Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors

Hans Matter*

TRIPOS GmbH, Martin-Kollar-Str. 15, 81829 München, Germany

Received May 14, 1996[⊗]

The efficiency of the drug discovery process can be significantly improved using design techniques to maximize the diversity of structure databases or combinatorial libraries. Here, several physicochemical descriptors were investigated to quantify molecular diversity. Based on the 2D or 3D topological similarity of molecules, the relationship between physicochemical metrics and biological activity was studied to find valid descriptors. Several compounds were selected using those descriptors from a database containing diverse templates and 55 biological classes. It was evaluated whether the obtained subsets represent all biological properties and structural variations of the original database. In addition, hierarchical cluster analyses were used to group molecules from the parent database, which should have similar biological properties. Using various sets of structurally similar molecules, it was possible to derive quantitative measures for compound similarities in relation to biological properties. A similarity radius for 2D fingerprints and molecular steric fields was estimated; compounds within this radius of another molecule were shown to have comparable biological properties. This study demonstrates that 2D fingerprints alone or in combination with other metrics as the primary descriptor allow to handle global diversity. In addition, standard atom-pair descriptors or molecular steric fields can be used to correlate structural diversity with biological activity. Hence, the latter two descriptors can be classified as secondary descriptors useful for analog library design, while 2D fingerprints are applicable to design a general library for lead discovery. Based on these findings, an optimally diverse subset containing only 38% of the entire IC93 database was generated using 2D fingerprints. Here no structure is more similar than 0.85 to any other (Tanimoto coefficient), but all biological classes were selected. This reduction of redundancy led to a child database with the same physicochemical diversity space, which contains the same information as the original database.

Introduction

Today pharmaceutical companies are applying combinatorial chemistry and high-throughput screening as new methods in the lead-finding process. Ideally, new leads will be discovered using mass screening, which are subsequently refined using the arsenal of existing medicinal chemistry techniques. A screening project requires chemical compound databases as the starting point assembled from natural sources, synthetic products, or combinatorial libraries.[1] A useful mass-screening program depends not only on the size of a library but also on its diversity.[2] Hence it is important to select compounds which do not contain redundant information in order not to waste time and resources.[3]

To select an ensemble of nonredundant compounds for biological screening,[4] similarity considerations are widely used.[5] Structurally similar molecules should have similar physicochemical and biological properties, following the *similar property principle*.[6] It should be possible to predict unknown properties for a molecule based on known values for similar molecules. The important question to ask is, which physicochemical measure of similarity correlates with biological properties? Such a descriptor would allow to select representative compounds which cover the entire property space of the parent database.

In the present study the following 2D and 3D molecular descriptors are evaluated, which have been suggested as similarity metrics for rational compound selections: UNITY 2D, 3D, and flexible fingerprints, 2D atom-pair fingerprints, 2D topological indices, 3D molecular steric fields, and molecular weight (as reference). Additionally 2D and 3D spatial autocorrelation functions and WHIM indices are evaluated. The latter descriptors contain information about the whole molecular structure in terms of size, shape, symmetry, and atom distribution. While the use of molecular steric fields requires a superposition rule, which is not a trivial problem for a diverse database, a superposition is not required for WHIM indices or 3D autocorrelation functions. On the basis of preliminary results, modified flexible 3D fingerprints and 2D atom-pair fingerprints were also compared to standard methods. For those metrics 2D or 3D distances are computed between pairs of pharmacophoric groups. The descriptor performances for compound selections were examined using two validation tests:

(a) The ability of these descriptors to predict the biological activity of similar molecules was studied using published series of molecules. This addresses the question of how similar two molecules must be to select only one of them for a representative subset. Because this study was done on datasets being active only in a

* Address for correspondence: Hoechst AG, Computational Chemistry, Central Pharma Research, Building G 838, D-65926 Frankfurt am Main, Germany. Tel: ++49-69-305-84329. Fax: ++49-69-331399.

⊗ Abstract published in *Advance ACS Abstracts,* March 1, 1997.

S0022-2623(96)00352-4 CCC: $14.00

single assay, this result reflects the performance of those descriptors for local diversity. Based on these findings, the estimation of a similarity radius for the most useful descriptors was possible. Within the similarity radius of a bioactive compound there should be a high probability that any other compound belongs to the same activity class. Hence a descriptor can be characterized using this similarity radius allowing to select minimally redundant compounds.

(b) To get insight into the global diversity performance, the sampling properties of these descriptors on a structurally and biologically diverse public database were compared. Several selected subsets were examined to check whether they cover all biological classes present in the original database. Several clustering results were investigated to check whether biologically similar compounds were grouped into similar clusters.

Finally these results are used to design an optimally diverse compound subset, which is defined as a dataset without redundant structures but which still covers the biological property space of the entire database. Using a valid molecular descriptor it should be possible to reduce the number of compounds in a database without missing any significant biological information.

Methods

General Methods. All calculations were done using the program SYBYL, versions 6.2 and 6.22.[7] Database manipulations were done using the UNITY 2.5 database management tools[8] in connection with the module SELECTOR to analyze databases. Automation of procedures was done using the SYBYL Programming Language (SPL) or UNIX shell scripts.

Computation of Descriptors. 1. 2D Fingerprints. 2D Fingerprints contain information about the presence of molecular fragments in a binary format. Because of the large number of fragments in a database, it is not possible to assign one bit to only a single fragment. In contrast, each fragment is projected using a pseudorandomization algorithm into a bitstring of limited size. Additionally, the presence of specific functional groups, rings or atoms is encoded in 60 of the total 988 bits.[9] The similarity of fingerprints is computed using the Tanimoto coefficient,[10] which is defined by the number of bit sets in both individual bitstrings normalized by the number of bit sets in common. This translates into a number between 0.0 and 1.0 for no to perfect similarity.

2. Atom-Pair Fingerprints. Atom-pair fingerprints are 2D topological descriptors, which count the distance between two atoms as the shortest path of bonds.[11] This fingerprint is constructed similar to flexible 3D fingerprints: each bit corresponds to the presence of a specific pair of atomic classes separated by a predefined number of bonds.

3. Topological 2D Descriptors. The following four topological 2D indices denoted as HDisq indices were used. 1. Electrotopological state values are generated for every one- and two-atom fragment in a molecule by summing up fragment-specific values.[12] The descriptors are in essence structure-specific weighted one- and two-atom fragment counts. 2. Molecular connectivity indices encode the connectivity pattern within a molecule.[13] These indices quantify features like type and number of atoms and bonds and the extent and position of branching and rings. 3. Molecular shape indices are used to quantify 2D molecular shape based on a graphical representation of a molecule.[14] 4. Topological symmetry indices are used to quantify molecular symmetry in terms of counts of topologically equivalent atoms within a molecule.[15]

During the analyses, the large number of 340 highly correlated descriptors was reduced using a principal component analysis (PCA[16−18]). The first eight eigenvalues as eigenvectors of the covariance matrix were used.

4. Autocorrelation Functions. 2D and 3D spatial autocorrelation functions containing electrostatic or lipophilic properties were calculated.[19,20] They represent the 2D topology or 3D spatial distances of a molecular structure and can be computed using the following equation:

$$A(d) = \sum p_i p_j$$

with $A(d)$ being the autocorrelation coefficient as the sum over all atom pairs i,j, which are separated by d bonds (for the 2D topological function) or the Cartesian distance d (for the 3D function). p_i refers to an atomic property, i.e., partial charge on atom i based on the Gasteiger and Marsili method[21] or Crippen's partial atomic lipophilicities.[22] Thus, one obtains a series of coefficients for different topological or Cartesian distances d (*autocorrelation vector*). For each vector, 12 coefficients corresponding to a range of d from 1 to 12 were computed.

5. Flexible 3D Fingerprints. Flexible 3D fingerprints contain information about spatial relationships, in particular the ability of a structure to adopt a conformation with a specific Cartesian distance between two atoms or functional groups. Distances between predefined features are stored in a bitstring.

6. Molecular Shape Descriptors. Molecular steric fields based on the comparative molecular field analysis (CoMFA) technique[23,24] were used as 3D molecular shape descriptors. This method requires a superposition rule for database structures. For individual quantitative structure−activity relationship (QSAR) datasets in the local diversity study, the published conformations and alignments were used. Subsequently the interaction energies between a probe atom and every structure are computed at surrounding points of a predefined grid, typically using a volume-dependent lattice with 2 Å grid spacing and a carbon atom as a probe. For descriptor validation the pairwise correlation coefficient between steric fields is calculated. The lack of an obvious superposition rule for the database used for the global diversity studies is one of the major drawbacks for this descriptor. Hence, after generation of an extended 3D structure, the principal axes for the given molecule were calculated and all molecules were reoriented using these principal axes. Further studies to solve the alignment problem for structurally diverse databases in combination with steric fields are in progress.[25]

7. WHIM Indices. To overcome this inherent alignment problem, WHIM indices (weighted holistic invariant molecular indices)[26] were investigated, which are invariant to rotations and translations. WHIM indices contain information about the 3D structure in terms of size, shape, symmetry, and atom distribution derived from Cartesian coordinates. After applying a weighting scheme to a particular atom (none, van der Waals, or volume scaling, denoted as WHIM_1, WHIM_2, or WHIM_3), a weight-centered coordinate matrix is created for each molecule. Then a principal component analysis is carried out for this matrix, and a new transformed set of atomic coordinates is obtained by projecting the old coordinate system axes onto the three principal axes. Finally, 12 WHIM descriptors for each weighting scheme are computed, based on (a) three eigenvalues representing the variance of each molecule matrix related to the *molecular size*, (b) three eigenvalue proportions representing a measure for the molecular shape (i.e., extended versus spherical), (c) three symmetry descriptors calculated from the third-order moments of the PCA scores, and (d) three kurtosis descriptors calculated from the fourth-order moments of the scores, related to atomic distribution and density around the origin and the principal axis.

Compound Selections. Two approaches for compound selections were used. The faster approach is called a maximum dissimilarity method:[27] every new selected member is maximally dissimilar from the previously selected set. The selection stops when a maximum number of compounds has been selected or when no new molecule can be selected without being too similar to already existing members.

Cluster analyses as a second method offer more specific control by assigning every structure to a group. Here, hierarchical clustering[28,29] was applied, which does not require any assumption about the number of clusters to be generated. Four different methods are used to compute distances between clusters: (a) single, the distance between the closest pair of data points in both clusters, (b) complete, the distance between

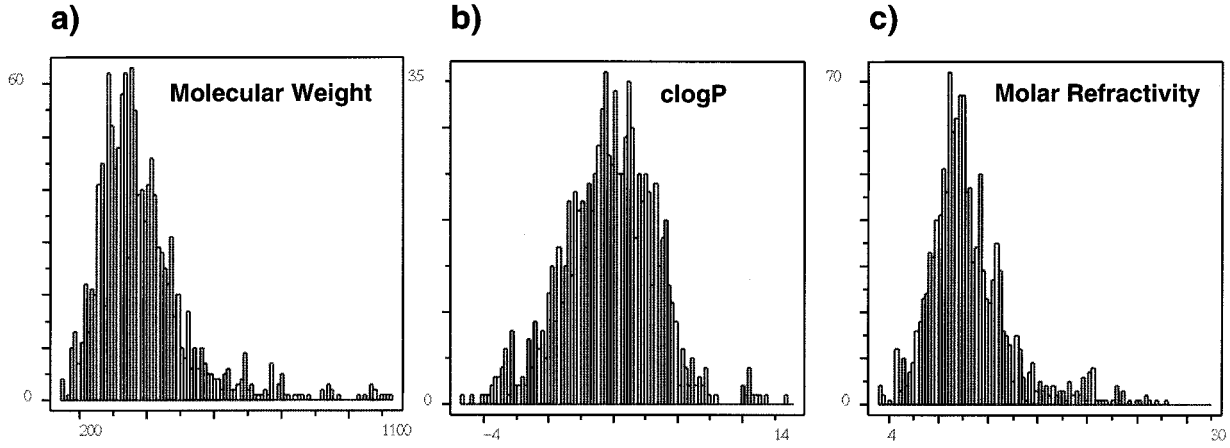

Figure 1. Physicochemical characterization of the database used for global diversity investigations. A total of 1283 bioactive compounds in 55 biological classes from the IndexChemicus 93 database are investigated. Histograms were generated for (a) molecular weight, (b) clog *P* as logarithm of the estimated octanol−water partition coefficient, and (c) molar refractivity.

the most distant pair of data points in both clusters, (c) average, the average of all pairwise data points between two clusters, and (d) median, the distance between two cluster centroids. The latter option was used based on initial studies, but the difference to all other options is small. Subsequently the structure closest to the center of a cluster is selected as the representative structure.

Results and Discussion

Global Diversity−Clustering of the IC93 Database. All 1283 biologically active compounds from the IndexChemicus 93 database (IC93)[30] were used for this study; 77 biological classes were defined according to the biological activity strings extracted from this database (cf. Supporting Information). Some compounds are active in more than one biological class, while some classes were only populated by a few members. It should be noted that this database contains for some classes heterogeneous bioactive molecules, which might act on more than one receptor, which is a potential source of uncertainty. It is likely that a subdivision into more classes corresponding to biological receptors could improve the classification results. However, this database was the only public data collection available for this purpose, when this investigation was started.

Initial selections were analyzed using this grouping, while for later investigations, classes with very similar biological activities were grouped together, leading to 55 classes. Again, details of this grouping are given in the Supporting Information, where all 77 classes are listed. Since the results for both biological classifications in preliminary studies were similar, it was decided to use the classification leading to a lower total number of groups and less groups with only a few members.

Initial 3D structures were generated using CONCORD.[31] The mean Tanimoto coefficient for this database is 0.91 (SD 0.11). This value is computed as average using the 2D fingerprint of each compound and the Tanimoto coefficient to its nearest neighbor. This reveals a high degree of redundant structures in the database. The distribution of other physicochemical descriptors (molecular weight, computed octanol water−partition coefficient (clog *P*), and molar refractivity[32]) is shown in Figure 1.

For each descriptor 120 clusters were generated using an appropriate cut in the corresponding dendrogram, and the central compound from each cluster was selected. The clustering results were analyzed in two different ways: (a) the percentage of biological classes covered by each subset of 120 compounds was computed (coverage analysis) and (b) the separation of active from inactive compounds for a specific target in different clusters was investigated (cluster separation analysis).[33,34] For this analysis, an *active cluster* for a particular target is defined as a cluster containing at least one active compound for this target. The *active cluster subset* now is defined as the total number of structures in all active clusters for a single target (active and inactive compounds). Then the proportion p of active structures only in this subset is computed and compared to the proportion (p_0) of active structures for this target in the total database. If 10 active clusters are found with 80 active and 20 inactive compounds in total, the proportion p is $80/100 = 0.8$ for this target. If the entire database contains 1000 compounds, p_0 is $80/1000 = 0.08$. Any increase in p compared to p_0 shows a trend to separate active and inactive structures.

A comparison of some selection methods is shown in Figure 2. Here the identification numbers of the biological classes are given on the *x*-axis versus the populations for the entire IC93 database (b) and various subsets on the *y*-axis, generated using maximum dissimilarity methods (a), random selection (c), or hierarchical clustering using 2D fingerprints (d). Using random selections, only 55% of the biological classes are covered (from 40.8% to 65.8%), while the maximum dissimilarity selection covers 83.6% of all classes (cf. Table 1). All maximum dissimilarity selections used here are based on 2D fingerprints for compound selections.

These investigations were extended to the following descriptors (given with abbreviations): 1. Fp/Ap, 2D and atom-pair fingerprints, 2. 2D fingerprints, 3. all, 2D fingerprints, atom-pair fingerprints, HDisq descriptors, and pharmacophoric flexible 3D fingerprints, 4. Fp/Ap/Mod3D, 2D, atom-pair, and pharmacophoric flexible 3D fingerprints, 5. Fp/Mod3D, 2D and pharmacophoric flexible 3D fingerprints, 6. HDisq(PCA), topological HDisq indices (eight principal properties), 7. Mod.Flex3D, pharmacophoric flexible 3D fingerprints, 8. CoMFA ster, steric molecular fields, 9. StdFlex3D, standard flexible 3D fingerprints, 10. molecular weight, and 11. atom-pair fingerprints. To include pharmacophore information for 3D flexible fingerprints, distances between pharmacophoric groups were encoded in a bitstring for hydrogen bond donors and acceptors,

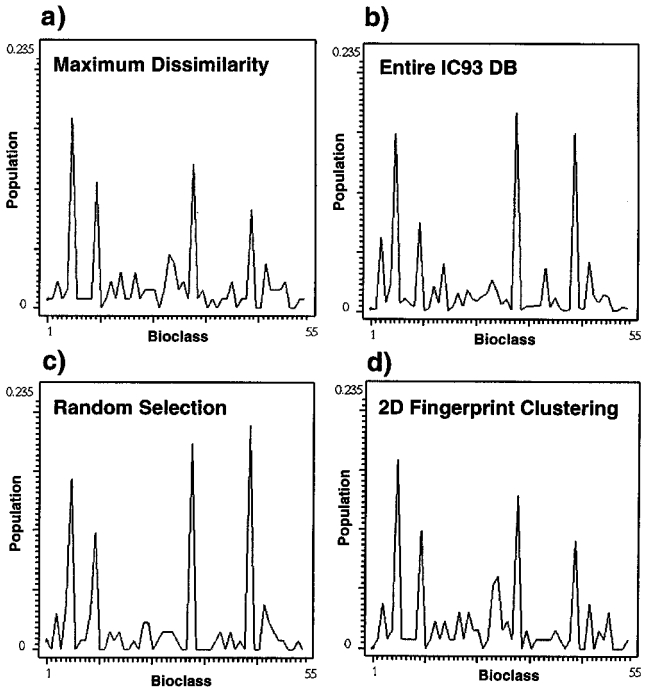

Figure 2. Normalized populations for all 55 biological classes of the IC93 database. The biological classes are shown on the *x*-axis; the population are given on the *y*-axis (normalized between 0 and 1): (a) subset of 120 compounds obtained using maximum dissimilarity selection and 2D fingerprints, (b) entire IC93 database, (c) subset of 120 compounds obtained using random selection (only one of 100 random selection is displayed), and (d) subset of 120 compounds obtained using hierarchical clustering and 2D fingerprints.

Table 1. Selection and Cluster Analysis of 120 Compounds from the IC93 Database: Population Analysis[a]

selection method/descriptor	% noncovered	% covered
random100[b]	44.43	55.57
maximum dissimilarity	16.40	83.60
2D fingerprints	21.80	78.20
atom pairs	52.70	47.30
HDisq(PCA)	29.10	70.90
CoMFA ster	49.10	50.90
StdFlex3D	47.30	52.70
Mod.Flex3D	41.80	58.20
mol weight	30.90	69.10
all[c]	23.60	76.40
Fp/Ap[c]	21.80	78.20
Fp/Ap/Mod3D[c]	23.60	76.40
Fp/Mod3D[c]	21.80	78.20

[a] The percentage of noncovered and covered biological classes is given; populations are given in percent. [b] Percentage obtained as an average from 100 random selections (from 34.2% to 50.2%) with a standard deviation of 4.8. [c] Scaling using AUTOSCALE method.

charged centers, and aromatic and hydrophobic centers.[35] The results are listed in Tables 1 and 2 for the population analysis and the cluster separation analysis. Additionally, the data are visualized in Figure 3.

The standard 2D fingerprints perform best when used with a hierarchical cluster analysis (78.2% classes selected in a coverage analysis) or with maximum dissimilarity methods (83.6%). Combined descriptors containing 2D fingerprints generally show a performance similar to 2D fingerprints alone, while no other descriptor led to comparable results. Atom-pair, molecular steric field, or standard flexible 3D descriptors led to coverage rates of only 47.3%, 50.9%, and 52.7%, respectively. Using the modified flexible 3D fingerprints, an increase to 58.2% can be observed, while 2D

Table 2. Selection and Cluster Analysis of 120 Compounds Using the IC93 Database: Cluster Separation Analysis[a]

selection method	mean[d]	median[d]	SD[d]
2D fingerprints	0.52	0.50	0.30
atom pairs	0.09	0.07	0.06
HDisq(PCA)	0.31	0.20	0.29
CoMFA ster	0.14	0.06	0.20
StdFlex3D	0.12	0.10	0.08
Mod.Flex3D	0.14	0.11	0.10
mol weight	0.11	0.08	0.10
all[a]	0.51	0.52	0.32
Fp/Ap[a]	0.54	0.56	0.33
Fp/Ap/Mod3D[a]	0.46	0.47	0.31
Fp/Mod3D[a]	0.40	0.27	0.31
WHIM_1	0.11	0.06	0.17
WHIM_2	0.13	0.07	0.18
WHIM_3	0.10	0.05	0.14
ACF_2D	0.15	0.02	0.26
ACF_3D	0.14	0.02	0.28
Fp/WHIM_1[b]	0.13	0.05	0.20
Fp/WHIM_2[b]	0.13	0.08	0.17
Fp/WHIM_3[b]	0.10	0.06	0.14
Fp/ACF_2D[b]	0.33	0.16	0.36
Fp/ACF_3D[b]	0.43	0.28	0.37
Fp/WHIM_2/ACF_3D[b]	0.14	0.06	0.18
p_0[c]	0.018	0.005	0.03

[a] The proportion of active structures on a single biological target in active clusters is measured and averaged over 55 biological classes. Scaling using method AUTOSCALE for hierarchical clustering. [b] No scaling for hierarchical clustering. [c] p_0: proportion of active structures for the entire database. [d] The statistical mean, median, and standard deviation values are obtained by averaging individual properties over 55 biological classes.

topological descriptors further increase this rate to 70.9%.[36]

In the cluster separation analysis, flexible and rigid 3D descriptors lead to lower averaged proportions than 2D fingerprints (0.52). In general, the combination of 2D and 3D descriptors leads to a better separation than 3D descriptors. These values are computed as averages over all 55 classes based on the proportions of active structures in active clusters for each individual class. A proportion of 1 indicates the ability of a descriptor to completely separate active from inactive compounds.

Replacing standard by pharmacophoric 3D fingerprints led to a small improvement of the proportion. In accordance with other 3D descriptors, molecular shape descriptors do not show any remarkable cluster separation tendency. Obviously the alignment based on molecular principal axes for extended conformations is not sufficient as a superposition rule for this heterogeneous database, as already discussed above. However, selections based on molecular steric fields in a database with a known superposition rule are adequate to achieve meaningful results (see below). As expected, molecular weight is inappropriate as the primary descriptor. Detailed proportions for all 55 classes for selected descriptors are shown in Figure 4 (cf. Supporting Information).

Subsequently, additional 2D and 3D molecular descriptors were investigated alone or in combinations. The abbreviations acf2D and acf3D indicate 2D and 3D spatial autocorrelation functions using electrostatic or lipophilic atomic properties. Atom-pair descriptors were modified in the following way: Using the program DISCO[37] pharmacophoric points were identified and bond distances between pairs of pharmacophoric points were used to set bits in a fingerprint. The results from the corresponding cluster separation analysis are given in Figure 5 and the Supporting Information.

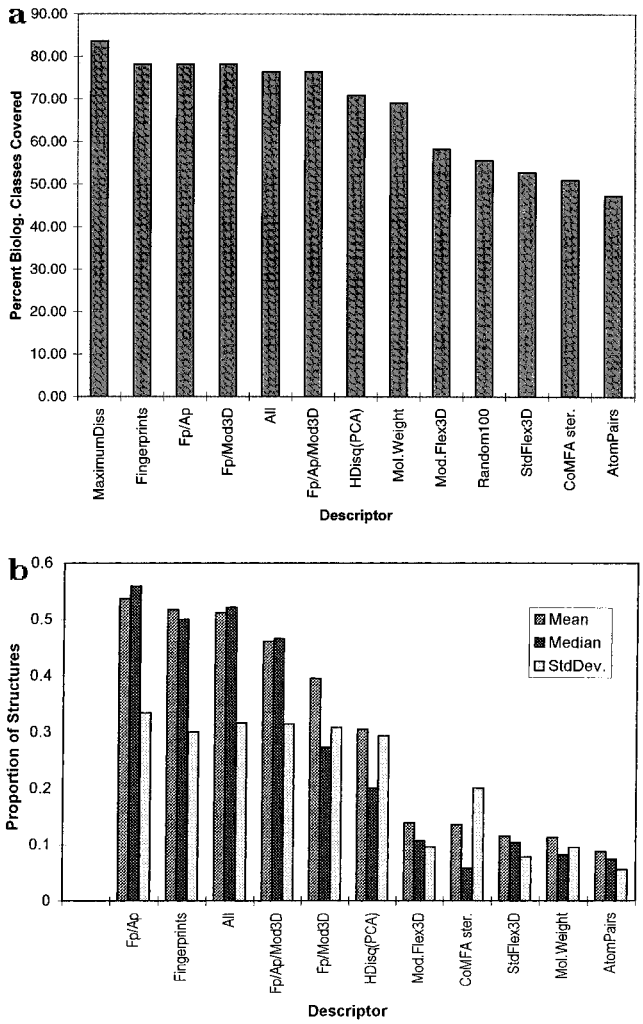

Figure 3. (a) Coverage of 55 biological classes of the IC93 database. Each descriptor displayed on the *x*-axis was used to select 120 diverse structures. On the *y*-axis the percentage of classes covered is shown. (b) Correspondence between biological activity and cluster formation based on a cluster separation analysis. Using hierarchical cluster analyses and various descriptors, the averaged proportions of active structures in active clusters are computed.

All 2D and 3D descriptors showed lower averaged proportions of active structures in active clusters than 2D fingerprints. It is remarkable that the alignment-independent WHIM descriptors do not significantly improve the results when compared to molecular steric fields, while a positive trend can be seen for 2D and 3D autocorrelation functions. It could be shown that 3D metrics are useful for addressing this type of question only in combination with a valid 2D descriptor. The modified pharmacophoric atom-pair descriptors perform slightly better than the original descriptor: a significant improvement is remarkable only for the biological classes 13, 18, 33, 41, and 44. Combining 2D fingerprints with different atom-pair descriptor implementations does not improve these results.

Only small differences were observed when comparing different hierarchical cluster methods using 2D fingerprints (median, complete, average, single). The lowest proportion (0.48) is obtained using the method single, while the best proportion of 0.52 was obtained using the method median. Hence this method was used for all clustering studies. An extensive investigation of different clustering techniques was recently described by Martin et al.[34]

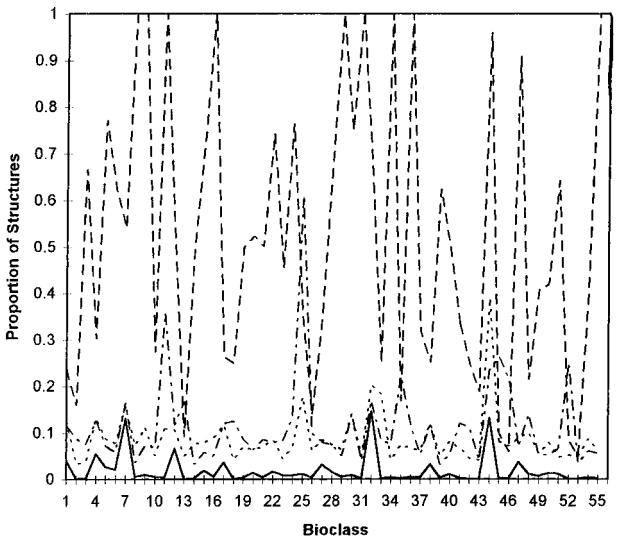

Figure 4. Detailed proportion of active structures in active clusters for all 55 biological classes of the IC93 database for 2D fingerprints (– – –), atom-pair descriptors (···), and molecular weight (–·–) in comparison to the ratio of active compounds for a single target to the total number of compounds (p_0) (—).

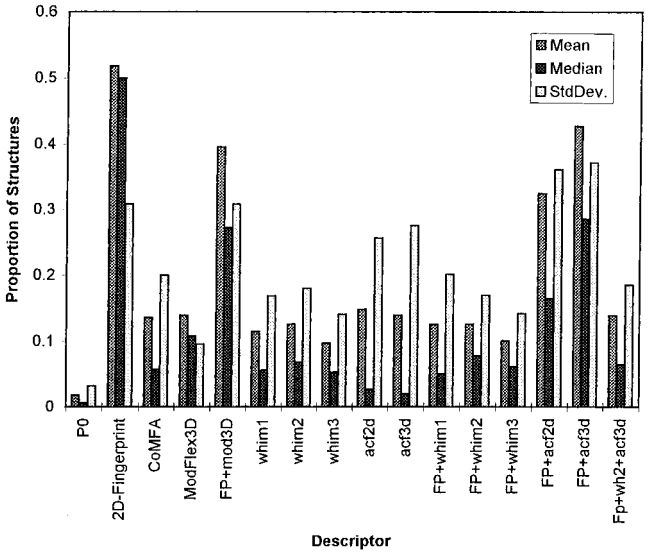

Figure 5. Correspondence between biological activity and cluster formation based on a cluster separation analysis for a set of alternative descriptors. For details, see Figure 3b.

Local Diversity—Radius of Similarity. A molecular descriptor useful for library design should have a well-defined similarity radius.[38] Two compounds, which are more similar than defined by this radius, should have similar biological properties; otherwise any descriptor-based design would be similar to random selection. Hence the relationship between structural similarity and biological activity was studied for molecular steric fields and 2D fingerprints. Structurally and biologically diverse QSAR datasets containing 729 molecules with known alignment rules were used for this investigation (Table 3). The datasets used as follows: 1. a set of 14 tropanes binding to the dopamine transporter;[39] 2. a set of 19 analogs of the insecticidal alkaloid ryanodine, which induces calcium release in muscle by binding to a muscle-specifc receptor;[40] 3. a set of 38 1,4-benzodiazepines binding with high affinities and selectivities to the diazepam-insensitive (DI) isoform of the benzodiazepine receptor,[41,42] data for the binding to the diazepam-sensitive isoform (DS) and the

Table 3. Local Structural Diversity for Different Biological Target Series[a]

	max slope		range			
dataset	fingerprint	CoMFA	fingerprints	CoMFA	bioactivity	total no. of compds
1. tropanes	0.07	0.07	0.32	0.16	2.40	14
2. ryanodines	0.05	0.02	0.60	0.30	3.20	19
3. Bzr(log DS)	0.06	0.04	0.80	0.40	3.50	38
3. Bzr(log DI)	0.05	0.06	0.80	0.40	3.20	38
3. Bzr(DI/DS)	0.05	0.05	0.80	0.40	4.00	38
4. benzodiazepines	0.04	0.02	0.55	0.40	3.70	40
5. steroids	0.04	0.04	0.55	0.35	2.80	48
6. HIV-1 inhibitors (100)	0.02	0.05	0.47	0.40	2.80	100
7. ACE inhibitors	0.10	0.06	0.85	0.80	7.70	138
8. DHFR inhibitors	0.07	0.04	0.55	0.50	5.30	256
sum						729
mean	0.05	0.04	0.63	0.41	3.86	
SD	0.020	0.016	0.18	0.16	1.6	

[a] Details about used datasets and the biological target properties are given in the text and the corresponding references. For each dataset the pairwise distance of the physicochemical descriptor was plotted on the *x*-axis versus the pairwise distance of the biological target property on the *y*-axis. The maximum slope was extracted based on a diagonal from $x,y = 0,0$ to $\max(y),\max(y)$. For 2D fingerprints the similarity given as 1 − Tanimoto coefficient is noted, while for CoMFA molecular steric fields the field correlation coefficient is reported.

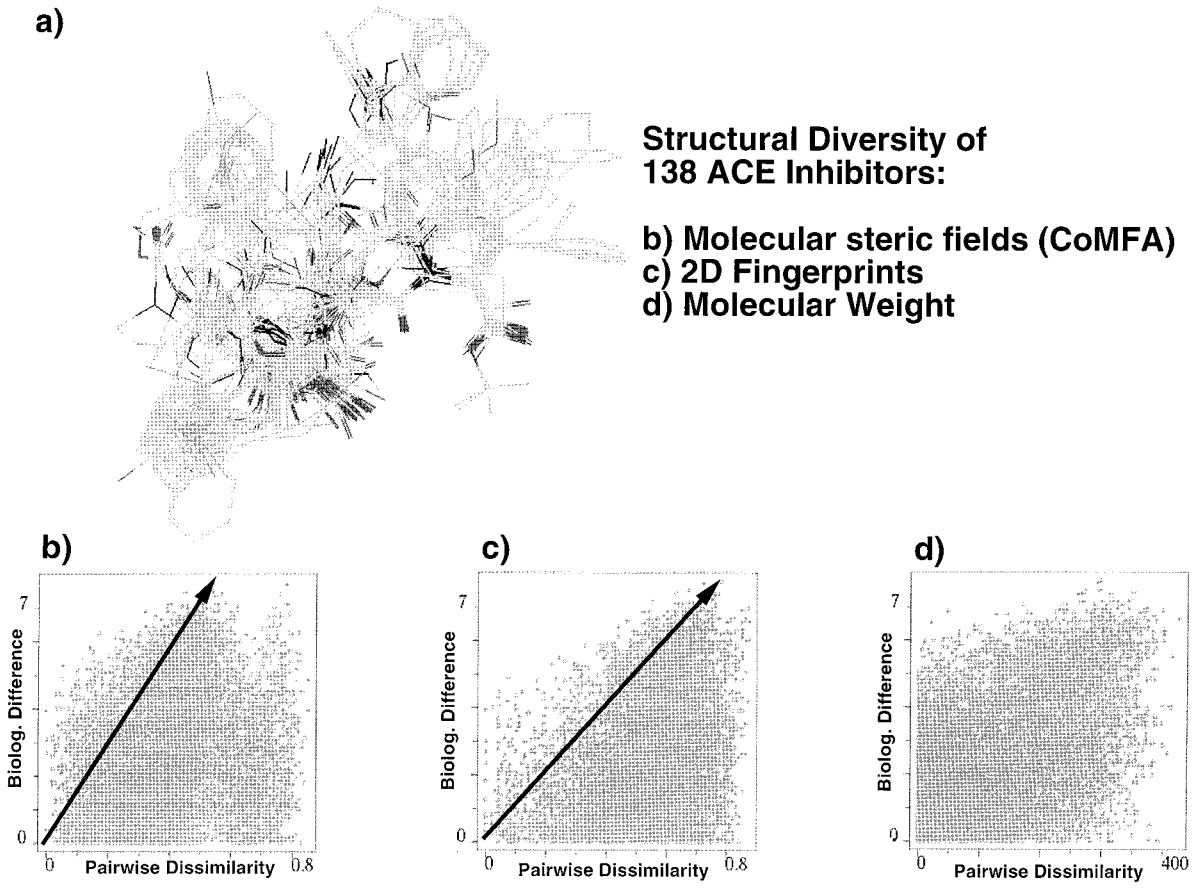

Figure 6. Comparison of pairwise biological distances versus various molecular descriptor differences for a dataset consisting of 138 angiotensin-converting enzyme (ACE) inhibitors with 9453 data points [$n(n-1)/2$]: (a) molecular geometries and the superposition rule used to derive the molecular steric field similarities in accordance with literature data, (b) molecular steric fields, (c) 2D fingerprints, and (d) molecular weight.

selectivity for the high-affinity binding to the DI site (DI/DS) were used; 4. a set of 40 benzodiazepines with inhibition of fluoronitrazepam radioligand binding to bovine cortical membranes;[43] 5. a set of 48 steroids with binding affinities to the progesterone receptor;[44] 6. a set of 100 2-heterosubstituted statine derivatives inhibiting human immunodeficiency virus type-1 proteinase;[45] 7. a set of 138 inhibitors of angiotensin-converting enzyme ACE;[46] and 8. a set of 256 triazines as inhibitors of dihydrofolate reductase.[47,48]

Those diverse datasets were selected because the reported biological activity covers more than 2.5 orders of magnitude. For each dataset the dissimilarity for each pair of molecules was computed using 2D fingerprints (1 − Tanimoto coefficient) and steric fields (1 − field correlation coefficient). In addition, the absolute differences of the biological activities, molecular weights, and random numbers (as reference) were computed. Subsequently, scatter plots showing the descriptor differences on the *x*-axis versus the biological differences on the *y*-axis were generated.[49] A graph for a valid molecular descriptor reveals a characteristic shape (Figure 6), which allows to derive a maximum change of the biological activity per change in the descriptor.

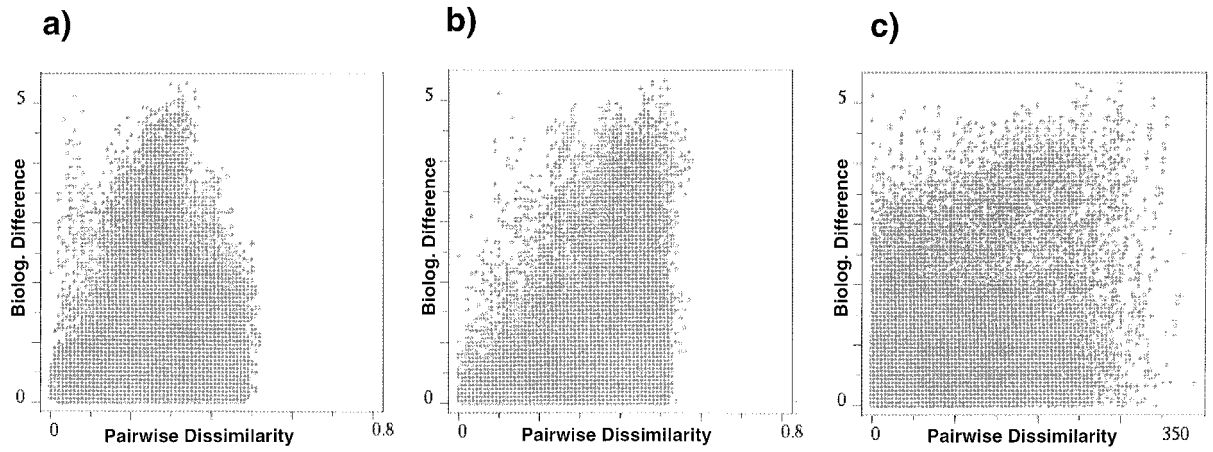

Figure 7. Comparison of pairwise biological distances versus molecular descriptor differences for a dataset of 256 triazines as inhibitors of dihydrofolate reductase: (a) molecular steric field, (b) 2D fingerprints, and (c) molecular weight.

Small changes in the physicochemical descriptor should lead to only small changes in the biological properties.

This concept will be illustrated using 138 diverse ACE inhibitors (Figure 6a).[46] For this set, three graphs are displayed in Figure 6 with the pairwise differences of molecular steric fields (b), 2D fingerprints (c), and molecular weight (d) versus the absolute pairwise differences of the biological activities on the *y*-axis. The arrows in Figure 6b,c indicate a linear gradient with some discontinuities. Both graphs for 2D fingerprints and steric fields reveal maximum gradients with slopes of 0.10 or 0.06 (Table 3), while for molecular weight such a gradient could not be derived, suggesting that this is not a valid descriptor. Similar plots are shown in Figure 7 for 256 triazines as inhibitors of dihydrofolate reductase.[47,48] Again, only for 2D fingerprints and molecular steric fields is it possible to derive this maximum gradient.

Table 3 summarizes the results for all datasets; averaged maximal slopes of 0.05 for 2D fingerprints and 0.04 for steric fields were obtained. Using these data, a similarity radius was estimated in the following way: the averaged maximal slopes were multiplied by 3 (corresponding to a tolerance factor of 3 orders of magnitude for biological activity), and values of 0.15 and 0.12 were obtained for 2D fingerprints and steric fields, respectively. Hence, if two molecules have a Tanimoto coefficient larger than 0.85 (1 − similarity radius) or a steric field similarity larger than 0.88, the biological activity of the first molecule is similar to that of the second one (within a tolerance of 3 orders of magnitude). These findings are in agreement with other observations.[33,50,51] This concept now allows to reduce the redundancy of a database by rejecting structurally similar molecules based on this similarity radius.

Subsequently all other descriptors were investigated using this validation concept and the set of 138 ACE inhibitors (Table 4). For each descriptor the maximum slopes were derived from the pairwise distance plots based on a hypothetical diagonal from $x,y = 0,0$ to x,y=max(y),max(y) (x,y-coordinates from the datapoint with the highest y-value, given in column max slope in Table 4). For each slope the corresponding similarity radius was computed based on the same tolerance value as used before. To monitor the distribution of points in each individual scatter plot, the results from a χ^2 statistical test are summarized in Table 4. These data are obtained considering the number of points within the lower right triangle (i.e., below the maximum slope diagonal) versus the expected number in this area for an equal distribution. A large χ^2 value indicates that the null hypothesis is rather unlikely and the corresponding descriptor is able to predict biological activities for similar compounds. For all 12 descriptors listed in Table 4, the pairwise distance plots used to generate these results are shown in Figures 8 and 9.

Table 4. Local Diversity and Validation Studies for Various Descriptors Using the ACE Inhibitor Dataset[a]

descriptor	max_slope[e]	sim rad[e]	ULTP[b]	χ^2 [c]
2D fingerprints	0.10	0.30	12.54	3555.9
atom pairs	0.10	0.30	12.96	2593.8
Mod.Flex3D	0.09	0.27	26.09	1080.8
HDisq(PCA)[d]	2.22	6.66	29.75	775.3
CoMFA	0.06	0.18	29.87	766.1
random number	0.04	0.11	31.79	626.9
Mod_AP	0.07	0.22	40.04	187.6
ACF_2D[d]	0.48	1.45	41.07	150.8
ACF_2Dlipo[d]	0.35	1.05	41.48	137.2
WHIM_2[d]	6.66	20.00	42.48	106.9
ACF_3Dlipo[d]	1.22	3.66	44.69	53.3
StdFlex3D	0.07	0.22	45.15	44.5
WHIM_1[d]	0.85	2.54	50.56	0.6
ACF_3D[d]	1.32	3.95	51.11	2.3
mol weight	38.10	114.31	51.25	2.3
WHIM_3[d]	5.68	17.09	66.86	537.4

[a] Sorted by ascending values in column ULTP. ACF_2Dlipo and ACF_3Dlipo indicate 2D or 3D autocorrelation vectors with Crippen's lipophilicity parameters as atomic properties p_i. Pairwise descriptor differences are given as 1 − Tanimoto coefficient or correlation coefficent (CoMFA fields), except where otherwise stated. [b] ULTP: percent of all $n(n-1)/2$ data points in the *u*pper *l*eft *t*riangle. [c] χ^2: χ^2 statistical test using the number of points in the lower left triangle versus the expected number of points in this area for an equal distribution (i.e., null hypothesis). [d] Pairwise differences given as eucledian distances. [e] Maximum slope computed based on a diagonal from $x,y = 0,0$ to max(y),max(y). The corresponding similarity radius (sim rad) is estimated using a biological activity tolerance of 3 orders of magnitude.

2D fingerprints show the best separation of data points between the upper left and the lower right triangle. Only 12.54% of all datapoints are found in the upper diagonal area, which is consistent with the high χ^2 value of 3555.9. Atom-pair descriptors (Figure 8a) also show such a tendency: here only 12.96% of the datapoints are located in the upper diagonal area (χ^2 value of 2593.8). The replacement of standard by modified pharmacophoric atom-pair descriptors (Figure 8b) decreases the χ^2 value to 187.6. A possible rationale for this fact is that standard atom-pair descriptors are atom-based metrics, while the pharmacophores used to

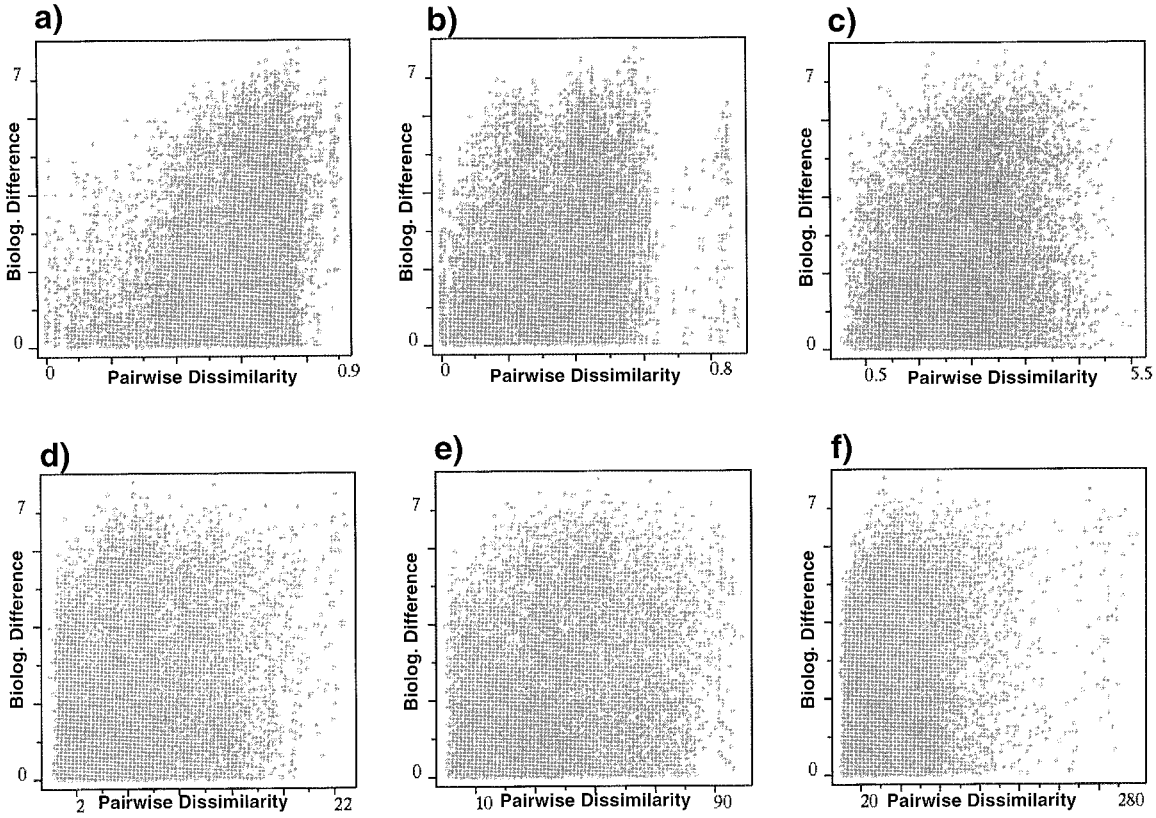

Figure 8. Comparison of pairwise biological distances versus various molecular descriptor differences for 138 ACE inhibitors. The descriptors investigated are as follows: (a) atom-pair descriptors, (b) modified (pharmacophoric) atom-pair descriptors, (c) HDisq, (d) WHIM_1 (no weighting), (e) WHIM_2 (van der Waals weighting), and (f) WHIM_3 (volume weighting). For further details, see text and Table 4.

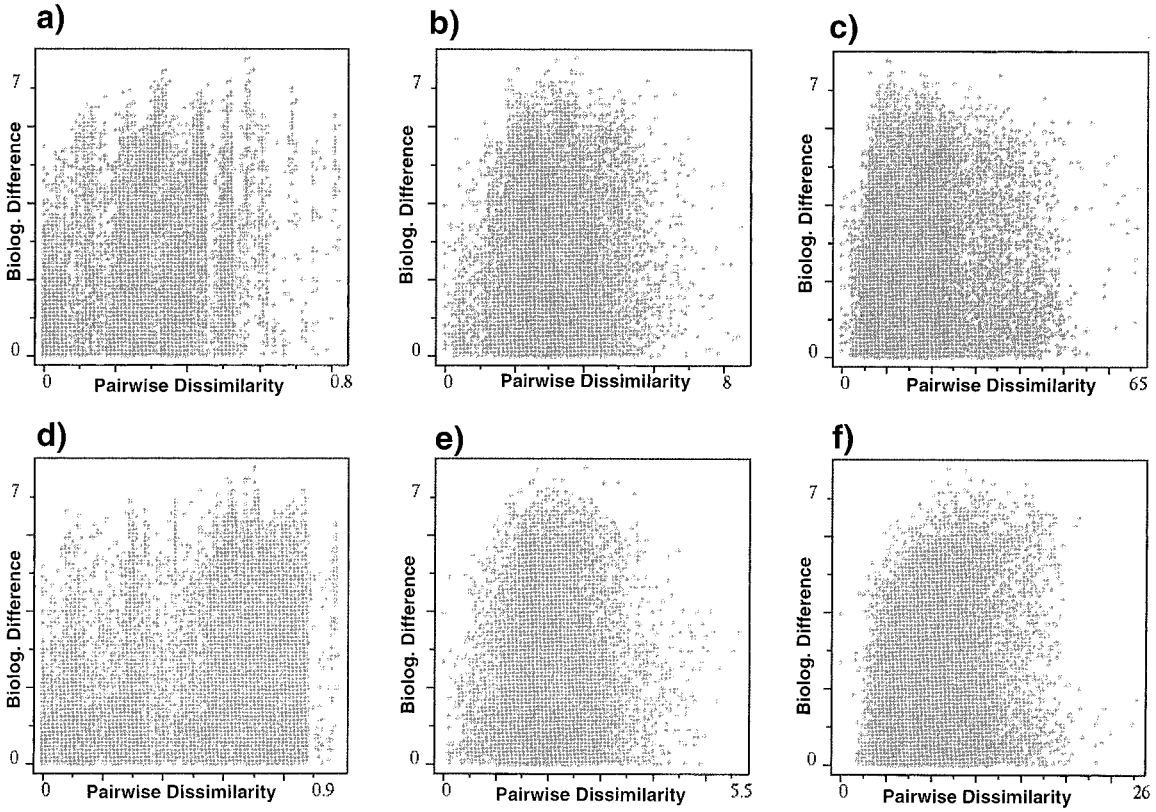

Figure 9. Comparison of pairwise biological distances versus various molecular descriptor differences for 138 ACE inhibitors. The descriptors investigated are as follows: (a) standard flexible 3D fingerprints, (b) 2D spatial autocorrelation functions, (c) 3D spatial autocorrelation functions, (d) modified pharmacophoric flexible 3D fingerprints, (e) 2D lipophilic spatial autocorrelation functions, and (f) 3D lipophilic spatial autocorrelation functions. For details, see text and Table 4.

derive the modified descriptors are distributed over groups of multiple atoms thus leading to an uncertainty in the definition of a bond distance. However, replacing the standard atom-based flexible 3D fingerprints (Fig-

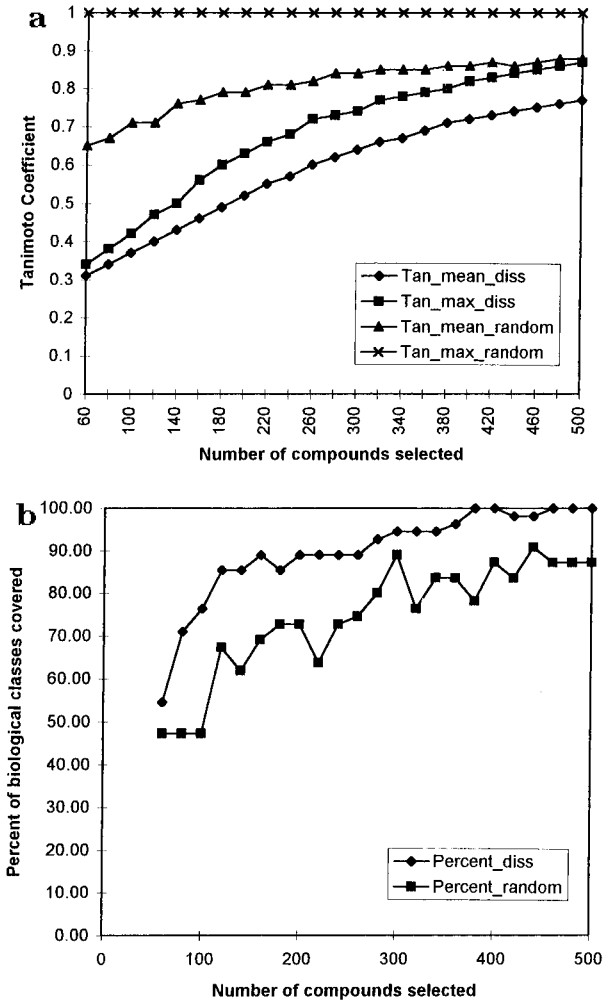

Figure 10. Selection of compound subsets from the IC93 database using different methods; random or maximum dissimilarity (denoted as random or diss) selections based on 2D fingerprints: (a) comparison of the mean and maximum Tanimoto coefficient (denoted as mean or max) for both selections (*y*-axis) versus the number of compounds selected within each subset on the *x*-axis and (b) comparison of the percentage of biological classes covered from the IC93 database (*y*-axis) versus the number of compounds selected within each subset (*x*-axis).

ure 9a) with modified pharmacophore-based flexible 3D fingerprints (Figure 9d) increases the χ^2 value from 44.5 to 1080.8. Here the bond distance is replaced by the Cartesian distance between feature centroids. It should be noted that 2D spatial autocorrelation functions with electrostatic (Figure 9b) and lipophilic (Figure 9e) atomic properties show a better separation tendency than the corresponding 3D autocorrelation descriptors (Figure 9c,f).

Design of Optimally Diverse Database Subsets. Finally several compound subsets with 60−500 members were selected from the IC93 database using 2D fingerprints and maximum dissimilarity or random selections. For each subset the Tanimoto coefficients from each compound to its nearest neighbor were computed and further analyzed. The mean and maximum values for each subset are plotted in Figure 10 against the subset population to monitor the degree of redundant compounds in each compound collection. The abbreviations used are as follows: Tan_mean_diss as mean Tanimoto coefficient for the maximum dissimilarity selection and Tan_mean_random for the random selection, Tan_max_diss as maximum Tanimoto coefficient for the maximum dissimilarity selection and Tan_max_random for the random selection.

An increase in the subset population led to an increased mean Tanimoto coefficient for either random or maximum dissimilarity subsets. However for a given number of compounds, the maximum dissimilarity subset is more diverse than the random selection subset. The maximum Tanimoto coefficient steadily increases for the maximum dissimilarity subsets, while it is close to 1 for the randomly chosen subset. This suggests that more diverse subsets are obtained using the first method, while a random selection samples redundant compounds.

Additionally, the subsets obtained using maximum dissimilarity selections cover more biological classes of the IC93 database than the corresponding random subsets. In Figure 10b, the percentage of covered biological classes is plotted on the *y*-axis versus the number of structures for the random subset (percent_random) or maximum dissimilarity subset (percent_diss). When selecting more than 380 structures using maximum dissimilarity methods, all biological classes are covered except a single biological class at levels of 420 and 440 structures. This class 41 with four members is not selected because of its similarity to class 35 (0.80 as lowest pairwise similarity). In contrast, there are more classes not selected in the corresponding random subsets. Maximum dissimilarity subsets with more than 460 structures corresponding to a maximum Tanimoto coefficient of 0.85 (Figure 10a) can be selected without missing any biological information. This Tanimoto coefficient corresponds to the 2D fingerprint similarity radius as derived in the previous section.

Thus a diverse subset for the 1283 biologically active structures from IC93 can be obtained, when selecting 487 compounds (38%) using 2D fingerprints and a similarity radius of 0.85. Now diverse compounds are selected, until no new compound can be found without violating this similarity radius. This subset covers all 55 biological classes (cf. Figure 10b), while 10 random selections of 487 compounds did not cover 13.1% of all classes (as average). Hence such a subset should be called an *optimally diverse* subset: all classes are covered using the lowest possible number of diverse compounds.

Conclusion

The use of valid molecular descriptors is an essential problem in the design of combinatorial libraries or selection of nonredundant compounds from databases. The present study addresses the question of which descriptor is appropriate to group molecules according to their biological properties. A randomly selected subset covers less biological classes than any descriptor-based rational selection. 2D fingerprint-based descriptors are very effective in selecting representative subsets of bioactive compounds. Molecules with similar biological properties are efficiently grouped. Thus, 2D substructure information is important for a valid descriptor.

Although an improvement of atom-pair and flexible 3D fingerprint descriptors was achieved using pharmacophore-based definition rules, none of these descriptors samples as efficiently as UNITY 2D fingerprints. From this study a hierarchy of various descriptors can be derived. 2D fingerprints can be classified as *first-order* descriptors, while descriptors based on pharmacophores

or molecular fields could be classified as *second-order* descriptors. They are useful in designing an analog library based on their local diversity properties, but they do not allow to select diverse compounds from a library in a lead discovery project. Molecular weight, cost of reactants, and clog *P* can then be classified as tertiary descriptors to further refine an initial selection.

The estimation of a similarity radius for valid descriptors was based on datasets for 10 biological targets. Comparing several randomly selected subsets with subsets obtained using maximum dissimilarity methods based on 2D fingerprints also shows that a similarity radius of 0.85 leads to a complete coverage of all biological classes in the IC93 database.

A diverse compound selection with a mean Tanimoto coefficient lower than this similarity radius reduces not only the number of compounds but also the information present in the parent database. Thus, an optimal diverse selection can be defined as a selection without redundant structures but with all the information from the original database. The design of an optimally diverse subset starting from the IC93 database leads to a selection of 487 compounds corresponding to 38% of this database, which indeed covers all biological classes, while no structure is more similar than 0.85 to any other structure within this subset. Thus the rejection of 62% redundant structures increases efficiency within a biological screening project.

While the results from the local diversity analysis show that 2D fingerprints, molecular steric fields, and atom-pair descriptors are useful metrics to design an analog library, for designing an analog library, the analysis of their global diversity performances clearly has shown that only 2D fingerprints alone or in combination with one of the secondary descriptors are sufficient to separate biologically active from inactive compounds.

Acknowledgment. The comments of Dr. T. Pötter (BAYER), Dr. A. Zaliani (Italfarmaco), Dr. R. Cramer III, D. Patterson, Dr. A. Ferguson, Dr. P. Hecht, and Dr. R. Clark (TRIPOS) are gratefully acknowledged. The author thanks Dr. K. Koehler (IRBM, Pomezia), Dr. R. Krömer (Sandoz, Vienna), Dr. S. DePriest, Dr. M. Bohl, and Dr. S. Güssregen (TRIPOS) for making structures available.

Supporting Information Available: Additional 11 figures (S1−S11) with proportions of active structures in active clusters for the IC93 database given for each individual biological class, with a comparison of pairwise distances versus various molecular descriptor differences for HIV protease inhibitors, and two tables (T1, T2) with an overview of the biological classes from the IC93 database and the numeric results from the comparison of random versus maximum dissimilarity selections (17 pages). Ordering information is given on any current masthead page.

References

(1) For recent reviews, see: (a) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251. (b) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1399. (c) Madden, D.; Krchnak, V.; Lebl, M. Synthetic combinatorial libraries: Views on techniques and their applications. *Perspect. Drug Discovery Des.* **1995**, *2*, 269−285. (d) Ellman, J. A. Design, Synthesis and Evaluation of Small-Molecule Libraries. *Acc. Chem. Res.* **1996**, *29*, 132−143. (e) Gordon, E. M.; Gallop, M. A.; Patel, D. V. Strategy and Tactics in Combinatorial Organic Synthesis. Application to Drug Discovery. *Acc. Chem. Res.* **1996**, *29*, 144−154.

(2) Moos, W. H.; Green, G. D.; Pavia, M. R. Recent Advances in the Generation of Molecular Diversity. *Annu. Rep. Med. Chem.* **1993**, *28*, 315−324.

(3) Ferguson, A. M.; Patterson, D. E.; Garr, C.; Underiner, T. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, *1*, 65−73.

(4) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity; Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(5) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman and Hall: London, 1995.

(6) Maggiora, G. M.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990; pp 99−117.

(7) SYBYL Molecular Modelling Package, versions 6.2 and 6.22; Tripos Inc., 1699 S. Hanley Rd, St. Louis, MO 63144.

(8) UNITY Chemical Information Software, version 2.5; Tripos Inc., 1699 S. Hanley Rd, St. Louis, MO 63144.

(9) For details to compute fingerprints, see: UNITY Chemical Information Software, version 2.5, Reference Guide pp 45−58; Tripos Inc., 1699 S. Hanley Rd, St. Louis, MO 63144.

(10) (a) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987. (b) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Inter-molecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.

(11) Sheridan, R. P.; Nachbar, R. B.; Bush, B. L. Extending the trend vector: The trend matrix and sample-based partial least squares. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 323−340.

(12) (a) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82. (b) Gombar, V. K.; Enslein, K. Quantitative Structure-Activity Relationship (QSAR) Studies using Electronic Descriptors Calculated from Topologial and Molecular Orbital (MO) Methods. *Quant. Struct.-Act. Relat.* **1990**, *9*, 321−325.

(13) Kier, L. B.; Hall, L. H. *Molecular Connectivity and Drug Research*; Academic Press, New York, 1976.

(14) (a) Gombar, V. K.; Jain, D. V. S. Quantification of Molecular Shape and Its Correlation with Physicochemical Properties. *Indian J. Chem.* **1987**, *26A*, 554−555. (b) Kier, L. B. Index of Molecular Shape from Chemical Graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.

(15) *SYBYL 6.2, Molecular Spreadsheet Manual*; TRIPOS, Inc.: St. Louis, MO, 1995; pp 234−235.

(16) (a) Dillon, W. R.; Goldstein, M. *Multivariate Analysis: Methods and Applications*; Wiley: New York, 1984. (b) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley: New York, 1980. (c) Cramer, R. D., III. BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State. *J. Am. Chem. Soc.* **1980**, *102*, 1837−1849. (d) Stahle, L.; Wold, S. Multivariate Data Analysis and Experimental Design in Biomedical Research. In *Progress in Medicinal Chemistry*; Ellis, G. P., West, G. B., Eds.; Elsevier: New York, 1988; pp 292−338. (e) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johanson, E.; Lindberg, W.; Sjöström, M. Multivariate Data Analysis in Chemistry. In *Chemometrics: Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; NATO ISI Series C 138; D. Reidel Publ. Co.: Dordrecht, Holland, 1984; pp 17−96.

(17) *SYBYL 6.2, Ligand-Based Design Manual*; TRIPOS, Inc.: St. Louis, MO, 1995; pp 220−225 for details of the implementation and references cited therein.

(18) Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. *J. Chemom.* **1994**, *8*, 111−125.

(19) (a) Moreau, G.; Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360. (b) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures, Application to SAR Studies. *Nouv. J. Chim.* **1980**, *4*, 757−764. (c) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *Eur. J. Med. Chem. Chim. Ther.* **1984**, *19*, 66−70.

(20) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Molecular Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(21) (a) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228. (b) Marsili, M.; Gasteiger, J. Charge Distribution from Molecular Topology and π Orbital Electronegativity. *Croat. Chem. Acta* **1980**, *53*, 601−614. (c) Gasteiger, J.; Marsili, M. Prediction of Proton Magnetic Resonance Shifts: The Dependence of Hydrogen Charges Obtained by Iterative Partial Equalization. *Org. Magn. Reson.* **1981**, *15*, 353−360. (d) Details of the implementation are given in *Sybyl 6.2 Theory Manual*; Tripos: St. Louis, MO, 1995; p 67.

(22) (a) Ghose, A.; Crippen, G. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. 1. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577. (b) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occuring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(23) Cramer, R. D., III; Patterson, D. E.; Bunce, J. E. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(24) Clark, M.; Cramer, R. D., III; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative Molecular Field Analysis (CoMFA). 2. Towards its use with 3D-Structural Databases. *Tetrahedron Comput. Methodol.* **1990**, *3*, 47−59.

(25) Clark, R. D.; Cramer, R. D., III. Personal communication.

(26) (a) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D structures. Theory. *J. Chemom.* **1994**, *8*, 263−272. (b) Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221−229. (c) Todeschini, R.; Bettiol, C.; Giurin, G.; Gramatica, P.; Miana, P.; Argese, E. Modeling and prediction by using WHIM descriptors in QSAR studies: submitochondrial particles (SMP) as toxicity biosensors of chlorphenols. *Chemosphere* **1996**, *33*, 71−79.

(27) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing Drug Screening Programs using Molecular Similarity Methods. In *QSAR: Quantitative Structure-Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.: New York, 1989; pp 173−176.

(28) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.

(29) *SYBYL 6.2, Ligand-Based Design Manual*; TRIPOS, Inc.: St. Louis, MO, 1995; pp 246−255 and references cited therein.

(30) Index Chemicus - Subset from 1993; Institute for Scientific Information, Inc. (ISI), 3501 Market St, Philadelphia, PA 19104.

(31) Pearlman, R. S.; Balducci, R.; Rusinko, A.; Skell, J. M.; Smith, K. N. CONCORD, program version 3.2.1; Tripos Inc., St. Louis, MO 63144.

(32) Leo, A. Programs CLOGP and CMR; BioByte Corp., Claremont, CA. Also available from Tripos Inc., St. Louis, MO 63144.

(33) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and Cluster Analysis Applied to Molecular Diversity. ACS Meeting, Anaheim, CA, 1995; abstract COMP3.

(34) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(35) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.

(36) For 15% of the compounds no HDisq descriptors were obtained due to the limited fragment database.

(37) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. Available from Tripos Inc., St. Louis, MO 63144.

(38) Patterson, D. E.; Cramer, R. D., III; Ferguson, A. M.; Clark, R. D. Personal communication.

(39) Carroll, F. I.; Mascarella, S. W.; Kuzemko, M. A.; Gao, Y.; Abraham, P.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J. Synthesis, Ligand Binding, and QSAR (CoMFA and Classical) Study of 3β-(3′-Substituted phenyl)-, 3β-(4′-Substituted phenyl)-, and 3β-(3′,4′-Disubstituted phenyl)tropane-2β-carboxylic Acid Methyl Esters. *J. Med. Chem.* **1994**, *37*, 2865−2873.

(40) Welch, W.; Ahmad, S.; Gerzon, K.; Humerickhouse, R. A.; Besch, H. R., Jr.; Ruest, L.; Deslongchamps, P.; Sutko, J. L. Structural Determinants of High-Affinity Binding of Ryanoids to the Vertebrate Skeletal Muscle Ryanodine Receptor: A Comparative Molecular Field Analysis. *Biochemistry* **1994**, *33*, 6074−6085.

(41) Wong, G.; Koehler, K. F.; Skolnick, P.; Gu, Z.-Q.; Ananthan, S.; Schönholzer, R.; Hunkeler, W.; Zhang, W.; Cook, J. M. Synthesis and Computer-Assisted Analysis of the Structural Requirements for Selective, High-Affinity Ligand Binding to Diazepam-Insensitive Benzodiazepine Receptors. *J. Med. Chem.* **1993**, *36*, 1820−1830.

(42) Zhang, W.; Koehler, K. F.; Zhang, P.; Cook, J. M. Development of a Comprehensive Pharmacophore Model for the Benzodiazepine Receptor. *Drug Des. Discovery* **1995**, *12*, 193−248.

(43) (a) Hobbs DeWitt, S.; Kiely, J. S.; Stankovic, C. J.; Schroeder, M. C.; Reynolds Cody, D. M.; Pavia, M. R. "Diversomers": An approach to nonpeptide, nonoligomeric chemical diversity. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6909−6913. (b) Hobbs DeWitt, S.; Czarnik, A. W. Combinatorial Organic Synthesis using Parke-Davis's DIVERSOMER Method. *Acc. Chem. Res.* **1996**, *29*, 114−122.

(44) Loughney, D. A.; Schwender, C. F. A comparison of progestin and androgen receptor binding using the CoMFA technique. *J. Comput.-Aided. Mol. Des.* **1992**, *6*, 569−581.

(45) Kroemer, R. T.; Ettmayer, P.; Hecht, P. 3D-Quantitative Structure-Activity Relationships of Human Immunodeficiency Virus Type-1 Proteinase Inhibitors: Comparative Molecular Field Analysis of 2-Heterosubstituted Statine Derivatives - Implications for the Design of Novel Inhibitors. *J. Med. Chem.* **1995**, *38*, 4917−4928.

(46) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(47) Silipo, C.; Hansch, C. Correlation Analysis. Its Application to the Structure-Activity Relationship of Triazines Inhibiting Dihydrofolate Reductase. *J. Am. Chem. Soc.* **1975**, *95*, 6849−6861.

(48) (a) Kroemer, R. T.; Hecht, P. A new procedure for improving the predictiveness of CoMFA models and its application to a set of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 396−406. (b) Kroemer, R. T.; Hecht, P. Replacement of steric 6−12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 205−212.

(49) Within TRIPOS the first individual to use these pairwise distance plots for descriptor validations was D. E. Patterson.

(50) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.

(51) Martin, Y. C. Personal communication.

JM960352+