Design and Evaluation of Bonded Atom Pair Descriptors

Hany E. A. Ahmed, Martin Vogt, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received December 31, 2009

Atom pairs have been among the first systematically derived fragment-type topological descriptors and have been one of the origins of two-dimensional fingerprint searching. These descriptors continue to be popular and widely used to this date. Herein we introduce a new type of atom pair descriptors, bonded atom pairs, that exclusively capture short-range atom environment information and, thus, depart in their design from other topological descriptors that enumerate bond paths of varying length. Bonded atom pairs combine different types of structural information including element type, hybridization state, aliphatic/aromatic character, and cyclic/acyclic arrangement. Systematic design led to a set of 117 bonded atom pairs, all of which exist in synthetic compounds. A further expanded bonded atom pair set accounting for specific halogen atoms and including a total of 159 descriptors is also provided. Atom pair distribution and frequency analysis in sets of compounds having different selectivity reveals that both conventional and bonded atom pairs capture complementary structural information. In similarity searching, bonded atom pairs meet or exceed the performance of standard atom pairs and structural fragment fingerprints. The complementary nature of structural information captured by atom pairs of different design is also reflected by individual search calculations. Taken together, our findings indicate that bonded atom pairs extend the current repertoire of topological molecular descriptors.

INTRODUCTION

The assessment of molecular similarity plays a central role in chemoinformatics.[1] For similarity analysis, the way molecules are represented is a critical factor.[2] In order to correlate chemical and biological similarity, molecular representations must be capable of capturing activity-determining structural features. To these ends, structural fragment-type and topological descriptors have been among early two-dimensional (2D) representations of small molecules for similarity analysis and for database searching and are widely used to this date.[3] Current structural fragment representations are often based on dictionaries of predefined substructures[4] or obtained by enumerating fragments using structural rules.[5,6] Other popular fragment/topological descriptors include atom-centered fragments,[7] atom pairs,[8,9] topological torsions,[10] and various atom environment strings.[11]

Substructure searching has been one of the origins of similarity searching.[12] In 1973, Adamson et al. investigated the use of various types of atom-, bond-, and ring-centered fragments for substructure searching.[13] During the same year, Adamson and Bush already went beyond substructure matching and compared sets of molecular fragments in order to evaluate structural and also biological similarity, thereby laying a foundation for fingerprint similarity searching.[14]

Atom pair descriptors originally introduced by Carhart et al. represent a pioneering development in the molecular similarity field.[8] Together with the fragment dictionary fingerprints of Willett et al.,[15] atom pairs were utilized to generate the first substructure-type fingerprints for similarity searching. Since their introduction in 1985 atom pairs have been widely applied descriptors in molecular similarity analysis and in similarity searching. From a test compound, atom pairs are systematically extracted as all possible pairs of atoms and as the shortest bond distance (path length) between them. Atoms are assigned atom types that consist of three components, i.e., the element type, the number of neighboring non-hydrogen atoms, and the number of π electrons of the atom. Thus, depending on their complexity and topology, compounds produce varying numbers of atom pairs.

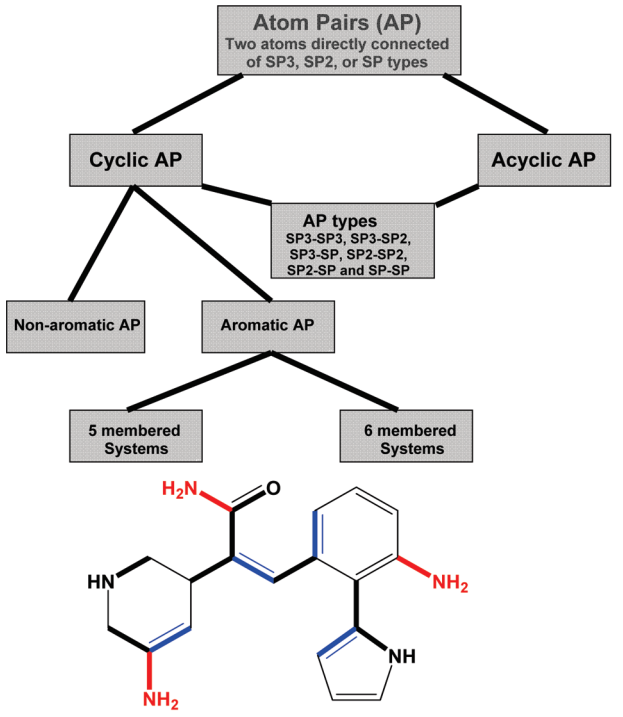
Here, we report atom pair descriptors of different designs and initial proof-of-concept evaluations. In order to form pairs, we exclusively focus on bonded atoms. Accordingly, these descriptors are termed "bonded atom pairs" (BAPs). Accordingly, these atom pairs do not contain bond path information. Rather, they combine information from element type, hybridization states, acyclic or cyclic structural environments, aliphatic or aromatic character, and ring size. Thus, BAPs capture differentiated short-range atom environment information. Similar to conventional atom pair descriptors, BAPs are primarily designed for similarity search applications. We have evaluated BAPs in feature frequency analysis, feature mapping, and similarity searching, in comparison to standard atom pairs and structural keys.

MATERIALS AND METHODS

Design of Bonded Atom Pair Descriptors. For each atom in a molecule, BAPs represent pairs with each of its (non-hydrogen) neighbors (e.g., for carbon atoms, there are one to three bonded atom pairs). For each atom in a pair, its

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

10.1021/ci900512g
Published on Web 03/16/2010



Atom Pairs with Description

C=O, C_{SP2}[acyclic]-O_{SP2}[acyclic]
C-N, C_{SP2}[acyclic]-N_{SP3}[acyclic]
C-N, C_{SP3}[non-aromatic cyclic]-N_{SP3}[non-aromatic cyclic]
C-N, C_{SP2}[non-aromatic cyclic]-N_{SP3}[acyclic]
C-C, C_{SP3}[non-aromatic cyclic]-C_{SP2}[non-aromatic cyclic]
C-C, C_{SP3}[non-aromatic cyclic]-C_{SP2}[acyclic]
C=C, C_{SP2}[non-aromatic cyclic]-C_{SP2}[non-aromatic cyclic]
C=C, C_{SP2}[acyclic]-C_{SP2}[acyclic]
C=C, C_{SP2}[6 membered aromatic cyclic]-C_{SP2}[6 membered aromatic cyclic]
C=N, C_{SP2}[aromatic cyclic]-N_{SP2}[aromatic cyclic]
C=C, C_{SP2}[5 membered aromatic cyclic]-C_{SP2}[5 membered aromatic cyclic]
C-C, C_{SP2}[5 membered aromatic cyclic]-C_{SP2}[6 membered aromatic cyclic]
C-N, C_{SP2}[aromatic cyclic]-N_{SP3}[acyclic]
C-C, C_{SP2}[aromatic cyclic]-C_{SP2}[acyclic]
C-C, C_{SP2}[acyclic]-C_{SP2}[acyclic]

Figure 1. Information content of bonded atom pairs. The design and information content of bonded atom pair descriptors is illustrated. For an exemplary compound, different types of BAPs are color coded, and their description is provided.

element type (C, O, N, S, P, and "H" (halogen) or F, Cl, Br, and I) and hybridization state (sp^3, sp^2, or sp) are reported, and information is provided whether the atom is part of an acyclic moiety or a ring system. For atoms in rings, it is reported whether the ring is aromatic or non-aromatic, and aromatic rings are further distinguished by ring size (i.e., five- or six-membered ring). As 2D descriptors, atom pairs do not capture stereochemcial information. The BAP design strategy is summarized in Figure 1, and exemplary BAPs are shown. On the basis of the applied design criteria, a total of 117 unique BAPs with generalized halogen atoms were generated, as reported in Supporting Information, Table S1. A further expanded set of BAPs explicitly accounting for different halogen atom types and containing 159 descriptors is provided in Supporting Information, Table S2. Thus, the number of possible BAPs is much smaller than the number of possible atom pairs for paths of up to seven bonds (see below).

Standard Atom Pairs. Atom pairs (APs), according to Carhart et al., were implemented for reference calculations. For all possible atom types and paths of up to seven bonds, APs were calculated using PowerMV,[16] giving rise to a total of 4662 possible pairs. In an AP, each atom is defined based on the number of bonded non-hydrogen atoms and on its π electrons. For example, in the pair "C(1,0)-07−C(2,1)", a carbon atom with one heavy atom neighbor and no π electrons is connected over a seven-bond path to another carbon atom with two heavy atom neighbors and one π electron.

Table 1. Selectivity Data Sets[a]

selectivity sets	compounds	SR (K_i or IC_{50} ratio)	Tc AvTc	Tc SD
AMPA Ion Channel[b]				
A/K	15	50.0−1255.6	0.43	0.20
A/N	48	55.7−3300	0.61	0.21
Kainate Ion Channel[b]				
K/A	20	51.6−50 000	0.54	0.26
K/N	11	51.0−2833.3	0.47	0.24
NMDA Ion Channel[b]				
N/A	32	56.7−2000	0.44	0.22
N/K	26	50.1−10 000	0.42	0.19
Non-selective sets				
AK	27	0.4−6.2	0.54	0.26
AN	41	0.7−9.1	0.47	0.25
KN	23	0.2−7.5	0.65	0.42

[a] SR stands for selectivity ratio. Six of nine selectivity sets consist of compounds that are selective for one target over another, and three of compounds that are non-selective for a target pair. In the Tanimoto coefficient (Tc) column, AvTc reports the average Tc values for intraset pairwise compound comparison using MACCS keys and SD, the corresponding standard deviations. [b] Indicates a selective set.

Compound Selectivity Sets. Standard APs and BAPs were tested for their ability to differentiate between compounds having different target selectivity. For this purpose, nine previously assembled[17] data sets were analyzed consisting of compounds having different selectivity against three ionotropic glutamate ion channels, including the 2-amino-3-(3-hydroxy-5-methyl-4-isoxazolyl) propionic acid (AMPA, A), the *N*-methyl-D-aspartic acid (NMDA, N), and the kainic acid (kainate, K) receptors. The composition of these compound sets is summarized in Table 1. Six compound sets designated with a slash (e.g., A/K) exclusively consist of compounds that are at least 50-fold more potent (i.e., selectivity ratio SR $\geq$ 50) for one target (A) over another (K), whereas the remaining three compound sets (e.g., AK) only contain compounds with less than a 10-fold potency difference for two targets (SR < 10) that are considered non-selective. Hence, for each target (e.g., A) there are six selectivity sets available, including two sets consisting of compounds selective for this target over the two others (A/K, A/N), two with compounds of inverse selectivity (K/A, N/A), and two with non-selective compounds (AK, AN), as reported in Table 1. Accordingly, for each target pair, three sets are available (e.g., A/K, K/A, AK). These compound data sets were assembled from original literature sources to enable the evaluation of compound selectivity using molecular similarity methods.[17]

Atom Pair Frequency Analysis and Mapping. In order to examine whether certain APs and BAPs might preferentially occur in compounds with different target selectivity, the frequency of occurrence of all atom pairs was analyzed. To establish a quantitative measure for the set specificity of a descriptor, on the basis of different occurrence frequencies,

Shannon entropy (SE)[18] calculations were carried out. Atom pairs with highest specificity only occur in a single target pair set, but none of the others. In an information-theoretic context, the SE value resulting from the distribution of such descriptors is zero. By contrast, the more evenly atom pairs are distributed over different selectivity sets, the larger the SE becomes. Thus, zero- or low-entropy descriptors (or features, following information-theoretic terminology) represent the set-specific or set-characteristic discriminatory features.

For entropy calculations, feature frequencies are converted into normalized probabilities. Therefore, we calculate the conditional probability $P(\mathrm{C}|\mathrm{F})$ for each compound set C under the condition that feature F is present by applying the Bayes theorem:[19]

$$P(\mathrm{C}|\mathrm{F}) = \frac{P(\mathrm{F}|\mathrm{C})P(\mathrm{C})}{\sum_{\mathrm{D}} P(\mathrm{F}|\mathrm{D})P(\mathrm{D})}$$

Without prior knowledge, one assumes that the prior probabilities $P(\mathrm{C})$ for each set are the same, which simplifies the formula:

$$P(\mathrm{C}|\mathrm{F}) = \frac{P(\mathrm{F}|\mathrm{C})}{\sum_{\mathrm{D}} P(\mathrm{F}|\mathrm{D})}$$

The Shannon entropy of $P(\cdot|\mathrm{F})$ is given by

$$\mathrm{SE}(\mathrm{F}) = \sum_{\mathrm{C}} P(\mathrm{C}|\mathrm{F}) \log_2 P(\mathrm{C}|\mathrm{F})$$

For three classes of selective, inverse selective, and non-selective compounds, the maximum possible entropy corresponding to equal distribution of a feature (and, hence, no specificity) is $\log_2 3 \approx 1.585$. For each selectivity set, atom pairs that occurred in at least two compounds were ranked according to ascending SE values. Discriminatory atom pairs were mapped onto the compound set they preferentially originated from.

Similarity Searching. Systematic similarity search calculations were carried out with reference compounds from each set consisting of selective compounds (i.e., a total of six sets; Table 1). For similarity searching, APs and BAPs were encoded in a keyed binary fingerprint format, and four fingerprints were used; AP-FP (4662 bit positions, each accounting for a unique AP), BAP-FP (117 bits, each accounting for a unique BAP), MACCS keys[4] (166 bits), and AP-BAP-FP (4779 bits, combining unique APs and BAPs).

In each case, 10 sets of 5 randomly chosen selective reference molecules were used for 10 independent search calculations, and the remaining selective compounds as well as inverse selective and non-selective compounds were added to a background database consisting of 100 000 molecules randomly selected from ZINC7.[20] The recall of selective, inverse selective, and non-selective compounds was monitored, and the results were averaged over 10 independent trials. As a search strategy, 1-nearest-neighbor (1-NN) calculations were carried out,[21] i.e., the largest Tanimoto coefficient (Tc)[22] value of a database compound relative to an individual reference compound was assigned to this compound as the final similarity value. Search calculations were carried out using the molecular operating environment.[23]

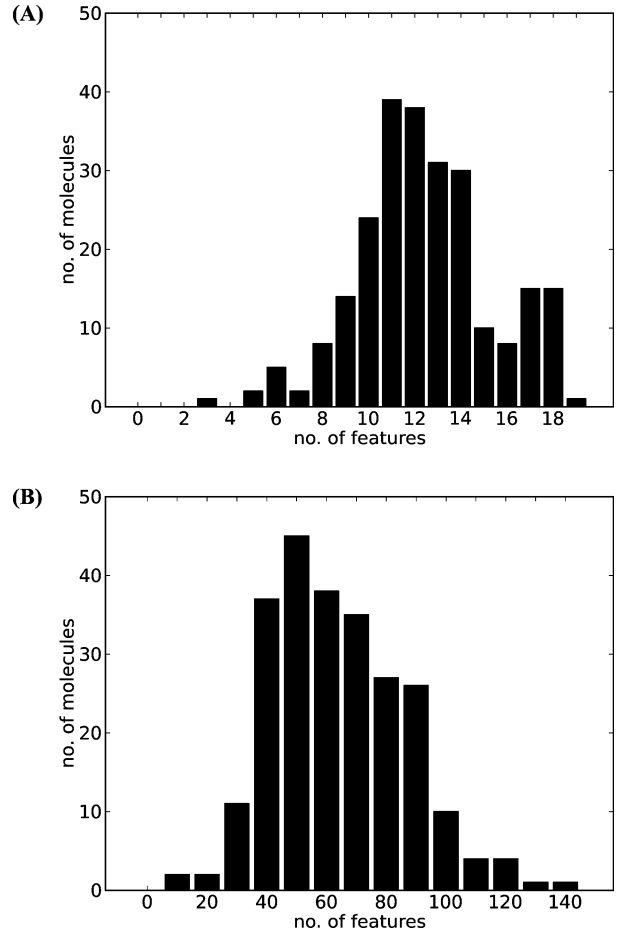


Figure 2. Distribution of atom pairs. Shown is a histogram with the distribution of atom pairs in all 243 selectivity set compounds analyzed herein. (A) BAP and (B) AP. For each test compound, the number of BAPs and APs generated was counted. For APs, feature numbers were binned using a bin size of 10.

RESULTS AND DISCUSSION

Bonded Atom Pairs. The design of BAPs has intentionally focused on capturing short-range atom environment information because many current topological descriptors predominantly consist of long-range features.[8−11] With information about hybridization states, aromatic character, and ring membership, BAPs contain information about planarity and geometric features of local atom environments. Standard APs also include neighboring atoms but contain less environment information. Moreover, the vast majority of APs span bond paths longer than one. In evaluating the new BAP descriptors, we put emphasis on chemical interpretability, in addition to benchmark calculations. Therefore, an atom pair frequency and a mapping analysis were carried out for sets of compounds having different target selectivity.

Feature Distribution and Molecular Size Effects. For the compound selectivity sets studied here, the distribution of BAPs and APs was recorded to better understand how many atom pairs are generated by typical active test compounds. The results are shown in Figure 2. These compounds produced between 3 and 19 BAPs and between

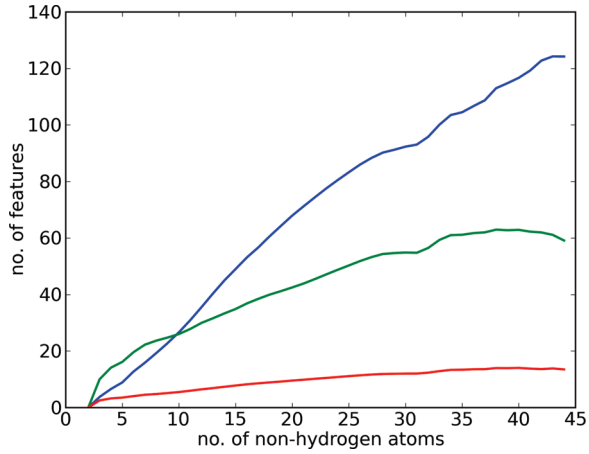


Figure 3. Molecular size dependence of feature distributions. The average number of BAPs (red), APs (blue), and MACCS (green) in compounds of increasing size, measured by the number of non-hydrogen atoms, is reported for a randomly selected subset of 100 000 ZINC compounds.

Table 2. Unique Bonded Atom Pairs with Largest Differences in Frequency of Occurrence in Target Pair Sets A/K, AK, and K/A[a]

BAP	A/K	AK	K/A	SE
[#7]:[#8$(a1aaaa1)]	33	4	0	0.47
c-!@O	40	11	0	0.76
[#6]:[#7$(a1aaaa1)]	73	33	0	0.9
C-!@P	13	7	0	0.94
[#7]:[#7$(a1aaaa1)]	47	30	0	0.96
P-!@O	13	11	0	0.99
[#6]:[#8$(a1aaaa1)]	33	7	5	1.11
a-[#G7]	20	19	5	1.4
[#6]:[#6$(a1aaaa1)]	60	19	30	1.42
c-!@C	93	63	70	1.56
C=@N	0	11	0	0
C-!@S	0	7	0	0
N-@N	0	7	0	0
N = !@O	7	30	0	0.69
C=@C	0	19	5	0.75
N-!@O	13	30	0	0.89
c-@C	20	41	0	0.91
N!@[C$(C@*)]	7	11	0	0.95
c-@N	33	41	0	0.99
O-!@S	7	7	0	1
C−C[#G7]	0	4	35	0.46
C-@A[#G7]	0	4	30	0.5
C-!@[#G7]	0	11	35	0.8
C-@O	0	15	25	0.95
[#6]:[#16$(a1aaaa1)]	7	4	25	1.15
C!@[N$(N@*)]	13	30	65	1.32
C!@[C$(C@*)]	20	30	55	1.46
C-@C	47	81	95	1.53
O!@[C$(C@*)]	47	70	85	1.54
C-!@N	73	59	95	1.56

[a] The top 10 target-pair unique BAPs on the basis of SE ranking are shown for each selectivity set. Reported are rounded frequency of occurrence values in percent, e.g. "100" means that 100% of the compounds in the set contained the specified atom pair (given in SMARTS[6] representation). Atom pairs are ranked in the order of ascending SE values calculated for frequency of occurrence distributions.

15 to 141 APs, hence, relatively small feature numbers compared to the size of the complete descriptor sets. In light of these findings, even a relatively small descriptor set, such as our 117 BAPs, should be highly discriminatory for many active compounds.

Table 3. Target Pair Sets A/N, AN, and N/A

BAP	A/N	AN	N/A	SE
c-!@S	4	0	0	0
C=@N	67	10	0	0.55
O-!@S	4	2	0	0.95
N-!@S	4	2	0	0.95
c-!@N	88	29	9	1.13
N-!@O	77	27	9	1.17
N = !@O	75	27	9	1.18
C!@[C$(C@*)]	71	27	13	1.26
[#6]:[#6$(a1aaaa1)]	75	44	22	1.43
C-!@O	88	78	31	1.47
N-@N	0	7	0	0
[#7]:[#7$(a1aaaa1)]	10	63	0	0.59
c-@C	13	73	3	0.8
C-@O	2	5	0	0.88
n-@C	19	41	0	0.89
C!@[N$(N@*)]	6	10	0	0.97
[#6]:[#7$(a1aaaa1)]	75	80	3.1	1.12
C=@C	17	22	16	1.57
C−C[#G7]	0	0	6	0
C-@A[#G7]	0	0	6	0
[#6]:[#16$(a1aaaa1)]	0	2	16	0.57
C-!@N	83	32	84	1.47
a-[#G7]	25	46	66	1.49
c-!@O	4	2	6	1.49
C-!@[#G7]	6	5	9	1.53
C-@C	90	59	97	1.55
N!@[C$(C@*)]	4	5	6.3	1.58
c-!@C	83	71	84	1.58

Because BAPs were primarily designed for similarity search applications, we also compared the molecular complexity or size dependence of their distribution with those of MACCS keys and APs. It is well-known that molecular complexity and size effects affect similarity searching using keyed fingerprint representations.[24] Simply put, the larger molecules are, the more bits are usually set on in a structural key or an atom pair fingerprint, which might lead to preferential detection of large database compounds.[24] Figure 3 reports the average numbers of BAPs, APs, and MACCS keys in ZINC compounds of increasing size. As can be seen, BAPs yield fewer features than those of MACCS and APs, consistent with the findings described above, but their numbers are relatively constant for molecules of increasing size, in contrast to MACCS keys and APs. Thus, BAP fingerprints would be much less affected by molecular size and by complexity effects than MACCS keys and, in particular, AP fingerprints.

Atom Pair Frequency Distributions. In order to search for atom pairs that might be specific or, at least characteristic, for compounds having similar selectivity, we determined the frequency of occurrence of BAPs and APs in the three selectivity sets of each target pair (e.g., A/K, AK, K/A) and ranked the atom pairs according to ascending SE values (i.e., zero entropy indicates set-specific and low entropy values characteristic features). For each target pair, only BAPs and APs were selected for further analysis that occurred with high frequency in one of the target sets but not the two others, as assessed by SE calculations. These atom pairs were most likely to reflect differences in target selectivity.

Our frequency distribution analysis, at the level of target pairs, identified subsets of BAPs (Tables 2−4) and also APs (Tables 5−7) that exclusively or preferentially occurred in each selectivity set of a target pair. As shown in Tables 2−4, the top 10 descriptors ranked by SE values contained several

Table 4. Target Pair Sets K/N, KN, and N/K

BAP	K/N	KN	N/K	SE
C!@[N$(N@*)]	27	0	0	0
N = !@O	55	26	0	0.91
N-!@O	55	30	0	0.94
c-!@N	73	30	12	1.26
O!@[C$(C@*)]	100	74	31	1.45
c-@N	82	78	27	1.45
n-@C	18	9	8	1.47
C-@N	100	83	54	1.54
[#6]:[#6$(a1aaaaa1)]	91	74	50	1.54
[#7]:[#8$(a1aaaa1)]	0	13	0	0
c-!@O	0	13	8	0.95
[#6]:[#8$(a1aaaa1)]	0	17	12	0.97
C=@C	55	65	0	0.99
C!@[C$(C@*)]	9	39	35	1.34
a-[#G7]	9	48	35	1.35
C-!@P	18	35	8	1.36
P-!@O	18	35	8	1.36
c-@C	45	70	19	1.42
C=@N	0	0	19	0
C-@O	0	0	12	0
N-@O	0	0	12	0
n-!@O	0	4	23	0.63
N!@[C$(C@*)]	0	4	12	0.85
n@a[#8]	0	22	38	0.94
[#7]:[#7$(a1aaaa1)]	18	30	50	1.47
C-!@N	55	26	58	1.51
[#6]:[#7$(a1aaaa1)]	36	48	73	1.52
c-!@C	36	39	65	1.53

Table 5. Unique Atom Pairs with Largest Differences in Frequency of Occurrence in Target Pair Sets A/K, AK, and K/A[a]

AP	A/K	AK	K/A	SE
C(2−0)_01_N(3−0)	13	0	0	0
C(3−1)_05_N(2−0)	27	7	5	1.21
C(3−1)_04_O(2−1)	33	15	5	1.26
C(2−0)_05_O(2−1)	20	7	5	1.33
C(2−0)_05_O(2−0)	20	4	15	1.35
C(3−1)_06_O(1−0)	80	33	60	1.5
C(2−0)_05_O(1−0)	60	33	35	1.53
C(2−1)_04_N(3−0)	7	30	0	0.69
O(1−1)_05_O(1−1)	7	22	0	0.78
C(1−0)_06_O(1−1)	7	15	0	0.89
C(2−1)_07_C(2−1)	0	19	10	0.93
C(1−0)_07_C(2−1)	7	11	0	0.95
C(2−0)_06_N(3−0)	7	30	5	1.14
C(2−1)_01_C(3−1)	13	26	5	1.33
C(2−1)_05_O(1−1)	20	59	45	1.46
C(4−0)_04_O(1−1)	0	4	25	0.55
C(2−1)_03_S(2−1)	0	4	20	0.63
C(2−1)_06_O(1−0)	0	15	70	0.67
C(3−1)_05_N(3−0)	13	48	60	1.38
C(3−0)_07_C(3−1)	40	30	90	1.42
C(2−0)_06_C(2−0)	53	37	80	1.52

[a] Target-pair unique APs are shown. Reported are rounded frequency of occurrence values in percent. Atom pairs are ranked in the order of ascending SE values calculated for frequency of occurrence distributions.

Table 6. Target Pair Sets A/N, AN, and N/A

AP	A/N	AN	N/A	SE
C(2−1)_06_O(2−0)	60	0	13	0.66
C(3−1)_01_N(4−0)	77	29	3	1.01
C(3−1)_07_O(1−0)	90	41	9	1.19
C(2−1)_01_N(3−0)	63	7	41	1.25
C(2−1)_06_O(1−0)	19	7	3	1.26
C(2−1)_07_N(2−0)	73	10	69	1.28
O(1−1)_04_N(2−0)	75	20	69	1.41
C(3−1)_03_N(2−0)	75	22	69	1.43
C(3−1)_07_C(3−1)	96	78	75	1.58
C(3−1)_01_C(1−2)	88	68	72	1.58
C(2−0)_07_C(3−0)	0	12	0	0
C(2−0)_05_N(2−1)	2	5	0	0.88
C(3−0)_04_C(2−1)	0	17	9	0.94
O(1−1)_01_O(2−1)	17	56	3	0.99
C(2−1)_07_C(2−1)	21	32	3	1.23
C(3−1)_01_Br	4	24	13	1.3
C(1−0)_07_C(2−1)	6	17	22	1.43
C(2−0)_04_C(2−0)	8	22	16	1.49
N(2−0)_07_Br	2	0	59	0.21
C(2−1)_04_S(2−1)	0	2	16	0.57
O(1−1)_05_O(1−1)	10	2	75	0.7
C(3−0)_06_C(2−1)	0	15	25	0.95
C(2−0)_04_N(2−0)	8	5	34	1.12
C(2−0)_06_C(2−1)	25	46	97	1.38
C(3−1)_04_N(2−0)	31	20	69	1.39

Table 7. Target Pair Sets K/N, KN, and N/K

AP	K/N	KN	N/K	SE
C(2−1)_07_C(2−1)	64	4	8	0.78
C(2−1)_07_N(3−0)	55	26	4	1.13
C(2−1)_05_N(3−0)	45	30	4	1.2
C(2−1)_05_O(1−1)	100	74	15	1.31
C(3−1)_06_N(2−0)	27	13	12	1.47
C(3−1)_01_O(2−1)	100	70	46	1.52
C(2−1)_07_C(3−1)	82	52	50	1.55
C(1−0)_05_C(2−0)	0	13	4	0.77
O(1−0)_06_O(1−1)	18	52	0	0.82
C(3−1)_07_O(2−0)	9	17	0	0.93
C(3−1)_03_P(4−1)	18	30	0	0.95
O(1−1)_03_P(4−1)	18	35	8	1.36
C(3−1)_03_Cl	9	43	31	1.37
C(3−1)_07_O(1−0)	64	78	50	1.56
C(3−0)_03_O(2−0)	0	0	19	0
C(3−0)_03_C(3−0)	0	4	35	0.5
C(3−0)_05_O(1−1)	0	9	23	0.85
C(3−1)_07_N(2−1)	0	4	12	0.85
O(2−0)_07_O(1−1)	9	0	15	0.95
C(3−0)_06_C(3−1)	0	22	35	0.96
C(2−0)_01_O(2−0)	0	17	27	0.97
C(3−1)_07_O(2−1)	0	9	12	0.99
C(2−0)_06_C(3−1)	36	30	62	1.52

specific or characteristic BAPs for each selectivity set, including multiple descriptors with zero entropy. Tables 5−7 show that characteristic APs, but fewer low-entropy descriptors, were also found for different selectivity sets. Furthermore, in all but one case (set A/N), only fewer than 10 unique APs were identified per set, although there were about 40 times more APs than BAPs available. Moreover, Tables 5−7 also show that most of the characteristic APs were separated by long bond paths (often 5−7) and, thus, represented long-range features.

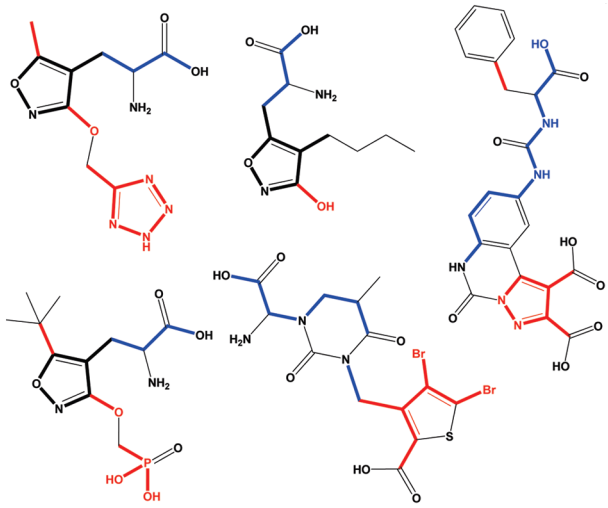
Taken together, these observations suggested that BAPs are of high chemical resolution and are responsive to characteristic features of immediate atom environments. Furthermore, BAPs revealed more structural differences between the compound sets analyzed here than APs.

Characteristic Atom Pairs. Distinct atom pair information was found to be characteristic for different selectivity sets. For example, for the A/K set, signature BAPs included a pair of sp^2-hybridized aromatic carbon atoms and a sp^2 aromatic carbon, oxygen, or nitrogen atom in five-membered rings or, alternatively, a sp^2 aromatic carbon atom bound to a sp^3 carbon or heteroatoms in

acyclic moieties. Signature APs for this set included sp^3 carbon atoms bound to sp^3 nitrogen atoms and sp^2 carbon atoms connected to oxygen atoms over a six-bond path. By contrast, for the inverse selective K/A set, BAP signatures were pairs of sp^3 carbon atoms and sp^3 oxygen or halogen atoms in cyclic or acyclic structures, and AP signatures sp^3 carbon atoms connected to sp^2 oxygen atoms through a four-bond path. Thus, atom pair signatures for the inverse selective A/K and K/A sets completely differed in their chemical nature, and similar examples were found for essentially all target pair selectivity sets. Thus, the analysis pointed at intuitive chemical differences between selectivity sets.

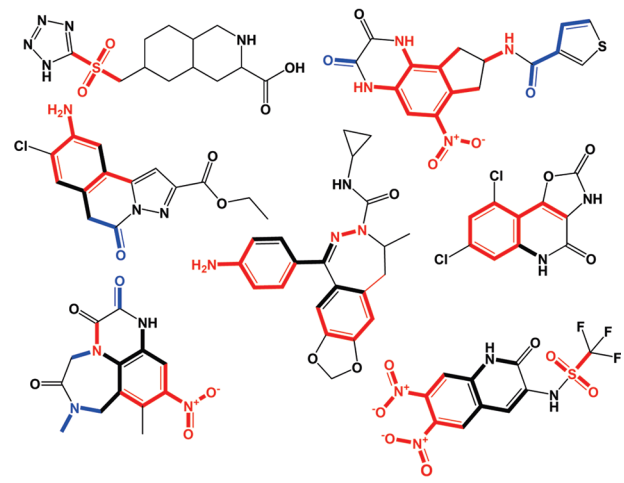
Mapping of Atom Pair Signatures. For a structural interpretation of statistical atom pair analysis, specific and characteristic BAPs and APs were also mapped on the

A



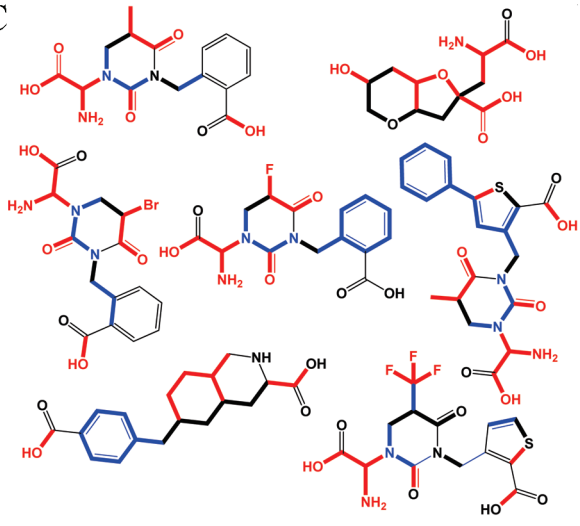
A/K BAPs	A/K APs
[#6]:[#6$(a1aaaa1)]	C(3-1)_06_O(1-0)
c-!@O	C(2-0)_05_O(1-0)
c-!@C	C(3-1)_04_O(2-1)
[#6]:[#7$(a1aaaa1)]	C(3-1)_05_N(2-0)
[#7]:[#7$(a1aaaa1)]	C(2-0)_05_O(2-0)
[#6]:[#8$(a1aaaa1)]	C(2-0)_05_O(2-1)
[#7]:[#8$(a1aaaa1)]	C(2-0)_01_N(3-0)
a-[#G7]	
C-!@P	
P-!@O	

B



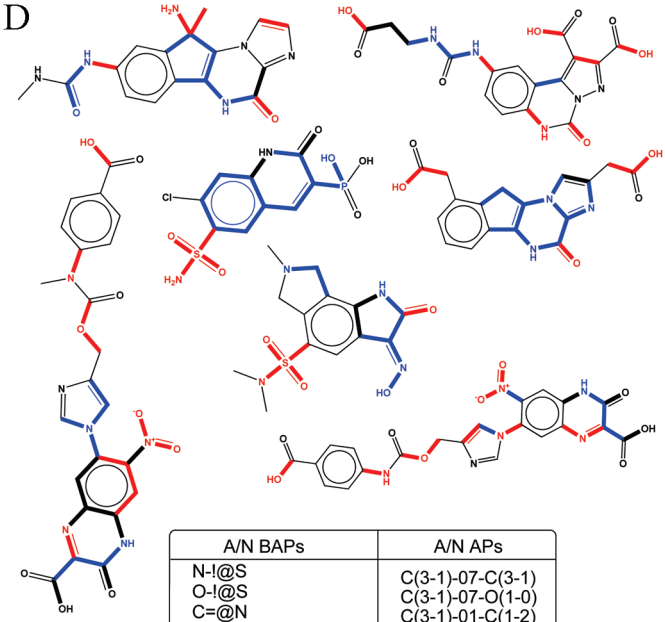
AK BAPs	AK APs
c-!@N	C(2-1)_05_O(1-1)
N!@[C$(C@*)]	C(2-0)_06_N(3-0)
c-@C	C(2-1)_04_N(3-0)
c-@N	C(2-1)_01_C(3-1)
[#6]:[#6$(a1aaaaa1)]	O(1-1)_05_O(1-1)
N-!@O	C(2-1)_07_C(2-1)
C=@C	C(1-0)_06_O(1-1)
N=!@O	C(1-0)_07_C(2-1)
C=@N	
O-!@S	
C-!@S	
N-@N	

C



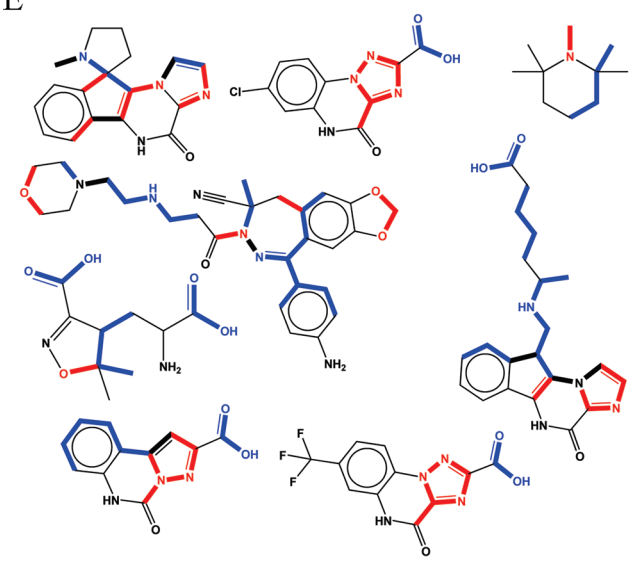
K/A BAPs	K/A APs
C-!@C	C(3-0)_07_C(3-1)
C-@C	C(2-0)_06_C(2-0)
C-!@O	C(2-1)_06_O(1-0)
C-!@N	C(2-0)_05_N(3-0)
C-!@[#G7]	C(3-1)_05_N(3-0)
C-@A[#G7]	C(3-1)_01_N(3-0)
C-@O	C(4-0)_04_O(1-1)
[#6]:[#16$(a1aaaa1)]	C(2-1)_03_S(2-1)
C!@[C$(C@*)]	
C!@[N$(N@*)]	
O!@[C$(C@*)]	

D



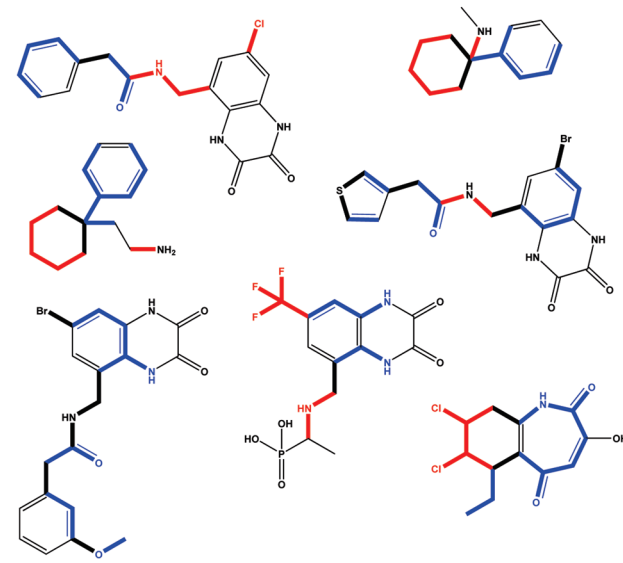
A/N BAPs	A/N APs
N-!@S	C(3-1)-07-C(3-1)
O-!@S	C(3-1)-07-O(1-0)
C=@N	C(3-1)-01-C(1-2)
N=!@O	C(2-1)-01-C(1-2)
[#6]:[#6$(a1aaaa1)]	C(3-1)-01-N(3-0)
N-!@O	C(3-1)-03-N(2-0)
C-!@O	O(1-1)-04-N(2-0)
C-!@C	C(2-1)-07-C(2-0)
c-!@S	C(2-1)-01-N(3-0)
c-!@N	C(2-1)-06-O(2-0)
c-@N	
C!@[C$(C@*)]	

E



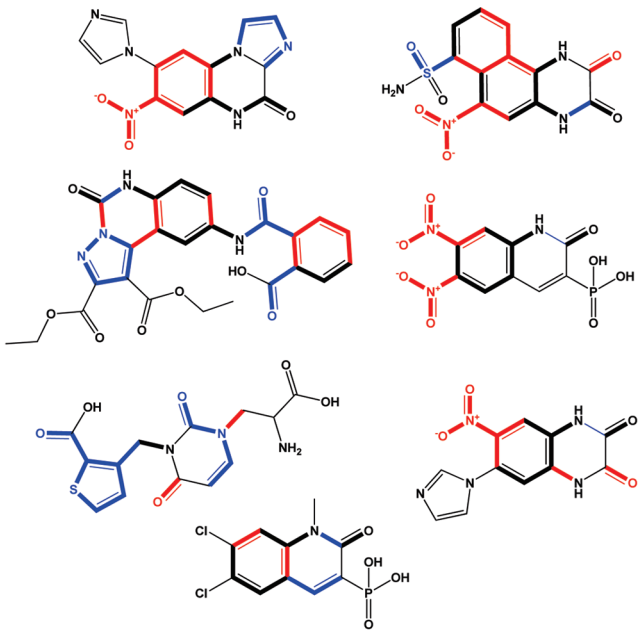
AN BAPs	AN APs
c-@C n-@C C!@[N$(N@*)] C=@C [#6]:[#7$(a1aaaa1)] [#7]:[#7$(a1aaaa1)] N-@N C-@O	O(1-1)_01_O(2-1) C(2-1)_07_C(2-1) C(2-0)_04_C(2-0) C(1-0)_07_C(2-1) C(3-0)_04_C(2-1) C(2-0)_07_C(3-0) C(2-0)_05_N(2-1)

F



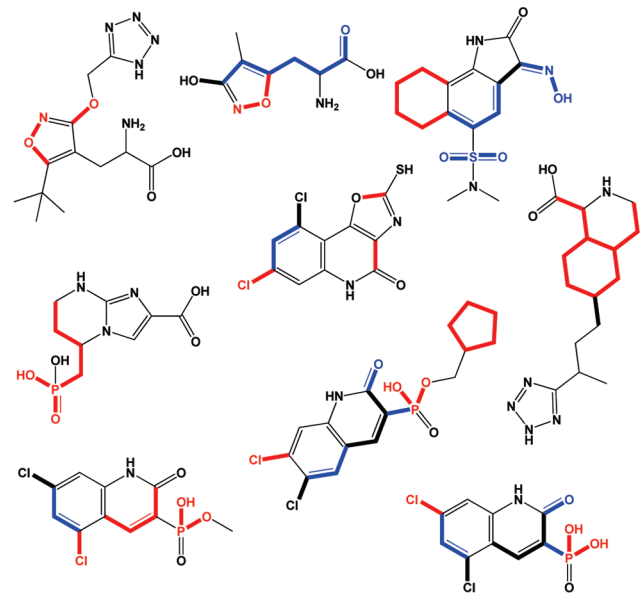
N/A BAPs	N/A APs
C-@C c-!@C C-!@N a-[#G7] [#6]:[#16$(a1aaaa1)] C-!@[#G7] N!@[C$(C@*)] c-!@O C-C[#G7] C-@A[#G7]	C(2-0)_06_C(2-1) C(2-0)_01_O(2-1) O(1-1)_05_O(1-1) C(3-1)_04_N(2-0) N(2-0)_07_Br C(2-0)_04_N(2-0) C(3-0)_06_C(2-1)

G



K/N BAPs	K/N APs
C-@N O!@[C$(C@*)] [#6]:[#6$(a1aaaaa1)] c-@N c-!@N N-!@O N=!@O C!@[N$(N@*)] n-@C	C(2-1)_05_O(1-1) C(2-1)_07_C(3-1) C(2-1)_07_N(3-0) C(2-1)_05_N(3-0) C(3-1)_06_N(2-0)

H



KN BAPs	KN APs
C-@C c-@C C=@C a-[#G7] C!@[C$(C@*)] C-!@P P-!@O [#6]:[#8$(a1aaaa1)] [#7]:[#8$(a1aaaa1)] c-!@O	C(3-1)_07_O(1-0) O(1-0)_06_O(1-1) C(3-1)_03_Cl O(1-1)_03_P(4-1) C(3-1)_03_P(4-1) C(3-1)_07_O(2-0)

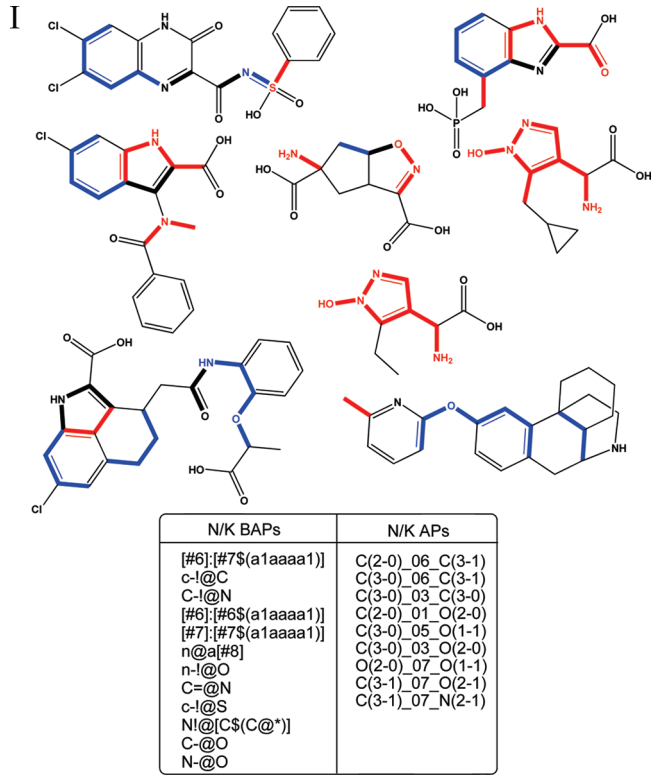


Figure 4. Mapping of atom pairs. Most discriminatory BAPs (red) and APs (blue) are mapped onto different compound selectivity sets. Overlapping BAPs and APs are shown in bold black. (A) Selectivity set A/K, (B) AK, (C) K/A, (D) A/N, (E) AN, (F) N/A, (G) K/N, (H) KN, and (I) N/K.

compounds they originated from. The results of these mapping studies are shown in Figure 4. Several observations were made. Specific or characteristic BAPs were often found to delineate coherent substructures, which was also observed for APs, albeit to a lesser extent (as discussed above, there were also fewer characteristic APs than BAPs). Furthermore, there was only limited overlap of substructures formed by BAPs and APs. However, in all cases, distinct substructures were formed by BAPs and APs (red and blue, respectively, in Figure 4). These observations were consistently made for all selectivity sets and indicated that BAPs and APs complemented each other in the compound set-specific structural information they captured. This complementary nature has been a key finding of feature mapping analysis.

Activity Determinants. We have also mapped BAPs for compounds with well-established SARs, i.e., series of inhibitors of tyrosine kinases Abl (Gleevec analogs) and Cdk2 (Purvalanol analogs).[25] BAPs were calculated for all compounds, and those occurring in at least two inhibitors were mapped, as shown in Figure 5. In Gleevec analogs, these BAPs delineated the molecular region encompassing the 4-(3-pyridyl) pyrimidine and aryl carboximide moieties (Figure 5A). In Purvalanol analogs, BAPs were found to map the purine or 2-aminopurine group (Figure 5B). These molecular regions in Gleevec and Purvalanol analogs are known to be critical for binding.[25]

Similarity Searching. In our similarity search calculations, we exclusively used target-selective compounds as reference molecules and monitored the recall of selective, non-selective, and inverse selective compounds. Maximizing the number of selective compounds and minimizing the number of non-selective and inverse selective compounds in a search was the general goal of these calculations. The results of the similarity search calculation are summarized in Figure 6. In general, both atom pair and structural key fingerprints produced satisfactory recovery rates for selective compounds of up to ~80% in database selection sets of ~100 molecules and much lower rates of maximal 20% for non- or inverse selective compounds. However, there were set-specific differences in search performance. For example, for the A/K set (Figure 6A), AP-FP recall rates were much lower than for BAP-FP or MACCS, and BAP-FP detected fewest inverse selective compounds. The complementary nature of structural features preferentially captured by BAPs and APs was also apparent in several cases. For example, for the K/A set in Figure 6B, AP-BAP-FP further increased the recall of AP-FP and BAP-FP. In addition, in search calculations using A/N- or N/A-selective (Figure 6C and D, respectively) reference molecules, AP-BAP-FP detected fewest non- and inverse selective compounds. Furthermore, the overall search performance of BAP-FP was promising. Only in one instance, set N/A in Figure 6D, BAP-FP recovery rates were considerably lower than for the other fingerprints. In this case, MACCS dominated the search. However, BAP-FP performed best in three cases including A/K, K/N, and N/K (Figure 6A, E, and F, respectively). Thus, although BAP-FP contained the smallest number of features (i.e., only 2.5% of the feature number of AP-FP), its search performance was at least comparable to the reference fingerprints.

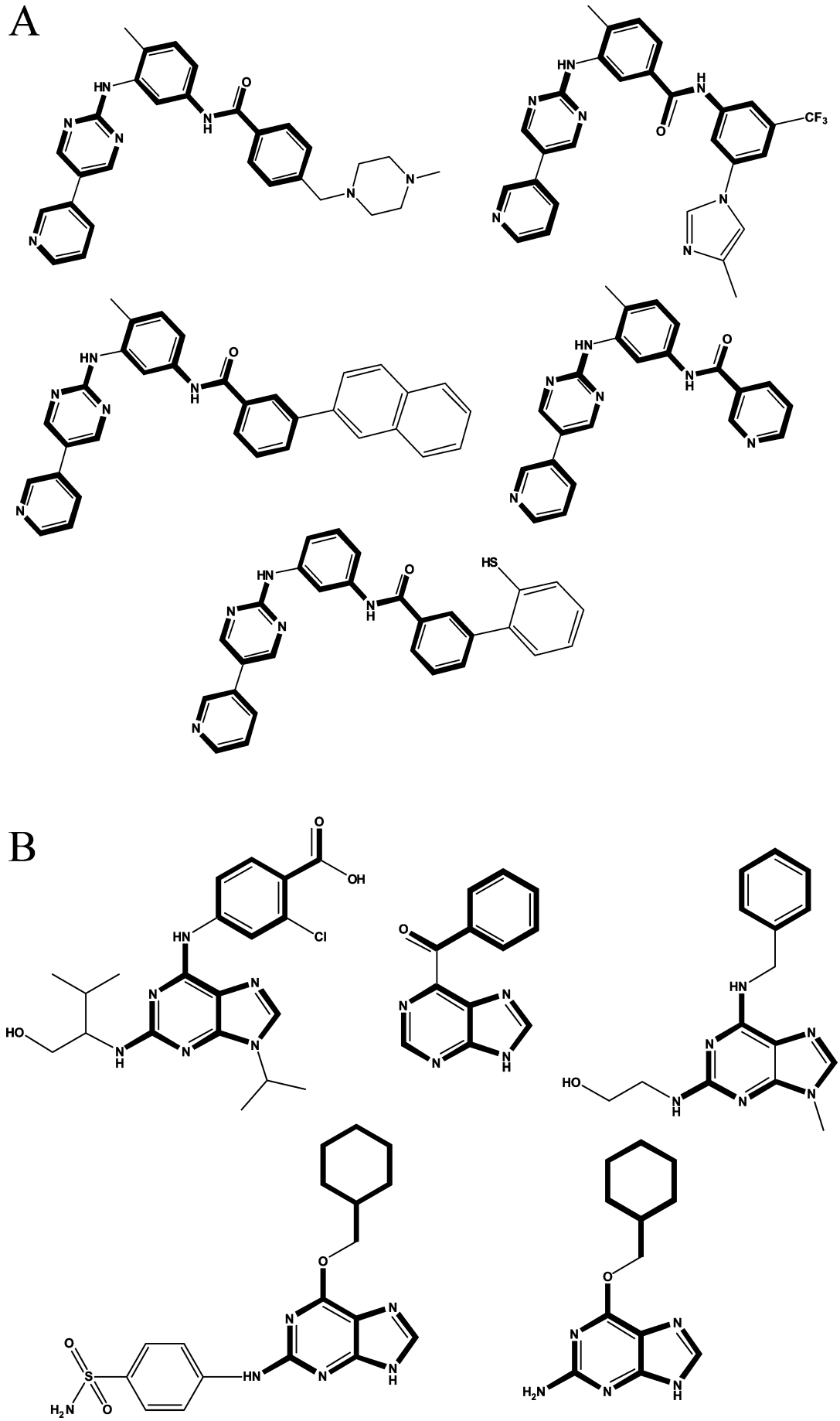
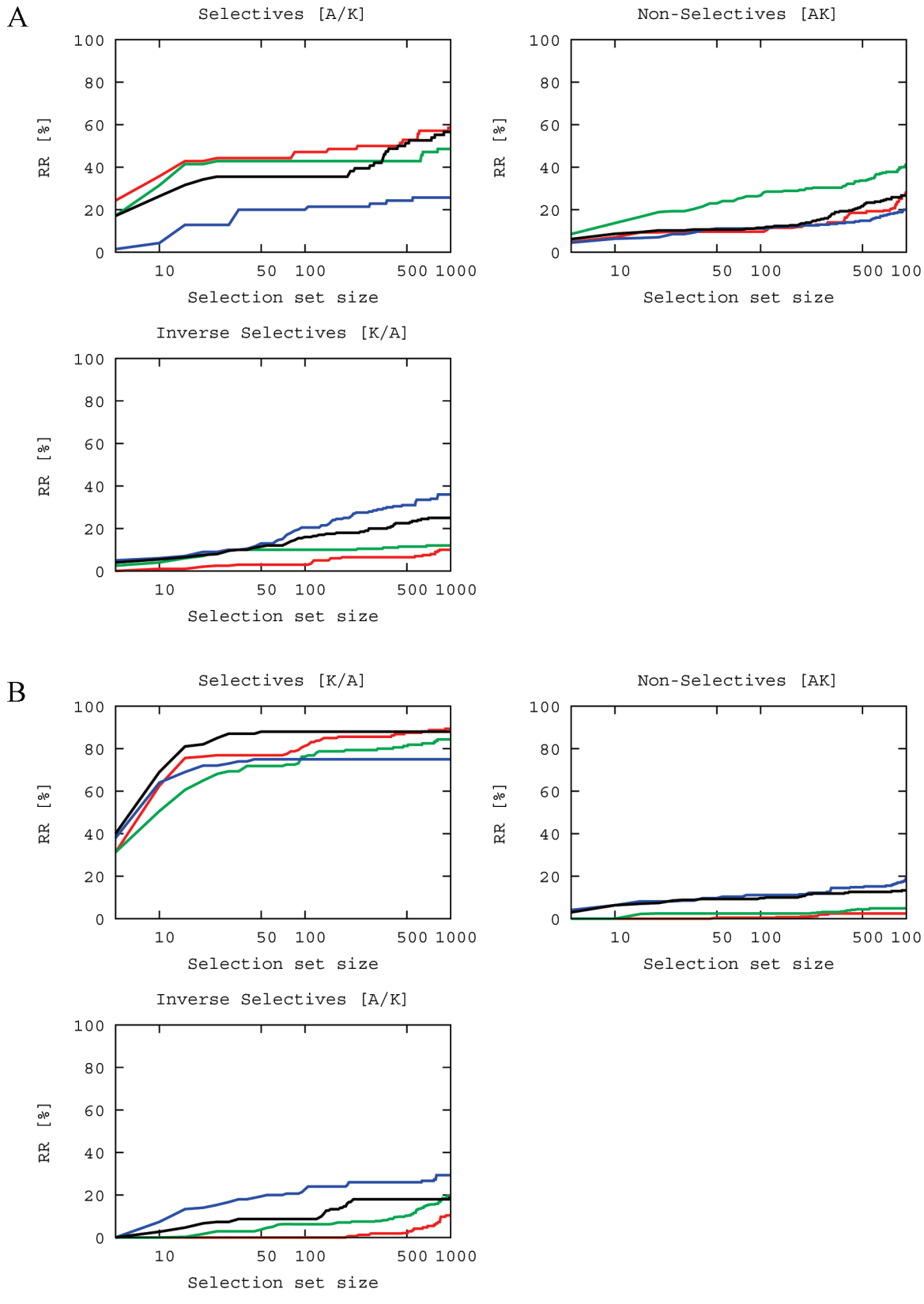
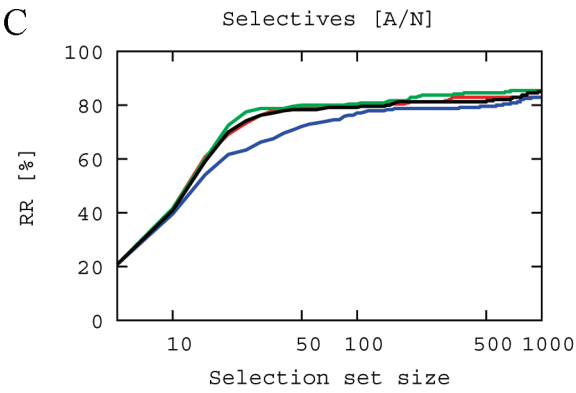


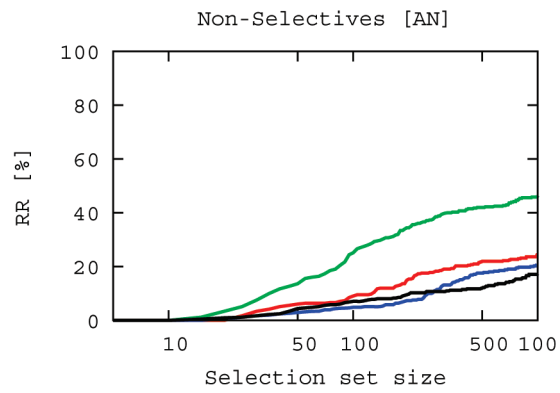
Figure 5. Mapping of activity determinants. BAPs were calculated and mapped for series of tyrosine kinase inhibitors. Shown are well-known inhibitors of (A) Abl and (B) Cdk2. Substructures delineated by more than one BAP are shown in bold.

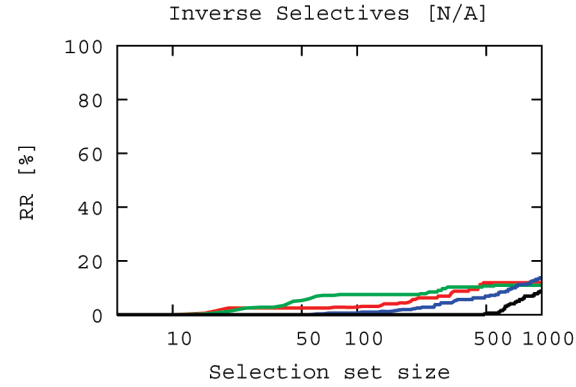
APs and BAPs utilize specific atom types, which implicitly accounts for many atomic properties (including, for example, size, hydrophobic/polar character or electronegativity differences). The only difference in the specification of atom types between APs and BAPs is that the set of 117 bonded atom pairs uses generalized halogen atoms, whereas APs contain specific halogen atom types. Therefore, we also generated a further expanded set of 159 BAPs with explicit halogen atom types (reported in Supporting Information, Table S2) and repeated similarity search calculations on the halogen-containing selectivity sets. We found that compound recovery rates achieved by the original and extended BAP sets varied by only less than 1%, indicating that the use of explicit halogen atom types was not critical in these cases.



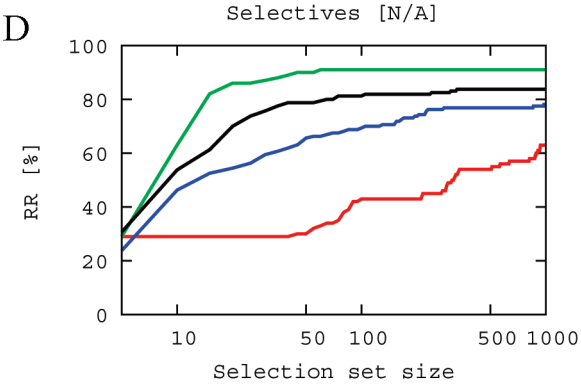
C

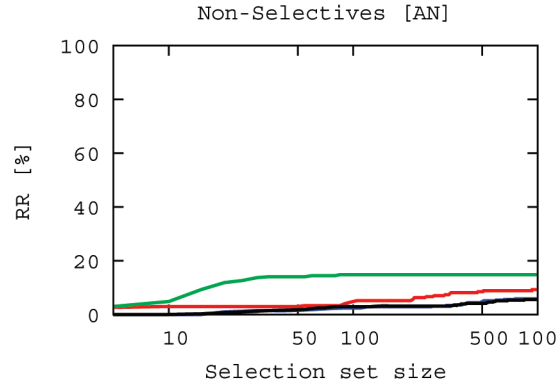

Selectives [A/N]
RR [%]
Selection set size
100
80
60
40
20
0
10
50
100
500
1000

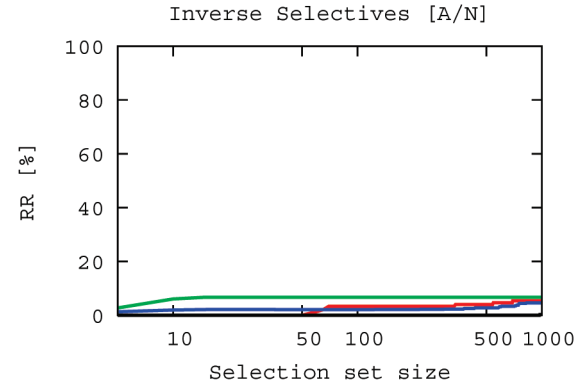

Non-Selectives [AN]
RR [%]
Selection set size
100
80
60
40
20
0
10
50
100
500
1000


Inverse Selectives [N/A]
RR [%]
Selection set size
100
80
60
40
20
0
10
50
100
500
1000

D


Selectives [N/A]
RR [%]
Selection set size
100
80
60
40
20
0
10
50
100
500
1000


Non-Selectives [AN]
RR [%]
Selection set size
100
80
60
40
20
0
10
50
100
500
1000


Inverse Selectives [A/N]
RR [%]
Selection set size
100
80
60
40
20
0
10
50
100
500
1000

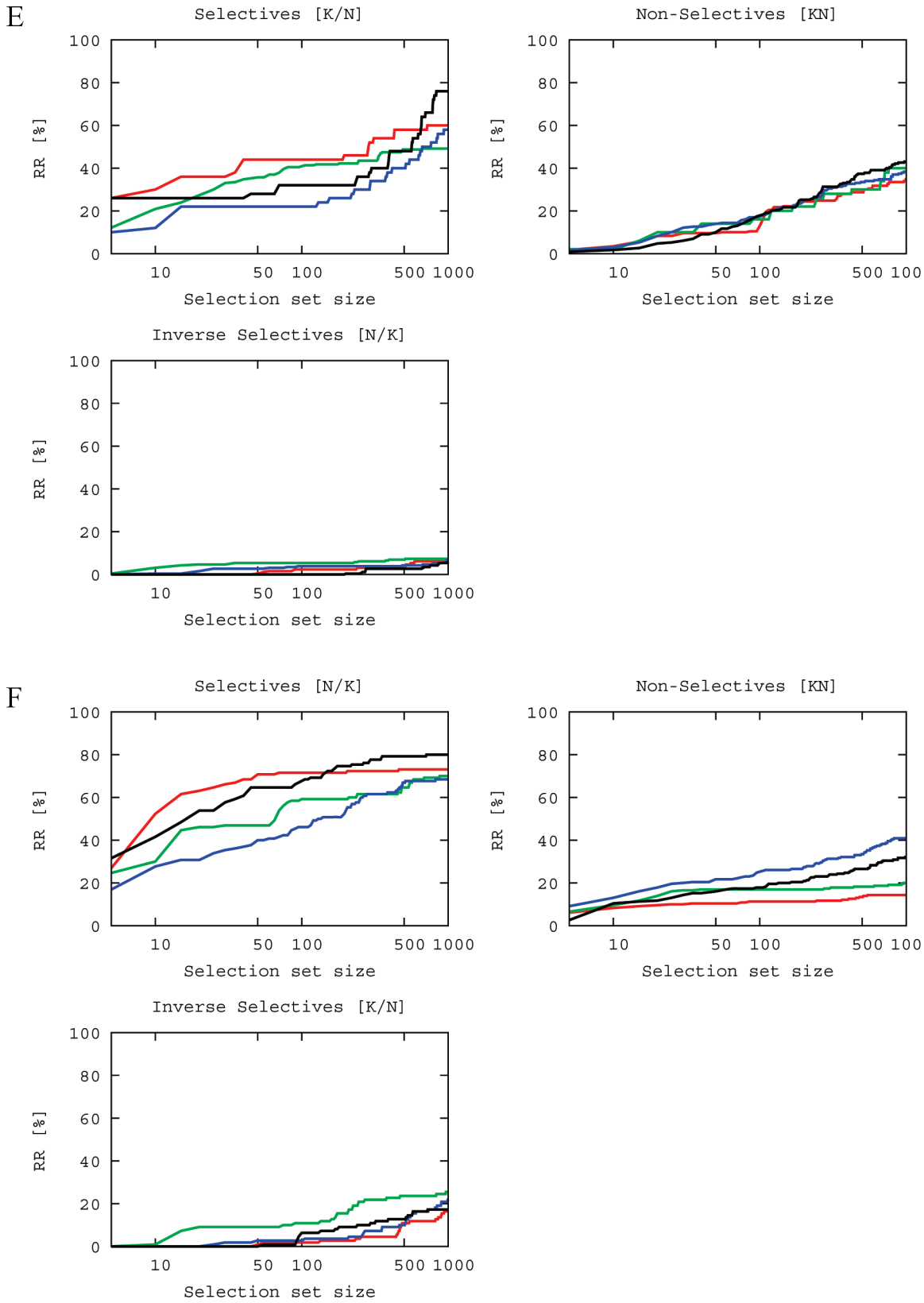


Figure 6. Similarity searching. Cumulative recall curves are displayed for different compound sets and 1-NN search calculations using BAP-FP (red), AP-FP (blue), MACCS (green), and AP-BAP-FP (black). Compound selection set size is given on a logarithmic scale. RR stands for recovery rate. (A) **A/K**, AK, K/A; (B) **K/A**, AK, A/K; (C) **A/N**, AN, N/A; (D) **N/A**, AN, A/N; (E) **K/N**, KN, N/K; and (F) **N/K**, KN, K/N. Selectivity sets from which reference molecules for the search trials were selected are shown in bold.

CONCLUSIONS

We have introduced bonded atom pairs that capture short-range chemical atom environment information. Bonded atom pairs depart in their design from many other fragment-type descriptors that enumerate bond pathways of varying length and predominantly account for long-range topological information. A total of 117 generally applicable unique bonded atom pairs have been generated

that can be utilized as a descriptor dictionary. Our analysis has shown that bonded atom pairs account for more compound set-specific information than those of conventional atom pairs, at least in the cases studied here. However, structural information captured by bonded and standard atom pairs was found to be complementary. In similarity searching, 117 bonded and 4662 conventional atom pairs were compared, and search performance was found to be at least comparable, suggesting that BAPs are rich in chemical information. Hence, these newly designed atom pairs add to the spectrum of currently available molecular descriptors, and our findings suggest that topological descriptors focusing on short-range atom environments should merit further investigation in similarity searching and compound classification.

ACKNOWLEDGMENT

The authors thank Eugen Lounkine for helpful discussions. H.E.A.A. is supported by a fellowship from the Egyptian Government (Al-Azhar University).

Supporting Information Available: Table S1 contains the complete set of 117 bonded atom pairs with generalized halogen atoms. Table S2 provides a further expanded set of 159 atom pairs with different halogen atom types. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
(2) Eckert, H; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
(3) Willet, P. Searching Techniques for Databases of Two- and Three-dimensional Structures. *J. Med. Chem.* **2005**, *48*, 1–17.
(4) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(5) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(6) *SMARTS*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008.
(7) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part II. Atom-Centered Fragments. *J. Chem. Soc. C* **1971**, 3702–3706.
(8) 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.
(9) 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.
(10) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
(11) Bender, A.; Mussa, H. Y.; Glen, R. C. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
(12) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.
(13) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153–157.
(14) Adamson, G. W.; Bush, J. A. A Method for the Automatic Classification of Chemical structures. *Inf. Storage Retr.* **1973**, *9*, 561–568.
(15) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(16) Liu, K.; Feng, J.; Young, S. R. PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 515–522.
(17) Ahmed, H. E. A.; Geppert, H.; Stumpfe, D.; Lounkine, E.; Bajorath, J. Methods for Computer-aided Chemical Biology. Part 4: Selectivity Searching for Ion Channel Ligands and Mapping of Molecular Fragments as Selectivity Markers. *Chem. Biol. Drug Des.* **2009**, *73*, 273–282.
(18) Shannon, C. E., Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1963.
(19) Duda R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley-Interscience: New York, 2001; pp. 20−83.
(20) 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.
(21) Hert, J.; Willet, 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.
(22) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(23) *Molecular Operating Environment (MOE), Version 2008.10;* Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2008.
(24) Wang, Y.; Bajorath, J. Balancing Complexity Effects in Fingerprint Similarity Searching. *J. Chem. Inf. Model.* **2008**, *48*, 75–84.
(25) Aronov, A.; Bemis, G. W. A Minimalist Approach to Fragment-Based Ligand Design Using Common Rings and Linkers: Application to Kinase Inhibitors. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 36–50.

CI900512G