (1.1 M)		ZINC (13.6 M)		GDB-17 (50 M)	
fingerprint	av	max	av	max	av	max	av	max
APfp (100)[b]	1.33	121	1.47	138	2.42	1365	52.61	10959
APfp (10)[b]	1.41	121	1.64	236	3.45	1365	140.04	79063
Xfp (100)[b]	1.05	31	1.09	28	1.17	182	1.04	51
Xfp (10)[b]	1.05	31	1.10	34	1.21	182	1.07	256
MQN	1.07	18	1.22	30	1.93	542	1.60	171
SMIfp	1.10	14	1.23	34	2.48	994	13.50	22956
Sfp	1.03	29	1.03	43	1.03	111	1.01	404
ECfp4	1.01	21	1.02	60	1.01	31	1.00	2

[a]The number of molecules per fingerprint bit combination (bin) is given as the average across all bins (av) and the number of molecules in the highest occupied bin (max). The 50 M subset of GDB-17 was analyzed. [b]APfp (100) is the 20-bit APfp expressed in percentages rounded to unity. Xfp (100) is the 55-bit Xfp expressed in percentages rounded to unity. APfp (10) and Xfp (10) are the corresponding fingerprints with values divided by 10 and rounded to unity.

The APfp and Xfp values were calculated and sorted to enable a similar fast searching for APfp and Xfp neighbors in ZINC and GDB-17. To simplify the hash function, the bit values originally expressed in percents were divided by 10 and rounded to the integer value. This operation reduced the size of the information to be stored, and more importantly the number of total sum values to be considered such that database searching would be accelerated. The distribution of compounds in individual bit-value combinations (bins) was only marginally affected by the rounding operation. The 20-bit APfp was similar to our previous scalar integer value fingerprints MQN and SMIfp by giving relatively high average and maximum bin occupancy values, reflecting the coarse nature of the fingerprint. On the other hand Xfp, although having only 55 bits and values rounded to the 10s, gave a very good resolution of molecules into individual bins, which was almost as good as for the 1024-bit binary fingerprints Sfp and ECfp4, reflecting the much more detailed nature of the fingerprint (Table 2).

The APfp and Xfp searchable ZINC (13.5 M cpds) and GDB-17 (50 M cpd random subset) were obtained from the APfp (10)/Xfp (10) data containing bit-values rounded to the tens. Searches for a typical set of 10,000 nearest neighbors were

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

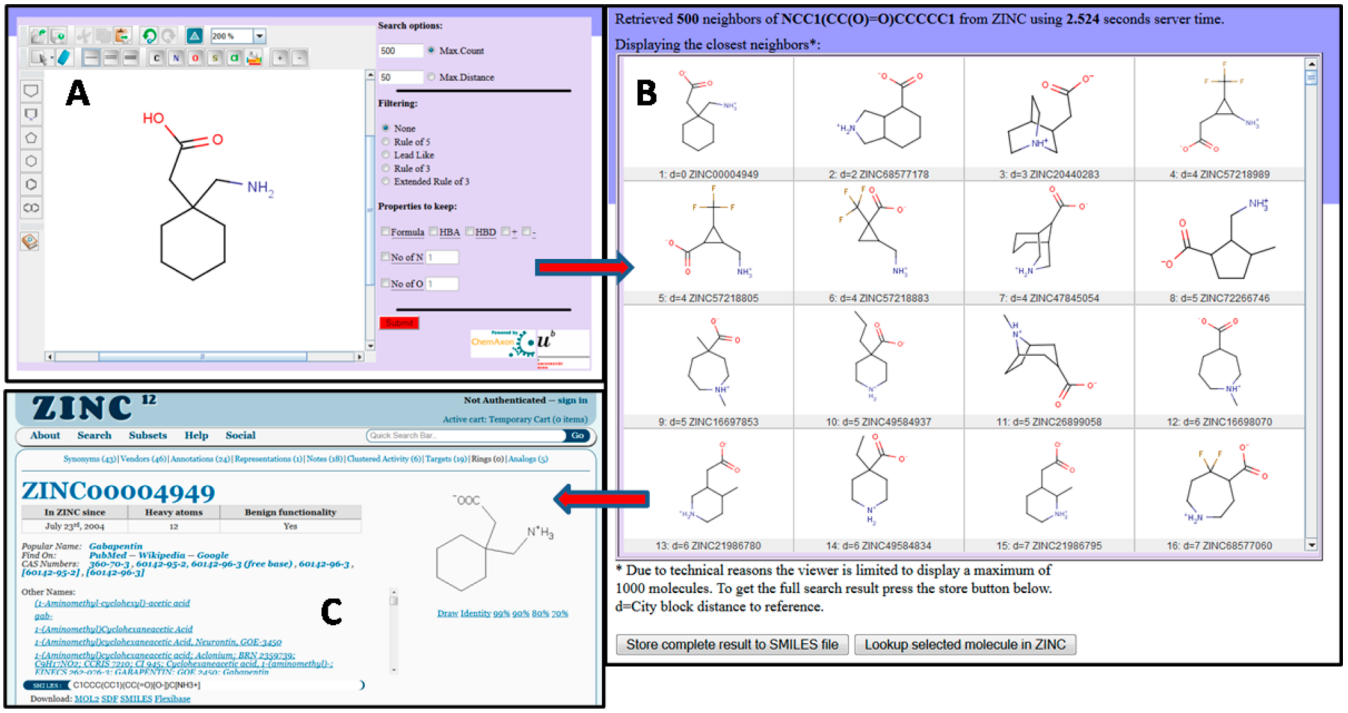


Figure 10. Xfp-browser for ZINC. (A) Molecule drawing window to set the query molecule, shown with gabapentin. (B) Results panel showing Xfp-nearest neighbors of gabapentin in ZINC. (C) ZINC Web site connected from the results panel using the "lookup selected molecules in ZINC" option.

completed on average within 2.4 s/11 s for ZINC and 7.5 s/39 s for GDB-17, with some variation in speed depending on molecule size and complexity (Table S11 in the SI). For searching ZINC (GDB-17), the APfp-browser offered an acceleration of 54-fold (46-fold) compared to scanning the entire list of precomputed APfp-values, and 2600-fold (1900-fold) compared to computing the APfp-values on the fly. The acceleration by the Xfp-browser was 16-fold (15-fold) over to scanning the entire list of precomputed Xfp-values of ZINC (GDB-17), and 2100-fold (1600-fold) over computing the Xfp-values on the fly. By comparison the average search times for the corresponding MQN-nearest neighbor searches in ZINC (GDB-17) were 0.6 s (2.7 s) average search time representing an acceleration of 270-fold (180-fold) over precomputed MQN-values and 90,000-fold (70,000-fold) over computing the MQN-values on the fly. The more modest search time accelerations achieved by the APfp- and Xfp-browsers compared to the MQN-browser probably reflects the higher dimensionality of these fingerprints, as well as the fact that the computation of APfp and Xfp values (1.7 and 6.5 h for ZINC, 3.9 and 16.8 h for GDB-17) was substantially faster than for MQN values (14.1 h for ZINC, 51.3 h for GDB-17, computation performed on a 6 × 1.8 GHz core with 16 GB RAM).

The APfp- and Xfp-browsers for ZINC and GDB-17 are available online at www.gdb.unibe.ch with a user interface comprising a drawing window and the option to search up to a preset number of CBD nearest neighbors or a preset CBD value. The display is limited to a maximum of 1000 molecules and the download to 10,000 molecules to avoid stalling of the Internet browser. Additional search criteria to focus search results include compliance to Lipinski's rule of five,[66] Oprea's lead-likeness,[67] and Congreve's rule of three and extended rule of three criteria,[68] as well as the possibility to lock the elemental formula (isomer search), the number of HBA, HBD, positive and negative charges, and to select for a preset number of N or O atoms. These options can be used in particular to implement simple pharmacophore criteria into the APfp search, and to enforce electrostatic charge information which is not encoded in these fingerprints. These interactive browsers provide a straightforward method to rapidly interrogate ZINC and GDB-17 for shape and pharmacophore analogs of any molecule of interest. The search for Xfp-nearest neighbors of the drug gabapentin in ZINC is shown as an example (Figure 10).

CONCLUSION

In summary, the above study showed that the simple atom pair fingerprint (APfp) counting atom pairs at topological distances up to 10−20 bonds in 2D-structures perceives significant features of the 3D-shape of molecules as encoded by the PMIfp, USR, and ROCS shape similarity scoring functions. The correlation of APfp to these 3D-measures was stronger than with other 2D-fingerprints, for both the correlation of distances between random pairs and in the recovery of 3D-shape analogs using 2D-fingerprints, and occurred across the entire shape triangle. Extending APfp with only four atomic properties gave a 55-dimensional pharmacophore fingerprint Xfp which performed better than other 2D-fingerprints in recovering active in DUD from the decoys and from the entire ZINC database. Xfp also showed good performance for recovering ROCS pharmacophore analogs from both CSD and ZINC. APfp and Xfp similarity searching was enabled in freely accessible web-based APfp and Xfp browsers for ZINC and GDB-17 requiring only seconds per search. These APfp- and Xfp-browsers should provide a useful support for drug discovery projects as rapid substitutes for 3D-shape and 3D-pharmacophore similarity searching. Several application examples using these browsers for ligand discovery projects are

dx.doi.org/10.1021/ci500232g | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

currently in progress in our laboratory and will be reported in due course.

METHODS

Databases. ChEMBL (https://www.ebi.ac.uk/chembl/), ZINC (https://docking.org/), subset of GDB17 (www.gdb.unibe.ch) and DUD (http://dud.docking.org/) databases were downloaded respectively from database Web sites. CSD was copied from a licensed CD to Dr. Jürg Hauser, University of Bern. Counter ions were removed, and ionization states of molecules were adjusted to pH 7.4, using an in-house built java program utilizing Java Chemistry library (JChem) from ChemAxon, Ltd., as a starting point.

APfp, Xfp and Other Fingerprints. APfp and Xfp were generated using an in-house built java program. Computation of shortest topological distances between atom pairs were achieved using the JChem tool TopologyAnalyserPlugin. APfp and Xfp were constructed by counting atom pairs at increasing topological distance, followed by normalization with the heavy atom count or the category atom count. The ratio values were then multiplied by 100 and rounded to the integer value.

MQN and SMIfp were calculated using the previously reported source code written in Java.[58,60] For the substructure fingerprint Sfp, a daylight type 1024-bit hash fingerprint was computed using JChem library. The extended connectivity fingerprint (ECfp4) was calculated with bond diameter of 4 and contained 1024 bits. The simplified fingerprints sMQN, sSMIfp, sSfp, and sECfp4 were obtained by converting each molecule to its corresponding saturated hydrocarbon graph by changing all heavy atoms to carbon and all bonds to carbon−carbon single bonds, and calculating the fingerprint (MQN, SMIfp, Sfp or ECfp4) for that compound.

3D-Shape Analysis. The 3D structures of molecules was taken from the experimental 3D-coordinates (all calculation with CSD molecules), determined using CORINA[69] (USR and PMIfp calculation on the entire ZINC), or the OpenEye Omega tool (ROCS Tanimoto Comboscore on selected compounds from ZINC).

PMIfp calculation were adopted from Sauer and Schwarz[4] and was written in Java. Molecules were oriented in their principal axis, and moments of inertia were then calculated for each of the principal axes using formulas:

$$I_x = mr_x^2$$

$$I_y = mr_y^2$$

$$I_z = mr_z^2$$

in which the squares of the radii around the axes are defined as $r_x^2 = y^2 + z^2$, $r_y^2 = x^2 + z^2$, and $r_z^2 = x^2 + y^2$. The moments of inertia I_x, I_y, and I_z were then sorted in ascending order to yield I_1, I_2, and I_3. Scaling these values to the molecular weight (MW) defines three dimensions of the PMIfp used in this paper. Division of I_1 and I_2 by the highest moment of inertia I_3 results in the values $P1 = I_1/I_3$ and $P2 = I_2/I_3$, which defines the two dimensions of nPMIfp. Plotting of the P1 and P2 for molecules results in typical triangular plot, in which rod- (0, 1), disc- (0.5, 0.5) and sphere- (1, 1) like molecules occupy three distinct edges of triangle (see main text).

USR (Ultrafast Shape Recognition) fingerprint calculation was facilitated by java source code obtained from the Chemistry Development Tool Kit (CDK) and based on the work published by Ballester et al.[53] Following the generation of 3D structures as above, Euclidean distances were computed for atoms in a molecule with respect to four chosen reference points (centroid (ctd), closet atom to centroid (cst), the furthest atom to ctd (fct), and farthest atom to fct (ftf)). Each of the four distance distributions is then represented by three statistical moments namely, average, standard deviation, and kurtosis. This results in 12 dimensional USR (4 * 3 moments) shape fingerprint for the molecule.

ASSOCIATED CONTENT

Supporting Information

Tables S1−S10 for AUC and EF values and Figures S1−S6 for ROC curves of the recovery of DUD actives from decoys and from ZINC. Table S11 for search time statistics. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jean-louis.reymond@dcb.unibe.ch. FAX: +41 31 631 80 57.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported financially by the University of Berne, the Swiss National Science Foundation and the NCCR TransCure. We thank OpenEye Scientific Software and ChemAxon Pvt. Ltd. for free academic and web licenses for their products.

REFERENCES

(1) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

(2) Renner, S.; Popov, M.; Schuffenhauer, A.; Roth, H. J.; Breitenstein, W.; Marzinzik, A.; Lewis, I.; Krastel, P.; Nigsch, F.; Jenkins, J.; Jacoby, E. Recent trends and observations in the design of high-quality screening collections. *Future Med. Chem.* **2011**, *3*, 751−766.

(3) Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2011**, *2*, 349−355.

(4) Sauer, W. H.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987−1003.

(5) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(6) Venhorst, J.; Núñez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* **2008**, *51*, 3222−3229.

(7) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: Choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49*, 678−692.

(8) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: A perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.

(9) Ebalunode, J. O.; Zheng, W. Molecular shape technologies in drug discovery: Methods and applications. *Curr. Top. Med. Chem.* **2010**, *10*, 669−679.

(10) Perez-Nueno, V. I.; Ritchie, D. W. Using consensus-shape clustering to identify promiscuous ligands and protein targets and to

choose the right query for shape-based virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 1233−1248.
(11) Kim, S.; Bolton, E. E.; Bryant, S. H. PubChem3D: Conformer ensemble accuracy. *J. Cheminform.* **2013**, *5*, 1−17.
(12) Wirth, M.; Volkamer, A.; Zoete, V.; Rippmann, F.; Michielin, O.; Rarey, M.; Sauer, W. H. Protein pocket and ligand shape comparison and its application in virtual screening. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 511−524.
(13) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: Increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52*, 6752−6756.
(14) Ritchie, T. J.; Macdonald, S. J.; Young, R. J.; Pickett, S. D. The impact of aromatic ring count on compound developability: Further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discovery Today* **2011**, *16*, 164−171.
(15) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−1508.
(16) Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.
(17) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(18) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(19) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(20) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.
(21) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(22) Olender, R.; Rosenfeld, R. A fast algorithm for searching for molecules containing a pharmacophore in very large virtual combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 731−738.
(23) Klebe, G. Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−594.
(24) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504−1519.
(25) Kolb, P.; Ferreira, R. S.; Irwin, J. J.; Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* **2009**, *20*, 429−36.
(26) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.
(27) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30−38.
(28) Reymond, J. L.; Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **2012**, *3*, 649−657.
(29) Yu, M. J. Druggable chemical space and enumerative combinatorics. *J. Cheminform.* **2013**, *5*, 19−30.
(30) Virshup, A. M.; Contreras-Garcia, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(31) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(32) Blum, L. C.; van Deursen, R.; Bertrand, S.; Mayer, M.; Burgi, J. J.; Bertrand, D.; Reymond, J. L. Discovery of alpha7-nicotinic receptor ligands by virtual screening of the chemical universe database GDB-13. *J. Chem. Inf. Model.* **2011**, *51*, 3105−3112.
(33) Burgi, J. J.; Awale, M.; Boss, S. D.; Schaer, T.; Marger, F.; Viveros-Paredes, J. M.; Bertrand, S.; Gertsch, J.; Bertrand, D.; Reymond, J. L. Discovery of potent positive allosteric modulators of the alpha3beta2 nicotinic acetylcholine receptor by a chemical space walk in ChEMBL. *ACS Chem. Neurosci.* **2014**, *5*, 346−359.
(34) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Discovery of NMDA glycine site inhibitors from the chemical universe database GDB. *ChemMedChem.* **2008**, *3*, 1520−1524.
(35) Luethi, E.; Nguyen, K. T.; Burzle, M.; Blum, L. C.; Suzuki, Y.; Hediger, M.; Reymond, J. L. Identification of selective norbornane-type aspartate analogue inhibitors of the glutamate transporter 1 (GLT-1) from the chemical universe generated database (GDB). *J. Med. Chem.* **2010**, *53*, 7236−7250.
(36) Garcia-Delgado, N.; Bertrand, S.; Nguyen, K. T.; van Deursen, R.; Bertrand, D.; Reymond, J.-L. Exploring a7-nicotinic receptor ligand diversity by scaffold enumeration from the chemical universe database GDB. *ACS Med. Chem. Lett.* **2010**, *1*, 422−426.
(37) Brethous, L.; Garcia-Delgado, N.; Schwartz, J.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Synthesis and nicotinic receptor activity of chemical space analogues of *N*-(3*R*)-1-azabicyclo[2.2.2]oct-3-yl-4-chlorobenzamide (PNU-282,987) and 1,4-diazabicyclo[3.2.2]nonane-4-carboxylic acid 4-bromophenyl ester (SSR180711). *J. Med. Chem.* **2012**, *55*, 4605−4618.
(38) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(39) Ebalunode, J. O.; Zheng, W. Unconventional 2D shape similarity method affords comparable enrichment as a 3D shape method in virtual screening experiments. *J. Chem. Inf. Model.* **2009**, *49*, 1313−1320.
(40) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103−1013.
(41) Kalaszi, A.; Szisz, D.; Imre, G.; Polgar, T. Screen3D: A novel fully flexible high-throughput shape-similarity search method. *J. Chem. Inf. Model.* **2014**, *54*, 1036−1049.
(42) Koutsoukas, A.; Paricharak, S.; Galloway, W. R.; Spring, D. R.; Ijzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model.* **2014**, *54*, 230−242.
(43) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(44) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(45) Khalifa, A. A.; Haranczyk, M.; Holliday, J. Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.* **2009**, *49*, 1193−1201.
(46) Awale, M.; Reymond, J. L. A multi-fingerprint browser for the ZINC database. *Nucleic Acids Res.* **2014**, DOI: 10.1093/nar/gku379.
(47) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic surface representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787−1796.
(48) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(49) Randic, M. Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607−613.
(50) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small molecule shape-fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673−684.

(51) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536−1548.
(52) Firth, N. C.; Brown, N.; Blagg, J. Plane of best fit: A novel method to characterize the three-dimensionality of molecules. *J. Chem. Inf. Model.* **2012**, *52*, 2516−2525.
(53) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.
(54) Schreyer, A. M.; Blundell, T. USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminform.* **2012**, *4*, 27−39.
(55) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(56) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(57) Hagadone, T. R. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521.
(58) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of organic molecules by molecular quantum numbers. *ChemMedChem.* **2009**, *4*, 1803−1805.
(59) van Deursen, R.; Blum, L. C.; Reymond, J. L. A searchable map of PubChem. *J. Chem. Inf. Model.* **2010**, *50*, 1924−1934.
(60) Schwartz, J.; Awale, M.; Reymond, J. L. SMIfp (SMILES fingerprint) chemical space for virtual screening and visualization of large databases of organic molecules. *J. Chem. Inf. Model.* **2013**, *53*, 1979−1989.
(61) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.
(62) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.
(63) Venkatraman, V.; Perez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079−2093.
(64) Blum, L. C.; van Deursen, R.; Reymond, J. L. Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637−647.
(65) Ruddigkeit, L.; Blum, L. C.; Reymond, J. L. Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2013**, *53*, 56−65.
(66) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(67) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743−3748.
(68) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.
(69) Sadowski, J.; Gasteiger, J. From atoms and bonds to 3-dimensional atomic coordinates - Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.