

Exploring Peptide-likeness of Active Molecules Using 2D Fingerprint Methods

Hanna Eckert and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität,
Dahlmannstrasse 2, D-53113 Bonn, Germany

Received March 5, 2007

Similarity searching for peptide-like small molecules is a difficult task because the amide backbone shared by these molecules tends to mask features that determine biological activity. We have investigated 2D fingerprints for their ability to differentiate between peptide-like molecules having different activity or to facilitate a peptidomimetic transition from molecules with strong peptide character to compounds having little or none. For these purposes, different compound activity classes were assembled consisting of molecules having strong, moderate, and weak peptide character. For the quantification of peptide character, a “peptide flavor” index was introduced. In systematic search calculations, an encouraging finding has been that most of the investigated 2D fingerprints were capable of distinguishing between peptide-like molecules having different activities. However, only two fingerprints of different design also displayed a strong tendency to detect molecules with decreasing peptide character. One of these search tools is a recently introduced property descriptor-based fingerprint that showed two additional advantages: its flexible design could be adjusted to increasingly recover molecules with little peptide-likeness, and in addition, its search performance was not affected by differences in molecular size.

1. INTRODUCTION

Chemical similarity searching^{1,2} is a standard tool in today's computational compound screening repertoire^{3–5} and aims at the enrichment of compound selection sets of varying size with active molecules. In a typical application, a small number (e.g., five) of known active reference molecules is used to identify novel molecules sharing the biological activity of interest. Usually, one does not aim at the identification of closely related active compounds or analogs but at active molecules having distinct structures.^{6,7} Peptides are generally poor drug candidates because they suffer from problems such as low oral bioavailability and short half-life, and it is often attempted to replace active peptides with molecules having little or no peptide character.^{8–11} Computational methods that were successfully applied to assist in such efforts have mostly been restricted to receptor-based^{12–15} approaches. By contrast, little data is currently available for ligand-based methods^{3,16} such as similarity searching. Therefore, we have set out to thoroughly investigate the behavior of 2D similarity fingerprints when applied to peptide-like molecules. These search tools are often the method of choice for virtual compound screening because of their computational efficiency and surprising effectiveness when, for example, compared to other more complex methods.^{5,17,18} Two-dimensional fingerprints are calculated from the 2D graph representation of molecules and usually represented as bit strings, where each bit monitors the presence or absence of a specific chemical feature or value range of a molecular descriptor. Representative state-of-the-art 2D fingerprint designs include hashed connectivity pathways,¹⁹ structural dictionary-based,^{20,21} and layered atom environment fingerprints.²²

To quantitatively assess the similarity of a database compound and an active reference molecule, a similarity measure such as the popular Tanimoto coefficient (T_c)¹ is calculated for comparison of their fingerprint representations. In typical similarity search situations, more than one active reference compound is available, and search performance can increase significantly if multiple reference compounds are used.²³ Therefore, several multiple template search strategies have been introduced,^{24–27} and nearest neighbor^{24,27} and centroid techniques²⁷ are currently most widely used.

In our current study, we focus on the evaluation of different 2D fingerprint designs and multiple template search strategies to detect active molecules having different degrees of peptide character when starting from distinctly peptide-like reference compounds. First, we analyzed the retrieval of any active peptide-like molecules to evaluate whether fingerprint searching can differentiate between peptide-like molecules having different activity or if search calculations are entirely dominated by amide backbone structures. Second, we investigated the more challenging task of detecting active molecules with decreasing peptide character. For these purposes, we designed eight activity classes that contained molecules of low, medium, and high peptide character and reference sets of molecules having strong peptide-likeness were assembled for systematic search calculations. The results obtained in our study suggest that 2D fingerprints can differentiate between peptide-like molecules of different activity but that only specific fingerprint designs can successfully detect molecules with decreasing degrees of peptide-likeness. In addition, fingerprints often prefer different search strategies, and the results are frequently biased by the molecular size-dependence of Tanimoto similarity calculations.²⁸

* To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

2. DESIGN OF THE STUDY

2.1. Quantifying the Peptide Flavor of Molecules. To classify small synthetic molecules according to different degrees of peptide character, we needed to derive a measure to objectively quantify their “peptide-likeness”. Therefore, we introduced the quantity peptide flavor (PF) that relates the number of amide or peptide-like bonds to the number of non-hydrogen atoms in test compounds. We determined the number of peptide-like bonds using the SMARTS¹⁹ string “O=[CX3H0][NX3H1]C”. This string captures the amide bond signature pattern O=CNC under the restriction that the first carbon atom must not be bonded to a hydrogen atom. Patterns matching the above SMARTS string that were part of a ring having less than nine ring atoms were not considered peptide-like bonds. In addition, overlapping SMARTS matches (e.g., within a pattern like CNC(=O)-NC) were only counted as single matches. Thus, the peptide flavor index PF was then defined by the following equation and the two special cases and rules discussed above

PF =

$$\begin{cases} 1 & \text{for HA} = 5, \text{ numPB} > 0 \\ (\text{numPB} \times 4) / (\text{HA} - 5) & \text{for HA} \neq 5 \\ 0 & \text{for HA} = 5, \text{ numPB} = 0 \end{cases}$$

HA stands for the number of heavy atoms, and numPB for the number of peptide-like bonds. PF produces values between 0 and 1 representing minimal and maximal peptide character, respectively. The factor four in the numerator of the quotient is applied because a peptide-like bond is composed of four heavy atoms. The subtrahend five is included in the denominator because five non-hydrogen atoms form the N- and C-termini of a peptide structure and do not contribute to a peptide bond. This might also apply to peptide-like small molecules. The two special cases 1 and 0 cover the possibility that a molecule has only five heavy atoms (which would lead to an illegal division by zero in the quotient). As an example for the calculation of PF, we consider a glycine tripeptide (Figure 1a) and three other compounds with increasing peptide character according to our PF scheme (Figure 1b). The glycine tripeptide contains three peptide-like bonds and a total of 17 heavy atoms, five of which are not part of the peptide bonds. The triglycine is a prototypic molecule for which the calculated PF value is 1 corresponding to maximal peptide-likeness. It should be noted that, for any peptide containing amino acids other than glycine, the PF value is smaller than 1 because of the presence of side chain atoms. This illustrates that PF was designed to only account for amide bond content as a measure of peptide-likeness and to classify small synthetic molecules. The peptide flavor index was not designed to describe peptides because it only accounts for conserved backbone elements, whereas peptides are distinguished by side-chain features. We also did not intend to include peptides in our similarity search calculations. Clearly, it would be very difficult to distinguish between natural peptides having different activity using 2D molecular representations. Our focus has exclusively been on synthetic compounds with peptide character. For small molecules, amide bond content was considered a meaningful measure of peptide character, which is illustrated in Figure 1b. The PF values for the three synthetic compounds vary between 0 (no amide bond

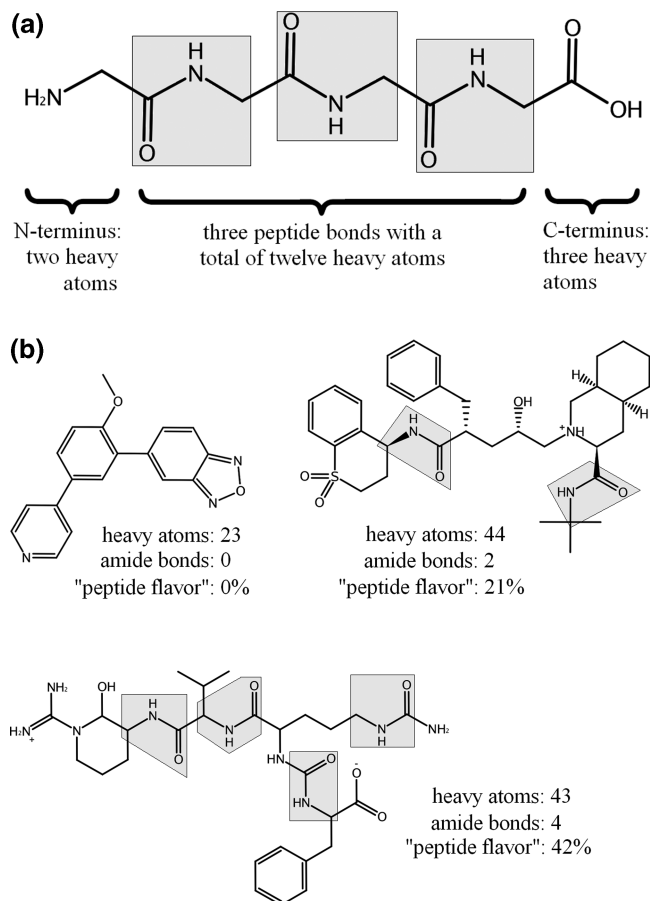


Figure 1. Defining peptide flavor. (a) A glycine tripeptide and (b) three small molecules are shown. Shaded areas contain atoms contributing to peptide-like bonds. Peptide flavor (PF) is calculated as the quotient of the number of non-hydrogen (heavy) atoms that are involved in the formation of the amide bonds (i.e., 12 in a and 0, 8, and 16, respectively, in b) and the total number of heavy atoms excluding five atoms that potentially form the amino and carboxy termini (i.e., 12 in a and 18, 39, and 38, respectively, in b). For the glycine tripeptide, the resulting PF value is 1.0 corresponding to 100% peptide flavor. For the other three compounds, the PF values are 0, 21, and 42%, respectively.

content) and 0.42, and the differences in peptide character can be intuitively appreciated.

2.2. Design of Compound Activity Classes. For our analysis, we used compound activity classes that had to meet two requirements. First, each activity class had to cover a broad PF range. Second, activity classes should not contain very similar structures or analog series, which would be expected to distort the analysis and perhaps lead to artificially good search results. Given these selection criteria, we analyzed numerous activity classes from the Molecular Drug Data Report²⁹ (MDDR) and ultimately selected eight classes that were subjected to the following assembly and refinement procedure. For each compound activity class, the corresponding MDDR activity index was identified; MDDR entries sharing this index were combined, and entries containing no structures, as well as molecules with a molecular weight of <200 or >1000, were omitted. Then, analog series were eliminated in a two-step procedure. In the first step, for each molecule containing at least two (separate or condensed) rings, core structures were isolated by automated removal of nonring substituents.³⁰ Subsequently, molecules were grouped by core structures, and for each subset, all molecules except the one with lowest

Table 1. Peptide-like Activity Classes^a

class code	biological activity	MDDR activity index	cpds	cpds PF \geq 25%	cpds PF < 25%	cpds PF \leq 10%	potential reference cpds
ACE	ACE inhibitors	31 410	108	27	81	45	23
CLG	collagenase inhibitors	78 371	98	54	44	21	53
CTP	cathepsin B inhibitors	78 413	63	40	23	3	40
HIV	HIV-1 protease inhibitors	71 523	144	37	107	73	36
MMT	matrix metalloproteinase inhibitors	78 432	220	36	184	140	35
NRI	neuronal injury inhibitors	12 452	1331	46	1285	1126	28
PRT	protease inhibitors	78 330	112	67	45	26	65
TNF	TNF inhibitors	02454	201	20	181	145	18

^a Reported are eight activity classes that were used as test sets for similarity search trials, their class codes, number of compounds per class (cpds), and peptide flavor (PF) distribution: cpds PF \geq 25%, cpds PF < 25%, and cpds PF \leq 10% stand for the number of compounds in each class having PF values greater or equal to 25%, smaller than 25%, or smaller or equal to 10%, respectively. Potential reference cpds reports how many compounds had PF values greater or equal to 25% and contained two or more peptide-like bonds. These compounds met our selection criteria for potential reference molecules.

molecular weight were removed. In the second step, for all remaining molecules, pairwise Tanimoto¹ similarities were calculated using the publicly available set of 166 MACCS structural keys,^{20,31} and molecules having a Tc value of 0.85 or greater to any other compound within the activity class were iteratively removed until all pairwise Tc values were smaller than 0.85. The resulting compound sets contained between 63 and 1331 molecules and are reported in Table 1. Figure 2 shows three exemplary compounds for each activity class and illustrates the presence of a PF spectrum within each class.

For the assembly of reference sets of compounds with distinct peptide character for similarity searching, molecules were only accepted if they contained at least two peptide-like bonds and had a PF value of equal to or greater than 0.25 (25%). For the eight activity classes, between 23 and 65 molecules satisfied these criteria, as reported in Table 1. In each case, ten reference sets of five molecules each were randomly selected.

2.3. Background Database. As a source database for similarity searching, we decided to use the MDDR²⁹ because it consists of biologically active small molecules and covers a broad range of peptide-likeness, as shown in Figure 3. Therefore, we could determine, for example, if fingerprint search methods preferentially detected molecules with significant peptide character but different biological activities or retrieved molecules having similar activity but low peptide character. Before using the MDDR as background database, we removed entries containing no structural data and applied a general molecular weight filter [200, 1500]. Since our eight activity classes were specially designed subsets of existing MDDR classes, we removed the remaining active compounds in each case. The resulting MDDR source database contained approximately 148 400 molecules.

3. METHODS AND CALCULATIONS

3.1. Fingerprints. We selected five 2D fingerprints for our study that represented different state-of-the-art fingerprint design strategies: TGD^{32,33} (consisting of 420 bits), TGT³³ (1704 bits), Daylight¹⁹ (2048 bits), Molprint2D^{22,34} (no bit representation), and PDR-FP³⁵ (500 bits). TGD is a two-point pharmacophore-type fingerprint that assigns atoms to seven different pharmacophore features and uses 15 distance ranges to monitor distances between feature pairs. These

distances are determined as the shortest connecting path (number of bonds) in the 2D molecular graph representation. TGT is a corresponding three-point pharmacophore fingerprint that captures triangles of four atomic features using graph distances divided into six distance ranges. TGD and TGT are implemented in the molecular operating environment (MOE).³³ The Daylight fingerprint records molecular connectivity pathways of varying length. Each pathway is mapped on a characteristic bit pattern using a hash function and all bit patterns are joined by a logical OR to generate the fingerprint. We used the Daylight fingerprint version that consists of 2,048 bit positions and monitors connectivity pathways of lengths zero to seven. Molprint2D consists of sets of atom environments that are derived from the connectivity table of a molecule. Atom environments are directly represented as strings and are not assigned to specific bit positions. PDR-FP encodes value ranges of 93 molecular property descriptors that were selected from a large descriptor pool because they displayed a general tendency to adopt selective value ranges for a panel of different activity classes.³⁶ Value ranges of PDR-FP descriptors are divided into nonoverlapping intervals (assigned to bit positions) such that the same number of screening database compounds falls into each interval. This procedure is called “equipotent binning” and provides a basis for class-specific training of PDR-FP, as described previously.³⁵

3.2. Search Strategies. For similarity searching, sets of five reference compounds were used, and alternative screening strategies and similarity metrics were applied. For TGD, TGT, Daylight, and Molprint2D, we calculated Tc values (with a set-theoretic interpretation for Molprint2D), in combination with the one-nearest-neighbor^{24,27} (1-NN) and centroid²⁷ strategies, to create rankings of database compounds. PDR-FP was used with its unique “frequency” approach³⁵ for multiple reference compounds that is conceptually similar to the centroid method but only applicable to PDR-FP because it requires the presence of a constant fingerprint bit density. PDR-FP is currently the only fingerprint with this feature. The 1-NN and centroid techniques are widely used search strategies for conventional 2D fingerprint designs.^{24,27,37} When applying 1-NN, the similarity score of a database compound is defined as the maximum pairwise Tc obtained for this compound against each individual reference molecule. In contrast, the centroid

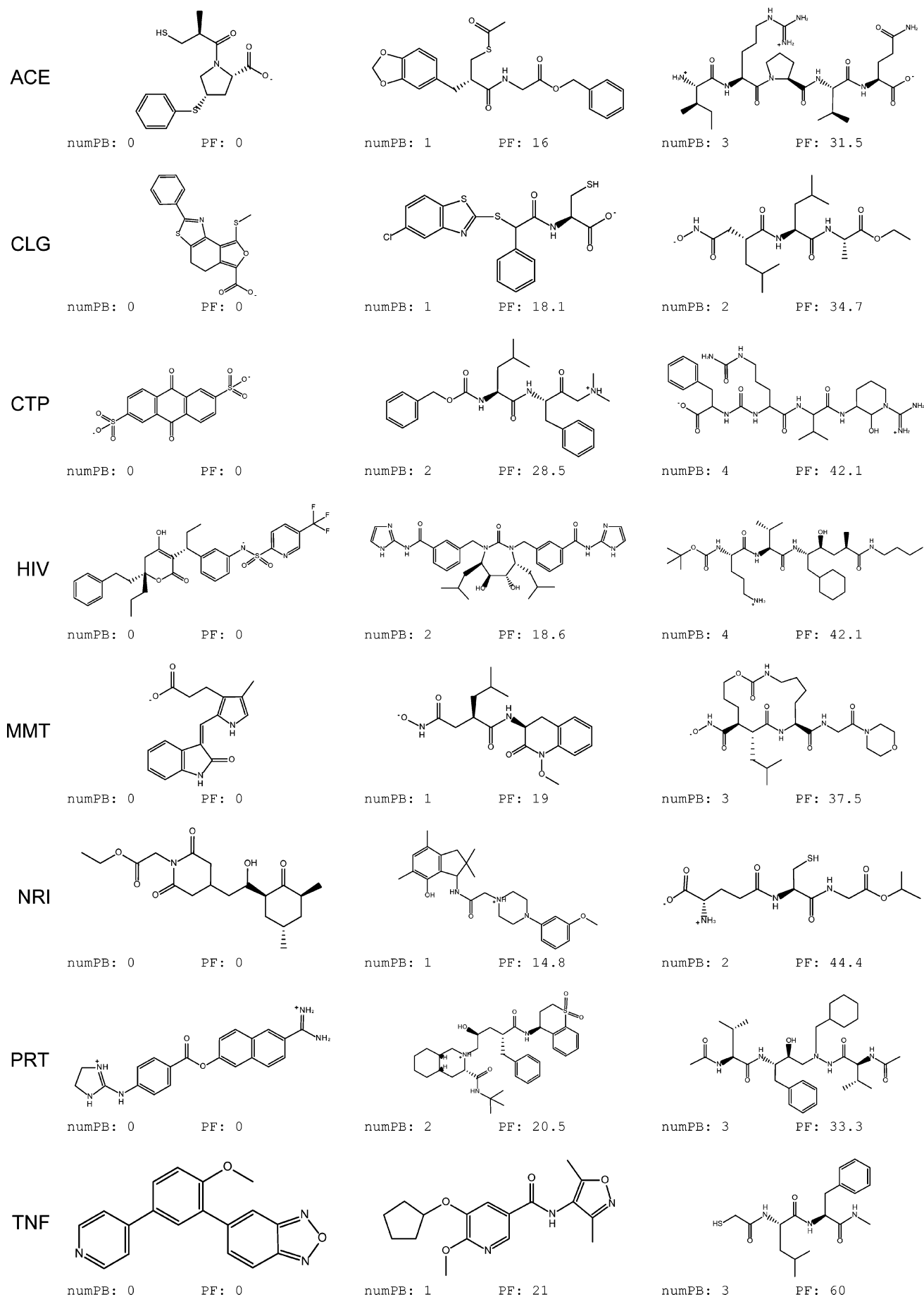


Figure 2. Representative structures. For each of the eight activity classes studied here, three representative structures are displayed side-by-side with increasing PF values (in %) and numbers of peptide-like bonds (numPB).

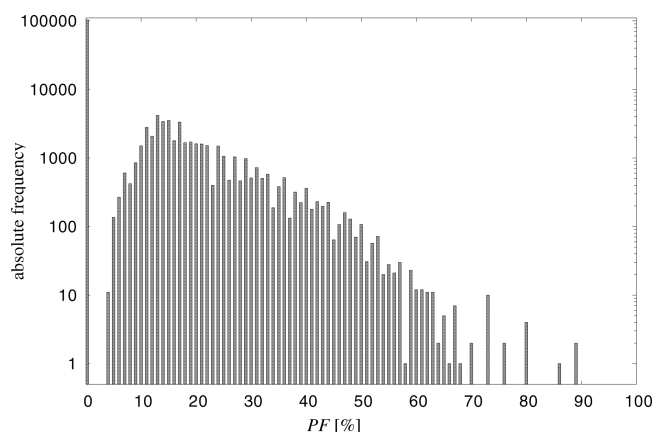


Figure 3. Distribution of PF values for MDDR database compounds. For the filtered MDDR database used in this study, absolute frequencies of PF values (in %) are reported on a logarithmic scale.

approach determines an average fingerprint of all reference molecules and compares it to the individual fingerprints of database molecules. The frequency approach designed for PDR-FP creates an activity class-specific search string by recording bit frequencies of fingerprint positions over all reference molecules. Bit positions with high-frequency values indicate activity-selective descriptor value ranges and are emphasized during similarity assessment. This is accomplished through the calculation of the dot product for the non-binary search string and the bit string of each database molecule that is then normalized to produce similarity values in the interval [0,1], analogous to Tc.

3.3. Virtual Screening trials and Performance Measures. For each activity class, we conducted ten similarity searches with different reference sets consisting of five molecules with distinct peptide character (PF = 25% or greater). The remaining active compounds (between 58 and 1326) were retained in the MDDR source database as potential hits. Database compounds were ranked according to their similarity values and the number of recovered active molecules was determined for the top 0.5% of each individual ranking and summed up over the ten different trials for each activity class. In addition, enrichment factors relative to random compound selection were calculated and averaged over the ten trials. To compare the results for different fingerprints, the better performing search strategy (1-NN or centroid) was determined for all fingerprints, except for PDR-FP, where the frequency method was applied. Compound recall and enrichment factors were compared for different PF categories (PF \geq 25%, PF < 25%, PF \leq 10%) and also taking only peptide-likeness but not specific activity into account.

4. RESULTS

4.1. General Search Performance. The results of our systematic similarity search trials are summarized in Table 2. Depending on the activity class, between 58 and 1326 active molecules were available as potential hits within \sim 148 400 database compounds. Cumulative compound recall and average enrichment factors were determined for the top-ranked 0.5% of the database. Molprint2D performed better in combination with the 1-NN strategy for six of eight classes, but the other fingerprints displayed no clear preference for 1-NN or centroid searching, although other studies

suggested that 1-NN would be generally superior to the centroid or other strategies.^{24,27} However, other findings have indicated that preferences for search strategies depend on both the fingerprint type and the structural diversity of targeted active compounds.³⁸

When the overall performances of the five fingerprints are compared, Molprint2D detected more active molecules (between 92 and 426) than the others for six of eight classes (ACE, CLG, MMT, NRI, PRT, TNF). PDR-FP achieved the best results for HIV (222 hits) and second-best ones for ACE, CLG, MMT, and TNF (66–398 molecules), and TGT had the best results for CTP (163 hits). Lowest sums of correctly recovered active compounds (53–196) were produced by the Daylight fingerprint in five of eight cases (CLG, CTP, MMT, NRI, TNF). Calculated enrichment factors were also class-dependent. Overall, the best results were found for CLG with enrichment factors between 40 and 87. For classes CTP, HIV, and PRT, factors between 19 and 52 were obtained; for classes ACE and MMT, factors between nine and 25 were obtained. For TNF, they were between five and 12, and for NRI the factors were only one to two. Thus, with the exception of NRI, the different fingerprints produced reasonable to good compound recall for the activity classes tested here.

4.2. Detecting Active Molecules with Strong Peptide Flavor. An important aspect of our study was the analysis of PF categories for correctly detected active compounds. Molecules belonging to the first category (PF \geq 25%) have distinct peptide character comparable to that one of the reference molecules. In this case, relative fingerprint performance displayed the same trends as discussed above: Molprint2D recovered the most active compounds for the same six classes as above; PDR-FP reached best results for HIV and TGT for CTP. In addition, the absolute performance of all fingerprints increased in part dramatically compared to the overall results. Enrichment factors between 42 and 146 were achieved for CLG, MMT, and TNF, factors of 29–87 for CTP, HIV, and PRT, 19–50 for ACE, and also 8–32 for NRI. This increase in performance indicated that in most cases, fingerprints recovered proportionally more compounds with strong than weak peptide character. Thus, for classes with a large fraction of compounds belonging to the PF \geq 25% category (CLG, CTP, HIV, PRT), the overall performance was good, whereas for classes with only small fractions of such compounds (NRI and TNF), fingerprint performance was generally poor. This indicated that 2D fingerprint methods are capable of differentiating between peptide-like molecules having different activity, which is an encouraging and previously unobserved finding, but that they have difficulties to relating molecules with different degrees of peptide character to each other.

4.3. Active Molecules with Weak Peptide Flavor. Therefore, we next determined the tendency of 2D fingerprints to detect compounds belonging to lower PF categories (<25%) than the reference compounds. In this case, significantly different results were obtained. PDR-FP performed best for six of eight activity classes (ACE, CLG, HIV, MMT, PRT, TNF) with cumulative recall of 23–79 compounds, and TGD recovered most molecules for the two other classes (CTP and NRI) and was the second-best for ACE, CLG, HIV, and PRT. Enrichment factors decreased for all fingerprint methods and activity classes, albeit to different extents.

Table 2. Recovery and Enrichment of Active Compounds^a

class code	fingerprint	best approach	all PF		PF \geq 25%		PF < 25%		PF \leq 10%	
			hits	EF	hits	EF	hits	EF	hits	EF
ACE	PDR-FP		66	12.2	26	19.3	40	9.9	10	4.4
	Molprint2D	1-NN	92	17.0	68	50.4	24	5.9	0	0.0
	TGD	centroid	62	11.5	34	25.2	28	6.9	0	0.0
	TGT	1-NN	60	11.1	38	28.1	22	5.4	0	0.0
	Daylight	1-NN	65	12.0	48	35.6	17	4.2	3	1.3
CLG	PDR-FP		398	81.2	319	118.1	79	35.9	15	14.3
	Molprint2D	centroid	426	86.9	394	145.9	32	14.5	2	1.9
	TGD	centroid	269	54.9	200	74.1	69	31.4	5	4.8
	TGT	centroid	260	53.1	221	81.8	39	17.7	0	0.0
	Daylight	centroid	196	40.0	150	55.5	46	20.9	15	14.3
CTP	PDR-FP		86	27.3	62	31.0	24	20.9	0	0.0
	Molprint2D	1-NN	132	41.9	118	59.0	14	12.2	0	0.0
	TGD	1-NN	123	39.0	92	46.0	31	27.0	0	0.0
	TGT	centroid	163	51.7	150	75.0	13	11.3	0	0.0
	Daylight	1-NN	60	19.0	57	28.5	3	2.6	0	0.0
HIV	PDR-FP		222	30.8	160	86.5	62	11.6	5	1.4
	Molprint2D	1-NN	194	26.9	152	82.1	42	7.8	2	0.5
	TGD	centroid	165	22.9	113	61.1	52	9.7	5	1.4
	TGT	centroid	144	20.0	118	63.8	26	4.9	0	0.0
	Daylight	1-NN	146	20.3	118	63.8	28	5.2	5	1.4
MMT	PDR-FP		214	19.5	135	75.0	79	8.6	28	4.0
	Molprint2D	centroid	269	24.5	210	116.6	59	6.4	5	0.7
	TGD	1-NN	156	14.2	100	55.5	56	6.1	24	3.4
	TGT	1-NN	145	13.2	107	59.4	38	4.1	12	1.7
	Daylight	centroid	97	8.8	75	41.7	22	2.4	3	0.4
NRI	PDR-FP		74	1.1	28	12.2	46	0.7	18	0.3
	Molprint2D	1-NN	113	1.7	73	31.7	40	0.6	7	0.1
	TGD	1-NN	92	1.4	27	11.7	65	1.0	17	0.3
	TGT	centroid	64	1.0	19	8.3	45	0.7	6	0.1
	Daylight	1-NN	64	1.0	30	13.0	34	0.5	19	0.3
PRT	PDR-FP		153	27.3	130	38.8	23	10.2	0	0.0
	Molprint2D	1-NN	223	39.8	212	63.3	11	4.9	0	0.0
	TGD	centroid	122	21.8	107	31.9	15	6.7	0	0.0
	TGT	centroid	191	34.1	177	52.8	14	6.2	0	0.0
	Daylight	1-NN	155	27.7	147	43.9	8	3.6	0	0.0
TNF	PDR-FP		106	10.5	76	76.0	30	3.3	0	0.0
	Molprint2D	1-NN	120	11.9	100	100.0	20	2.2	1	0.1
	TGD	1-NN	75	7.5	67	67.0	8	0.9	4	0.6
	TGT	1-NN	64	6.4	57	57.0	7	0.8	3	0.4
	Daylight	centroid	53	5.3	49	49.0	4	0.4	0	0.0

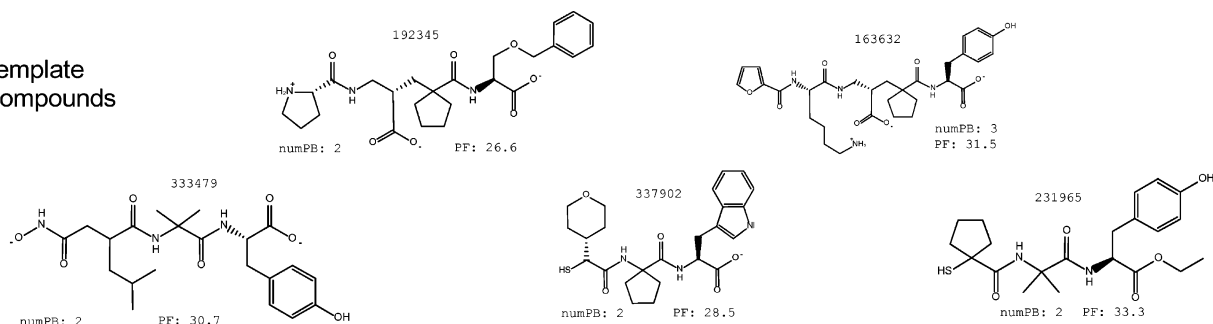
^a Hits is the sum of recovered molecules in all ten trials together, and EF stands for the average enrichment factor over these trials. All PF reports the results without distinguishing between different PF categories. PF \geq 25%, PF < 25%, and PF \leq 10% provide a more detailed overview and report hits and enrichment factors relative to the magnitude of PF. Best approach shows which multiple-template search strategy (1-NN or centroid) produced best results. Activity classes are designated according to Table 1.

Highest factors of between three and 36 were calculated for CLG and CTP, and for ACE, HIV, MMT, and PRT, enrichment factors of two to 12 were observed. For activity class TNF, only PDR-FP (3.3) and Molprint2D (2.2) produced enrichment factors that were better than random selection, and for NRI, all methods failed. Class NRI is characterized by the presence of extremely high structural diversity among compounds with low peptide character. When searching for active molecules in category PF \leq 10%, meaningful results were only achieved for four activity classes (ACE, CLG, HIV, MMT). In these cases, PDR-FP consistently performed best with enrichment factors between 1.4 and 14. Taken together, these calculations revealed that essentially only PDR-FP and TGD were capable of detecting similarity relationships between active compounds having significant differences in peptide character. Figure 4a–c shows examples of reference molecules and hits that were correctly identified with different fingerprint methods.

4.4. Total Recall of Compounds with Weak Peptide Flavor. In light of the results discussed above, we next

analyzed the peptide character of all molecules within the top-ranked 0.5% of the source database including our activity classes and other MDDR compounds (false-positives, decoys). Decoys are also considered in parts of our analysis to generally study the ability of 2D fingerprints to abstract from the features of strongly peptide-like molecules. These investigations complement the analysis of structure–activity relationships that are the primary focal point. Table 3 reports the average percentage of MDDR molecules that belong to the PF \leq 10% category. Two trends were observed. First, we found that TGD and PDR-FP displayed the overall strongest tendency to enrich selection sets with low PF molecules, in contrast to Molprint2D and TGT. For TGD, in combination with 1-NN, 7–24% of the recovered molecules belonged to this category, and for the centroid approach, the corresponding results were 5–16%. For PDR-FP, compound recall was between 4 and 17%. Second, in almost all cases and regardless of the fingerprint design, the 1-NN technique produced more low PF compounds than the centroid strategy. For the Daylight fingerprint, this trend was

(a)

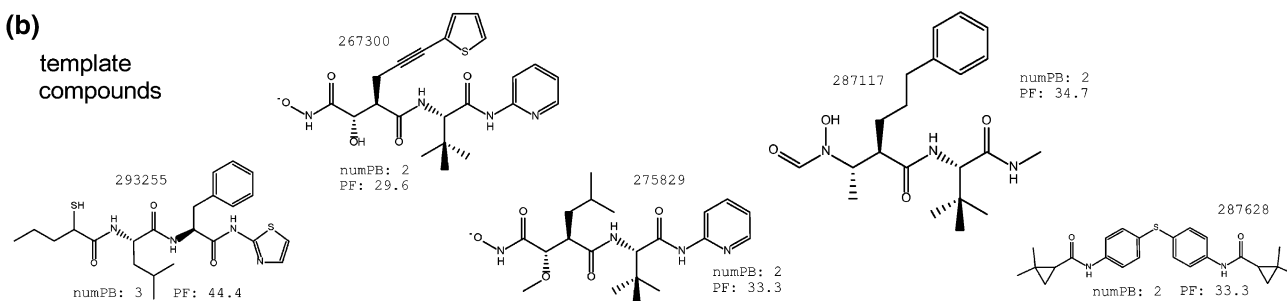
template
compounds

unique hits	PDR-FP	TGD	Molprint2D
 148956 numPB: 0 PF: 0	 173075 numPB: 0 PF: 0	 242964 numPB: 1 PF: 11.7	 207906 numPB: 1 PF: 14.8
 197971 numPB: 1 PF: 13.7	 197973 numPB: 1 PF: 13.7	 245087 numPB: 1 PF: 12.5	 330311 numPB: 2 PF: 27.5
 141048 numPB: 2 PF: 24.2	 330312 numPB: 3 PF: 44.4	 194486 numPB: 4 PF: 35.5	

compounds identified by more than one approach

TGT Daylight 243825 numPB: 5 PF: 43.4	Molprint2D TGT Daylight 207737 numPB: 2 PF: 23.5	Molprint2D TGD TGT Daylight 226040 numPB: 2 PF: 38
PDR-FP TGD 144681 numPB: 1 PF: 13.7	Molprint2D TGD TGT 163637 numPB: 2 PF: 20	PDR-FP Molprint2D TGD Daylight 338113 numPB: 2 PF: 32
PDR-FP TGD 333475 numPB: 1 PF: 13.7	Molprint2D TGD TGT 200581 numPB: 3 PF: 27.2	Molprint2D TGD TGT Daylight 195997 numPB: 2 PF: 34.7
PDR-FP Molprint2D Daylight 192348 numPB: 2 PF: 28.5	Molprint2D TGD TGT 200585 numPB: 4 PF: 32.6	PDR-FP Molprint2D TGD TGT Daylight 242410 numPB: 3 PF: 33.3
	Molprint2D TGD TGT 231503 numPB: 3 PF: 25	

(b)

 template
compounds


unique hits	PDR-FP	Molprint2D
<p>264882 numPB: 0 PF: 0</p> <p>278238 numPB: 1 PF: 14.2</p> <p>383055 numPB: 1 PF: 14.8</p>	<p>216580 numPB: 1 PF: 15.3</p> <p>266630 numPB: 2 PF: 33.3</p> <p>313681 numPB: 3 PF: 42.8</p>	<p>287119 numPB: 2 PF: 24.2</p> <p>267105 numPB: 2 PF: 26.6</p> <p>265783 numPB: 2 PF: 33.3</p>
		<p>Daylight</p> <p>388466 numPB: 0 PF: 0</p>
		<p>TGT</p> <p>279490 numPB: 3 PF: 37.5</p>

compounds identified by more than one approach

<p>Molprint2D Daylight</p> <p>297590 numPB: 1 PF: 13.3</p>	<p>PDR-FP Molprint2D Daylight</p> <p>269174 numPB: 1 PF: 14.8</p>	<p>PDR-FP Molprint2D TGT</p> <p>329660 numPB: 2 PF: 27.5</p>
<p>Molprint2D Daylight</p> <p>212500 numPB: 2 PF: 44.4</p>	<p>PDR-FP Molprint2D TGT</p> <p>269629 numPB: 2 PF: 27.5</p>	<p>PDR-FP Molprint2D TGT Daylight</p> <p>272734 numPB: 2 PF: 32</p>
<p>Molprint2D TGT Daylight</p> <p>269274 numPB: 2 PF: 32</p>	<p>TGT TGT Daylight</p> <p>287631 numPB: 2 PF: 33.3</p>	<p>PDR-FP Molprint2D TGT TGT Daylight</p> <p>270136 numPB: 2 PF: 26.6</p>

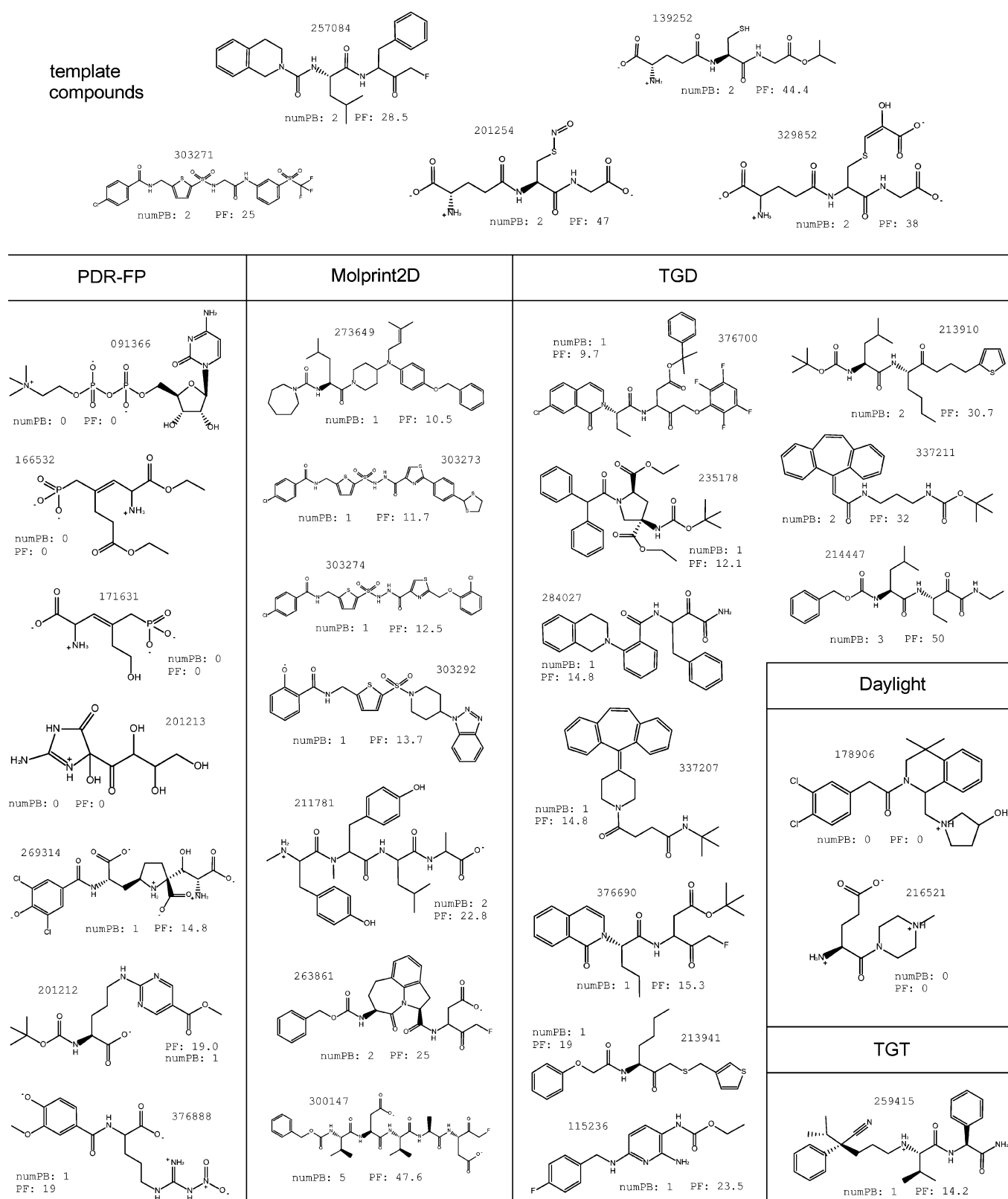


Figure 4. Peptide character of hits identified by fingerprint searching. For three different activity classes, (a) ACE, (b) MMT, and (c) NRI, examples of hits are shown that were correctly detected by different fingerprint methods. Each compound is labeled with its MDDR external registry number, the number of peptide-like bonds (numPB), and PF values (in %). For class ACE, all hits are shown. For MMT, all unique hits are reported, and only a representative subset of non-unique hits. For NRI, only unique hits are shown. In the NRI trial, the five different fingerprints produced only four non-unique hits (not displayed here) but 27 unique hits, suggesting their complementarity.

especially clear. On average, twice as many (~8 vs ~4%) molecules in its 1-NN selection set had weak peptide character (for the difficult class NRI, we observed 20 vs 6%).

How can we rationalize these findings? The first observation indicated that different fingerprint designs have different intrinsic preferences: low resolution or complexity finger-

Table 3. Top-Ranked Molecules with Weak Peptide Character^a

	ACE	CLG	CTP	HIV	MMT	NRI	PRT	TNF	av
PDR-FP	4	9	14	8	17	14	4	9	9.8
Molprint2D 1-NN	1	4	3	3	7	5	1	4	3.5
Molprint2D centroid	2	3	2	2	6	3	1	3	2.6
TGD 1-NN	7	18	16	18	24	18	10	14	15.6
TGD centroid	5	16	10	10	16	15	6	12	11.2
TGT 1-NN	1	3	3	1	6	3	1	4	2.8
TGT centroid	2	2	2	0	4	5	1	2	2.2
Daylight 1-NN	1	5	11	6	12	20	3	4	7.7
Daylight centroid	1	3	9	1	6	6	1	4	3.9

^a For each fingerprint method and activity class, the average percentage of all molecules within the top 0.5% of the correspondent MDDR ranking (correctly identified hits and decoys) is reported that belong to the PF \leq 10% category.

Table 4. Descriptor Statistic for Top-Ranked MDDR Compounds^a

descriptor	ACE			CLG			CTP			HIV		
	numPB	HA	PF	numPB	HA	PF	numPB	HA	PF	numPB	HA	PF
PDR-FP	4.8	56.8	34.0	1.9	33.3	26.9	1.5	33.1	21.1	4.3	66.5	27.8
Molprint2D	4.9	57.4	35.4	3.4	42.9	34.6	3.2	44.3	30.7	3.9	51.7	31.9
TGD	3.6	52.6	28.4	1.5	34.1	20.0	1.3	33.7	18.0	3.1	55.7	24.0
TGT	4.7	57.8	34.7	2.5	40.5	27.9	2.0	37.0	25.9	4.2	56.7	32.3
Daylight	5.0	58.1	36.5	4.0	47.0	36.6	2.8	41.8	27.6	3.9	52.1	30.7

descriptor	MMT			NRI			PRT			TNF		
	numPB	HA	PF	numPB	HA	PF	numPB	HA	PF	numPB	HA	PF
PDR-FP	1.4	31.4	21.5	1.5	31.5	22.2	4.3	61.0	30.5	1.7	32.8	24.9
Molprint2D	2.4	36.7	29.6	2.8	41.7	28.0	3.7	49.1	32.6	4.1	50.8	34.3
TGD	1.1	31.8	17.1	1.7	37.9	29.0	2.3	44.8	23.3	1.9	36.5	22.3
TGT	2.2	39.7	24.6	1.9	38.2	23.0	3.5	50.8	30.8	2.3	40.1	25.8
Daylight	3.3	42.1	32.9	2.4	39.0	24.8	4.2	50.5	35.4	4.0	47.7	35.8

^a Reported are average values for the number of peptide-like bonds (numPB), the number of heavy atoms (HA), and PF (in %) calculated from the top ranked 0.5% MDDR compounds (correctly identified hits and decoys). For Molprint2D, TGD, TGT, and Daylight, results are reported for the preferred search strategy according to Table 2.

prints like TGD and PDR-FP that predominantly take chemical properties into account should have a greater ability to abstract from chemical structure and amide components, which relates to “scaffold hopping”,^{6,7} the identification of molecules having different core structures. For this task, PDR-FP was previously shown to be often more adequate than other 2D fingerprints.³⁸ By contrast, high-resolution fingerprints like Molprint2D that facilitate a close-up structural view of molecules have a greater tendency to detect structural features. Thus, in the case of peptide-like molecules, it might be more difficult to depart from amide structures.

4.5. Alternative Search Strategies. The second observation was that the 1-NN strategy could better abstract from peptide backbones than the centroid approach when decoys were included in the analysis, as presented in Table 3. This can be explained by considering that the centroid strategy emphasizes fingerprint bit positions that are conserved in reference molecules. Usually, the centroid strategy is applied assuming that conserved fingerprint positions are likely to account for activity-relevant features. For peptide-like molecules, amide bond components are likely to be also emphasized, regardless of whether they are determinants of specific activity or not. However, when focused only on correctly identified hits, the centroid strategy often performed better than the 1-NN approach and even recovered more active molecules with weak peptide character. Strongly peptide-like active molecules tend to be recovered using the centroid strategy because they share activity-relevant features

in addition to having high amide content. Although weakly peptide-like active molecules have low amide content, they can be differentiated from inactive peptide-like or other compounds on the basis of activity-determining features. In contrast to the centroid strategy, the 1-NN approach considers each reference molecule separately and can thus not deduce any conserved activity-relevant features from multiple reference compounds. Its search performance entirely depends on the best individual matches between reference and database compounds.

4.6. Number of Peptide-like Bonds and Size of Preferred Molecules. We further analyzed the number of amide bonds and the size of molecules within the top-ranked 0.5% of the MDDR. Table 4 reports the number of peptide-like bonds, heavy atoms, and PF values of molecules detected in search calculations averaged over all trials (and all ranking positions). Overall, the results in Table 4 mirror the relative fingerprint performances reported in Table 3. TGD and PDR-FP were found to produce the lowest average PF values for all activity classes. This means that these two fingerprints were most capable of abstracting from amide bond features. For activity classes CLG, CTP, MMT, NRI, and TNF, these findings correlate with the presence of on average less than two (1.1–1.9) peptide-like bonds in retrieved molecules. Since reference molecules had to contain at least two peptide-like bonds, the PDR-FP and TGD fingerprints decreased the average amide content of detected compounds for the majority of activity classes. By contrast, the Molprint2D and Daylight fingerprints displayed the opposite tendency and

Table 5. Unique Hits^a

	ACE	CLG	CTP	HIV	MMT	NRI	PRT	TNF	sum
PDR-FP	24	47	19	52	51	45	37	18	293
Molprint2D	25	34	21	13	44	48	48	14	247
TGD	13	19	21	36	27	60	16	3	195
TGT	7	16	43	12	22	39	34	1	174
Daylight	9	11	7	15	4	27	29	1	103

^a For each fingerprint method and activity class, the sum of uniquely identified hits in ten trials is reported. For Molprint2D, TGD, TGT, and Daylight, results are reported for the preferred search strategy according to Table 2.

Table 6. Hit Overlap between Alternative Search Strategies^a

		ACE	CLG	CTP	HIV	MMT	NRI	PRT	TNF
TGD	av no. of hits	8.3	31.3	15.8	18.2	21.0	12.0	17.7	9.7
	unique hits (%)	56.7	52.5	44.0	46.8	67.2	76.2	64.0	40.3
	1-NN uniqueness	41.0	23.9	29.2	17.3	52.8	71.1	51.6	24.9
	centroid uniqueness	43.3	44.3	28.9	40.7	40.3	53.9	47.8	22.4
TGT	av no. of hits	7.1	31.0	18.3	17.2	19.1	8.8	22.9	8.2
	unique hits (%)	36.9	53.9	43.7	65.3	60.1	81.2	61.1	49.1
	1-NN uniqueness	27.4	23.9	15.8	40.6	47.0	56.4	32.6	36.6
	centroid uniqueness	24.1	44.6	36.4	59.0	38.5	69.4	44.9	26.4
Daylight	av no. of hits	9.3	26.9	8.7	18.0	12.5	9.8	21.1	6.9
	unique hits (%)	54.4	55.0	65.1	49.4	60.1	79.6	49.2	61.1
	1-NN uniqueness	42.1	39.6	54.0	36.1	41.3	70.1	30.5	33.6
	centroid uniqueness	37.8	38.9	50.7	28.5	47.8	62.7	34.7	50.1
Molprint2D	av no. of hits	11.4	44.9	16.2	25.3	29.4	12.9	30.3	13.1
	unique hits (%)	46.7	19.8	43.4	35.6	24.8	50.1	44.2	18.1
	1-NN uniqueness	33.2	6.2	30.5	15.3	10.3	43.6	25.2	10.1
	centroid uniqueness	29.3	15.4	26.6	26.9	17.6	20.2	33.5	10.2

^a av no. of hits stands for the sum of different hits identified by the 1-NN or centroid approach averaged over ten trials. Unique hits (%) gives the percentage of these hits uniquely detected by one of the two search strategies. 1-NN uniqueness displays the percentage of all 1-NN hits that were unique to the 1-NN strategy, and centroid uniqueness is the percentage of all centroid hits that were unique to the centroid strategy.

molecules selected with these fingerprints consistently contained on average more than two peptide-like bonds.

When the average size of the top-ranked molecules is considered, PDR-FP enriched selection sets with smallest molecules (31.4–33.3 heavy atoms) for five classes (CLG, CTP, MMT, NRI, and TNF). Compared to Molprint2D and Daylight, molecules selected with PDR-FP consisted of approximately ten fewer non-hydrogen atoms and where thus about one-fourth smaller. These findings could be attributed to the fact that PDR-FP is a size-independent fingerprint because, for any molecule, PDR-FP has a constant bit density of 93/500 bits that are set on (i.e., one for each encoded descriptor).³⁵ By contrast, calculation of Tanimoto similarity for standard fingerprints with variable bit density displays a statistical preference for the detection of increasingly large molecules.²⁸ This is the case because larger molecules generally set more fingerprint bits on and therefore have the tendency to produce higher Tc similarity values against other compounds. However, recognition of smaller-sized hits is generally desirable because newly identified active molecules need to be subjected to chemical optimization efforts, which typically lead to an increase in molecular weight because of the introduction of additional functional groups.

For two classes (HIV and PRT), PDR-FP detected the on-average largest molecules, which could be rationalized by analyzing the reference molecules. For these two classes, available reference molecules were considerably larger than for the others, and PDR-FP preferentially recovered molecules of comparable size. These results suggest that PDR-FP had a preference to select database compounds that were of similar size as the reference molecules, at least for peptide-like compounds.

4.7. Hit Overlap. We were also interested in analyzing the overlap of hits that were correctly identified by different fingerprint methods. Therefore, we determined the number of compounds that were exclusively recovered by one method and calculated the sum over all trials. The results are shown in Table 5. All fingerprints were found to produce unique hits. Although PDR-FP did not have the overall highest hit rates, as discussed above, it recovered the largest number of unique hits (between 18 and 52) over all classes, as also illustrated in Figure 4a–c. These molecules contributed to, on average, between 12% (CLG) and 60% (NRI) of all hits identified with PDR-FP. These findings suggest that the chemical space representation of the fingerprints evaluated here substantially differs. Such differences make the parallel application of alternative methods a promising virtual screening approach, as also concluded by others.³⁹

Because the 1-NN and centroid search strategies produced different results in our study, we also analyzed their hit overlap. Table 6 reports for all fingerprints, except PDR-FP, the average numbers of hits and average percentages of unique hits that were identified with the 1-NN, centroid, or both strategies. For TGD, TGT, and Daylight, percentages of hits uniquely identified with the 1-NN or centroid method were found to range from approximately 40–80%, whereas for Molprint2D approximately 20–50% unique hits were observed. Table 6 also shows that the 1-NN and centroid strategies generated comparable percentages of unique hits. Thus, there was also no significant difference between these alternative search strategies in the detection of unique hits. In fact, in light of the distribution of unique hits, these search strategies are complementary in nature, regardless of the chosen fingerprint, and should best be applied in concert.

Table 7. Comparison of PDR-FP and PDR-FP-491^a

	ACE	CLG	CTP	HIV	MMT	NRI	PRT	TNF
PDR-FP	3.8	9.5	14.0	7.7	17.0	14.3	3.6	9.1
PDR-FP-491	3.6	11.1	15.0	10.0	18.9	14.3	4.2	11.1

^a For PDR-FP and its modified version, the table reports the average percentage of all molecules (hits and decoys) within the top 0.5% of their MDDR rankings that belong to the PF \leq 10% category.

4.8. Modification of the PDR-FP Design. The original implementation of PDR-FP³⁵ encoded exactly 93 property descriptors that were binned according to their value distributions observed in the ZINC database.⁴⁰ The PDR-FP can be easily modified by alteration of its descriptor library or by adjustment of the descriptor value range binning scheme for another screening database. This is simply done by changing interval boundaries stored within the PDR-FP descriptor library. For the current analysis, we initially adjusted the interval boundaries according to the MDDR descriptor value distribution. Moreover, since we focused on peptide-like molecules, we produced a variant of PDR-FP by removing two of its descriptors that account for the number of oxygen and nitrogen atoms and are thus directly associated with amide bond content. The resulting PDR-FP variant consisted of only 491 instead of 500 bit positions, was termed “PDR-FP-491”, and was also tested in similarity search calculations. Table 7 reports average percentages of database compounds belonging to the PF \leq 10% category detected with PDR-FP-491. As we expected, this fingerprint recognized on average more low PF compounds than PDR-FP (with the exception of class ACE). Figure 5 shows an example where PDR-FP-491 correctly identified four more hits than PDR-FP with peptide flavors between 6.8 and 18.6%. Thus, simple modifications of the PDR-FP design tuned the search calculations toward further increasing recognition of molecules with weak peptide character. In other words, minor modifications further extended the range of peptide-like molecules with similar activity that could be recognized using this fingerprint.

5. DISCUSSION AND CONCLUSIONS

In this study, we systematically analyzed the ability of five different 2D fingerprints and two prominent search strategies (1-NN and centroid) to (a) differentiate between peptide-like molecules having different activity and (b) detect a spectrum of active molecules having different degrees of peptide character. For practical applications, the ability to depart from strongly peptide-like molecules and detect other compounds having similar activity is of particular relevance because properties of peptides are often undesired in lead generation. In our analysis, we attempted to simulate such transitions. Therefore, we assembled eight different activity classes that were characterized by high degrees of structural diversity (no intraclass pairwise Tc similarity above 0.85) and a wide variation of PF values. In systematic calculations, essentially all 2D fingerprints that we tested were able to distinguish between peptide-like molecules having different activities and produced meaningful enrichment factors. In this case, Molprint2D performed overall best, followed by PDR-FP.

When we then analyzed the ability of fingerprints to recover molecules with PF values lower than the reference

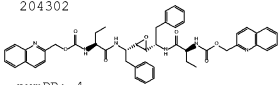
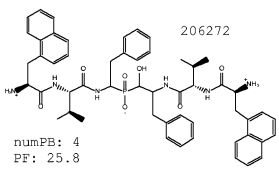
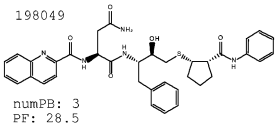
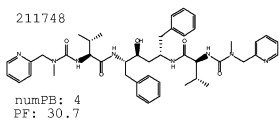
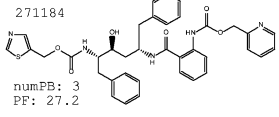
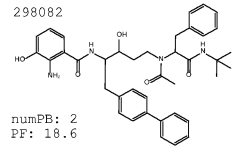
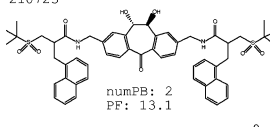
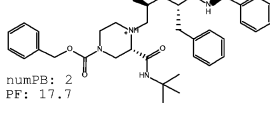
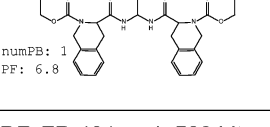
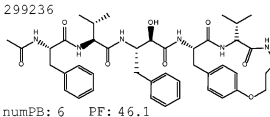
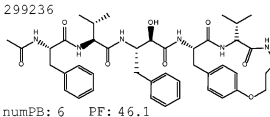
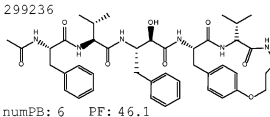
template compounds	unique PDR-FP-491 hits				
<p>204302</p>  <p>numPB: 4 PF: 28.5</p> <p>206272</p>  <p>numPB: 4 PF: 25.8</p> <p>198049</p>  <p>numPB: 3 PF: 28.5</p> <p>211748</p>  <p>numPB: 4 PF: 30.7</p> <p>271184</p>  <p>numPB: 3 PF: 27.2</p>	<p>298082</p>  <p>numPB: 2 PF: 18.6</p> <p>210723</p>  <p>numPB: 2 PF: 13.1</p> <p>199182</p>  <p>numPB: 2 PF: 17.7</p> <p>265970</p>  <p>numPB: 1 PF: 6.8</p> <tr> <td colspan="2">PDR-FP-491 and -500 hit</td></tr> <tr> <td></td><td> <p>299236</p>  <p>numPB: 6 PF: 46.1</p> </td></tr>	PDR-FP-491 and -500 hit			<p>299236</p>  <p>numPB: 6 PF: 46.1</p>
PDR-FP-491 and -500 hit					
	<p>299236</p>  <p>numPB: 6 PF: 46.1</p>				

Figure 5. Example of active compounds only identified with PDR-FP-491. Compared to PDR-FP, the PDR-FP-491 variant identified four more hits for HIV with PF values below 20%. In addition to the template compounds and unique PDR-FP-491 hits, an example of a shared hit with strong peptide character (PF = 46%) is also displayed. All molecules are labeled as described in the legend of Figure 4.

molecules (PF < 25%), the scenario changed and PDR-FP was generally superior, followed by TGD (a low-complexity 2D pharmacophore fingerprint). PDR-FP and TGD showed also the highest potential to principally detect molecules with low PF values (PF \leq 10%). This means that these fingerprints were most suited to facilitate a transition from peptide-like compounds to other small molecules. We also found that the 1-NN and centroid search strategies displayed comparable performance in our calculations on peptide-like molecules. This suggests that selecting a preferred search strategy depends on both the fingerprint design and the characteristics of activity classes that are studied. For example, high-resolution and structure-oriented fingerprints like Molprint2D in combination with the 1-NN strategy can lead to the recovery of many hits that are closely related to reference molecules. On the other hand, lower resolution search tools like TGD have a greater tendency to recover hits that differ in structure and chemical nature from reference sets. A potential tradeoff is that lower resolution fingerprints often produce larger false-positive rates because of the fuzziness of their molecular view.

The property descriptor-based PDR-FP fingerprint also produced meaningful enrichment factors, but recovered many molecules of moderate to low peptide character. From this point of view, it was perhaps most suitable for the treatment of diverse peptide-like molecules. The molecular size

independence of PDR-FP led to the identification of active molecules that were on average smaller than those detected by other fingerprint methods. Furthermore, PDR-FP had the smallest hit overlap with other 2D fingerprint methods studied here and was thus most complementary.

Even if no fingerprint was able to detect a significant amount of active molecules having PF values of zero (no single peptide-like bond), we found that in particular PDR-FP and TGD could recognize hits with substantially reduced peptide character. This ability is promising in the context of iterative similarity searching and screening to facilitate peptidomimetic transitions: active molecules with decreasing peptide character identified initially can serve as new reference molecules in subsequent rounds, thereby gradually reducing the peptide-likeness of hits until an acceptable low PF level is reached or a peptidomimetic discovered.

ACKNOWLEDGMENT

We are grateful to Daylight Chemical Information Systems for making the Daylight fingerprint available to us and to Andreas Bender for Molprint2D.

REFERENCES AND NOTES

- Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
- Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
- Green, D. V. Virtual screening of virtual libraries. *Prog. Med. Chem.* **2003**, *41*, 61–97.
- Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
- Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping". Validation of topomer similarity as a superior predictor of biological activities. *J. Med. Chem.* **2004**, *47*, 6777–6791.
- Schneider, G.; Schneider, P.; Renner, S. Scaffold-hopping: How far can we jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.
- Huang, D.; Luthi, U.; Kolb, P.; Edler, K.; Cecchini, M.; Audetat, S.; Barberis, A.; Caflisch, A. Discovery of cell-permeable non-peptide inhibitors of β -secretase by high-throughput docking and continuum electrostatics calculations. *J. Med. Chem.* **2005**, *48*, 5108–5111.
- Huang, N.; Nagarsekar, A.; Xia, G.; Hayashi, J.; MacKerell, A. D., Jr. Identification of non-phosphate-containing small molecular weight inhibitors of the tyrosine kinase p56 Lck SH2 domain via in silico screening against the pY + 3 binding site. *J. Med. Chem.* **2004**, *47*, 3502–3511.
- Thaisrivongs, S.; Janakiraman, M. N.; Chong, K.-T.; Tomich, P. K.; Dolak, L. A.; Turner, S. R.; Strohbach, J. W.; Lynn, J. C.; Horng, M.-M.; Hinshaw, R. R.; Watenpaugh, K. D. Structure-based design of novel HIV protease inhibitors: Sulfonamide-containing 4-hydroxy-coumarins and 4-hydroxy-2-pyrones as potent non-peptidic inhibitors. *J. Med. Chem.* **1996**, *39*, 2400–2410.
- Debnath, A. K.; Radigan, L.; Jiang, S. Structure-based identification of small molecule antiviral compounds targeted to the gp41 core structure of the human immunodeficiency virus type 1. *J. Med. Chem.* **1999**, *42*, 3203–3209.
- Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.
- Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.
- Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.
- Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Rev. Drug Discovery* **2004**, *3*, 935–949.
- Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.
- Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.
- Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.
- James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA.
- McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
- Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.
- Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
- Hert, J.; Willett, P.; Wilton, D. J. New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
- Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
- Godden, J. W.; Stahura, F. L.; Xue, L.; Bajorath, J. Searching for molecules with similar biological activity: Analysis by fingerprint profiling. *Pac. Symp. Biocomput.* **2000**, *5*, 566–575.
- Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Comput. Sci.* **2001**, *41*, 746–753.
- Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target protein. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.
- Molecular Drug Data Report (MDDR)*, version 2005.2; MDL Elsevier: San Leandro, CA, 2005.
- Xue, L.; Bajorath, J. Distribution of molecular scaffolds and R-groups isolated from large compound databases. *J. Mol. Model.* **1999**, *5*, 97–102.
- MACCS structural keys*; MDL Elsevier: San Leandro, CA, 2002.
- Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- MOE (Molecular Operating Environment)*, version 2005.06; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2005.
- MOLPRINT 2D. <http://www.molprint.com> (accessed Feb 2006).
- Eckert, H.; Bajorath, J. Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2515–2526.
- Eckert, H.; Bajorath, J. Determination and mapping of activity-specific descriptor value ranges for the identification of active compounds. *J. Med. Chem.* **2006**, *49*, 2284–2293.
- Whittle, M.; Gillet, V. J.; Willett, P. Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: A comparison of similarity coefficients. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1840–1848.
- Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem* **2007**, *12*, 225–233.
- Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. *J. Med. Chem.* **2005**, *48*, 240–248.
- Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

CI700086M