Journal of
Chemical Information
and Modeling

Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets

Hans Matter, and Thorsten Ptter

J. Chem. Inf. Comput. Sci., **1999**, 39 (6), 1211-1225• DOI: 10.1021/ci980185h • Publication Date (Web): 29 October 1999

Downloaded from http://pubs.acs.org on March 12, 2009

Journal of Chemical Information and Computer Sciences is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets

Hans Matter*,† and Thorsten Pötter

Tripos GmbH, Martin-Kollar-Strasse 15, D-81829 München, Germany, and Bayer AG, Alfred-Nobel Strasse,
D-40789 Monheim, Germany

Received August 2, 1999

The performance of two important 2D and 3D molecular descriptors for rational design to maximize the structural diversity of databases is investigated in this publication. Those methods are based either on a 2D description using a binary fingerprint, which accounts for the absence or presence of molecular fragments, or a 3D description based on the geometry of pharmacophoric features encoded in a fingerprint (pharmacophoric definition triplets, PDTs). Both descriptors in combination with maximum dissimilarity selections, complete linkage hierarchical cluster analysis, or sequential dissimilarity selections were compared to random subsets as reference. This comparison is based on their ability to cover representative biological classes from parent databases (*coverage analysis*) and the degree of separation between active and inactive compounds for a biological target from hierarchical clustering (*cluster separation analysis*). While the similarity coefficients (Tanimoto, cosine) show only a minor influence, the number of conformations to generate the 3D PDT fingerprint lead to remarkably different results. PDT fingerprints derived from a lower number of conformers perform significantly better, but they are not comparable to a 2D fingerprint-based design. When 2D and 3D descriptors are combined with weighting factors > 0.5 for 2D fingerprints, a significant improvement of coverage and cluster separation results is observed for a small number of PDT conformers and medium sized subsets. Some combined descriptors outperform 2D fingerprints, but not for all subset populations. Applying sequential dissimilarity selection to PDT descriptors reveals that its performance is dependent on the initial ordering of compounds, while presorting according to 2D fingerprint diversity does not improve results. Finally the relationship between biological activity and similarity was investigated, showing that PDTs quantify smaller structural differences due to the large number of bits in the fingerprint.

1. INTRODUCTION

High-throughput screening and combinatorial chemistry are nowadays changing research in the chemical and pharmaceutical industry. For successful, but efficient, discovery of lead compounds, the use of rational design strategies for representative compound subsets is indispensable.[1] Even with the advent of miniaturization strategies, appropriate compound subsets are important to speed up lead finding by handling more assays in a given time. Several methods for diversity selection have been proposed.[2] In general designed subsets were shown to perform significantly better than randomly picked compounds in retrospective analyses.[3] In particular, 2D fingerprints are appropriate for designing subsets representing all biological properties of parent databases.[4] They were shown to perform better than many other common 2D or 3D descriptors.[5]

The concept of molecular diversity[6,7] is based on the *similar property principle*,[8] which states that structurally similar molecules should reveal similar physicochemical and biological properties. Thus it is possible to predict target properties for a molecule using known values for similar compounds. This should also allow one to select representative compounds covering the properties of the parent database or combinatorial library.[9] One interesting question is, whether the knowledge of the molecular 3D structure or the geometry of key features offers advantages for design.[10] Although 3D molecular structures are often important for explaining structure−activity relationships, most classifications into similar and dissimilar ones are still based on a 2D description. Previous investigations[5] of 3D descriptors included alignment-independent WHIM indices,[11] CoMFA steric fields,[12,13] flexible UNITY 3D fingerprints, and 3D spatial autocorrelation functions.[14,15] Although significant enhancements to CoMFA steric fields were introduced,[16] those still require a common framework for superposition. However, none of these descriptors led to better results than 2D fingerprints in terms of covering biological properties of a database by smaller subsets.

Publications by Mason et al.,[17,18] Brown and Martin et al.,[19,20] and Davies[21] have highlighted interesting properties of pharmacophoric triplets (PDTs) as novel 3D descriptors for selecting representative subsets. Today commercial software products are available[22,23] for pharmacophore-based library design. Successful applications of enhanced four-point pharmacophore descriptors have also been reported.[24] Those properties and the availability of three-point pharmacophore descriptors prompted us for a comparative study to group

* To whom all correspondence should be addressed.

† Current address: Chemical Research, Hoechst Marion Roussel GmbH, Building G 838, D-65926 Frankfurt am Main, Germany. Phone: ++49-69-305-84329. Fax: ++49-69-331399. E-mail: hans.matter@hmrag.com.

10.1021/ci980185h CCC: $18.00
Published on Web 10/29/1999

active and inactive compounds and select representative subsets for biological screening using PDTs and 2D fingerprints as reference.

The term “pharmacophore definition triplets” refers to a set of three pharmacophoric features, like acceptor atom/acceptor atom/hydrophobic point. Each possible triangle geometry for such a triplet disregarding its specification order is encoded in a fingerprint. Individual bits are referring to different triangle geometries formed between pharmacophoric points.

These descriptors were evaluated using maximum dissimilarity selection and complete linkage hierarchical clustering and compared to random subsets. To monitor the descriptors’ performance, two approaches were used: (a) a *coverage analysis* investigates the sampling of biological classes from parent databases in smaller subsets; (b) a *cluster separation analysis* is used to assess the degree of separation between active and inactive compounds for a particular biological class from hierarchical clustering. Furthermore the influence of the number of conformers and the similarity coefficient (Tanimoto[25] or cosine coefficients as dimensionless metrics; see ref 26 for a comparison) on the coverage and cluster separation results is investigated in detail. Finally it is evaluated, whether the combination of 2D fingerprints and PDTs into a single descriptor with different weighting factors can improve subset selection or clustering performance.

Compound selections and classifications were done using reference databases previously used to compare rational and random approaches:[3] a public database containing 1283 compounds active in 55 biological classes with several diverse templates and a database encompassing 334 compounds from 11 different structure−activity series. For these, database cross-checking between different classes was done and activities were determined in a single laboratory. These data are obtained in defined assay systems and not compiled from literature, which might be a potential source of uncertainty for the first database.

The choice of 2D fingerprints as reference is based on their comparison to other 2D or 3D descriptors.[4,5] It was found that compound subsets without any compound closer than 0.85 to another one (Tanimoto coefficient) are able to span the biological property space of a database. Each biological class is still populated by one or more bioactive compounds. Any removal of redundant structures should result in a subset spanning the same physicochemical diversity space and retaining the biological information from the parent database.[27]

2. METHODS

All calculations and database manipulations were done using the programs SYBYL[28] and UNITY.[29] In general, chemical structures are represented in the SYBYL line notation (SLN).[30] Automation of design and analysis procedures was done using the SYBYL programming language (SPL), UNIX shell scripts, and PERL scripts.

2.1. Two-Dimensional Fingerprints. Two-dimensional fingerprints, computed using UNITY,[29] contain information about the presence of molecular fragments in a binary format. For each structure, a list of all possible fragments of a particular length is generated and converted into a bitstring. Due to the large number of existing fragments in a database, it is not possible to assign one individual bit to only a single fragment. Hence, the following procedure is used: the SLN for each fragment generated is mapped to a unique integer in the range of $0-2^{31}$ using a cyclic redundancy check algorithm.[31] Each integer is then projected into this size-limited bitstring by a procedure known as “hashing”, setting one or multiple bits to “1”.[32] For each feature, multiple occurrences set more neighboring bits to “1”. This way of storing molecular information allows one to quantify the similarity of two molecules based on similarity coefficients, like the Tanimoto or cosine coefficients.[25,26] Both coefficients are based on the number of bit positions set in both individual bitstrings for both molecules normalized by the number of bits set in common, while they differ in the applied scaling. The Tanimoto coefficient is widely used in database analysis, as it has certain properties making the work with larger data sets very efficient. Here we used both similarity coefficients for comparison. A similarity coefficient of 0 means that both structures have no “1” bits in common and there is no intersection between both sets of fragments. In contrast, a value of 1 indicates that both fingerprints are identical.

2.2. Pharmacophore Definition Triplets. Pharmacophore definition triplets (PDTs) as 3D descriptors offer an alternative way to quantify molecular diversity by encoding spatial relationships within pharmacophoric pattern in molecules. For this study the Sybyl 6.3 implementation was used. The 2D fingerprint descriptor is modified such that each individual bit in a binary fingerprint now refers to a geometry in pharmacophoric space. The setting of an individual bit to 1 indicates the presence of a specific triangle geometry: a set of three pharmacophoric points separated by three particular distances. In general, five pharmacophoric feature definitions were used: acceptor atoms, acceptor sites, donor atoms, donor sites, and hydrophobic centers. While donor and acceptor atoms are part of the molecule, site points refer to interaction points located on a “virtual” receptor, defined by geometrical criteria.[33] The pharmacophoric feature definitions reflect biologically relevant physicochemical conditions and accommodate tautomeric potential. A set of 27 distance bins is specified from 2.5 to 15 Å in steps of 0.5 Å, which leads in total to a PDT fingerprint of 307.020 bits encoding triangle geometries. A single PDT fingerprint per molecule is stored for maximum dissimilarity-based selections and hierarchical cluster analysis, while for sequential selection, a cumulative fingerprint is used as union for all molecules in the subset.

Any PDT fingerprint is computed for conformational ensembles to account for molecular flexibility. Starting geometries were generated by 3D conversion using CONCORD.[34] Individual conformers were generated by random setting of rotatable dihedral angles, followed by refinement using the *directed tweak* algorithm[35] to release from steric overlap, bumps, and strain. This fast conformational analysis allows for processing of larger databases. Pharmacophore geometries from acceptable conformers were combined into a union fingerprint. The effect of the number of conformers on the sampling performance was also investigated. Hence, PDT descriptors were generated using different numbers of conformers for every molecule, namely, 10, 20, 50, 100, 200, 500, and 1000, respectively. If for rigid molecules the predefined number of conformers could not be generated, the fingerprint was computed from the maximum possible number of conformations.

Furthermore it was assessed, whether a fusion of 2D fingerprints and PDTs into a single descriptor can improve the subset selection or clustering performance. Different weighting factors were applied to 2D fingerprints and PDTs before computing the average similarity coefficients. Those weighting factors were changed from 0.2/0.8 for 2D fingerprints/PDTs to 0.8/0.2 in increments of 0.1. Here 10 or 100 conformers were used to generate PDT fingerprints.

2.3. Compound Selection and Analysis. Compound selections for 2D fingerprints and 3D pharmacophoric triplets were done using the maximum dissimilarity algorithm[36−39] and Tanimoto or cosine coefficients, respectively. A recent comparative study highlights the properties of several algorithms for dissimilarity-based selections.[40] In the present implementation new compounds are successively selected such that they are maximally dissimilar from the previously selected subset. This process is terminated either when a maximum number of compounds is chosen or when no further molecule can be selected without being too similar to one of the already selected members. After randomly selecting a seed, every new compound is chosen to be maximally dissimilar from all previous members. The first three compounds are rejected after the fourth selection, but they are allowed for later picking. The mean similarity coefficient is computed as average from coefficients for every structure to its nearest neighbor. For all PDT and combined 2D/3D descriptor-based dissimilarity selections, the same seed compound as that for 2D fingerprints is utilized. The success of maximum dissimilarity selections is evaluated by the coverage of biological classes from the original database in smaller subsets (*coverage analysis*).

Hierarchical cluster analysis[41,42] was used as an alternative method for molecular descriptor assessment and validation, as it offers more specific control by assigning every molecule to a group of compounds. Hierarchical clustering does not require any assumption about a final number of clusters to be generated; small clusters with very similar elements are nested within larger clusters containing more dissimilar structures. There is no a priori guideline which method is appropriate for a particular data set, while some techniques perform better for grouping similar compounds.[19] Here complete linkage clustering was applied using the Tanimoto or cosine coefficient; i.e., intercluster distances are computed using the most distant pair of elements in both clusters, leading to compact clusters and a lower number of singletons.

For analysis it was evaluated whether compounds of similar chemical structure and biological activity are grouped (*cluster separation analysis*). The degree of separation between actives and inactives for a particular target was determined from various cluster levels, generated from the final dendrograms. For this cluster separation analysis,[5,19] an *active cluster* is defined as a cluster containing at least one active compound for a particular target. This allows one to define an *active cluster subset* as the total number of structures in all active clusters for one target (combined actives and inactives). Then the proportion p of active structures only in this active cluster subset is computed and compared to the proportion of active structures in the entire database. If 10 active clusters are found with 80 active and 20 inactive compounds, the proportion p is $80/100 = 0.8$ for this target. If the entire database contains 1000 compounds, then the proportion of active structures in the entire database is $80/1000 = 0.08$. Any increase in p compared to that number indicates a trend to separate active and inactive compounds. The proportion p was averaged over all biological classes and plotted versus the increasing number of clusters at different levels of the complete dendrogram. Singletons were excluded from the analysis, as their proportion of 1.0 skew the results.[19]

Alternatively a sequential selection for PDTs was investigated. This is a computationally efficient procedure based on a composite PDT database fingerprint. Here a new structure is selected, if its PDT fingerprint is more diverse than a given Tanimoto coefficient threshold to the composite fingerprint of the already selected hitlist. Resulting subsets were evaluated using the coverage analysis.

Probability calculations were used to compare random selections to the rational approach, as earlier described.[3] Assuming a particular statistical distribution, it is possible to compute the probability p to find n_1 hits by n selections in a database with a total of N compounds and N_1 hits for a particular target. This allows one to evaluate how many target classes are covered by a purely random selection of n compounds.

3. RESULTS AND DISCUSSION

3.1. Characteristics of the Databases. Two databases from diverse sources with different characteristics were investigated. The first database IC93 represents a collection of 1283 biologically active molecules as a subset from the *IndexChemicus* 1993 database.[43] This database was divided into 55 biological classes according to the biological indication area, specified as a string in the original database. Compounds with similar biological activities were grouped into the same class for all subsequent analyses.[3,5] The second database BAYER contains 334 compounds; it was retrospectively generated on the basis of quantitative structure−activity series for 11 diverse biological assays. One important criterium in the selection of these quantitative structure−activity relationship (QSAR) series was the different size and similarity of individual series. Some physicochemical and structural properties of both databases are summarized in ref 3. Inactive compounds were not added, as every compound is assumed to be inactive in all but one biological assay, thus providing negative information to evaluate selection performances of diversity descriptors.

3.2. Maximum Dissimilarity Based Selections of Diverse Subsets. For the IC93 database, various subsets with 60−500 members (in steps of 20) were selected using 2D fingerprints or 3D PDTs (100 conformers) and Tanimoto or cosine coefficients, respectively. For analysis the coverage of biological classes is plotted in Figure 1 versus subset population; numerical results are given in Table 1. The coverage of biological classes is reported in percent on the basis of 55 classes for the IC93 database. Theoretical values[3] for random selections are also listed in Table 1 (*Random_theo*).

Two-dimensional fingerprint-based selections perform significantly better than 3D PDTs, while there is an almost similar behavior for both similarity coefficients. The performance of 3D PDTs is significantly better than random

Table 1. Random versus Maximum Dissimilarity Selection Using 2D Fingerprints or PDTs for Designing Representative Subsets of the IC93 Database (1283 Compounds; 55 Target Classes)[a]

NComp	FP_Tanimoto	FP_Cosine	PDT_Tanimoto	PDT_Cosine	Random_theo
60	61.82	58.18	47.27	49.09	45.47
80	65.45	72.73	54.55	56.36	51.93
100	76.36	76.36	60.00	58.18	57.05
120	81.82	81.82	61.82	61.82	61.29
140	83.64	87.27	67.27	63.64	64.78
160	85.45	87.27	69.09	69.09	67.73
180	85.45	87.27	72.73	69.09	70.36
200	89.09	89.09	74.55	72.73	72.73
220	89.09	89.09	74.55	74.55	74.73
240	89.09	89.09	74.55	74.55	76.60
260	89.09	89.09	74.55	74.55	78.13
280	89.09	89.09	76.36	78.18	79.71
300	90.91	89.09	76.36	80.00	81.04
320	96.36	96.36	78.18	80.00	82.27
340	96.36	96.36	80.00	80.00	83.40
360	96.36	96.36	81.82	80.00	84.38
380	98.18	98.18	81.82	80.00	85.44
400	98.18	98.18	81.82	81.82	86.25
420	98.18	98.18	81.82	81.82	87.15
440	100.00	98.18	81.82	83.64	87.85
460	100.00	100.00	83.64	83.64	88.53
480	100.00	100.00	87.27	85.45	89.20
500	100.00	100.00	87.27	85.45	

[a] The percentage of biological classes in IC93 covered by a subset is reported: NComp, number of compounds in a subset; FP_Tanimoto, 2D fingerprints and the Tanimoto coefficient; FP_Cosine, 2D fingerprints and the cosine coefficient; PDT_Tanimoto, 3D PDTs (100 conformers) and the Tanimoto coefficient; PDT_Cosine, 3D PDTs (100 conformers) and the cosine coefficient; Random_theo, theoretical random selection.

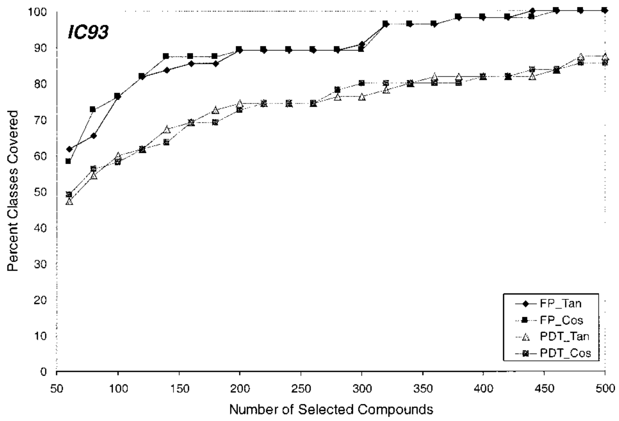


Figure 1. Maximum dissimilarity selection using 2D fingerprints and 3D PDTs for designing representative subsets of the IC93 database. The percent biological classes covered are plotted versus subset sizes (*coverage analysis*).

Table 2. Maximum Pairwise Similarity Coefficients (Tanimoto or Cosine) for Maximum Dissimilarity Selections in the IC93 Database Using 2D Fingerprints or 3D PDTs[a]

NComp	FP_Tanimoto	FP_Cosine	PDT_Tanimoto	PDT_Cosine
60	0.34	0.52	0.07	0.17
80	0.39	0.55	0.10	0.23
100	0.43	0.61	0.12	0.27
120	0.47	0.64	0.16	0.31
140	0.50	0.68	0.19	0.34
160	0.55	0.71	0.21	0.37
180	0.59	0.75	0.23	0.40
200	0.64	0.78	0.25	0.43
220	0.66	0.80	0.27	0.45
240	0.68	0.82	0.29	0.47
260	0.72	0.84	0.31	0.49
280	0.73	0.85	0.33	0.51
300	0.75	0.86	0.34	0.53
320	0.76	0.87	0.36	0.54
340	0.78	0.88	0.38	0.56
360	0.80	0.89	0.39	0.57
380	0.81	0.89	0.41	0.58
400	0.82	0.90	0.41	0.60
420	0.83	0.91	0.44	0.62
440	0.84	0.92	0.44	0.62
460	0.85	0.92	0.46	0.64
480	0.86	0.93	0.47	0.65
500	0.87	0.93	0.49	0.66

[a] No pair of compounds in a subset is more similar than the maximum pairwise similarity coefficient. See Table 1 for further details.

selections only for smaller subsets (60−200, Table 1), while for larger ones (280−480) a lower performance than expected for a random approach is observed. When focusing on small subsets with 60 members, the 2D fingerprint-based method samples 62/58% (Tanimoto/cosine) of all biological classes, while a 45% coverage is expected for a random selection. The PDT performance of 47/49% reveals a better performance than random. Selecting more than 440 structures using 2D fingerprints covers all biological classes, while only 88% are covered using a random subset. Remarkably, for a PDT-based selection only 81/84% of all classes are covered, thus showing the slightly lower performance of this descriptor compared to a random approach.

In Table 2 the maximum pairwise Tanimoto or cosine coefficients for maximum dissimilarity selections in the IC93 database using 2D fingerprints or 3D PDTs are summarized. Here one of the main differences between both descriptors is obvious: While the maximum similarity coefficient for 2D fingerprints in a subset with 60 diverse compounds is 0.34/0.52 (Tanimoto/cosine coefficients), these maximum similarities are reduced to 0.07/0.17 for PDTs, respectively. In addition the maximum similarity for 500 compounds is 0.87/0.93 for 2D fingerprints, while for PDTs the highest similarities are 0.49/0.66. A PDT fingerprint encodes more subtle structural information than a conventional 2D fingerprint using a much larger number of bits. While this allows

Table 3. Increasing Number of Conformers for Subset Selection Using 3D PDTs and the Maximum Dissimilarity Method for the IC93 Database[a]

NComp	PDT-10t	PDT-10c	PDT-20t	PDT-20c	PDT-50t	PDT-50c	PDT-100t	PDT-100c	PDT-200t	PDT-200c	PDT-500t	PDT-500c	PDT-1000t	PDT-1000c
60	47.27	52.73	45.45	43.64	43.64	41.82	47.27	49.09	47.27	45.45	41.82	47.27	43.64	47.27
80	58.18	58.18	56.36	50.91	56.36	54.55	54.55	56.36	56.36	50.91	52.73	52.73	56.36	52.73
100	65.45	67.27	60.00	60.00	60.00	60.00	60.00	58.18	61.82	58.18	56.36	56.36	60.00	56.36
120	69.09	72.73	69.09	69.09	61.82	65.45	61.82	61.82	63.64	61.82	60.00	58.18	63.64	60.00
140	69.09	74.55	70.91	70.91	65.45	67.27	67.27	63.64	65.45	65.45	63.64	63.64	65.45	65.45
160	76.36	76.36	74.55	74.55	69.09	69.09	69.09	69.09	67.27	67.27	65.45	63.64	67.27	65.45
180	78.18	76.36	74.55	74.55	72.73	69.09	72.73	69.09	69.09	69.09	65.45	67.27	67.27	67.27
200	80.00	78.18	76.36	76.36	74.55	72.73	74.55	72.73	70.91	70.91	67.27	69.09	70.91	69.09
220	80.00	78.18	78.18	78.18	74.55	72.73	74.55	74.55	72.73	70.91	69.09	72.73	70.91	69.09
240	80.00	78.18	80.00	78.18	74.55	74.55	74.55	74.55	74.55	70.91	70.91	74.55	72.73	69.09
260	80.00	78.18	81.82	78.18	74.55	76.36	74.55	74.55	74.55	72.73	74.55	76.36	72.73	70.91
280	80.00	78.18	81.82	78.18	76.36	76.36	76.36	78.18	78.18	72.73	76.36	78.18	76.36	74.55
300	80.00	80.00	81.82	78.18	78.18	76.36	76.36	80.00	80.00	72.73	78.18	78.18	76.36	76.36
320	81.82	80.00	81.82	80.00	81.82	78.18	78.18	80.00	80.00	76.36	78.18	78.18	78.18	76.36
340	81.82	81.82	81.82	80.00	81.82	80.00	80.00	80.00	81.82	81.82	78.18	78.18	78.18	78.18
360	81.82	81.82	83.64	81.82	81.82	81.82	81.82	80.00	81.82	81.82	78.18	78.18	80.00	78.18
380	81.82	83.64	83.64	81.82	81.82	81.82	81.82	80.00	83.64	81.82	80.00	80.00	81.82	80.00
400	81.82	83.64	83.64	83.64	83.64	83.64	81.82	81.82	83.64	81.82	81.82	81.82	83.64	80.00
420	83.64	83.64	85.45	83.64	83.64	83.64	81.82	81.82	83.64	83.64	83.64	81.82	83.64	83.64
440	83.64	85.45	85.45	83.64	83.64	83.64	81.82	83.64	83.64	83.64	85.45	85.45	83.64	83.64
460	83.64	85.45	87.27	85.45	83.64	85.45	83.64	83.64	87.27	85.45	87.27	85.45	85.45	83.64
480	85.45	85.45	87.27	85.45	83.64	85.45	87.27	85.45	87.27	85.45	89.09	85.45	87.27	85.45
500	85.45	87.27	89.09	85.45	85.45	85.45	87.27	85.45	87.27	85.45	89.09	87.27	89.09	85.45

[a] The percentage of biological classes in IC93 covered by a subset is given. The column headers indicate the number of conformers (10, 20, 50, 100, 200, 500, and 1000) and the similarity coefficient (t, Tanimoto; c, cosine) for subset selection.

one to detect very small differences, there are still too many dissimilar triangle geometries populated even in two very similar molecules. Thus this descriptor does not use the full dynamic range for the pairwise similarity coefficients (between 0 and 1). Most pairwise similarities fall in a narrow range, which might cause a less clear similarity ranking.

Similar coverage performances are observed when altering the number of conformers for PDT descriptors. Figure 2 summarizes the conformational dependence of 3D PDT fingerprints for compound selections using the maximum dissimilarity algorithm, while numerical results are reported in Table 3. None of these PDT descriptors shows coverage similar to 2D fingerprints. Interestingly, for both similarity coefficients a lower number of conformers (10 or 20) performs better for smaller subset sizes (100−320). For very small subsets and the Tanimoto coefficient (60, 80) or large subsets (>340) with both similarity coefficients, all PDT descriptors behave almost similarly with a slight preference for a lower number of conformers. In general, PDT derived descriptors with a lower number of conformations and smaller subset sizes (<340) perform significantly better than a random selection, while this difference disappears for larger subsets and all PDT descriptors and similarity coefficients.

Thus extensive conformational sampling to account for molecular flexibility might introduce additional noise into PDT fingerprints, especially for a higher number of conformers. More pharmacophore triangle geometries in a bitstring are set by such a larger number of acceptable conformations. In contrast, triangle geometries for single low-energy conformers also reveal low performances as reported earlier.[19,20] This opens into the general conformational flexibility problem associated with 3D descriptors. Without detailed studies to investigate the influence of additional conformational analysis protocols and parameters on coverage and cluster separation results, it cannot be decided whether there

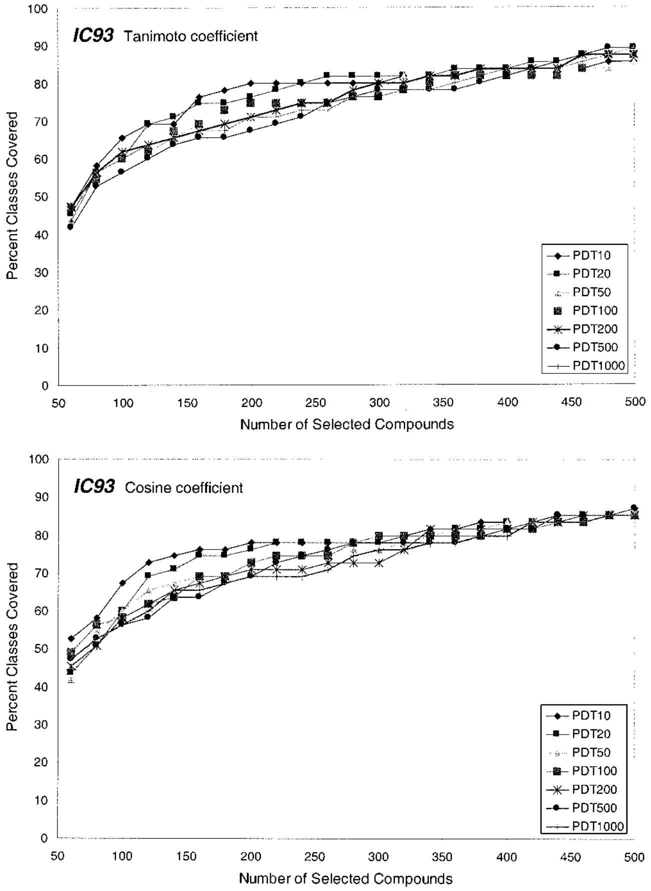


Figure 2. Conformational dependence of PDT descriptors for maximum dissimilarity selections using Tanimoto (a, top) or cosine coefficients (b, bottom). The percent biological classes covered are plotted versus subset sizes (*coverage analysis*). The number of conformers is indicated in the legend: *PDT10* indicates the use of 10 conformations per molecule.

is a general limitation of 3D descriptors or one can improve those descriptors by better, computationally expensive conformational analysis procedures plus adequate energy thresholds to generate relevant ensembles. Such a detailed analysis of the conformational flexibility problem was beyond the scope of this study.

The entire biological property space of the IC93 parent database is better represented by subsets designed using 2D fingerprints, while many biological classes are not represented in PDT derived or random subsets. For the IC93 database selected subsets with ~460 structures (38%, maximum Tanimoto coefficient 0.85) still represent all biological classes. While such a reduction does not necessarily translate to a higher hit rate after screening, it allows one to retrieve additional hits (entire biological activity islands) by similarity searches using valid descriptors around initial actives in a second step.

3.3. Hierarchical Cluster Analysis for Compound Classifications. Another useful approach for validation and comparison of chemistry space descriptors is to test the extent to which those descriptors group compounds of similar chemical structure and biological activity.[5,19] For compound classification a complete linkage hierarchical cluster analysis was applied to each descriptor/similarity coefficient combination. For each descriptor and cluster analysis, the resulting dendrogram was cut at different levels to generate between 60 and 500 individual clusters (in increments of 20). Cutting at lower levels produces more clusters with a higher similarity between all members. These classifications allow one to evaluate whether compounds of similar chemical structure and biological activity are grouped into similar clusters (*cluster separation analysis*). For each cluster level of each individual descriptor/coefficient combination, the average proportion p over all biological classes is plotted in Figure 3 versus the number of clusters at a certain level, while numerical results are reported in Table 4.

Again 2D fingerprints perform significantly better than PDTs, while both similarity coefficients behave similarly. In Figure 3a the average proportions for PDTs (100 conformers) are compared to those for 2D fingerprints. When 60 clusters are generated from the dendrograms, an average proportion p of 0.33/0.39 (Tanimoto/cosine) is observed for 2D fingerprints, while for PDTs a proportion of only 0.12/0.14 is computed. For comparison the mean proportion of active structures in the entire database is 0.018. When the number of clusters is increased, these differences between both descriptors are more significant (300 clusters: 0.88/0.88 for 2D fingerprints, but 0.58/0.56 for PDTs).

For a better assessment of the descriptors' abilities to group active molecules, additional 10 hierarchical cluster analyses were carried out using random numbers as descriptors. Their results were analyzed similarly; random proportions over all classes are averaged, listed in Table 4 (*RandomAv*), and plotted in Figure 3 for comparison.

Any increase in the proportion p of actives in the active cluster subset for valid molecular descriptors can arise from two different origins.[19] When more clusters are generated than active molecules are present in a data set by appropriately partitioning the hierarchical cluster dendrogram, the actives may distribute at no more than one per cluster. Such a distribution is very likely using random numbers for clustering. Cutting the dendrogram at even lower levels will not increase the number of active clusters but reject more and more inactives from the active cluster subset by further partitioning. Hence, the proportion p must increase with an increasing number of clusters. This is revealed by Figure 3, where an increasing number of clusters from the same dendrogram are connected. As expected, the average random proportion p also increases with increased partitioning of the dendrogram, i.e., when more clusters are generated, although no grouping of active compounds has taken place. Thus the

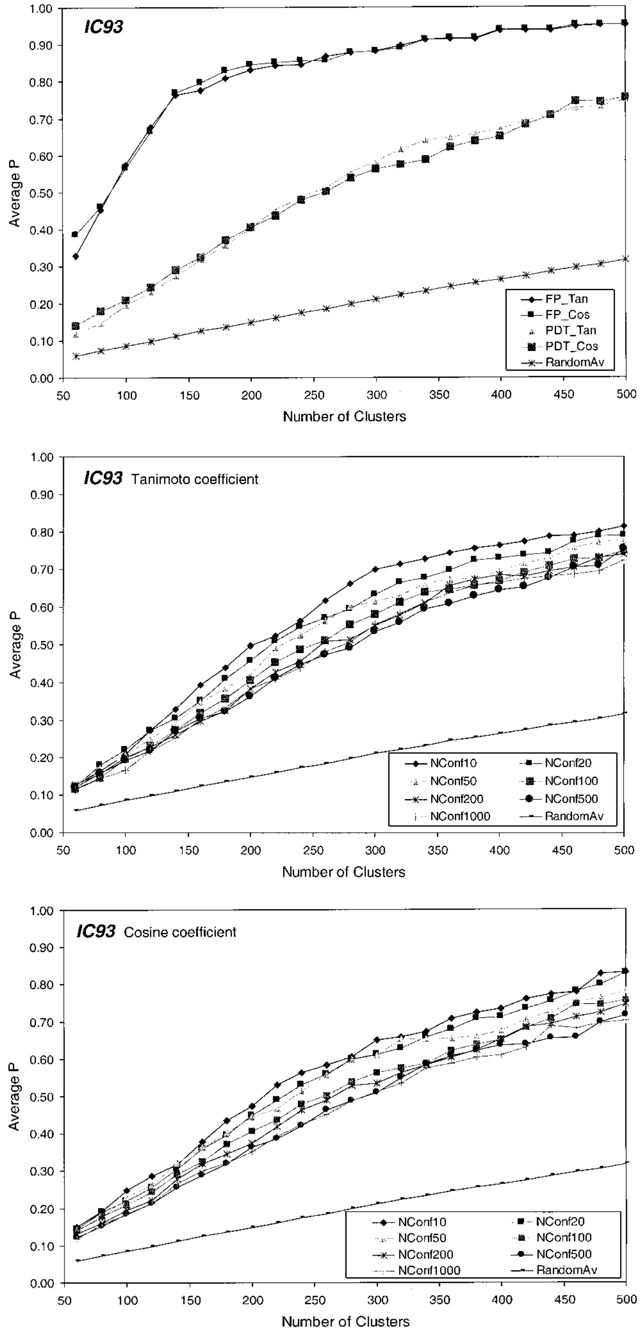


Figure 3. Hierarchical cluster analysis of the IC93 database. The average proportion p on the y-axis is plotted versus the number of clusters generated at different levels of the dendrogram: (a, top) Comparison between 2D fingerprints and PDTs generated using 100 conformers (Tanimoto or cosine coefficients are both used); (b, middle) comparison between various numbers of conformers to generate the PDT fingerprints (Tanimoto coefficient); (c, bottom) comparison between various numbers of conformers to generate the PDT fingerprints (cosine coefficient). *RandomAv* indicates an average over 10 individual cluster analyses using random numbers.

Table 4. Hierarchical Cluster Analysis of the IC93 Database Using 2D Fingerprints or 3D PDTs

(a) Comparison between 2D Fingerprints and PDTs (100 Conformers)[a]

NCluster	FP_Tanimoto	FP_Cosine	PDT_Tanimoto	PDT_Cosine	RandomAv
60	0.33	0.39	0.12	0.14	0.059
80	0.45	0.46	0.15	0.18	0.073
100	0.58	0.57	0.20	0.21	0.086
120	0.68	0.66	0.23	0.24	0.099
140	0.76	0.77	0.27	0.29	0.111
160	0.78	0.80	0.32	0.33	0.125
180	0.81	0.83	0.36	0.37	0.136
200	0.83	0.85	0.41	0.41	0.149
220	0.84	0.85	0.45	0.44	0.161
240	0.85	0.86	0.49	0.48	0.175
260	0.87	0.86	0.51	0.50	0.186
280	0.88	0.88	0.55	0.54	0.199
300	0.88	0.88	0.58	0.56	0.212
320	0.90	0.89	0.62	0.58	0.223
340	0.91	0.91	0.64	0.59	0.233
360	0.91	0.92	0.65	0.62	0.246
380	0.91	0.92	0.66	0.64	0.256
400	0.94	0.94	0.67	0.65	0.265
420	0.94	0.94	0.69	0.68	0.275
440	0.94	0.94	0.71	0.71	0.287
460	0.95	0.95	0.73	0.75	0.296
480	0.95	0.95	0.73	0.74	0.306
500	0.95	0.95	0.75	0.76	0.316

(b) Increasing Number of Conformers To Generate PDT Descriptor for Hierarchical Clustering Based on the Tanimoto Coefficient[b]

NCluster	NConf10	NConf20	NConf50	NConf100	NConf200	NConf500	NConf1000
60	0.13	0.12	0.12	0.12	0.13	0.12	0.12
80	0.17	0.18	0.17	0.15	0.16	0.16	0.15
100	0.21	0.22	0.21	0.20	0.20	0.19	0.17
120	0.27	0.27	0.25	0.23	0.23	0.22	0.22
140	0.33	0.31	0.29	0.27	0.26	0.27	0.25
160	0.39	0.35	0.35	0.32	0.30	0.31	0.30
180	0.44	0.41	0.39	0.36	0.33	0.32	0.33
200	0.50	0.46	0.42	0.41	0.38	0.36	0.38
220	0.53	0.51	0.49	0.45	0.43	0.41	0.41
240	0.57	0.55	0.53	0.49	0.46	0.45	0.44
260	0.62	0.57	0.57	0.51	0.51	0.47	0.48
280	0.66	0.60	0.60	0.55	0.52	0.49	0.51
300	0.70	0.63	0.62	0.58	0.55	0.54	0.55
320	0.72	0.67	0.63	0.62	0.58	0.56	0.59
340	0.73	0.68	0.66	0.64	0.61	0.60	0.61
360	0.74	0.70	0.68	0.65	0.66	0.61	0.64
380	0.76	0.72	0.68	0.66	0.67	0.63	0.66
400	0.77	0.73	0.70	0.67	0.69	0.65	0.67
420	0.78	0.74	0.72	0.69	0.68	0.66	0.68
440	0.79	0.74	0.73	0.71	0.70	0.68	0.68
460	0.79	0.78	0.76	0.73	0.71	0.71	0.69
480	0.80	0.79	0.77	0.73	0.73	0.71	0.70
500	0.82	0.79	0.78	0.75	0.74	0.76	0.72

(c) Increasing Number of Conformers to Generate PDT Descriptor for Hierarchical Clustering Based on the Cosine Coefficient[c]

NCluster	NConf10	NConf20	NConf50	NConf100	NConf200	NConf500	NConf1000
60	0.15	0.15	0.14	0.14	0.13	0.12	0.12
80	0.19	0.19	0.17	0.18	0.16	0.15	0.16
100	0.25	0.22	0.23	0.21	0.20	0.18	0.18
120	0.29	0.26	0.26	0.24	0.22	0.21	0.21
140	0.32	0.30	0.32	0.29	0.28	0.26	0.27
160	0.38	0.36	0.36	0.33	0.32	0.29	0.30
180	0.44	0.40	0.40	0.37	0.35	0.32	0.32
200	0.47	0.45	0.45	0.41	0.38	0.36	0.35
220	0.53	0.49	0.47	0.44	0.42	0.39	0.39
240	0.56	0.53	0.51	0.48	0.46	0.42	0.42
260	0.58	0.56	0.56	0.50	0.49	0.46	0.45
280	0.61	0.60	0.60	0.54	0.53	0.49	0.49
300	0.65	0.61	0.61	0.56	0.54	0.51	0.51
320	0.66	0.63	0.65	0.58	0.56	0.55	0.53
340	0.67	0.66	0.65	0.59	0.58	0.59	0.58
360	0.71	0.68	0.66	0.62	0.60	0.61	0.59
380	0.72	0.71	0.66	0.64	0.62	0.62	0.60

Table IV (Continued)

(c) Increasing Number of Conformers to Generate PDT Descriptor for Hierarchical Clustering Based on the Cosine Coefficient[c]

NCluster	NConf10	NConf20	NConf50	NConf100	NConf200	NConf500	NConf1000
400	0.73	0.71	0.68	0.65	0.65	0.64	0.61
420	0.76	0.73	0.71	0.68	0.69	0.64	0.63
440	0.77	0.76	0.73	0.71	0.69	0.66	0.69
460	0.78	0.78	0.75	0.75	0.71	0.66	0.68
480	0.83	0.80	0.76	0.74	0.72	0.70	0.70
500	0.83	0.83	0.78	0.76	0.75	0.72	0.70

[a] The average proportion for 55 target classes from the IC93 database monitors the ability of a descriptor to group active compounds: NClusters, number of clusters formed; FP_Tanimoto, 2D fingerprints and the Tanimoto coefficient; FP_Cosine, 2D fingerprints and the cosine coefficient; PDT_Tanimoto, 3D PDTs (100 conformers) and the Tanimoto coefficient; PDT_Cosine, 3D PDTs (100 conformers) and the cosine coefficient; RandomAv, averaged proportions over 10 cluster analyses using random numbers as descriptors. [b] The column headers indicate the number of conformers used to derive the descriptor. See part a for details. [c] See a and b for details.

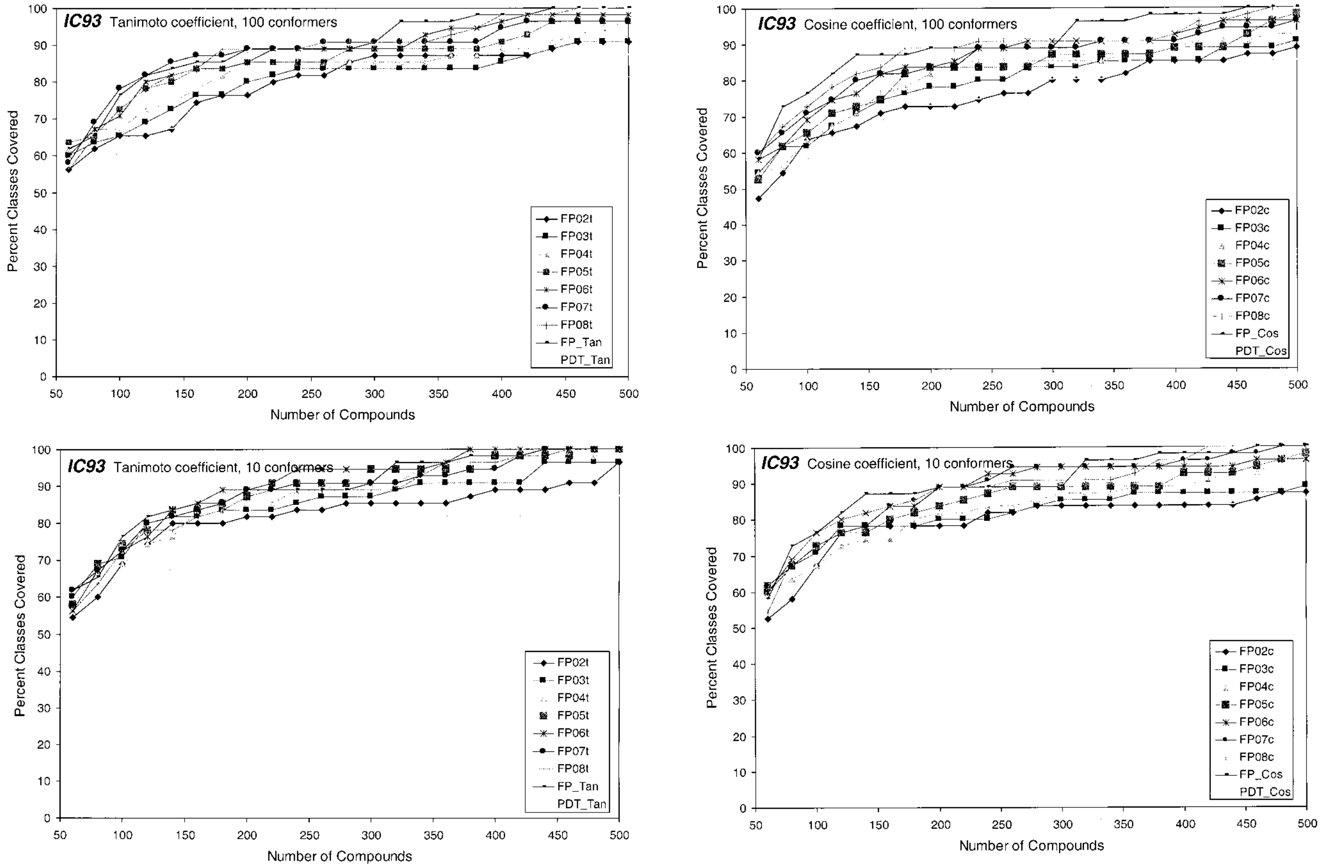


Figure 4. Combination of 2D fingerprints and PDT descriptors for maximum dissimilarity selections using (a, top left) the 100 conformers and Tanimoto coefficient; (b, top right) 100 conformers and the cosine coefficient; (c, bottom left) 10 conformers and the Tanimoto coefficient; (d, bottom right) 10 conformers and the cosine coefficient. The percent biological classes covered are plotted versus subset sizes (*coverage analysis*). The individual weighting factor for 2D fingerprints is indicated in the figure's legend: *FP02t* indicates a weighting of 0.2/0.8 for 2D fingerprints/PDTs using the Tanimoto coefficient as the similarity coefficient.

averaged random proportions reflect the lowest possible performance of hierarchical clustering. A further increase in this proportion will only occur if a particular descriptor ranks a pair of actives higher than similar pairs of actives and inactives. Indeed, a certain ability to group active compounds is found for all descriptors, as the average proportions are always higher than averaged random proportions in Figure 3.

Average proportions for an increasing number of conformers to generate PDT fingerprints are plotted in Figure 3b for Tanimoto coefficients and Figure 3c for cosine coefficients versus the number of clusters. The results are similar to those in section 3.2. For none of the PDT descriptors obtained for different numbers of conformers, an average proportion *p* comparable to 2D fingerprints is observed. Again lower numbers of conformers (10 or 20, indicated in Figure 3 as *NConf10*, *NConf20*) perform better for all numbers of clusters. While only 10 conformers and 300 clusters generated led to an average proportion *p* of 0.70/0.65 (Tanimoto/cosine), an increase to 100 conformers reduces this proportion to 0.58/0.56. In contrast, using 1000 conformers for PDT fingerprints further reduces this average proportion *p* to 0.55/0.51. Similar observations can be made for all numbers of clusters generated from the corresponding hierarchical cluster analyses. Thus a more detailed consideration of conformational flexibility to generate PDT fingerprints led to reduced

Table 5. Combining 2D Fingerprint and PDT Descriptors for Maximum Dissimilarity Selection[a]

NComp	FP02t	FP02c	FP03t	FP03c	FP04t	FP04c	FP05t	FP05c	FP06t	FP06c	FP07t	FP07c	FP08t	FP08c
(a) 100 Conformers														
60	56.36	47.27	60.00	54.55	61.82	54.55	63.64	52.73	60.00	58.18	58.18	60.00	56.36	60.00
80	61.82	54.55	63.64	61.82	67.27	56.36	65.45	61.82	67.27	61.82	69.09	65.45	65.45	67.27
100	65.45	63.64	65.45	61.82	67.27	63.64	72.73	65.45	70.91	69.09	78.18	70.91	72.73	72.73
120	65.45	65.45	69.09	67.27	72.73	67.27	78.18	70.91	80.00	74.55	81.82	74.55	78.18	78.18
140	67.27	67.27	72.73	70.91	74.55	70.91	80.00	72.73	81.82	76.36	85.45	80.00	81.82	81.82
160	74.55	70.91	76.36	74.55	78.18	76.36	83.64	74.55	83.64	81.82	87.27	81.82	83.64	83.64
180	76.36	72.73	76.36	76.36	81.82	78.18	83.64	81.82	83.64	83.64	87.27	81.82	89.09	89.09
200	76.36	72.73	80.00	78.18	85.45	81.82	85.45	83.64	85.45	83.64	89.09	83.64	89.09	89.09
220	80.00	72.73	81.82	78.18	85.45	85.45	85.45	83.64	89.09	85.45	89.09	83.64	89.09	89.09
240	81.82	74.55	83.64	80.00	85.45	85.45	85.45	83.64	89.09	89.09	89.09	89.09	89.09	90.91
260	81.82	76.36	83.64	80.00	85.45	85.45	85.45	83.64	89.09	89.09	90.91	89.09	90.91	90.91
280	85.45	76.36	83.64	83.64	85.45	85.45	89.09	83.64	89.09	90.91	90.91	89.09	90.91	90.91
300	87.27	80.00	83.64	83.64	85.45	85.45	89.09	87.27	89.09	90.91	90.91	89.09	90.91	90.91
320	87.27	80.00	83.64	83.64	85.45	85.45	89.09	87.27	89.09	90.91	90.91	89.09	90.91	90.91
340	87.27	80.00	83.64	85.45	85.45	85.45	89.09	87.27	92.73	90.91	90.91	90.91	90.91	90.91
360	87.27	81.82	83.64	85.45	87.27	87.27	89.09	87.27	94.55	90.91	90.91	90.91	92.73	90.91
380	87.27	85.45	83.64	85.45	87.27	87.27	89.09	87.27	94.55	90.91	90.91	90.91	94.55	90.91
400	87.27	85.45	85.45	85.45	89.09	89.09	90.91	89.09	96.36	92.73	94.55	90.91	98.18	92.73
420	87.27	85.45	87.27	85.45	89.09	90.91	92.73	89.09	98.18	94.55	96.36	92.73	98.18	96.36
440	89.09	85.45	89.09	89.09	90.91	90.91	96.36	89.09	98.18	96.36	96.36	94.55	98.18	96.36
460	90.91	87.27	90.91	89.09	92.73	90.91	96.36	92.73	98.18	96.36	96.36	94.55	100.00	98.18
480	90.91	87.27	90.91	89.09	94.55	92.73	96.36	96.36	98.18	96.36	96.36	94.55	100.00	100.00
500	90.91	89.09	90.91	90.91	96.36	92.73	96.36	98.18	98.18	96.36	96.36	96.36	100.00	100.00
(b) 10 Conformers														
60	54.55	52.73	60.00	60.00	58.18	60.00	58.18	61.82	56.36	60.00	61.82	61.82	56.36	54.55
80	60.00	58.18	69.09	67.27	67.27	63.64	69.09	67.27	67.27	69.09	67.27	67.27	63.64	69.09
100	69.09	67.27	70.91	70.91	69.09	67.27	74.55	72.73	72.73	76.36	72.73	72.73	72.73	76.36
120	74.55	76.36	80.00	78.18	74.55	72.73	78.18	76.36	76.36	80.00	80.00	76.36	78.18	80.00
140	80.00	78.18	81.82	78.18	76.36	74.55	83.64	76.36	83.64	81.82	81.82	78.18	78.18	81.82
160	80.00	78.18	81.82	78.18	83.64	74.55	83.64	80.00	85.45	83.64	83.64	83.64	81.82	83.64
180	80.00	78.18	83.64	78.18	83.64	80.00	85.45	81.82	89.09	83.64	85.45	85.45	83.64	85.45
200	81.82	78.18	83.64	80.00	85.45	81.82	87.27	83.64	89.09	89.09	89.09	89.09	87.27	89.09
220	81.82	78.18	83.64	80.00	85.45	81.82	90.91	85.45	90.91	89.09	89.09	89.09	89.09	89.09
240	83.64	81.82	85.45	80.00	89.09	83.64	90.91	87.27	94.55	92.73	90.91	90.91	89.09	89.09
260	83.64	81.82	87.27	81.82	89.09	83.64	90.91	89.09	94.55	92.73	90.91	94.55	89.09	90.91
280	85.45	83.64	87.27	83.64	90.91	85.45	90.91	89.09	94.55	94.55	90.91	94.55	89.09	90.91
300	85.45	83.64	87.27	85.45	90.91	87.27	94.55	89.09	94.55	94.55	90.91	94.55	89.09	90.91
320	85.45	83.64	89.09	85.45	90.91	87.27	94.55	89.09	94.55	94.55	90.91	94.55	89.09	90.91
340	85.45	83.64	90.91	85.45	94.55	87.27	94.55	89.09	94.55	94.55	92.73	94.55	92.73	90.91
360	85.45	83.64	90.91	87.27	94.55	89.09	94.55	89.09	96.36	94.55	92.73	94.55	96.36	92.73
380	87.27	83.64	90.91	87.27	94.55	89.09	94.55	89.09	100.00	94.55	94.55	94.55	96.36	96.36
400	89.09	83.64	90.91	87.27	98.18	89.09	98.18	92.73	100.00	94.55	94.55	96.36	96.36	96.36
420	89.09	83.64	90.91	87.27	98.18	90.91	98.18	92.73	100.00	94.55	98.18	96.36	98.18	100.00
440	89.09	83.64	96.36	87.27	98.18	92.73	98.18	92.73	100.00	94.55	100.00	98.18	98.18	100.00
460	90.91	85.45	96.36	87.27	98.18	92.73	98.18	94.55	100.00	96.36	100.00	98.18	100.00	100.00
480	90.91	87.27	96.36	87.27	98.18	96.36	100.00	96.36	100.00	96.36	100.00	100.00	100.00	100.00
500	96.36	87.27	96.36	89.09	98.18	96.36	100.00	98.18	100.00	96.36	100.00	100.00	100.00	100.00

[a] The percentage of covered biological classes in IC93 by a subset with a particular population is given. Column headers indicate the weights for 2D fingerprints for combined descriptors and the similarity coefficient (t, Tanimoto; c, cosine). The PDT weight is 1 − FP_weight. NComp, number of compounds in selected subset.

ability to group active and inactive compounds into similar classes.

3.4. Combining 2D and PDT Fingerprints for Selection and Clustering. Another study was carried out to combine the performance of 2D fingerprints with the interesting properties of PDT fingerprints. Although 2D fingerprint descriptors outperformed many other descriptors, they only led to identification of chemically similar classes in a similarity search, driven by 2D topology. It is hardly possible to identify candidates from nonrelated scaffolds in a 2D fingerprint-based search, while navigating from analogues of one chemical scaffold to another series should be possible with 3D descriptors, as they rank similarities of molecules by spatial properties and interactions. Thus, both descriptors were combined by averaging appropriately weighted pairwise similarity coefficients. The different weighting factors were changed from 0.2/0.8 (2D fingerprints/PDTs) to 0.8/0.2 in increments of 0.1. The weighting factors for PDTs are defined as (1−2D_fingerprint_weight). The resulting combined pairwise similarity coefficient matrices were used for maximum dissimilarity-based compound selection and hierarchical cluster analysis using the Tanimoto and cosine coefficients. Conformational flexibility was taken into account by using 10 or 100 conformers for generating the PDT descriptor.

Figure 4 summarizes the maximum dissimilarity results for differently weighted combinations of 2D fingerprints and 3D PDT descriptors. Weighting factors for 2D fingerprints are indicated in the corresponding figure legends: *FP02t* refers to weighting factors of 0.2/0.8 for 2D fingerprints/PDTs and the Tanimoto coefficient. Numerical results are reported in Table 5.

In general higher weighting factors for 2D fingerprints led to higher coverage of biological classes for all subset sizes and both similarity coefficients plotted in Figure 4 on the x-axis. For larger subset sizes (>300) and Tanimoto coefficients (Figure 4a), 2D fingerprints perform best. In contrast a 0.7/0.3 weighted descriptor outperforms 2D fingerprints alone for smaller subset sizes (<300). Thus it is possible to increase the performance of 2D fingerprints by a combined descriptor strategy for smaller subset sizes. This is not seen, when interpreting results obtained using the cosine coefficient and 100 conformers (Figure 4b), as here 2D fingerprints perform as good as or better than combined descriptors.

When combined descriptors are computed using a lower number of conformers for PDT fingerprints, the overall performance increases, as expected from studies with varied numbers of conformers for PDTs alone. Individual results for only 10 conformers are plotted in Figure 4c (Tanimoto coefficient) and Figure 4d (cosine coefficient). In the range between 180 and 300 members per subsets the combined descriptors clearly outperform 2D fingerprints. Weighting factors of 0.6 and 0.7 perform particularly good, thus clearly showing that an improvement over 2D fingerprints is possible using such a combined 2D/3D descriptor. While 85% of all biological classes are covered for 2D fingerprints and 180 members, a combination with a weighting factor of 0.6 led to a coverage rate of 89% using the Tanimoto coefficient. When investigating the cosine coefficient, this increased performance of combined descriptors starts at larger subset sizes (>220) but clearly outperforms 2D fingerprints. For 260 subset members, a coverage of 93% for a weighting factor of 0.6 is observed, while 2D fingerprints alone only led to 89% using the cosine coefficient. Thus both similarity coefficients for combined descriptors and weighting factors of 0.6/0.4 or 0.7/0.3 (2D fingerprints/PDTs) perform better as 2D fingerprints alone for medium sized subsets (200 to 300 members) and a small number of conformers. Those descriptors perform less efficiently, using a higher number of conformers for generating the PDT fingerprint, probably because of conformational averaging problems outlined above.

The interpretation of hierarchical clustering results for various combined descriptors and the Tanimoto coefficient led to a similar picture for PDT fingerprints based on 10 and 100 conformers, respectively. In general, higher weighting factors for 2D fingerprints led to an increase of the averaged proportion p. Furthermore a lower number of conformers also increases the average proportion. For smaller numbers of clusters (<200) 2D fingerprints perform best for all descriptor combinations and numbers of conformers, while for larger numbers of clusters 0.7/0.3 and 0.8/0.2 weighted 2D fingerprint/PDT descriptors show similar performances compared to 2D fingerprints, when investigating the results for 100 conformers. For each cluster level of each combined descriptor, the average proportion p over 55 biological classes is plotted in Figure 5 versus the number of clusters generated at a certain cluster level. Abbreviations such as *FP02* in the figure's legend refers to weighting factors such as 0.2/0.8 for 2D fingerprints/PDTs. Numerical results from the hierarchical cluster analysis are summarized in Table 6. The combination of 3D information into standard 2D fingerprints does not improve average proportions p at different cluster levels for 100 conformers (Figure 5a); while only 10 conformers are used, a remarkable improvement for the average proportion is observed for subset populations between 200 and 360 compounds (Figure 5b). It clearly can be seen that combined descriptors with weighting factors between 0.5 and 0.8 for 2D fingerprints outperform 2D fingerprints alone for those subset populations. In contrast, for larger subset sizes and both numbers of conformers, 2D fingerprints alone perform similarly to those combined descriptors.

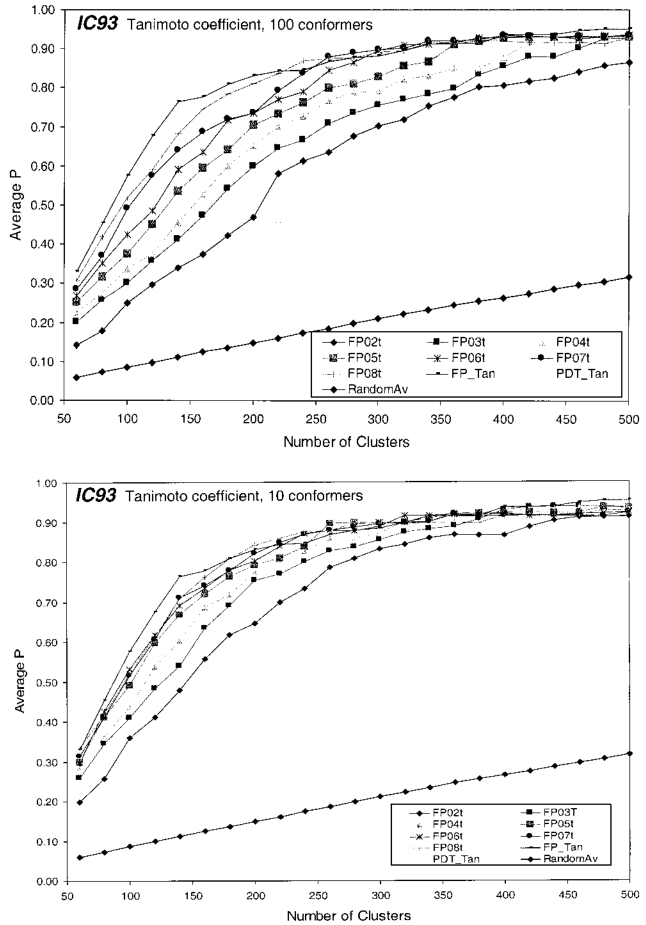


Figure 5. Combination of 2D fingerprints and PDT descriptors for hierarchical cluster analysis of the IC93 database generated using the Tanimoto coefficient: (a, top) 100 conformers for PDT generation; (b, bottom) 10 conformers for PT generation. The average proportion p on the y-axis is plotted versus the number of clusters generated at different levels of the dendrogram. The individual weighting factor for 2D fingerprints is indicated in the figure's legend: *FP02t* indicates a weighting of 0.2/0.8 for 2D fingerprints/PDTs; *RandomAv* indicates an average over 10 individual cluster analyses using random numbers as descriptors.

3.5. Sequential Dissimilarity Selection of Diverse Compound Subsets. In another study a sequential selection algorithm available in commercial software products[22,23] was used, where any selection is based on the dissimilarity of a candidate molecule to the single composite fingerprint as the union of all molecules of the previous selections. PDT results in terms of coverage of biological classes are plotted in Figure 6a (*PDT_ORIG*) for IC93 and Figure 6b for the BAYER database, while 2D fingerprint results using maximum dissimilarity selection are included for reference (*FP_MAXDISS* in Figure 6). These results are summarized in Table 7. This approach led to lower coverage rates than those obtained by random selections. The BAYER database

Table 6. Combining 2D Fingerprint and PDT Descriptors for Hierarchical Cluster Analysis[a]

NCluster	FP02t	FP03t	FP04t	FP05t	FP06t	FP07t	FP08t
			(a) 100 Conformers				
60	0.14	0.20	0.23	0.25	0.27	0.29	0.31
80	0.18	0.26	0.28	0.32	0.35	0.37	0.42
100	0.25	0.30	0.34	0.38	0.43	0.49	0.52
120	0.30	0.36	0.38	0.45	0.49	0.58	0.59
140	0.34	0.41	0.46	0.54	0.59	0.64	0.68
160	0.38	0.47	0.53	0.60	0.64	0.69	0.75
180	0.42	0.54	0.60	0.64	0.72	0.72	0.79
200	0.47	0.60	0.65	0.71	0.74	0.74	0.81
220	0.58	0.65	0.70	0.73	0.77	0.79	0.84
240	0.61	0.67	0.73	0.76	0.79	0.84	0.87
260	0.64	0.71	0.77	0.80	0.85	0.88	0.88
280	0.68	0.74	0.79	0.81	0.87	0.89	0.88
300	0.70	0.76	0.79	0.83	0.90	0.90	0.89
320	0.72	0.77	0.82	0.86	0.91	0.90	0.90
340	0.75	0.78	0.83	0.87	0.91	0.92	0.92
360	0.78	0.80	0.85	0.91	0.92	0.92	0.92
380	0.80	0.84	0.85	0.92	0.93	0.92	0.92
400	0.81	0.86	0.88	0.93	0.93	0.93	0.92
420	0.82	0.88	0.92	0.93	0.93	0.93	0.92
440	0.83	0.88	0.93	0.93	0.93	0.93	0.92
460	0.84	0.90	0.93	0.93	0.93	0.93	0.92
480	0.86	0.93	0.93	0.93	0.93	0.93	0.92
500	0.87	0.93	0.94	0.93	0.94	0.93	0.93
			(b) 10 Conformers				
60	0.20	0.26	0.29	0.30	0.30	0.31	0.34
80	0.26	0.35	0.36	0.41	0.43	0.41	0.42
100	0.36	0.41	0.44	0.49	0.53	0.52	0.52
120	0.41	0.48	0.54	0.60	0.62	0.61	0.61
140	0.48	0.54	0.60	0.67	0.69	0.71	0.71
160	0.56	0.64	0.69	0.72	0.73	0.74	0.76
180	0.62	0.69	0.72	0.76	0.78	0.78	0.81
200	0.65	0.75	0.78	0.79	0.80	0.82	0.84
220	0.70	0.77	0.81	0.81	0.84	0.85	0.86
240	0.73	0.80	0.83	0.84	0.87	0.87	0.88
260	0.79	0.83	0.86	0.90	0.88	0.88	0.88
280	0.81	0.84	0.86	0.90	0.88	0.89	0.90
300	0.83	0.85	0.87	0.90	0.89	0.89	0.90
320	0.84	0.87	0.90	0.90	0.92	0.90	0.90
340	0.86	0.88	0.90	0.91	0.92	0.90	0.90
360	0.87	0.89	0.92	0.92	0.92	0.92	0.90
380	0.87	0.91	0.92	0.92	0.92	0.91	0.90
400	0.87	0.93	0.92	0.92	0.92	0.91	0.92
420	0.89	0.94	0.93	0.92	0.91	0.92	0.92
440	0.90	0.94	0.93	0.92	0.92	0.91	0.92
460	0.91	0.94	0.93	0.92	0.92	0.91	0.92
480	0.91	0.94	0.94	0.94	0.92	0.91	0.93
500	0.91	0.93	0.93	0.93	0.92	0.92	0.93

[a] Tanimoto coefficient used as similarity coefficient. The average proportion for 55 target classes from the IC93 database is reported. Individual column headers indicate the weights for 2D fingerprints for combined descriptors. The PDT weight is 1 − FP_weight. NClusters, number of clusters formed.

with 11 different biological classes confirmed those results, which are a consequence of the sequential order of candidate evaluation, leading to an enrichment by the first molecules of a data set.

This led us to evaluate another combined selection strategy. Before sequential selection was applied on the basis of PDT fingerprints, the initial data set was ordered following 2D fingerprint-based dissimilarity. This corresponds to the *ChemDiverse* implementation,[22] where this sorting is based on a modified version of 2D atom-pair descriptors.[44] The new PDT-based selection (*PDT_CUT85* in Figure 6, Table 7) with 60 members covers 58% biological classes for the IC93 database. A corresponding subset with 15 members now represents 73% of all biological classes of the BAYER database, similar to 2D fingerprint-based selections. Thus small subsets are dominated by 2D fingerprint sorting and not 3D pharmacophoric key diversity, while no improvement of 2D fingerprint selections is found. When selecting more than 440 structures of the IC93 database or 40 structures of the BAYER database, respectively, all biological classes are covered using 2D fingerprints, while for IC93 78% and for BAYER 91% are covered using PDT-based sequential selection. For those studies a cutoff Tanimoto coefficient of 0.85 was used for adding new compounds to the temporary selection list. Results for other cutoff values (0.80, 0.90, and 0.95 for IC93 and 0.85, 0.90 for BAYER) are also presented in Table 7 and Figure 6 (IC93: *PDT_CUT85*, *PDT_CUT90* with 0.85 and 0.90 as cutoffs; BAYER: *PDT_CUT85* with a cutoff of 0.85, respectively). In general the subset *PDT_CUT85* with a cutoff value of 0.85 performs best. For smaller subsets from IC93 (<220), this selection performs better than a random approach, while this is reversed for for larger subsets, suggesting that the sequential selection strategy with PDT descriptors is not appropriate. Although presorting and PDT selection led to improved results, none of those subsets was superior compared to 2D fingerprints and maximum dissimilarity techniques.

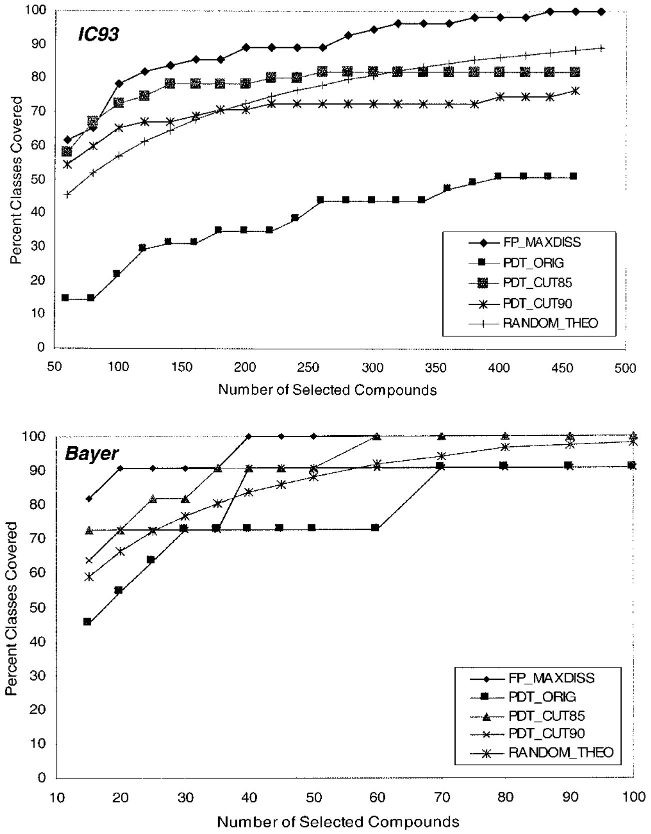


Figure 6. Percent biological classes covered from (a, top) the IC93 database and (b, bottom) the BAYER database, plotted on the *y*-axis versus the subset population (*x*-axis) for selection using various methods: 2D fingerprints and maximum dissimilarity selection (*FP_MAXDISS*), theoretical random selections (*RANDOM_THEO*), and different implementations of the PDT selection method (*PDT_ORIG, PDT_CUT85* and *PDT_CUT90*).

3.6. Neighborhood Relationship for PDT Fingerprints. The relationship between structural similarity and biological activity for PDT fingerprints was investigated using two QSAR data sets with 138 nonpeptidic angiotensin-converting enzyme ACE inhibitors[45] and 58 dipeptidic ACE inhibi-

Table 7. Performance of 2D Fingerprints, PDTs, and Random Selections for Designing Subsets Covering the Biological Properties of the IC93 and Bayer Database[a]

(a) IC93 Database

NComp	FP_MAXDISS	PDT_ORIG	PDT_CUT80	PDT_CUT85	PDT_CUT90	PDT_CUT95	RANDOM_THEO
60	61.82	14.55	56.36	58.18	54.55	54.55	45.47
80	65.45	14.55	69.09	67.27	60.00	56.36	51.93
100	78.18	21.82	74.55	72.73	65.45	60.00	57.05
120	81.82	29.09	76.36	74.55	67.27	60.00	61.29
140	83.64	30.91	78.18	78.18	67.27	63.64	64.78
160	85.45	30.91	80.00	78.18	69.09	63.64	67.73
180	85.45	34.55	80.00	78.18	70.91	63.64	70.36
200	89.09	34.55	80.00	78.18	70.91	63.64	72.73
220	89.09	34.55	80.00	80.00	72.73	63.64	74.73
240	89.09	38.18	81.82	80.00	72.73	63.64	76.60
260	89.09	43.64	81.82	81.82	72.73	63.64	78.13
280	92.73	43.64	83.64	81.82	72.73	63.64	79.71
300	94.55	43.64	83.64	81.82	72.73	65.45	81.04
320	96.36	43.64	83.64	81.82	72.73	65.45	82.27
340	96.36	43.64	83.64	81.82	72.73	67.27	83.40
360	96.36	47.27	83.64	81.82	72.73	67.27	84.38
380	98.18	49.09	83.64	81.82	72.73	67.27	85.44
400	98.18	50.91	83.64	81.82	74.55	67.27	86.25
420	98.18	50.91	83.64	81.82	74.55	67.27	87.15
440	100.00	50.91	83.64	81.82	74.55	67.27	87.85
460	100.00	50.91	83.64	81.82	76.36	67.27	88.53
480	100.00						89.20

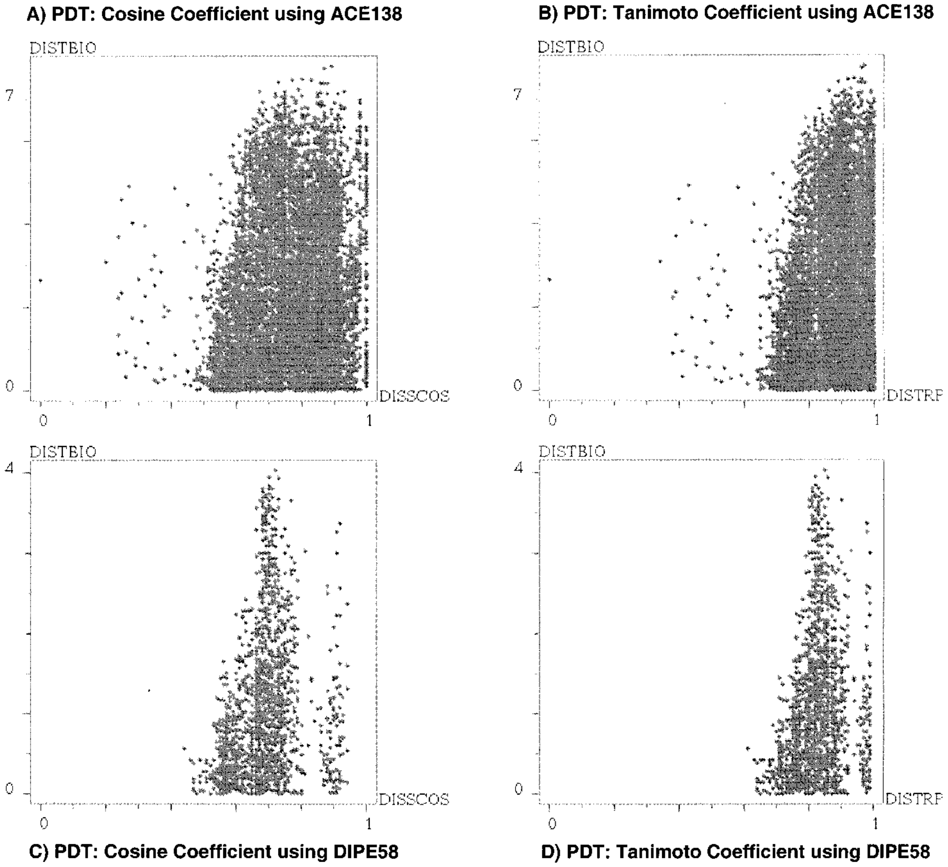
(b) Bayer Database

NComp	FP_MAXDISS	PDT_ORIG	PDT_CUT85	PDT_CUT90	RANDOM_THEO
15	82	45	73	64	59
20	91	55	73	73	67
25	91	64	82	73	72
30	91	73	82	73	77
35	91	73	91	73	81
40	100	73	91	91	84
45	100	73	91	91	86
50	100	73	91	91	88
60	100	73	100	91	92
70	100	91	100	91	94
80	100	91	100	91	97
90	100	91	100	91	97
100	100	91	100	91	98

[a] Individual values indicate the percentage of biological classes covered by one or more compounds in a subset.

tors.[46,47] The PDT fingerprint dissimilarity (1 − similarity coefficient) for Tanimoto or cosine coefficients and 100 conformers was computed for each pair of molecules of a data set,[4a,5] resulting in a data table with pairwise dissimilarities and absolute differences of biological activities. Scatter plots were used to study the descriptor differences on the *x*-axis versus the biological differences on the *y*-axis (Figure 7). Such a graph for a valid molecular descriptor should reveal a characteristic shape, which allows one to derive a maximum change of the biological activity per change in the descriptor.[4a] Following the similarity principle, any small physicochemical descriptor difference should correlate to only small changes in biological properties.[7,8] Hence, only a low number of data points is expected for a valid descriptor in the upper left triangle region of this graph. Any point in this upper left triangle corresponds to a pair of molecules, which is similar in terms of the molecular descriptor, but which reveals different biological properties.

For both data sets two graphs are generated using the Tanimoto (Figure 7b,d) or the Cosine coefficient (Figure 7a,c). The absolute difference of biological activities for each pair of compounds is plotted on the *y*-axis versus the descriptor differences. For both data sets neighborhood plots with a shape characteristic for valid descriptors are obtained. The upper left triangles for all descriptor/similarity coefficient combinations are essentially empty, indicating that PDT descriptors are able to a certain extent to group structurally similar molecules for this biological activity. This is in agreement with results from hierarchical clustering, as both techniques reveal that structurally similar molecules have similar activity. However, there are no points in Figure 7 indicating small structure descriptor changes, as found for other valid descriptors.[4a,5] While for 2D fingerprints with only ca. 1000 bits encoding chemical information very small descriptor differences (i.e. large Tanimoto coefficients > 0.90) are observed; for PDTs with 307 020 bits even for structurally very similar dipeptides no pair of molecules with a similarity coefficient > 0.5 is observed. Even the introduction of an additional methyl group from glycine to alanine causes the similarity coefficient to significantly decrease. Thus, more structural differences are encoded in this representation, leading to those smaller Tanimoto coefficients. Those greater distances do not imply that the PDT descriptors are less valid than 2D fingerprints, but in combination with any of the commonly used similarity coefficients, the full dynamic range for structural comparison


A) PDT: Cosine Coefficient using ACE138

B) PDT: Tanimoto Coefficient using ACE138

C) PDT: Cosine Coefficient using DIPE58

D) PDT: Tanimoto Coefficient using DIPE58

Figure 7. Comparison of pairwise absolute biological differences versus PDT fingerprint dissimilarities for two data sets: (a, top left) 138 non-peptidic ACE inhibitors and the cosine coefficient; (b, top right) 138 non-peptidic ACE inhibitors and the Tanimoto coefficient; (c, bottom left) 58 dipeptidic ACE inhibitors and the cosine coefficient; (d, bottom right) 58 dipeptidic ACE inhibitors and the Tanimoto coefficient.

between 0 and 1 is not used. This might cause a less clear similarity ranking. Significant information to compare two molecules might be hidden in noise, when both molecules are compared using their 3D pharmacophoric patterns. Furthermore the relevant bioactive conformation is not known, which might lead to the accumulation of additional noise.

4. CONCLUSION

The choice of valid molecular descriptors is an essential problem for designing representative subsets from virtual libraries and chemical databases. The novelty and interesting properties of 3D descriptors based on pharmacophore geometries led to this evaluation of pharmacophoric definition triplets for selection of representative subsets and for grouping active compounds into structurally similar classes. Their performance was studied using maximum dissimilarity methods, hierarchical cluster analysis, and sequential dissimilarity selections and compared to 2D fingerprints and random subsets or randomly generated clusters as reference. As any valid descriptor should follow the similarity principle, the degree of separation between active and inactive compounds for a single biological class is important to monitor. Furthermore any design should remove redundant compounds, but not biologically informative molecules. Hence their ability to select representative subsets covering biological classes to a certain degree is another quality criterium.

All methods lead to similar results: 2D fingerprints perform significantly better than 3D PDTs, while a lower number of conformers to generate PDT fingerprints significantly improves performances, suggesting that extensive conformational sampling to account for flexibility introduces additional noise into the PDT fingerprint. It is not clear whether this is a limitation of flexible 3D descriptors or can be addressed by more sophisticated conformational sampling techniques. For smaller subsets in maximum dissimilarity selection, PDTs perform significantly better than a random approach, while for larger subsets, both perfomances are almost similar. Interestingly generating a 2D/3D descriptor by combining 2D fingerprints and 3D PDT fingerprints with different weighting factors led to some combinations with significantly improved performance. Higher weighting factors for 2D fingerprints and lower number of conformers for PDT fingerprints in those fused descriptors improve performances. Remarkably, some combined descriptors with weighting factors between 0.5 and 0.8 for 2D fingerprints outperform 2D fingerprints for smaller subsets, while for larger subsets the performance is similar to 2D fingerprints. The detailed analysis of the relationship between biological activity and structural similarity for two smaller data sets revealed that a lot of important characteristic similarity information is lost or hidden in noise, when a pair of molecules is compared using their pharmacophoric patterns only.

The present study suggests that 2D information by itself or in combination with the 3D PDT fingerprints is indispensable for a successful diversity-based library design, compound selection, and classification, at least for the evaluated data sets. Obviously not the entire structural information encoded in a PDT fingerprint is relevant for molecular comparisons, while those descriptors have been shown to be valid to a certain extent. It might add a much deeper view into the general problem of 3D versus 2D descriptors when the outlined validation strategy is applied to other novel 3D descriptors, recently described in the literature.[24,48,49] It is also suggested that for successful descriptor design any combination of a 2D topological plus a 3D geometrical approach might be useful.

ACKNOWLEDGMENT

The authors would like to thank B. Wendt and R. D. Cramer (Tripos) for many useful discussions about descriptor validations.

REFERENCES AND NOTES

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

(2) (a) 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. (b) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discuss. Des.* **1997**, *7/8*, 31−49. (c) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying Diversity. In *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; Gordon, E. M., Kerwin, J. F., Jr., Eds.; Wiley: New York, 1998; pp 369−385. (d) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380. (e) Willett, P., Ed. *Computational Methods for the Analysis of Molecular Diversity*, Vol. 8; Kluwer/ESCOM: Dordrecht, The Netherlands, 1997.

(3) Pötter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.

(4) (a) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059. (b) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and cluster analysis applied to molecular diversity, American Chemical Society Meeting, Anaheim, CA, 1995, COMP3. (c) 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.

(5) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(6) 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.

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

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

(9) (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. *Persp. 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.

(10) Van Drie, J. H.; Lajiness, M. S. Approaches to virtual library design. *Drug Discuss. Today* **1998**, *3*, 274−283.

(11) (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.

(12) Cramer, R. D.; 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.

(13) Clark, M.; Cramer, R. D.; 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.

(14) (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.* **1984**, *19*, 66−70.

(15) 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.

(16) (a) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069. (b) Clark, R. D.; Ferguson, A. M.; Cramer, R. D. Bioisosterism and molecular diversity. *Perspect. Drug Discovery Des.* **1998**, *9−11* (3D QSAR in Drug Design: Ligand/Protein Interactions and Molecular Similarity), 213−224.

(17) Ashton, M. J.; Jaye, M. C., Mason, J. S. New Perspectives in Lead Generation. II. Evaluating Molecular Diversity. *Drug Discovery Today* **1996**, *2*, 71−78.

(18) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(19) 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.

(20) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(21) Davies, K. Using pharmacophore diversity to select molecules to test from commercial catalogs, including DIVERSet and HTS Chemicals. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; Chaiken, I. M., Janda, K. D., Eds.; American Chemical Society: Washington, D.C., 1996; pp 309−316.

(22) ChemDiverse, available from Chemical Design Ltd.: Oxon, U.K., 1996.

(23) *Selector*; Tripos, Inc.: St. Louis, MO, 1996.

(24) Mason, J. S. New Pharmacophore-Based Methods for Molecular Similarity Applications and to Design Diverse and Biased Libraries. *J. Mol. Graph.* **1998**, *16*, 51.

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

(26) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.−Activ. Relat.* **1996**, *14*, 501−506.

(27) Matter, H.; Lassen, D. Compound Libraries for Lead Discovery. *Chim. Oggi* **1996**, *14*, 9−15.

(28) *SYBYL Molecular Modelling Package*, Version 6.3; Tripos, Inc.: St. Louis, MO, 1996.

(29) *UNITY Chemical Information Software*, Version 2.6; Tripos, Inc.: St. Louis, MO, 1996.

(30) Ash, S.; Cline, M. A.; Homer, W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(31) Knuth, D. E. *Sorting and searching*; Addison-Wesley: Reading, MA.

(32) For more details to compute fingerprints, see e.g.: *UNITY Chemical Information Software*, Version 2.6; Tripos: Inc.: St. Louis, MO, 1996.

(33) 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.

(34) Pearlman, R. S.; Balducci, R.; Rusinko, A.; Skell, J. M.; Smith, K. N. *CONCORD*, program version 3.2.1; Tripos., Inc.: St. Louis, MO, 1996.

(35) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.

(36) 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.

(37) 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.

(38) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discuss. Des.* **1997**, *7/8*, 65−84.

(39) Holliday, J. D.; Ranade, S. S.; Willett, P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(40) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph.* **1997**, *15*, 372−385.

(41) Murtagh, F. *Multidimensional Clustering Algorithms. COMPSTAT Lectures. 4*. Physica-Verlag: Vienna, 1985.

(42) 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.

(43) *Index Chemical Database*-Subset from 1993; Institute for Scientific Information, Inc. (ISI): Philadelphia, PA, 1993.

(44) 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.

(45) 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.

(46) Wold, S.; Eriksson, L.; Hellberg, S.; Jonsson, J.; Sjöström, M.; Skagerberg, B.; Wikström, C. Principal property values for six nonnatural amino acids and their application to a structure−activity relationship for oxytocin peptide analogues. *Can. J. Chem.* **1987**, *65*, 1814−1820.

(47) Cushman, D. W.; Cheung, H.-S.; Sabo, E. F.; Ondetti, M. A. **1981**, in *Proceedings of the A. N. Richards Symposium, May 8−9, 1980*; Horowitz, Z. P., Ed.; Urban & Schwarzenberg: Baltimore, MD, 1980; pp 3−25.

(48) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 339−353.

(49) Martin, Y. C.; Brown, R. D.; Danaher, E. A.; DeLazzer, J., Lico, I. 3D descriptors that outperform substructures in diversity analysis. *Book of Abstracts, 214th ACS National Meeting*, Las Vegas, NV, 1997; American Chemical Society: Washington, DC, 1997; CINF-046.

CI980185H