Chemical Similarity Using Physiochemical Property Descriptors†

Simon K. Kearsley,* Susan Sallamack, Eugene M. Fluder, Joseph D. Andose, Ralph T. Mosley, and Robert P. Sheridan

Department of Molecular Design and Diversity, Merck Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065

Received August 10, 1995⊗

Similarity searches using topological descriptors have proved extremely useful in aiding large-scale screening. We describe alternative forms of the atom pair (Carhart et al. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.) and topological torsion (Nilakantan et al. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.) descriptors that use physiochemical atom types. These types are based on binding property class, atomic log *P* contribution, and partial atomic charges. The new descriptors are meant to be more "fuzzy" than the original descriptors. We propose objective criteria for determining how effective one descriptor is versus another in selecting active compounds from large databases. Using these criteria, we run similarity searches over the Derwent Standard Drug File with ten typical druglike probes. The new descriptors are not as good as the original descriptors in selecting actives if one considers the average over all probes, but the new descriptors do better for several individual probes. Generally we find that whether one descriptor does better than another varies from probe to probe in a way almost impossible to predict a priori. Most importantly, we find that different descriptors typically select very different sets of actives. Thus it is advantageous to run similarity probes with several types of descriptors.

INTRODUCTION

Similarity searches are now a standard tool for drug discovery.[1,2] The idea behind such searches is that, given a compound with a desired biological activity, compounds that are "similar" to it in structure are likely to have a similar activity. In practice, an investigator might know the structure of an active compound found in the patent literature or from random screening. He uses the active as a "probe" to search over a database of sample-available compounds and find those which are most similar. He then submits the compounds for testing in the appropriate biological assay. A useful feature of similarity searches, as opposed to substructure searches, is that the user need not specify what part(s) of the probe are important. Such knowledge is lacking early in a drug-discovery project, when one knows only one or two actives.

Since most chemical databases consist only of connection tables, i.e., atoms and bonds without three-dimensional information, molecular similarity measurements are usually based on topological features. Many such measurements are descriptor-based. In descriptor-based similarity searching, connection tables in the database are parsed into chemical descriptors, and these descriptors are stored in a separate descriptor database. At search time, the probe is similarly parsed. The similarity between the probe and each database entry is calculated by comparing the list of descriptors in the probe with the list for the entry. Since the information of which descriptor corresponds to which molecular feature is lost, one cannot obtain an equivalence between features in the probe and a database entry. However, since it is computationally inexpensive to compare lists of descriptors, searches can be done quickly.

Descriptors can be global properties (e.g., hydrophobicity, chemical connectivity indices, etc.) or the presence or absence of particular substructures. An ideal set of descriptors would be general enough so that chemically equivalent groups would be perceived as similar but specific enough so that inequivalent groups would not be. One set of useful substructure descriptors with which we have experience are those developed at Lederle Laboratories, the atom pair[3] and topological torsion.[4] These descriptors have two very desirable properties. First, they are easily calculated from connection tables. Second, they are general enough to be common among compounds in diverse chemical classes but specific enough in the aggregate to be able to distinguish between closely related analogs. Similarity searches using these descriptors are typically able to discover active compounds which are in different chemical classes than the probe.[3−5]

One drawback of these descriptors for similarity search applications is that they are very specific. They distinguish atom types on the basis of element, number of non-hydrogen neighbors, and number of π electrons. This does not allow for the perception of many types of physiochemical equivalences. For example, benzoic acid would not be perceived as very similar to phenyltetrazole even though carboxylates and tetrazoles are both anions. To remedy this, we devised alternative representations of these descriptors that use "fuzzy", i.e., less specific, atom types. In this paper we try three properties to define the types: binding property class, atomic log *P* contribution, and partial atomic charges. The hope is that similarity searches with such descriptors would be more likely to discover active compounds very different from the probe in terms of its chemical drawing representation. We describe the incorporation of these new descriptors into our in-house TOPOSIM system such that the user can calculate similarity based on any combination of descriptors. Finally, we show for several typical probe molecules the relative utility of the descriptors in selecting active compounds from a large database of druglike compounds.

* To whom correspondence should be addressed.

† Keywords: similarity searching, property descriptors, atom pairs, topological torsions, Derwent Standard Drug File.

⊗ Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0095-2338/96/1636-0118$12.00/0

METHODS

Definitions of Descriptors. Our descriptors are extensions of the atom pair described by Carhart et al.[3] and the topological torsion described by Nilakantan et al.[4] Atom pairs are substructures of the form

atom type i −(distance) − atom type j

where (distance) is the distance in bonds between atom i and atom j along the shortest path. A molecule with *n* atoms will have $n(n-1)/2$ atom pairs, although generally not all of the atom pairs will be unique. The topological torsion is of the form

atom type i − atom type j − atom type k − atom type m

where i, j, k, and m are consecutively bonded atoms. There is a topological torsion for every unique set of four bonded atoms. In the original descriptors (which in this paper we will refer to as the *ap* and *tt*, respectively), "atom type" includes element, number of neighbors, and number of π electrons. Figure 1 shows an example of a molecule parsed into *ap*'s and *tt*'s.

Our new descriptors are constructed analogously to the *ap* and *tt*'s except that the atom type is defined differently. In binding property pairs (*bp*) and binding property torsions (*bt*), the atoms are assigned to one of seven binding property classes for non-hydrogen atoms: 1 = cations, 2 = anions, 3 = neutral hydrogen bond donors, 4 = neutral hydrogen bond acceptors, 5 = polar atoms (both donor and acceptor, e.g., hydroxy oxygen), 6 = hydrophobic atoms, 7 = other. Figure 1 shows the sample molecule parsed into *bp*'s and *bt*'s.

In calculating hydrophobic pairs and torsions (*hp*,*ht*) and charge pairs and torsions (*cp*, *ct*), the atom types are more problematical since hydrophobic contribution and charge are continuous rather than discrete entities. We therefore use a set of overlapping "buckets" numbered 1−7 to represent ranges of these properties: 1 = −0.50 or below, 2 = −0.75 to −0.25, 3 = −0.50 to 0.00, 4 = −0.25 to 0.25, 5 = 0.00 to 0.50, 6 = 0.25 to 0.75, 7 = 0.50 or above. Each atom can be in two buckets, a lower bucket and an upper bucket. For instance, an atom with a charge of −0.33 would be in lower bucket 2 and upper bucket 3. There are two atom pair descriptors listed for each pair of atoms i and j in the molecule, one with

atom type i = lower bucket i
atom type j = lower bucket j

and one with

atom type i = upper bucket i
atom type j = upper bucket j

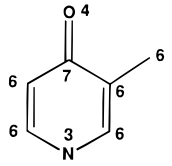
Similarly there are two topological torsions listed for each set of bonded i-j-k-m atoms, one with

atom type i = lower bucket i
atom type j = lower bucket j
atom type k = lower bucket k
atom type m = lower bucket m

and one with

atom type i = upper bucket i
atom type j = upper bucket j
atom type k = upper bucket k
atom type m = upper bucket m.

For most of the descriptors, only non-hydrogen atoms are used, the exceptions being *cp* and *ct*, for which polar hydrogens count as atoms. How binding property classes, atomic log *P* contributions, and atomic charges are assigned is discussed in more detail in the next section.



	unique ap	frequency
1	CX3.-(2)-CX2.	3
2	CX3.-(1)-CX2.	2
3	OX1.-(3)-CX2.	2
4	NX2 -(1)-CX2.	2
5	CX3.-(3)-CX2.	1
6	CX2.-(2)-CX2.	1
7	CX2.-(1)-CX2.	1
8	CX2.-(3)-CX2.	1
9	CX3.-(1)-CX3.	1
10	NX2 -(3)-CX3.	1
11	NX2 -(2)-CX3.	1
12	OX1.-(2)-CX3.	1
13	OX1.-(4)-NX2	1
14	NX2 -(2)-CX2.	1
15	OX1.-(1)-CX3.	1
16	OX1.-(2)-CX2.	1
17	CX3.-(1)-CX1	1
18	CX2.-(4)-CX1	1
19	CX3.-(2)-CX1	1
20	OX1.-(3)-CX1	1
21	CX2.-(3)-CX1	1
22	CX2.-(2)-CX1	1
23	NX2 -(3)-CX1	1
		28 total

	unique bp	frequency
1	4-(3)-6	3
2	6-(1)-6	3
3	6-(2)-6	3
4	6-(3)-6	3
5	6-(2)-7	3
6	3-(1)-6	2
7	3-(2)-6	2
8	4-(2)-6	2
9	6-(1)-7	2
10	3-(4)-4	1
11	3-(3)-6	1
12	3-(3)-7	1
13	4-(1)-7	1
14	6-(4)-6	1
		28 total

	unique tt	frequency
1	OX1.-CX3.-CX3.-CX2.	1
2	OX1.-CX3.-CX3.-CX1	1
3	CX3.-CX3.-CX2.-CX2.	1
4	CX2.-CX3.-CX3.-CX2.	1
5	CX2.-CX3.-CX3.-CX1	1
6	NX2 -CX2.-CX3.-CX3.	1
7	OX1.-CX3.-CX2.-CX2.	1
8	CX3.-CX2.-NX2 -CX2.	1
9	NX2 -CX2.-CX2.-CX3.	1
10	NX2 -CX2.-CX3.-CX1	1
11	CX2.-NX2 -CX2.-CX2.	1
		11 total

	unique bt	frequency
1	4-7-6-6	3
2	6-6-7-6	3
3	3-6-6-7	2
4	6-3-6-6	2
5	3-6-6-6	1
		11 total

Figure 1. Sample molecule parsed into descriptors. The original atom pairs (*ap*) and topological torsions (*tt*) are shown as well as binding property pairs (*bp*) and binding property torsions (*bt*). In *ap* and *tt*, atoms are distinguished by the element, number of non-hydrogen neighbors, and number of π electrons. For instance "OX1." is an oxygen with one neighbor and one π electron (e.g., a carbonyl oxygen). The *ap* "NX2-(4)-OX1." is an sp^3 nitrogen with two neighbors four bonds away from a carbonyl oxygen. The *tt* "NX2-CX2.-CX2.-CX3." represents four consecutively bonded atoms. In binding property descriptors, atom types are assigned to one of seven binding property classes (1 = cation, 2 = anion, 3 = neutral H-bond donor, 4 = neutral H-bond acceptor, 5 = polar, 6 = hydrophobic, 7 = other). The classes of the atoms in the sample molecule are indicated. The *bp* and *bt* descriptors are constructed analogously to *ap* and *tt*.

Construction of Descriptor Databases from Connection Tables. We typically generate descriptors starting from connection tables in MACCS format.[6] An entry in MACCS typically needs to be preprocessed before descriptors are generated. If there is more than one fragment per entry (salts, mixtures, etc.), each fragment is treated separately. Fragments with less than seven non-hydrogen atoms are deleted; this removes most small counter ions while retaining larger molecules of cocrystallization. Fragments containing undefined R-groups or "superatoms" are also deleted.

Because of inconsistencies in drawing molecules (e.g., different ways of representing dative bonds) and artificial asymmetries (e.g., both oxygens in a carboxylate should be

sp^2 even though the bonds are drawn differently), we must take steps to fix the chemical representation of each connection table. Also, we would like to assign formal charges to the molecules as they would exist in their ionization state at physiological pH. (The formal charges already in MACCS may be unreliable for this purpose and are ignored.) To accomplish this end, we use the program PATTY[7] to locate particular substructures and assign intermediate types to atoms in the substructures. The library of substructures is available as supporting information, and the PATTY language is explained in ref 7. The intermediate type includes information about the binding property class, the hybridization, and the formal charge. For instance, both oxygen atoms in the substructure "*-C(=O)-O&X1" (carboxylate) are assigned the intermediate type "32", indicating an anion with one π electron and a formal charge of $-1/2$.

Subsequent programs use the intermediate type to further prepare the molecule to be parsed into descriptors. The physicochemical class is parsed directly from the intermediate type. For the calculation of partial charges, explicit hydrogens are added to heteroatoms based on element, hybridization, and formal charge. For instance, a tertiary amine with a formal charge of $+1$ would receive one hydrogen. The method of Gasteiger and Marsili[8] is then applied to the molecules to get the final partial charges.

The algorithm for calculating atomic log *P* is modified from Klopman and Wang.[9] Their original method is used for estimating total molecular log *P* as the sum of contributions from 39 molecular fragments, which may contain 1−4 non-hydrogen atoms. Our approach is to first assign the contribution of each atom by the value for the corresponding single-atom fragment. For instance a carbon would start with an atomic log *P* of 0.320. If an atom is in one of the larger fragments, the atom gets an additional contribution divided by the size of the fragment. For instance if the carbon was in the group N=C(-X)-X, one fourth of the group value −0.150 would be added to the carbon. In this implementation, the atomic log *P*'s sum to the molecular log *P*'s as calculated by Klopman and Wang.

After the preparation steps above, the descriptors are calculated and stored in a randomly accessible database called a *topobase*. For each entry we store: a molecular identifier, an estimated molecular log *P*, and the number of each type of the eight descriptors. For each descriptor type we list the unique descriptors present in the molecule and their counts. To save space, the unique descriptors are identified by two-byte integers. For the property types, the mapping to integers is straightforward. For instance, the *bp* descriptor "1-(12)-7" can be written as "010712". However, for *ap* and *tt* there are too many possible atom types, so each unique descriptor must be arbitrarily assigned a unique integer. For instance the *ap* "NX2-(3)-CX3." might be assigned "012345" if it were the 12 345th unique atom pair encountered during the construction of a set of topobases. We store and update an auxiliary file of the mapping of *ap*'s and *tt*'s to integers so that all topobases can use the same mapping.

In our implementation topobases are stored as mapped files on a VAX 8000 using the VMS operating system. Mapped files allow for random access of any given molecule by its identifier and allow direct mapping of the data into program core for rapid reading.

Definition of Similarity. Throughout we will use the index of similarity used by Carhart et al.[3] The similarity of molecules A and B is

$$\mathrm{Sim}_{AB} = \frac{\sum_k \min(f_{Ak}, f_{Bk})}{0.5[\sum_k f_{Ak} + \sum_k f_{Bk}]}$$

where f_{Ak} is the count of descriptor k in molecule A. The index k goes over the union of unique descriptors in A and B. Sim_{AB} ranges from 0.0 (nothing in common) to 1.0 (identity). It should be noted that because of the denominator molecular size counts as part of the similarity index.

Calculation of "Fuzziness". Given that our new descriptors are meant to be less specific, i.e., more "fuzzy", than the originals, it would be useful to be able to calculate relative fuzziness. We propose two approaches. First, if descriptors are specific, any two of them from the same molecule are likely to be different. If they are fuzzy, any two descriptors are likely to be the same. We can monitor this for a single molecule by the ratio

$$R = \frac{\text{total no. of descriptors}}{\text{no. of unique descriptors}}$$

R can be regarded as a descriptor-based measure of intramolecular symmetry. For the *ap* descriptor in the molecule in Figure 1, for instance, $R = 28/23 = 1.22$. The fuzziness value for a given descriptor can be taken as the median of *R* over a large sample of molecules.

A second approach is to monitor how similar any two molecules are likely to be. The more fuzzy the descriptor, the more two molecules will have in common and the more similar they will appear. The fuzziness of a given descriptor is taken as the median similarity for a large sample of pairs of molecules.

For both measures it is important to use the same sample of molecules when comparing descriptors so the distribution of intramolecular symmetry (which affects the first measure) and size (which affects the second) will be constant.

How Searches Are Run. We run similarity searches with our in-house system TOPOSIM. During a search of a topobase, TOPOSIM calculates for each database entry the similarity for each of the eight descriptors. Within TOPOSIM the user has the option to calculate a final score for each entry as a user-defined linear combination of the individual similarities. For instance the scores for single descriptors might be

ap score = 1.0* *ap* similarity + 0.0* *bp* similarity + 0.0 * *hp* similarity + ...

bp score = 0.0* *ap* similarity + 1.0* *bp* similarity + 0.0 * *hp* similarity + ...

etc.

For this study we define scores for virtual descriptors called *combination descriptors* as the mean of the similarities for two single descriptors. For instance, the combination descriptor *ap* + *bp* is generated by the combination

ap + *bp* score = 0.5 **ap* similarity + 0.5 **bp* similarity + 0.0 **hp* similarity + ...

Sorting of Scores. Once all the scores are calculated for a topobase, they are sorted from high to low score. If there is more than one fragment per molecule, only the highest scoring fragment is kept. Ranks are then assigned: the molecule with the highest score is rank 1, the next highest rank 2, etc. We use only the ranks of the compounds in this study, since the distribution of absolute scores varies from one descriptor to another.

Combination descriptors (see previous section) are one way to use two descriptors simultaneously. An alternative way, which we call *minimum rank sorting*, is a postprocess that merges the sorted lists of individual descriptors. For instance, the virtual minimum rank descriptor mr(*ap*,*bp*) is generated from the *ap* and *bp* lists. We define a new score for each compound in the database as its rank in the *ap* list or its rank in the *bp* list, whichever is smaller. The compounds are then sorted by the new score, and new ranks are assigned, the smallest score being rank 1, the next rank 2, etc. This is analogous to the common situation where the investigator does separate searches with different descriptors and submits the union of the top-scoring compounds from each.

Measures of Merit for Similarity Searches. We propose two measures to determine whether one set of descriptors is better than another based on how well similarity to a particular probe correlates with a similar activity. The measures depend on a simulated screening experiment. Imagine a database with N compounds that contains *Nactive* actives. For a given set of descriptors and a given probe, calculate the rank for all molecules as described above. Next, "assay" the compounds in order of ascending rank. We can graph the total number of actives found versus the total number of compounds tested to see how rapidly actives are found. There are two limiting cases. If similarity to the probe were a perfect predictor of activity, all actives would be at the front of the list, and the curve would start out with a slope of 1 and then break to a horizontal line once *Nactive* compounds were tested. If similarity were a very poor predictor, actives would be randomly distributed throught the list and accumulate according to their frequency in the database. The curve would have a nearly constant slope of *Nactive*/*N*. Actual curves, as we will see, somewhat resemble hyperbolic curve; they start with a steep initial rise, then level off. Our measures are as follows.

(1) How many compounds must be tested until half the actives are found. We call this number A50, and it is analogous with IC50's of binding assays or K_M of enzyme assays. The smaller A50, the better the similarity method in a global sense. This measure is more relevant to assays where a very large number of compounds can be tested. A50 can be alternatively expressed as a *global enhancement*, the ratio of the A50 expected for the random case (N/2) over the actual A50. For instance, for A50 = 3000 in a database of 30 000 compounds, the global enhancement would be 30 000/(2 × 3000) = 5.0.

(2) How many actives are found after testing an arbitrary small fraction of the total database. For instance the number of actives at 300 compounds tested could be called A@300. The larger A@300, the better. (This is analogous to the "initial slope" in an enzyme assay.) In some assays, where one never tests more than a small percent of a large database, this measure could be more useful than A50. A@300 can be expressed as an *initial enhancement*: how many more actives there are than expected by chance. For instance for A@300 = 150 and a total number of 450 actives in the 30 000 compounds database, the expected number of actives is (300/30 000) × 450 = 4.5, and the enhancement is 150/4.5 = 33.3.

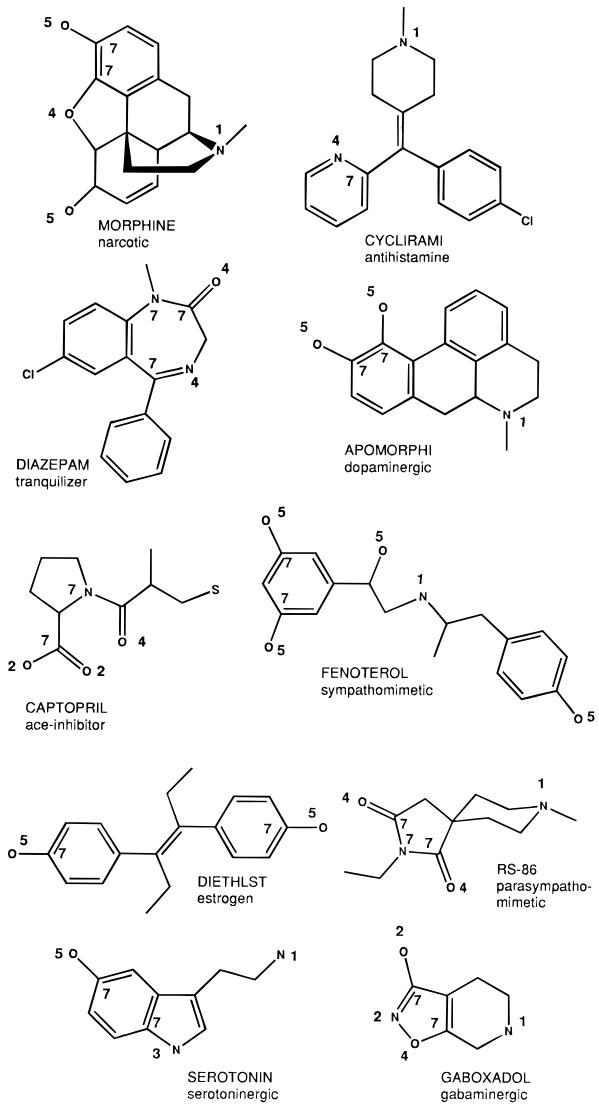


Figure 2. Chemical structures of probes used in this study. The binding property type for each atom is shown where it is not "6". The Derwent Standard Drug File names are used to label the molecules. The corresponding generic chemical names are: MORPHINE, morphine; CYCLIRAMI, cycliramine; DIAZEPAM, diazepam; APOMORPHI, apomorphine; CAPTOPRIL, captopril; DIETHYST, diethylstilbesterol; FENOTEROL, fenoterol; RS-86, RS-86; SEROTONIN, serotonin; and GABOXADOL, gaboxadol.

Database Used in this Study. In order to measure the merit of the descriptors we need to have a database of molecules for which we know the biological activities. For this purpose, we use the Derwent Standard Drug File (SDF),[10] which is a licensed database of druglike molecules compiled from the patent literature. There are ~43 000 connection tables in the MACCS-compatible Version 6.0. Most structures have one or more key words in the "therapeutic category" field. We will assume that a molecule is active in the dopaminergic therapeutic area, for instance, if it contains the key word "DOPAMINERGICS" in this field. Since not every compound has been tested in every area, one cannot assume the converse—that a compound without this key work is inactive. Thus for any given keyword there

Table 1. Activities from SDF Used in This Study

activity	keywords	comments	no. actives
narcotic	("NARCOTIC" and "ANALGESIC") or "OPIOID"	opiate agonists and antagonists	478
antihistamine	"ANTIHISTAMINES-H1"	histamine-H1 antagonists	408
tranquilizer	"TRANQUILIZER" or "BENZODIAZEPINE-AGONIST"	mostly benzodiazepine agonists	399
dopaminergic	"DOPAMINERGICS"	dopamine agonists	268
ace-inhibitor	"ANGIOTENSIN ANTAGONISTS"	mostly ACE inhibitors and analogs of angiotensin	218
estrogen	"ESTROGENS"	estrogen agonists	215
sympathomimetic	"SYMPATHOMIMETICS-BETA"	beta-adrenergic agonists	188
parasympatho-mimetic	"PARASYMPATHOMIMETICS"	muscarinic and nicotinic acetylcholine agonists	139
serotoninergic	"SEROTONINERGICS"	serotonin agonists	70
gabaminergic	"GABAMINERGICS"	GABA agonists	58

are probably some "false inactives". There are ~450 distinct key words in the SDF. Some key words, for example, "DOPAMINERGIC", involve only one mechanism or receptor. Others, for example "HYPOTENSIVES", involve many mechanisms. There has been no effort on the part of the compilers of this version SDF to maintain a list of synonymous key words or to subdivide a single therapeutic area by mechanism, so we have had to use our judgment in deciding what key words are relevant for a particular activity.

We generated the SDF topobase from the MACCS database. The topobase contains 37 005 fragments from 35 635 molecules.

Choice of Example Probes for Similarity Searches. Although it is possible to define a probe as a composite of two or more molecules, we confined ourselves to single molecule probes. Chemical structures of the probes used in this study (named by the SDF external registry number) with the corresponding activity are shown in Figure 2. Table 1 shows how the activities were constructed from key words in SDF. The probes were arbitrarily selected under two constraints: (1) The majority of the actives in the therapeutic area of the probe should work by the same mechanism as the probe. (Given the limitations of the database, there is no way to ensure that *all* actives that work by the same mechanism; thus some actives are "false actives" in this respect.) (2) There should be >50 actives to ensure reasonable statistics.

It happens that most of the active compounds in therapeutic areas that meet the constraints have cationic centers. We feel these probes are fairly representative of small druglike molecules.

RESULTS

Fuzziness of Descriptors. We calculated the median value for the descriptor ratio *R* over all the fragments in the SDF using each of the descriptors. We also calculated the median pairwise similarity for a 10 000 randomly selected pairs of compounds. The results are presented as two series, pairs and torsions:

	ap	*bp*	*hp*	*cp*	*tt*	*bt*	*ht*	*ct*
median *R*	2.17	4.68	5.43	6.20	1.68	3.75	4.26	4.56
median pairwise sim.	0.15	0.36	0.39	0.43	0.04	0.26	0.31	0.35

The order of fuzziness is the same for both series and for both measures: charge > hydrophobic > binding property ≫ original. Thus, we are able to demonstrate that the physiochemical property descriptors are indeed fuzzy compared to the original descriptors. Pairs are always more fuzzy than the corresponding torsions (e.g., *ap* > *tt*, *bp* > *bt*, etc.).

Measures of Merit for Similarity Searches. Figure 3 shows as an example the graph of the accumulation of actives versus rank for the CYCLIRAMI. Table 2 lists the measures

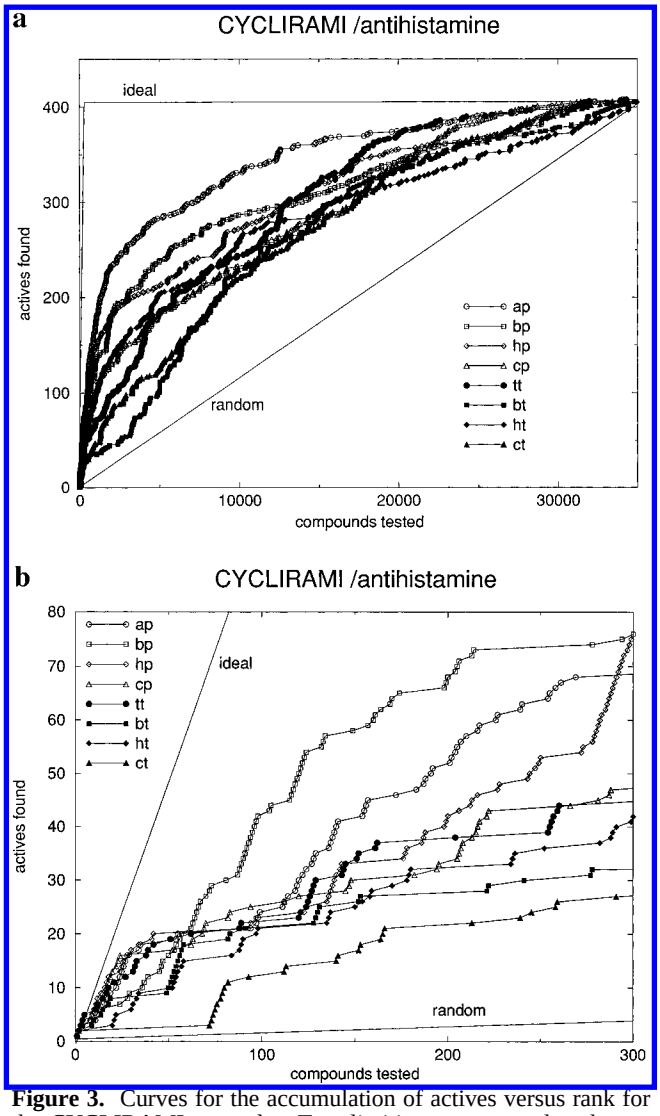


Figure 3. Curves for the accumulation of actives versus rank for the CYCLIRAMI example. Two limiting cases are also shown: "ideal" where all the actives would be at the front of the list, and "random" where all the actives would be evenly distributed throughout the list. a. the curve over the entire database and b. the curve for the first 300 molecules tested.

of merit for all probes. Since the list of actives for each probe inevitably contains false inactives and false actives, the global and initial enhancements reported Table 2 arelikely *underestimated* relative to an ideal list that does not contain such "noise". However, in this paper we are not interested in absolute enhancements but in comparisons of those enhancements among descriptors. The comparisons should be valid because, for any probe, every set of descriptors sees the same level of noise.

We can get an idea of the overall utility of the descriptors, at least for this set of probes and this database, by taking the mean global enhancement and initial enhancement over

Table 2. Measures of Merit for Various Descriptors

descriptor	A50	global enhancement	A@300	initial enhancement	descriptor	A50	global enhancement	A@300	initial enhancement
				MORPHINE/narcotic					
ap	2835	6.3	140	34.8	*ap+cp*	3502	5.1	138	34.3
bp	4750	3.8	115	28.6	*tt+bt*	953	18.7	183	45.5
hp	7490	2.4	101	25.1	*tt+ct*	6604	2.7	186	46.3
cp	5193	3.4	104	25.9	*ap+tt*	1641	10.9	177	44.0
tt	3108	5.7	190	47.3	mr(*ap*,*bp*)	2834	6.3	143	35.5
bt	1490	12.0	121	30.0	mr(*ap*,*cp*)	3320	5.4	136	33.8
ht	2053	8.7	116	28.8	mr(*tt*,*bt*)	1228	14.5	152	37.8
ct	8732	2.0	112	27.8	mr(*tt*,*ct*)	3915	4.6	151	37.6
ap+bp	2862	6.2	139	34.6	mr(*ap*,*tt*)	1144	15.5	175	43.5
				CYCLIRAMI/antihistamine					
ap	1438	12.4	68	19.8	*ap+cp*	1618	11.0	70	20.3
bp	3051	5.8	76	22.1	*tt+bt*	6104	2.9	48	14.0
hp	3699	4.8	75	21.8	*tt+ct*	5889	3.0	38	11.1
cp	7110	2.5	47	13.7	*ap+tt*	3008	5.9	70	20.3
tt	6109	2.9	44	12.8	mr(*ap*,*bp*)	1767	10.1	87	25.3
bt	8864	2.0	32	9.3	mr(*ap*,*cp*)	1845	9.7	71	20.6
ht	4974	3.6	42	12.2	mr(*tt*,*bt*)	7808	2.3	48	14.0
ct	8901	2.0	27	7.9	mr(*tt*,*ct*)	6663	2.7	51	14.8
ap+bp	1304	13.7	98	28.5	mr(*ap*,*tt*)	2157	8.3	54	15.7
				DIAZEPAM/tranquilizer					
ap	2935	6.1	102	30.4	*ap+cp*	4159	4.3	92	27.3
bp	4443	4.0	60	17.9	*tt+bt*	3314	5.4	114	33.9
hp	5172	3.4	44	13.0	*tt+ct*	4085	4.4	106	31.5
cp	7483	2.4	52	15.5	*ap+tt*	6772	2.6	112	33.3
tt	3308	5.4	113	33.6	mr(*ap*,*bp*)	3062	5.8	84	25.0
bt	5120	3.5	94	28.0	mr(*ap*,*cp*)	3983	4.5	79	23.5
ht	7658	2.3	63	18.8	mr(*tt*,*bt*)	3844	4.6	110	32.7
ct	9406	1.9	47	14.0	mr(*tt*,*ct*)	5341	3.3	93	27.7
ap+bp	3316	5.4	94	28.0	mr(*ap*,*tt*)	3159	5.6	113	33.6
				APOMORPHI/dopaminergic					
ap	3067	5.8	48	21.2	*ap+cp*	1448	12.3	57	25.2
bp	2982	6.0	56	24.8	*tt+bt*	1370	13.0	62	27.4
hp	6311	2.8	31	13.7	*tt+ct*	1069	16.7	54	23.9
cp	1800	9.9	51	22.6	*ap+tt*	2338	7.6	39	17.3
tt	3470	5.1	28	12.4	mr(*ap*,*bp*)	2924	6.1	57	25.3
bt	2791	6.4	52	23.0	mr(*ap*,*cp*)	2443	7.3	54	24.0
ht	6283	2.8	41	18.1	mr(*tt*,*bt*)	2630	6.8	56	24.9
ct	2074	8.6	45	19.9	mr(*tt*,*ct*)	2287	7.8	43	19.1
ap+bp	2130	8.4	60	26.5	mr(*ap*,*tt*)	2347	7.6	48	21.3
				CAPTOPRIL/ace-inhibitor					
ap	5798	3.1	42	22.9	*ap+cp*	15063	1.1	29	15.8
bp	18705	1.0	18	9.8	*tt+bt*	1416	12.6	82	44.7
hp	22322	0.8	20	10.9	*tt+ct*	2596	6.9	54	29.4
cp	22101	0.8	17	9.3	*ap+tt*	1446	12.3	75	40.8
tt	541	32.9	79	43.0	mr(*ap*,*bp*)	7703	2.3	33	18.0
bt	4321	4.1	40	21.8	mr(*ap*,*cp*)	8775	2.0	33	18.0
ht	2000	8.9	39	21.2	mr(*tt*,*bt*)	923	19.3	60	32.7
ct	10943	1.6	27	14.7	mr(*tt*,*ct*)	962	18.5	52	28.3
ap+bp	12366	1.4	30	16.3	mr(*ap*,*tt*)	898	19.8	64	34.9
				DIETHYLST/estrogen					
ap	10600	1.7	37	20.4	*ap+cp*	3131	5.7	46	25.4
bp	1483	12.0	67	37.0	*tt+bt*	4195	4.2	51	28.2
hp	1456	12.2	49	27.0	*tt+ct*	2910	6.1	56	30.9
cp	1013	17.6	84	46.4	*ap+tt*	12460	1.4	45	24.9
tt	15196	1.2	44	24.3	mr(*ap*,*bp*)	1319	13.5	73	40.3
bt	2725	6.5	33	18.2	mr(*ap*,*cp*)	1259	14.2	78	43.1
ht	2350	7.6	43	23.8	mr(*tt*,*bt*)	3855	4.6	49	27.1
ct	1064	16.7	53	29.3	mr(*tt*,*ct*)	1543	11.6	55	30.4
ap+bp	2504	7.1	38	21.0	mr(*ap*,*tt*)	13365	1.3	50	27.6
				FENOTEROL/sympathomimetic					
ap	489	36.4	83	52.4	*ap+cp*	1343	13.3	68	43.0
bp	2435	7.3	56	35.4	*tt+bt*	289	61.7	96	61.8
hp	6264	2.8	38	24.0	*tt+ct*	224	79.5	101	63.9
cp	4571	3.9	40	25.3	*ap+tt*	284	62.7	95	60.1
tt	354	50.3	91	57.5	mr(*ap*,*bp*)	552	32.3	73	46.2
bt	1438	12.4	52	32.9	mr(*ap*,*cp*)	595	29.9	68	43.0
ht	1396	12.8	46	29.1	mr(*tt*,*bt*)	371	48.0	85	53.8
ct	637	28.0	78	49.3	mr(*tt*,*ct*)	310	57.5	92	58.2
ap+bp	801	22.2	76	48.0	mr(*ap*,*tt*)	334	53.3	88	55.7
				RS-86/parasympathomimetic					
ap	6490	2.7	9	7.7	*ap+cp*	2567	6.9	29	24.8
bp	3654	4.9	23	19.7	*tt+bt*	12185	1.5	14	12.0
hp	10177	1.8	7	6.0	*tt+ct*	10537	1.7	8	6.8
cp	4411	4.0	24	20.5	*ap+tt*	7028	2.5	9	7.7
tt	14357	1.2	7	6.0	mr(*ap*,*bp*)	4622	3.9	14	12.0
bt	13306	1.3	7	6.0	mr(*ap*,*cp*)	3671	4.9	16	13.7
ht	15187	1.2	10	8.5	mr(*tt*,*bt*)	12264	1.5	10	8.6
ct	10417	1.7	10	8.5	mr(*tt*,*ct*)	11023	1.6	8	6.8
ap+bp	4204	4.2	31	26.4	mr(*ap*,*tt*)	6221	2.9	5	4.3

Table 2 (Continued)

descriptor	A50	global enhancement	A@300	initial enhancement	descriptor	A50	global enhancement	A@300	initial enhancement
				SEROTONIN/serotoninergic					
ap	2240	8.0	17	28.8	*ap*+*cp*	1660	10.7	18	30.5
bp	5304	3.4	13	22.0	*tt*+*bt*	2438	7.3	27	45.8
hp	9846	1.8	9	15.3	*tt*+*ct*	2056	8.7	25	42.4
cp	2494	7.1	18	30.5	*ap*+*tt*	2073	8.6	21	35.6
tt	2254	7.9	23	39.0	mr(*ap*,*bp*)	3380	5.3	16	27.2
bt	2883	6.2	25	42.4	mr(*ap*,*cp*)	2932	6.1	18	30.6
ht	9807	1.8	14	23.8	mr(*tt*,*bt*)	3807	4.7	25	42.5
ct	2914	6.1	20	33.9	mr(*tt*,*ct*)	2875	6.2	22	37.3
ap+*bp*	3066	5.8	15	25.5	mr(*ap*,*tt*)	2586	6.9	21	35.6
				GABOXADOL/gabaminergic					
ap	6991	2.5	13	26.6	*ap*+*cp*	2504	7.1	16	32.8
bp	3387	5.3	15	30.7	*tt*+*bt*	5336	3.3	6	12.3
hp	4024	4.4	11	22.5	*tt*+*ct*	5045	3.5	10	20.5
cp	5254	3.4	13	26.6	*ap*+*tt*	8558	2.1	7	14.3
tt	16130	1.1	3	6.1	mr(*ap*,*bp*)	2078	8.6	15	30.8
bt	3574	5.0	5	10.2	mr(*ap*,*cp*)	2874	6.2	15	30.8
ht	6714	2.7	3	6.1	mr(*tt*,*bt*)	6624	2.7	3	6.2
ct	3749	4.8	11	22.5	mr(*tt*,*ct*)	7021	2.5	10	20.5
ap+*bp*	1715	10.4	19	38.9	mr(*ap*,*tt*)	10814	1.6	8	16.4

all the probes:

	ap	*bp*	*hp*	*cp*	*tt*	*bt*	*ht*	*ct*
mean glob. enhancement	8.5	5.4	3.7	5.5	11.4	5.9	5.2	7.3
mean init. enhancement	26.6	25.0	18.0	23.7	28.3	22.3	19.1	22.8

All descriptors give enhancements $\gg 1$, indicating general usefulness. The original descriptors appear to be better on the average than the property descriptors. Hydrophobic descriptors are consistently the worst of the property descriptors.

These average numbers are somewhat misleading, however. An inspection of Table 2 shows that the effectiveness of a given individual descriptor varies greatly with the probe and the measure of merit. For MORPHINE, *ap* is the best pair descriptor, and *bt* is the best torsion descriptor for global enhancement; *ap* and *tt* are the best for initial enhancement. For DIAZEPAM and FENOTEROL *ap* and *tt* give the best results with both measures. For CYCLIRAMI only *ap* does well in global enhancement, and *ap*, *bp*, and *hp* do well in initial enhancement. For APOMORPHI property descriptors *cp* and *ct* do best in global enhancement, and *bp* and *bt* do best in initial enhancement. For CAPTOPRIL, *tt* shows much better performance than any other descriptor. For DIETHYLST, *cp* and *ct* show the best initial and global enhancement. For RS-86, only *bp* and *cp* show reasonable global and initial enhancements. For SEROTONIN, *ap* and *tt* show the best global enhancements, while *ap*, *cp*, *tt*, and *bt* show the best initial enhancements.

The explanation seems to lie in the structural requirements for activity. By inspecting the appropriate sets of actives it is easy to explain in retrospect why one descriptor would do better than another in some of the extreme cases. Sometimes specific descriptors are best. Most of the ace-inhibitors, including CAPTOPRIL, contain proline. The proline residue is specified best by *tt*, and so no other descriptor finds actives as effectively as *tt*. Most sympathomimetic actives have a secondary amine cation and a catechol, as does FENOTEROL; the *ap* and *tt* can specify the valence and aromaticity for these groups, and the property descriptors cannot. In other cases, fuzzy descriptors are best. Many of the estrogen actives are steroids. Steroids do not resemble DIETHYLST if one looks at the valence and hybridization of the atoms, as do the *ap* and *tt*, but do resemble that probe if one looks only at the polarity of the atoms as do all the property descriptors. Many of the parasympathomimetic actives have quaternary amines as cations and aromatic nitrogens or ether oxygens as hydrogen bond acceptors. These are recognized as similar to the tertiary amine and carbonyl oxygens in RS-86 by the property atom pairs but not by *ap*.

Combination Descriptors and Minimum Rank Descriptors. The idea behind the combination descriptors and minimum rank descriptors is that, since one cannot predict a priori how well one descriptor will do for a particular probe, using two or more descriptors simultaneously might increase the chance that reasonably good results could be obtained. Since there are 28 possible two-descriptor combinations, we could not look at all of them, so we selected five that we thought would be representative. Four of the five combine an original descriptor (*ap* or *tt*) with a corresponding property descriptor (*bp* or *cp*), so that a middle ground might be reached between specificity and fuzziness. The fifth descriptor combines the original *ap* and *tt*. The measures of merit for the selected combination descriptors and the equivalent minimum rank descriptors are also listed in Table 2.

Naively, one might expect the enhancement of a combination descriptor to be roughly halfway between the enhancements of the component descriptors, but about half the time a combination descriptor will do better than both (e.g., for CYCLIRAMI the initial enhancement for *ap* + *bp* is 28.5 versus 19.8 and 22.1 for *ap* and *bp*, respectively), or at least not much worse than the better component descriptor (e.g., for MORPHINE initial enhancement for *ap* + *bp* is 35.6 versus 35.9 and 29.5 for *ap* and *bp*). There are some examples where the combination descriptor does give an enhancement halfway between those of the individual descriptors. These are mostly from those probes where one of the component descriptors is very poor relative to the other (e.g., for CAPTOPRIL the initial enhancement for *tt* + *bt* is 12.6 versus 32.9 and 4.1 for *tt* and *bt*). A similar situation exists for the minimum rank descriptors.

Table 3 lists the mean values over all probes for the combination and minimum rank descriptors compared to the component descriptors. Both the combination descriptors and minimum rank descriptors on the average do as well or

Table 3. Measures of Merit for Pairs of Descriptors Averaged Over All Probes

		pair of descriptors				
		ap,bp	*ap,cp*	*tt,bt*	*tt,ct*	*ap,tt*
mean global enhancement	combination	8.5[a]	7.8	13.1	13.3	11.7
	min rank	9.4[b]	9.0	10.9	11.6	12.3
	component	8.5,5.4[c]	8.5,5.5	11.4,5.9	11.4,7.3	8.5,11.4
mean initial enhancement	combination	29.4	27.9	32.6	30.7	29.8
	min rank	28.6	28.1	28.0	28.1	28.9
	component	26.5,25.0	26.5,23.6	28.2,22.2	28.2,22.8	26.5,28.2

[a] For the combination descriptor $ap + bp$. [b] For the minimum rank descriptor mr(*ap*,*bp*). [c] For the individual descriptors *ap* and *bp*.

Table 4. Correlation between Ranks of Active Compounds for Pairs of Descriptors

	ap	*bp*	*hp*	*cp*	*tt*	*bt*	*ht*	*ct*		*ap*	*bp*	*hp*	*cp*	*tt*	*bt*	*ht*	*ct*
	MORPHINE/narcotic									DIETHYLST/estrogen							
ap	1.00	0.94	0.92	0.93	0.67	0.88	0.46	0.78	*ap*	1.00	0.33	0.37	0.30	0.85	0.47	0.40	0.24
bp		1.00	0.96	0.97	0.56	0.82	0.35	0.71	*bp*		1.00	0.85	0.88	−0.01	0.52	0.27	0.39
hp			1.00	0.96	0.57	0.74	0.40	0.68	*hp*			1.00	0.82	0.05	0.51	0.52	0.49
cp				1.00	0.55	0.79	0.34	0.72	*cp*				1.00	−0.02	0.43	0.24	0.44
tt					1.00	0.68	0.61	0.80	*tt*					1.00	0.35	0.31	0.10
bt						1.00	0.60	0.82	*bt*						1.00	0.70	0.76
ht							1.00	0.59	*ht*							1.00	0.65
ct								1.00	*ct*								1.00
	CYCLIRAMI/antihistamine									FENOTEROL/sympathomimetic							
ap	1.00	0.76	0.72	0.66	0.78	0.57	0.48	0.55	*ap*	1.00	0.65	0.59	0.48	0.26	0.50	0.59	0.52
bp		1.00	0.89	0.75	0.60	0.80	0.61	0.62	*bp*		1.00	0.82	0.81	0.22	0.44	0.63	0.68
hp			1.00	0.63	0.52	0.7	0.83	0.59	*hp*			1.00	0.84	0.39	0.21	0.80	0.68
cp				1.00	0.45	0.44	0.26	0.69	*cp*				1.00	0.32	0.16	0.58	0.78
tt					1.00	0.61	0.40	0.41	*tt*					1.00	0.21	0.39	0.42
bt						1.00	0.68	0.55	*bt*						1.00	0.39	0.33
ht							1.00	0.49	*ht*							1.00	0.66
ct								1.00	*ct*								1.00
	DIAZEPAM/tranquilizer									RS-86/parasympathomimetic							
ap	1.00	0.66	0.64	0.60	0.90	0.59	0.52	0.67	*ap*	1.00	0.35	0.39	0.38	0.54	0.61	0.48	0.52
bp		1.00	0.62	0.83	0.68	0.84	0.39	0.69	*bp*		1.00	0.63	0.84	0.19	0.45	0.32	0.38
hp			1.00	0.58	0.58	0.44	0.83	0.51	*hp*			1.00	0.62	0.30	0.52	0.71	0.34
cp				1.00	0.59	0.63	0.36	0.71	*cp*				1.00	0.23	0.37	0.38	0.52
tt					1.00	0.69	0.51	0.72	*tt*					1.00	0.54	0.48	0.44
bt						1.00	0.37	0.72	*bt*						1.00	0.65	0.66
ht							1.00	0.47	*ht*							1.00	0.57
ct								1.00	*ct*								1.00
	APOMORPHI/dopaminergic									SEROTONIN/serotoninergic							
ap	1.00	0.85	0.80	0.82	0.51	0.43	0.66	0.60	*ap*	1.00	0.85	0.74	0.72	0.74	0.75	0.65	0.72
bp		1.00	0.91	0.90	0.27	0.50	0.79	0.56	*bp*		1.00	0.81	0.76	0.49	0.69	0.55	0.61
hp			1.00	0.83	0.16	0.40	0.86	0.50	*hp*			1.00	0.67	0.47	0.65	0.78	0.58
cp				1.00	0.21	0.38	0.68	0.70	*cp*				1.00	0.50	0.64	0.52	0.73
tt					1.00	0.40	0.15	0.46	*tt*					1.00	0.75	0.67	0.75
bt						1.00	0.64	0.54	*bt*						1.00	0.65	0.85
ht							1.00	0.47	*ht*							1.00	0.66
ct								1.00	*ct*								1.00
	CAPTOPRIL/ace-inhibitor									GABOXADOL/gabaminergic							
ap	1.00	0.84	0.81	0.81	0.24	0.74	0.83	0.61	*ap*	1.00	0.78	0.77	0.71	0.34	0.67	0.14	0.22
bp		1.00	0.98	0.97	−0.04	0.70	0.72	0.52	*bp*		1.00	0.96	0.96	0.36	0.80	0.22	0.29
hp			1.00	0.98	−0.01	0.60	0.69	0.51	*hp*			1.00	0.92	0.39	0.77	0.29	0.15
cp				1.00	−0.03	0.62	0.66	0.56	*cp*				1.00	0.29	0.70	0.19	0.43
tt					1.00	0.18	0.25	0.18	*tt*					1.00	0.40	0.15	−0.21
bt						1.00	0.74	0.43	*bt*						1.00	0.39	0.22
ht							1.00	0.57	*ht*							1.00	0.12
ct								1.00	*ct*								1.00

slightly better than the better component descriptor, especially for the initial enhancement. The combination descriptors and the minimum rank descriptors seem roughly equivalent.

Correlation of Ranks between Descriptors. Besides looking at how well different descriptors do at selecting active compounds, one can also look at how they rank the actives. Table 4 summarizes the pairwise correlation of ranks for actives. If different descriptors were merely expressing the same chemical features in a different "notation", one would expect to see correlations near 1. Some correlations in the table are reasonably high, but many are not. Whether any two descriptors will give high correlations varies from probe to probe with no apparent pattern. Also, for any given probe we have not found any relationship of the correlation between two descriptors and other types of comparison between the descriptors.

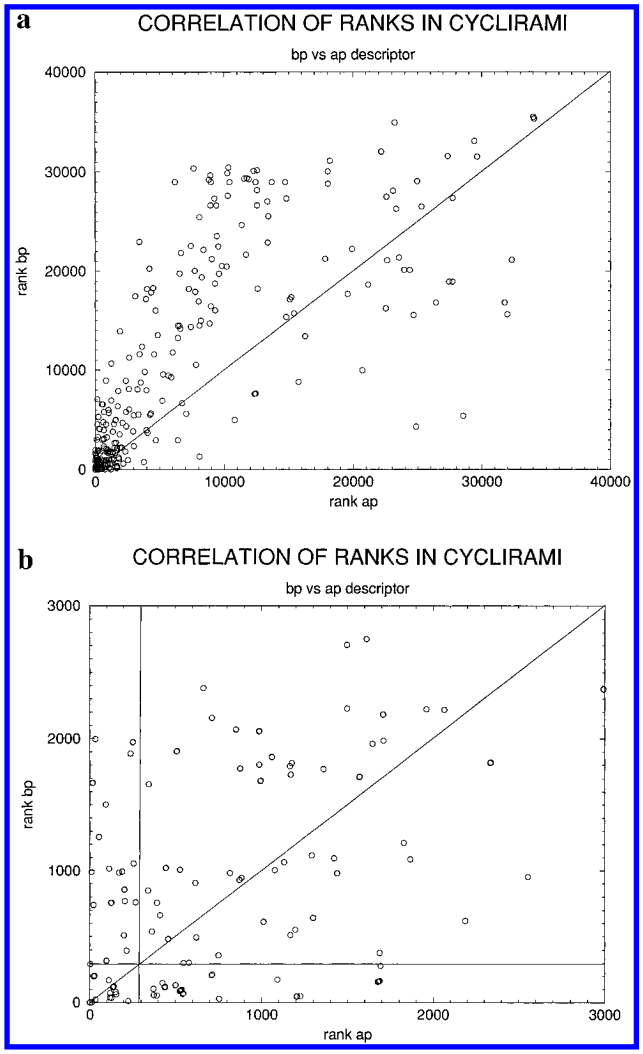


Figure 4. The correlation of rank for the *ap* and *bp* descriptors for CYCLIRAMI: a. The scatterplot over the entire database. Each circle represents a compound in SDF with antihistamine activity. b. Closeup of the origin of a. The rank cutoff at 300 is indicated.

Even for cases where a correlation coefficient in Table 4 appears high, when one looks at the correlation in a scatterplot, it is startling how much scatter there is. Figure 4 shows this for the *ap* and *bp* descriptors for CYCLIRAMI. In this typical example there is a reasonably high overall correlation of the ranks ($r = 0.76$), but the antihistamine actives do not fall near the diagonal line, as would be expected if the descriptors ranked actives the same. The fact that most of the actives fall above the diagonal in Figure 4a is consistent with the fact that the global enhancement for *ap* is better than *bp* for CYCLIRAMI. There are very many active compounds near the axes, as is more easily seen in Figure 4b. These are actives that would be considered highly similar to the probe by one descriptor but not very similar by another. This means that even if the *ap* is better than *bp* in selecting actives, it is not safe to ignore *bp*; there are still new actives to be found.

For any given compound, the difference in ranks between any two descriptors, for instance the *ap* and *bp*, can be quantitated as log(rank *ap*/ rank *bp*). Figure 5 shows the structures of some of the actives that have the most positive or negative values. Since the *ap* descriptor is very specific for valence and hybridization, the compounds that have low ranks on *ap* tend to be close analogs of CYCLIRAMI. The compounds with low ranks on *bp*, a descriptor which ignores these details in favor of physiochemical equivalence, have more variations (e.g., substitution of ammonium for tertiary amine in DIPHENHMI).

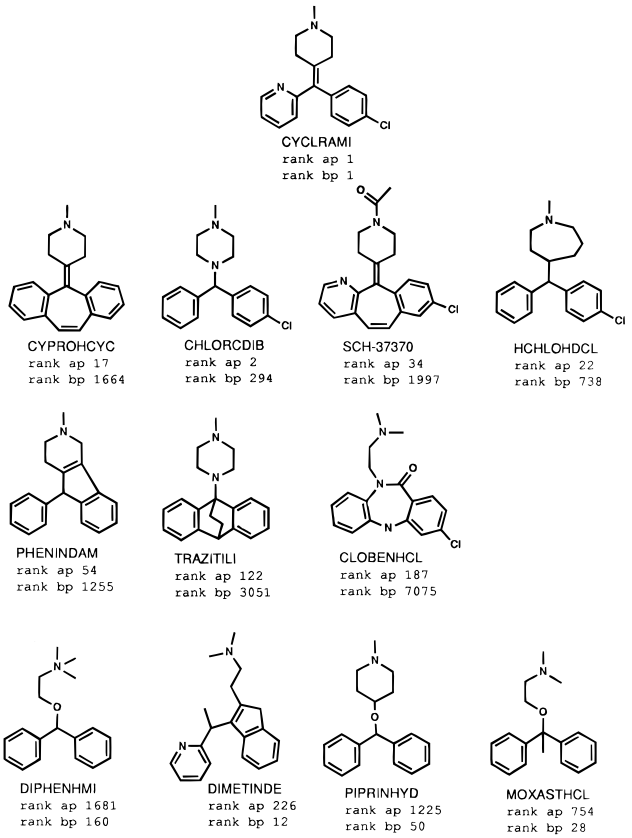


Figure 5. Selected antihistamines that have very different ranks by the *ap* and *bp* descriptors using CYCLIRAMI as the probe.

The observation that the ranks of actives are very sensitive to the atom type definition applies to almost any two pairs of descriptors for almost all the probes.

DISCUSSION

Similarity searches have been performed with a large variety of descriptor types.[1,2] In this paper we concentrate on variations of the atom pair and topological torsion. The new fuzzy descriptors on the average have lesser enhancements than the original descriptors, but we have retained all eight descriptors in our current version of TOPOSIM for reasons discussed below. It is not unusual for a Merck scientist to run similarity probes using all eight descriptors and combinations thereof and then to take a union of the highest scoring compounds from each search.

We feel it is important to retain and use a variety of descriptors as long as it can be shown that a descriptor has an average enhancement much better than 1. The first reason is that it is very hard to predict a priori whether a particular descriptor will do well in selecting active compounds for a particular probe. In retrospect, perhaps we should have expected this. Various receptors have varying levels of permissiveness and chemical groups on drug molecules that appear equivalent to one receptor may appear very different to another, and so the degree of fuzziness in the ideal descriptor varies from receptor to receptor. Of course, the results will also depend on what compounds are in the database being searched.

The second reason, and perhaps the most important reason in respect to the pharmaceutical industry, has to do with how

differently any two descriptors rank the same set of active compounds. In very many cases we found that active molecules that would be in the front of one list would be far down on another. It is clear that different descriptors are not just expressing the same chemical features in a different way; they actually capture very different features. This is probably why combination descriptors prove so useful. In practical similarity searches, where one takes a relatively small number of top-scoring compounds, the result is that different descriptors will seem to select different subsets of actives. This is very desirable because at the beginning of a drug discovery project one is using similarity probes to generate as diverse set of actives as possible, and one is willing to use a descriptor with a lesser enhancement in order to obtain this diversity.

A third reason for retaining a variety of descriptors is that the more fuzzy ones can become more useful as a project progresses and one gets a more inclusive idea of the physiochemical features in active molecules. For instance, one may find that active compounds should contain anions and not specifically carboxylates or tetrazoles or hydrophobes and not specifically sp^3 carbon. With the proper fuzzy descriptor, one can run a properly general similarity search.

ACKNOWLEDGMENT

We would like to thank the following people for motivating the work: Arthur Patchett, Martha Quesada, Debbie Allison, Pat Combs, Richard Norton, Mary Jo Zaborowski, Dennis Underwood, and in general the molecular modelers in the department of Molecular Systems. Dr. Norman Bauman, at Lederle Laboratories, first suggested the "global enhancement" as a measure of merit.

Supporting Information Available: PATTY script for intermediate types (3 pages). This material is contained in many libraries on microfiche, immediately follows this article in the microfilm version of the journal, can be ordered from the ACS, and can be downloaded from the Internet; see any current masthead page for ordering information and Internet access instructions.

REFERENCES AND NOTES

(1) Willet, P. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press, Ltd., John Wiley & Sons: New York, 1987.

(2) *Concepts and Applications of Molecular Similarity*, Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

(3) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. "Atom pairs as molecular features in structure-activity studies: definition and application". *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 64−73.

(4) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. "Topological torsions: a new molecular descriptor for SAR applications. comparison with other descriptors". *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 82−85.

(5) Sheridan, R. P.; Venkataraghavan, R. "New methods in computer-aided drug design". *Acc. Chem. Res*. **1987**, *20*, 322−329.

(6) Molecular Access System; a product of Molecular Design Ltd.: San Leandro, CA.

(7) Bush, B. L.; Sheridan, R. P. "PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases." *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 756−762.

(8) Gasteiger, J.; Marsili, M. "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges". *Tetrahedron* **1980**, *36*, 3219−3228.

(9) Klopman, G.; Wang, S. "A computer automated structure evaluation (CASE) approach to calculation of partition coefficient". *J. Comput. Chem*. **1991**, *12*, 1025−1032.

(10) Standard Drug File Version 6.0; developed and distributed by Derwent Publications Ltd.: London, England, 1991.

CI950274J